Kapitel 0
Was ist Information?

Wir leben bekanntlich im Informationszeitalter, der „Rohstoff Information“ ist ein wesentlicher Wirtschaftsfaktor, und auch für unser Zusammenleben ist Information so wichtig, daß seit einigen Jahrzehnten viele im Gefolge des amerikanischen Mathematikers NORBERT WIENER (1894–1964) und des amerikanischen Soziologen D. BELL (1919–2011) von einer Informationsgesellschaft reden. Was aber ist Information? Und wie kann man sie messen?

Erstaunlicherweise gibt es auf keine dieser beiden Fragen eine allgemein akzeptierte Antwort.

§ 1: Ein Beispiel

Es ist Wahlkampfzeit, und viele Politiker bemühen sich, unsere Stimmen zu bekommen. Deshalb lädt auch Amadeus Wohlgeraten von der Partei für Gesundheit und Wohlstand (PGW) zu einer Informationsveranstaltung, wo er sich und seine Ziele vorstellen möchte. Welche Information erhalten die Teilnehmer dieser Veranstaltung?

Aus der Sicht von Amadeus Wohlgeraten sollen sie lernen, daß nur er und seine Partei sich wirklich für ihre Interessen einsetzen und daß nur sie im Falle eines Wahlsiegs Gesundheit und Wohlstand für alle Bürger bringen werden; die Gegenparteien haben nur Krankheit und Armut zu bieten. Das betrachtet er als die wesentliche Information in seiner Rede.

Seine Anhänger, die alles das schon längst wissen, warten auf die griffigen Slogans, die die PGW für solche Zwecke entwerfen ließ, um an den

richtigen Stellen ihre Begeisterung zu zeigen; als einzige Information nehmen sie mit nach Hause, daß sie die Stimmung im Saal dominierten.

Der Redakteur der Lokalzeitung mußte im Laufe seines Lebens schon viel zu viele Wahlversammlungen besuchen; er kann sich bereits im Voraus ziemlich genau denken, worum es in der Rede gehen wird. Ihn interessiert nur, ob Amadeus Wohlgeraten auf dem Weg zum Podium stolpert, ob es lustige Versprecher gibt oder, idealerweise, eine Saalschlacht; diese Information genügen, um seinem bereits eine Woche zuvor geschriebenen Bericht die endgültige Form zu geben.

Das Modehaus, das zu den Sponsoren der Partei für Gesundheit und Wohlbefinden zählt, ist vertreten durch den für die Kundenzeitschrift zuständigen Mitarbeiter. Was er über die gesunden und günstigen Stoffqualitäten der angebotenen Waren schreiben wird, ist natürlich unabhängig vom Verlauf der Versammlung; er möchte aber zumindest noch erwähnen, welchen Anzug mit optimal passender Krawatte Amadeus Wohlgeraten für diesen Abend aus dem großen Angebot des Modehauses ausgewählt hat.

Der Vorsitzende des örtlichen Vereins der Kaulquappenfreunde kann sich nicht erklären, wie sich jemand mit Politik beschäftigen kann, obwohl es noch so viele offene Fragen über Kaulquappen gibt. Trotzdem muß er alle Wahlveranstaltungen besuchen, denn wer auch immer die Wahl gewinnt, wird seine Kaulquappenpolitik möglicherweise danach ausrichten, ob er sich von den Kaulquappenfreunden unterstützt fühlt oder nicht. Den Vortrag faßt er kurz als „übliches Politikergeschwätz“ zusammen; doch in der nächsten Ausgabe des *Kaulquappenfreunds* kann er in den *Informationen aus dem Vorstand* stolz vermelden, daß er mit dem PGW-Kandidaten Amadeus Wohlgeraten über die wichtige Rolle der Kaulquappenzucht für Gesundheit und Wohlstand der Bevölkerung gesprochen habe.

Die Partei für Sorgenfreies Wohlbefinden, einer der Hauptkonkurrenten der Partei für Gesundheit und Wohlstand, möchte Wohlgeratens Rede natürlich genau analysieren; da jegliche Art von Ton- und Videoaufnahmen verboten sind, schicken sie einen Stenographen, der die Rede Wort für Wort mitschreibt. Da dieser zum Mitdenken keine Zeit hat, ist für

ihn der Informationsgehalt der Veranstaltung einfach die Folge der zu notierenden Worte.

Der Organisator einer Wahlwette möchte wissen, wie sich die Wahlchancen von Amadeus Wohlgeraten durch die Veranstaltung verändern, so daß er die Wettquoten gegebenenfalls neu festlegen kann. Die Information, die er mitnimmt, ist eine neue Schätzung für die Wahrscheinlichkeit eines Siegs der Partei für Gesundheit und Wohlbefinden, basiert auf die bekannten Umfrageergebnisse und seine Einschätzung von Stimmungswandeln im Saal.

Der Mathematiker, der sich mit Information beschäftigt, darf sich nicht darauf beschränken, dieses Geschehen einfach in einen mehr oder weniger nützlichen Formalismus zwingen; er möchte *quantitativ* beschreiben, was hier geschehen ist; zumindest für den Wettanbieter sollte es sogar möglich sein, die gewonnene Information direkt in Euro und Cent umzurechnen.

Angesichts der Vielzahl von Interessen der Beteiligten wird es dabei sicherlich nicht reichen, die an diesem Abend vermittelte Information durch eine einzige Zahl zu beschreiben; bei jedem einzelnen müssen wir sowohl sein Vorwissen als auch seinen Umgang mit dem Gesagten berücksichtigen. Was bei ihm ankommt, wurde möglicherweise bereits einigen Verarbeitungsschritten entworfen, z.B. weil er dank des Lärmpegels nur einen Teil der Rede hören kann, und auch er verarbeitet die ankommende Information weiter (War von Kaulquappen die Rede?), bevor ein Teil des Ergebnisses in sein Gedächtnis wandert. Schon jetzt können wir einen wesentlichen Aspekt dieser Informationsverarbeitung festhalten: Wie auch immer wir Information quantitativ fassen werden muß offensichtlich gelten, daß sie durch diese Verarbeitung höchstens abnehmen kann. Das heißt allerdings nicht unbedingt, daß sie dadurch weniger nützlich werden *muß:* Eine ungeordnete Sammlung von mehreren Millionen Datensätzen ist oft deutlich weniger nützlich als ein Satz von daraus abgeleiteten statistischen Kenngrößen. Die Information darüber ist zwar natürlich in den Datensätzen enthalten, sie zu extrahieren kann aber aufwendig sein. Auch mit solchen Fragen muß sich ein Mathematiker beschäftigen.

Am einfachsten zu fassen ist wohl noch die Information aus Sicht des Stenographen: In erster Näherung könnten wir einfach zählen, wie viele Zeichen er zu Papier gebracht hat. Aber selbst das ist nicht wirklich wohldefiniert, denn Stenographen arbeiten schließlich auch mit Kürzeln, die verwendet werden können, aber nicht müssen, und wenn sich Amadeus Wohlgeraten zu oft wiederholt haben sollte, erfand der Stenograph vielleicht auch noch ad hoc neue Kürzel für einige besonders häufige Phrasen. Die Frage, wie weit der dabei optimieren kann, führt uns zur SHANNONschen Informationstheorie und, in letzter Konsequenz, zur algorithmischen Informationstheorie.

Um das Vorwissen des Lokalredakteurs ins Spiel zu bringen, können wir die Information, die er bereits vor der Veranstaltung hatte, vergleichen mit seinem Informationsstand danach; die Differenz ist die neu gewonnene Information. Dies führt uns auf den Begriff der bedingten Information.

Im Falle des Buchmachers müssen wir zwei Wahrscheinlichkeitsverteilungen miteinander vergleichen: Die Siegwahrscheinlichkeiten der einzelnen Kandidaten so, wie er sie vor der Veranstaltung einschätzte, und die entsprechenden Zahlen, nach denen er künftig seine Prämien berechnet. Die gewonnene Information aus seiner Sicht ist somit eine Art Distanz zwischen zwei Wahrscheinlichkeitsverteilungen, die wir später als KULLBACK-LEIBLER-Distanz formalisieren werden; der finanzielle Wert der Information läßt sich über die Erwartungswerte für seinen Gewinn bezüglich der beiden Verteilungen quantifizieren.

Die Analysten in der Zentrale der Partei für Sorgenfreies Wohlbefinden werten nicht nur den (nach der Veranstaltung in ihren Computer übertragenen) Bericht des Stenographen aus, sondern zahlreiche weitere Berichte von ähnlichen Veranstaltungen. Sie müssen einerseits diese Berichte nach Gemeinsamkeiten gruppieren, um so einen Überblick über die gegnerische Strategie zu bekommen; andererseits müssen sie aber auch Ausreißer finden, die sich vielleicht als Wahlkampfmunition eignen könnten. Da sie auch die entsprechenden Daten anderer politischer Gegner auswerten müssen, haben sie viel zu tun und wollen ihre Arbeit möglichst automatisieren. Die mathematischen Verfahren, die

sie dabei anwenden können, werden uns im zweiten Teil der Vorlesung beschäftigen.

§ 2: Information in sprachlicher Sicht

Die Etymologie des Wortes *Information* trägt leider nur wenig zum Verständnis dieses Begriffs bei: Das lateinische *informatio* enthält den Wortstamm *forma*, Form oder Gestalt, und informatio wurde im Sinne von Bildung oder Unterricht gebraucht. Bei der Aufnahme des Worts in die deutsche Sprache im 15. bis 16. Jahrhundert verschob sich die Bedeutung zu *Nachricht* oder *Unterrichtung* (über einen Sachverhalt), und diesen Sinn hat das Wort heute auch in anderen modernen Sprachen.

Information hat also etwas mit der Übermittlung von Nachrichten zu tun; eine mathematische Theorie der Information muß sich daher insbesondere auch mit der Struktur von Nachrichten beschäftigen. Dafür interessiert sich selbstverständlich nicht nur die Mathematik; schon in der Logik von ARISTOTELES (384–322) finden sich Überlegungen, die in diese Richtung gehen, und auch die Grammatiker befassen sich schon seit weit über Tausend Jahren mit entsprechenden Fragen.

Einen wesentlichen Schritt in Richtung auf eine mathematische Beschreibung von Sprache leistete 1879 der Philosoph FRIEDRICH LUDWIG GOTTLOB FREGE (1848–1925) mit seiner *Begriffsschrift*, in der er die mathematische Logik in ihrer heutigen Form begründete. Sein Versuch, die gesamte Mathematik auf Logik zu reduzieren, scheiterte zwar, führte aber zur Entwicklung alternativer Ansätze sowohl zur Grundlegung der Mathematik als auch zur allgemeinen Untersuchung formaler Systeme und schließlich auch natürlicher Sprachen.

Vor allem durch die Arbeiten des amerikanischen Philosophen CHARLES WILLIAM MORRIS (1901–1979) entstand ab Ende der Dreißigerjahre die *Semiotik* als allgemeine Lehre von den Zeichen und ihrer Verwendung. Er unterscheidet drei Aspekte:

1. Die Syntax, in der es um die Zeichen selbst und die Regeln für ihre Aneinanderreihung geht.

2. Die Semantik, die sich mit der Bedeutung von Zeichenfolgen beschäftigt.

3. Die Pragmatik, in der es um deren *Gebrauch* geht: Dazu zählen beispielsweise unterschiedliche Bedeutungsebenen (Kopf, Haupt, Rübe), aber auch unterschiedliche Ziele, die mit einem Wort oder Satz erreicht werden sollen: Der Ausruf *Feuer!* etwa kann zwar bedeuten, daß es brennt, kann aber beim Militär auch der Befehl zum Schießen sein und bei einem Raucher die Bitte, ihm die Zigarette anzuzünden.

Die klassische Informationstheorie beschränkt sich auf rein syntaktische Aspekte; bei der Suche nach Information stehen dagegen semantische Aspekte im Vordergrund. Pragmatik spielt bei der mathematischen Behandlung von Information bislang keine Rolle.

Sobald wir von praktischen Anwendungen der Information reden, kommt allerdings ein neuer, bislang noch nicht erwähnter Gesichtspunkt ins Spiel: Information kann einen teilweise sogar beträchtlichen wirtschaftlichen Wert darstellen. Informationen über den Zustand eines Landes oder einen Unternehmens beeinflussen beispielsweise die Preise von Aktien, und je nachdem wie früh oder spät jemand darauf reagiert, kann er viel Geld gewinnen oder verlieren.

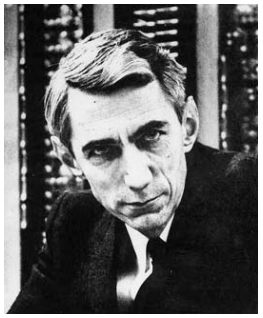
Der Wert solcher Informationen kann allerdings nicht objektiv beziffert werden: Die Nachricht, daß in einer südafrikanischen Goldmine eine neue stark erzführende Ader entdeckt wurde, ist wertlos für jemanden, der kein Geld für Aktienkäufe hat oder grundsätzlich nur in Europa investiert; für einen südafrikanischen Investor dagegen kann die Information einen beträchtlichen Wert haben.

Auch für ihn kann eine entsprechende Meldung allerdings völlig wertlos sein, etwa weil er sie schon seit Tagen kennt und längst darauf reagiert hat. Wenn wir vom Wert einer Information sprechen wollen, kann es sich daher immer nur um den Wert für eine bestimmte Person mit bestimmten Interessen handeln; insbesondere ist deren Vorwissen ein wichtiger, wenn auch bei weitem nicht der einzige Aspekt. Wir werden daher eine ganze Reihe weiterer Maße benötigen, um auch solche Situationen mathematisch zu beschreiben.

Kapitel 1

Shannons Informationstheorie

Die bekannteste quantitative Definition von Information geht zurück auf CLAUDE SHANNON; Bücher mit Titeln wie *Informationstheorie* befassen sich meist ausschließlich damit. Die inhaltliche Interpretation von Information spielt hier keinerlei Rolle, es geht nur um ihre sichere Übermittlung. Sicherheit bezieht sich dabei sowohl auf den Schutz vor Übertragungsfehlern (durch fehlererkennende und -korrigierende Codes) als auch auf die Geheimhaltung (Kryptographie).



CLAUDE ELWOOD SHANNON (1916–2001) wurde in Petoskey im US-Bundesstaat Michigan geboren; 1936 verließ er die University of Michigan mit sowohl einem Bachelor der Mathematik als auch einem Bachelor der Elektrotechnik, um am M.I.T. weiterzustudieren. Seine 1938 geschriebene Diplomarbeit *A symbolic analysis of relay and switching circuits* bildet die Grundlage der digitalen Informationsverarbeitung auf der Grundlage der hier entwickelten Schaltlogik; seine Dissertation 1940 befaßte sich mit Anwendungen der Algebra auf die MENDELschen Gesetze. Danach arbeitete er bis 1956 bei den Bell Labs, wo er während des zweiten Weltkriegs insbesondere über die Sicherheit kryptographischer Systeme forschte. Seine *Mathematical theory of cryptography* wurde aus Geheimhaltungsgründen erst 1949 zur Veröffentlichung freigegeben. Seine wohl bekannteste Arbeit ist die 1948 erschienene *Mathematical theory of communication,* in der er die fehlerfreie Übertragung von Nachrichten über einen gestörten Kanal untersuchte. Von 1956 bis zu seiner Emeritierung 1978 lehrte er am M.I.T., das er dadurch zur führenden Universität auf dem Gebiet der Informationstheorie und Kommunikationstechnik machte. Zu seinen zahlreichen Arbeiten zählt auch eine über die mathematische Theorie der Jongliermuster, anhand derer Jongleure eine Reihe neuer Muster gefunden haben; auch konstruierte er mehrere Jonglierroboter.

§ 1: Die Entropie einer Quelle

Gerade weil der SHANNONsche Informationsbegriff der in den Wissenschaften am weitesten verbreitete ist, müssen wir uns als allererstes klar werden, was er nicht ist: Es geht nicht darum, den Informationsgehalt einer einzelnen Nachricht zu messen.

Im 1949 erschienenen Buch *The mathematical theory of communication,* das SHANNONs gleichnamige Arbeit von 1948 zusammen mit einer ausführlichen Einleitung von WARREN WEAVER enthält, schreibt letzterer zu Beginn von §2.2:

> The word *information,* in this theory, is used in a special sense that mus not be confused with its ordinary usage. In particular, *information* must not be confused with meaning.
>
> In fact, two messages, one of which is heavily loaded with meaning and the other of which is pure nonsense, can be exactly equivalent, from the present viewpoint, as regards information. It is this, undoubtedly, that Shannon means when he says that "'the semantic aspects are necessarily irrelevant to the engineering aspects."' But this does not mean that the engineering aspects are necessarily irrelevant to the semantic aspects.
>
> To be sure, this word information in communication theory relates not so much to what you *do* say, as in what you *could* say.

SHANNON betrachtet Nachrichten also stets vor dem Hintergrund einer Auswahl; was er messen will, sind die Wahlmöglichkeiten des Senders und die Ungewißheit des Empfängers *vor* Übermittlung der Nachricht.

Ein solcher Ansatz kann nur funktionieren, wenn sowohl für den Sender als auch den Empfänger klar ist, welche Nachrichten grundsätzlich übertragen werden *könnten.* Zu diesem Zweck geht SHANNON aus von einem festen *Alphabet* A. Darunter versteht er irgendeine endliche Menge, deren Elemente zwar als Buchstaben bezeichnet werden, die aber auch elektrische Signale, ASCII-Zeichen, Ereignisse und vieles andere sein können.

Der Sender wird modelliert durch eine Nachrichten*quelle,* die eine Folge von Buchstaben des Alphabets A produziert. In den seltensten Fällen werden dabei alle Buchstaben mit der gleichen Häufigkeit vorkommen;

in SHANNONs Ansatz ist daher jedem Buchstaben $x_i \in A$ eine Häufigkeit p_i zugeordnet. Natürlich müssen alle $p_i \geq 0$ sein und ihre Summe gleich eins; bei einem Alphabet aus n Buchstaben liegt das Tupel der Wahrscheinlichkeiten also in der Menge

$$\Delta_n = \left\{ (p_1, \ldots, p_n) \;\middle|\; p_i \geq 0 \text{ für alle } i \text{ und } \sum_{i=1}^{n} p_i = 1 \right\}.$$

Mathematisch gesehen ist eine Nachrichtenquelle also eine diskrete Zufallsvariable, die Werte aus A annimmt.

Ausgangspunkt für die Quantifizierung von Information ist der *mittleren* Informationsgehalt eines Buchstabens. Da dieser Informationsgehalt sicherlich nicht von den Namen der Buchstaben abhängt, können wir H einfach als eine Funktion der Buchstabenwahrscheinlichkeiten p_i betrachten; wir suchen also für jede natürliche Zahl n eine Funktion $H\colon \Delta_n \to \mathbb{R}_{\geq 0}$, so daß $H(p_1, \ldots p_n)$ der mittlere Informationsgehalt eines Buchstabens aus einem n-elementigen Alphabet ist. wobei $p_1, \ldots, p_n$ die Häufigkeiten der einzelnen Buchstaben sind.

Eine solche Funktion sollte nach SHANNON vernünftigerweise die folgenden Bedingungen erfüllen:

1. H ist stetig, denn natürlich sollen kleine Änderungen an den p_i nicht zu sprunghaften Änderungen am Informationsgehalt führen.
2. Die Funktion $L(n) = H\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$ ist monoton wachsend, d.h. wenn wir eine Quelle haben, die die Buchstaben ihres Alphabets mit gleicher Wahrscheinlichkeit ausstößt, steigt der Informationsgehalt pro Buchstabe mit der Buchstabenanzahl. Beispielsweise liefert ein Meßfühler mehr Information, wenn er eine größere Auflösung hat.

Etwas technischer und schwerer zu verstehen ist SHANNONs dritte Forderung: Vor jeder Übertragung eines Buchstabens steht der Sender vor einer Wahl. Wenn er diese Wahl in mehrere Teilentscheidungen zerlegt, soll sich dadurch nichts an der Gesamtinformation ändern. Konkret: Ist C eine Teilmenge des Alphabets A, so kann der Sender in einem ersten Schritt entweder ein Element von $A \setminus C$ auswählen oder sich dafür entscheiden, ein Element aus C zu senden. Im letzteren Fall muß er dann in einem zweiten Schritt konkretisieren, welches der Elemente aus C er senden will. Wenn wir der Einfachheit halber annehmen, daß $A \setminus C$

die ersten m der n Elemente von A enthält und die Summe der Wahrscheinlichkeiten für die Elemente aus C gleich p^* ist, soll dann also gelten

3. $H(p_1, \ldots, p_n) = H(p_1, \ldots, p_m, p^*) + p^* H\left(\frac{p_{m+1}}{p^*}, \ldots, \frac{p_n}{p^*}\right)$, falls $p^* = \sum_{i=m+1}^{n} p_i > 0$ ist. (Der Faktor p^* vor dem zweiten Summanden kommt daher, daß nur mit Wahrscheinlichkeit p^* überhaupt eine zweite Entscheidung getroffen wird, und die Nenner in den Argumenten sind notwendig, da der Buchstabe a_i mit $i > m$ die Wahrscheinlichkeit p_i/p^* hat, falls bereits feststeht, daß ein Buchstabe aus C gesendet wird.)

Vielleicht hilft der folgende Spezialfall, diese Bedingung etwas besser zu verstehen:

Lemma: $A = \{a_1, \ldots, a_m\}$ und $B = \{b_1, \ldots, b_n\}$ seien zwei endliche Alphabete, wobei a_i mit Wahrscheinlichkeit p_i und b_j mit Wahrscheinlichkeit q_j auftrete. Gibt man dem Element $(a_i, b_j) \in A \times B$ die Wahrscheinlichkeit $p_i q_j$, so ist

$$H(\ldots, p_i q_j, \ldots) = H(p_1, \ldots, p_m) + H(q_1, \ldots, q_n),$$

bei zwei unabhängigen Zufallsvariablen addieren sich also die mittleren Informationsgehalte.

Beweis: Wir wenden Forderung 3 an auf die Teilmenge $C = \{a_m\} \times B$ von $A \times B$. Hier ist $p^* = \sum_{j=1}^{n} p_m q_j = p_m$, also folgt

$$H(\underbrace{\ldots, p_i q_j, \ldots}_{\substack{i=1,\ldots,m \\ j=1,\ldots,n}}) = H(\underbrace{\ldots, p_i q_j, \ldots}_{\substack{i=1,\ldots,m-1 \\ j=1,\ldots,n}}, p_m) + p_m H(q_1, \ldots, q_n).$$

Auf den ersten Summanden links können wir das gleiche Argument anwenden und die Paare mit a_{m-1} abspalten *usw.;* wir erhalten schließlich

$$\begin{aligned} H(\underbrace{\ldots, p_i q_j, \ldots}_{\substack{i=1,\ldots,m \\ j=1,\ldots,n}}) &= H(p_1, \ldots, p_m) + \sum_{i=1}^{m} p_i H(q_1, \ldots, q_n) \\ &= H(p_1, \ldots, p_m) + H(q_1, \ldots, q_n). \end{aligned}$$

∎

Satz: Zu jeder Familie von Funktionen $H\colon \Delta_n \to \mathbb{R}_{\geq 0}$, die obige drei Bedingungen erfüllt, gibt es eine reelle Zahl $a > 1$ so daß

$$H(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log_a p_i$$

ist, wobei $p_i \log p_i$ für $p_i = 0$ als Null interpretiert werden soll. Insbesondere ist H bis auf einen positiven Faktor eindeutig bestimmt.

Beweis: In einem *ersten Schritt* beschränken wir uns auf den Fall, daß alle p_i gleich sind, betrachten also für jedes $n \in \mathbb{N}$ nur den einen Wert $L(n) = H\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$.

A_1 bis A_m seien m voneinander unabhängige Quellen, die jeweils r verschiedene Buchstaben mit gleicher Wahrscheinlichkeit $1/r$ liefern; der Informationsgehalt jeder dieser Quellen ist also $L(r)$. Das Produkt $A_1 \times \cdots \times A_m$ enthält r^m tupel, die allesamt mit derselben Wahrscheinlichkeit $1/r^m$ auftreten; die Gesamtinformation ist also

$$H\left(\frac{1}{r^m}, \ldots, \frac{1}{r^m}\right) = L(r^m)\,.$$

Aus dem gerade bewiesenen Lemma folgt induktiv, daß dies die Summe der Informationsgehalte der Quellen A_i ist, d.h. $L(r^m) = mL(r)$.

Nun betrachten wir natürliche Zahlen r, s, m, n mit $r^m \leq s^n \leq r^{m+1}$; dann ist (unabhängig von der Basis des Logarithmus)

$$m \log r \leq n \log s \leq (m+1) \log r \quad \text{oder} \quad \frac{m}{n} \leq \frac{\log s}{\log r} \leq \frac{m+1}{n}\,.$$

Wegen der in Forderung zwei postulierten Monotonie von L gilt die Ungleichung $L(r^m) \leq L(s^n) \leq L(r^{m+1})$; wie wir gerade gesehen haben, können wir diese auch schreiben als

$$mL(r) \leq nL(s) \leq (m+1)L(r)\,.$$

Division durch $nL(r)$ macht daraus

$$\frac{m}{n} \leq \frac{L(s)}{L(r)} \leq \frac{m+1}{n}\,,$$

$L(s)/L(r)$ und $\log s/\log r$ liegen daher beide im Intervall $\left[\frac{m}{n}, \frac{m+1}{n}\right]$, so daß

$$\left|\frac{L(s)}{L(r)} - \frac{\log s}{\log r}\right| \leq \frac{1}{n}$$

sein muß. Da n beliebig groß gewählt werden kann, gilt dies für alle $n \in \mathbb{N}$, d.h.

$$\frac{L(s)}{L(r)} = \frac{\log s}{\log r} \quad \text{oder} \quad L(s) = \frac{L(r)}{\log r} \cdot \log s\,.$$

Somit ist $L(s)$ proportional zu einem Logarithmus, wobei die Proportionalitätskonstante $L(r)/\log r$ wegen der Monotonie sowohl von L als auch des Logarithmus positiv sein muß. Mithin gibt es eine reelle Zahl $a > 1$ mit $L(n) = \log_a n$ für alle $n \in \mathbb{N}$.

Im speziellen Fall von Quellen, die alle Buchstaben mit gleicher Wahrscheinlichkeit ausgeben, ist der Satz damit bewiesen.

Im *zweiten Schritt* verlangen wir von den Wahrscheinlichkeiten p_i nur noch, daß es sich dabei um positive rationale Zahlen handelt. Wir betrachten also ein Alphabet $A = \{a_1, \ldots, a_n\}$ aus n Buchstaben, deren i-ter die Wahrscheinlichkeit $p_i = g_i/g$ habe mit $g_i \in \mathbb{N}$. Da die Summe aller p_i gleich eins ist, muß dabei $\sum g_i = g$ sein.

Weiter betrachten wir ein Alphabet B aus g Buchstaben $b_1, \ldots, b_g$, die allesamt die gleiche Wahrscheinlichkeit $1/g$ haben. Diese Buchstaben verteilen wir auf n disjunkte Teilmengen $B_i \subseteq B$ derart, daß B_i aus g_i Buchstaben besteht. Durch n-fache Anwendung der dritten Forderung erhalten wir die Gleichung

$$H\left(\frac{1}{g}, \ldots, \frac{1}{g}\right) = H(p_1, \ldots, p_n) + \sum_{i=1}^{n} p_i H\left(\frac{1}{g_i}, \ldots, \frac{1}{g_i}\right)\,.$$

Die Funktionen mit lauter gleichen Argumenten können wir durch Lo-

garithmen ausdrücken und erhalten dann

$$\begin{aligned} H(p_1,\dots,p_n) &= H\Big(\frac{1}{g},\dots,\frac{1}{g}\Big) - \sum_{i=1}^{n} p_i H\Big(\frac{1}{g_i},\dots,\frac{1}{g_i}\Big) \\ &= \log_a g - \sum_{i=1}^{n} p_i \log_a g_i = \sum_{i=1}^{n} p_i \big(\log_a g - \log_a g_i\big) \\ &= -\sum_{i=1}^{n} p_i \log_a \frac{g_i}{g} = -\sum_{i=1}^{n} p_i \log_a p_i \,, \end{aligned}$$

wie behauptet.

Als *dritten Schritt* betrachten wir den allgemeinen Fall. Wir gehen also aus von einer Familie von Funktionen $H\colon \Delta_n \to \mathbb{R}$, die alle drei Forderungen SHANNONs erfüllt. Wie wir bereits wissen, ist dann

$$H(p_1,\dots,p_n) = -\sum_{i=1}^{n} p_i \log p_i \,,$$

falls wir für alle p_i positive rationale Zahlen einsetzen. Auf der rechten Seite steht eine Funktion, die für alle positiven reellen Werte der p_i stetig ist; die Funktion links muß nach SHANNONs erster Forderung stetig auf Δ_n sein. Da zwei stetige Funktionen, die für alle rationalen Werte aus einer offenen Menge übereinstimmen, dort gleich sind, gilt obige Gleichung im Innern von Δ_n, also für alle positiven reellen Werte der p_i.

Bleibt noch der Fall, daß eines oder mehrere der p_i verschwinden. In diesem Fall ist die rechte Seite nicht definiert, denn die Logarithmusfunktion hat an der Stelle Null einen Pol. Im Satz war vereinbart, daß wir für $p_i = 0$ den Term $p_i \log p_i$ als Null interpretieren; wenn wir zeigen können, daß dadurch die Funktion stetig auf Δ_n fortgesetzt wird, folgt Gleichheit auch in diesem Fall.

Offenbar genügt es, einen einzelnen Summanden zu betrachten; nach der Regel von DE L'HÔPITAL ist für den natürlichen Logarithmus

$$\lim_{p\searrow 0} p\log p = \lim_{p\searrow 0} \frac{\log p}{1/p} = \lim_{p\searrow 0} \frac{1/p}{-1/p^2} = \lim_{p\searrow 0}(-p) = 0 \,,$$

und da jeder andere Logarithmus proportional zum natürlichen ist, haben wir diesen Grenzwert auch für Logarithmen zu einer beliebigen Basis. Damit ist der Satz vollständig bewiesen. ∎

Damit sind wir allerdings noch nicht ganz fertig: Zwar wissen wir nun, daß jede Funktion, die SHANNONs drei Bedingungen genügt, die angegebene Form haben muß, wir wissen aber noch nicht, ob es überhaupt solche Funktionen gibt. Dazu müssen wir noch nachprüfen, daß die Funktionen

$$H(p_1,\dots,p_n) = -\sum_{i=1}^{n} p_i \log_a p_i$$

alle drei Bedingungen erfüllen.

Die Stetigkeit ist klar, da H nur durch Grundrechenarten und Logarithmen definiert ist. Auch mit der zweiten Bedingung gibt es keine Probleme, denn

$$L\Big(\frac{1}{n}\Big) = H\Big(\frac{1}{n},\dots,\frac{1}{n}\Big) = -\sum_{i=1}^{n} \frac{1}{n}\log\frac{1}{n} = \log n$$

ist eine monoton wachsende Funktion. Die dritte Bedingung schließlich ist erfüllt, denn für $m < n$ und $p^* = p_{m+1} + \dots + p_n$ ist

$$\begin{aligned} H(p_1,\dots,p_n) &= -\sum_{i=1}^{n} p_i \log_a p_i \\ &= -\sum_{i=1}^{m} p_i \log_a p_i - p^*\log_a p^* + p^*\log_a p^* - \sum_{i=m+1}^{n} p_i \log_a p_i \\ &= H(p_1,\dots,p_m,p^*) + \sum_{i=m+1}^{n} p_i(\log_a p^* - \log_a p_i) \\ &= H(p_1,\dots,p_m,p^*) - \sum_{i=m+1}^{n} p_i \log_a \frac{p_i}{p^*} \\ &= H(p_1,\dots,p_m,p^*) + p^* H\Big(\frac{p_{m+1}}{p^*},\dots,\frac{p_n}{p^*}\Big) \,. \end{aligned}$$

Damit haben wir für die Definition des Informationsgehalts nur noch die Freiheit, die Basis a des Logarithmus festzulegen; die traditionelle Wahl ist $a = 2$.

Definition: Die Entropie einer Quelle A mit einem m-buchstabigen Alphabet und Wahrscheinlichkeit p_i für das Auftreten des i-ten Buchstaben ist

$$H(A) = -\sum_{i=1}^{m} p_i \log_2 p_i \,.$$

Der Name *Entropie* ist ein Kunstwort, das der deutsche Physiker RUDOLF CLAUSIUS (1822–1888) in seiner Arbeit

> R. CLAUSIUS: Über verschiedene für die Anwendung bequeme Formen der Hauptgleichungen der Wärmetheorie, *Annalen der Physik und Chemie* **125** (1865), 353-400

einführte. Auf Seite 390 schreibt er:

> Sucht man für S einen bezeichnenden Namen, so könnte man . . . von der Größe S sagen, sie sey der *Verwandlungsinhalt* des Körpers. Da ich es aber für besser halte, die Namen derartiger für die Wissenschaft wichtiger Größen aus den alten Sprachen zu entnehmen, damit sie unverändert in allen neuen Sprachen angewandt werden können, so schlage ich vor, die Größe S nach dem griechischen Worte ἡ τροπή, die Verwandlung, die Entropie des Körpers zu nennen. Das Wort *Entropie* habe ich absichtlich dem Wort *Energie* möglichst ähnlich gebildet, denn die beiden Größen, welche durch diese Worte benannt werden sollen, sind ihren physikalischen Bedeutungen nach einander so nahe verwandt, daß eine gewisse Gleichartigkeit in der Benennung mir zweckmäßig zu seyn scheint.

Die Größe S, von der er hier spricht, hilft unter anderem bei der Erklärung, warum Wärme nie von einem kälteren zu einem wärmeren Körper fließen kann; wie LUDWIG BOLTZMANN (1844-1906) später gezeigt hat, kann sie auch mikroskopisch definiert werden durch eine Formel, die eng mit der hier zu definierenden SHANNONschen Entropie verwandt ist.

Als erstes Beispiel betrachten wir eine Zufallsvariable X, die alle Werte aus dem Alphabet A mit gleicher Wahrscheinlichkeit annimmt. Falls A aus n Buchstaben besteht, ist also $p(a) = 1/n$ für alle $a \in A$ und damit

$$H(X) = -\sum_{a\in A} p(a) = -\sum_{a\in A} \frac{1}{n}\log_2\frac{1}{n} = -\log_2\frac{1}{n} = \log_2 n\,.$$

Speziell im Fall einer Zweierpotenz $n = 2^r$ ist das gleich r und entspricht der Tatsache, daß man 2^r Objekte durch r Binärziffern eindeutig bezeichnen kann.

Im Falle eines Alphabets $A = \{a, b, c, d, e\}$ aus fünf Buchstaben und einer Zufallsvariablen Y, die diese mit Wahrscheinlichkeiten $p(a) = \frac{1}{2}$ und $p(b) = p(c) = p(d) = p(e) = \frac{1}{8}$ annimmt, ist

$$H(Y) = -\frac{1}{2}\log_2\frac{1}{2} - 4\cdot\frac{1}{8}\log\frac{1}{8} = \frac{1}{2} + \frac{4}{8}\cdot 3 = 2\,,$$

aber natürlich gibt es keine Möglichkeit, fünf Buchstaben mit nur zwei Binärziffern zu bezeichnen. Mit der Kodierung

$$a = 0,\quad b = 100,\quad c = 101,\quad d = 110\quad\text{und}\quad e = 111$$

kommen wir aber immerhin *im Durchschnitt* mit zwei Binärziffern aus, denn in der Hälfte aller Fälle haben wir a, wofür eine Ziffer ausreicht, und in der anderen Hälfte der Fälle brauchen wir drei Buchstaben, im Mittel also zwei. Für eine Zufallsvariable Z, die jedes Element von A mit Wahrscheinlichkeit $\frac{1}{5}$ annimmt, ist dagegen $H(Z) = \log_2 5 \approx 2{,}321928095$, und hier gibt es offensichtlich *keine* Kodierung, bei der wir im Durchschnitt $H(Z)$ Binärziffern brauchen, denn das arithmetische Mittel aus fünf natürlichen Zahlen muß ein Vielfaches von $\frac{1}{5}$ sein. Die nächstgrößere Zahl mit dieser Eigenschaft wäre 2,4, und das können wir tatsächlich erreichen, zum Beispiel mit der Kodierung

$$a = 00,\quad b = 01,\quad c = 10,\quad d = 110\quad\text{und}\quad e = 111\,.$$

SHANNONs Entropiebegriff steht also offensichtlich im Zusammenhang mit der mittleren Anzahl von Binärziffern, mit der wir die Buchstaben aus dem Alphabet kodieren können; wie genau dieser Zusammenhang aussieht, werden wir in Kürze untersuchen.

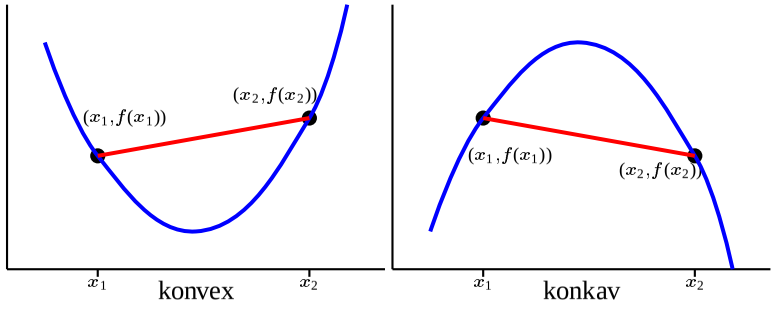
§ 2: Konvexität

SHANNONs drei Forderungen reichen zwar aus, um die Entropie (bis auf eine positive Konstante) eindeutig zu charakterisieren; der Begriff wäre aber nicht sonderlich nützlich, wenn wir nicht noch eine ganze Reihe weiterer Aussagen herleiten könnten. So erwarten wir beispielsweise, daß eine Quelle, die einen bestimmten ihrer Buchstaben mit einer sehr hohen Wahrscheinlichkeit produziert, einen kleineren mittleren Informationsgehalt hat als eine, bei der alle Buchstaben mit ungefähr gleicher Wahrscheinlichkeit vorkommen. Mit Aussagen dieser Art werden wir es auch noch in anderen Zusammenhängen zu tun haben; deshalb lohnt es sich, das Problem etwas allgemeiner anzugehen.

Definition: *a)* Eine *Teilmenge* $\Delta \subseteq \mathbb{R}^n$ heißt *konvex,* wenn zu je zwei Punkten $P, Q \in \Delta$ und jede reelle Zahl λ aus dem abgeschlossenen Intervall $[0, 1]$ auch der Punkt $(1-\lambda)P + \lambda Q$ in Δ liegt, wenn Δ also mit je zwei Punkten auch deren Verbindungsstrecke enthält.
b) Eine *Funktion* $f\colon \Delta \to \mathbb{R}$ auf einer konvexen Teilmenge $\Delta \subseteq \mathbb{R}^n$ heißt *konvex,* wenn für je zwei Punkte $P, Q \in \Delta$ und jedes $\lambda \in [0, 1]$ gilt:

$$f\big((1-\lambda)P + \lambda Q\big) \le (1-\lambda)f(P) + \lambda f(Q)\,,$$

wenn also der Graph von f über jeder Verbindungsstrecke zweier Punkte $P, Q \in \Delta$ unterhalb der Verbindungsstrecke der Punkte $\big(P, f(P)\big)$ und $\big(Q, f(Q)\big)$ liegt. Sie heißt *strikt konvex,* wenn dabei das Gleichheitszeichen nur für $\lambda = 0$ und $\lambda = 1$ gilt.
c) f heißt (strikt) *konkav,* wenn $-f$ (strikt) konvex ist.



Standardbeispiel einer konvexen Menge in $\mathbb{R}^n$ ist für uns die Menge

$$\Delta_n = \Big\{(p_1, \ldots, p_n) \;\Big|\; p_i \ge 0 \;\text{ für alle } i \text{ und } \sum_{i=1}^{n} p_i = 1\Big\}\,.$$

Sind $P = (p_1, \ldots, p_n)$ und $Q = (q_1, \ldots, q_n)$ zwei Punkte aus Δ_n, so ist

$$(1-\lambda)P + \lambda Q = \big((1-\lambda)p_1 + \lambda q_1, \ldots, (1-\lambda)p_n + \lambda q_n\big)\,.$$

Da alle p_i und q_i nichtnegativ sind, gilt für $\lambda \in [0, 1]$ dasselbe für die

Zahlen $(1-\lambda)p_i + \lambda q_i$, und

$$\sum_{i=1}^{n}\big((1-\lambda)p_i + \lambda q_i\big) = (1-\lambda)\sum_{i=1}^{n} p_i + \lambda \sum_{i=1}^{m} q_i = (1-\lambda) + \lambda = 1\,,$$

so daß auch $(1-\lambda)P + \lambda Q$ in Δ_n liegt.

Gerade bei der Definition einer konvexen Funktion erscheint es etwas seltsam, daß wir bei der Definition den Graphen nur über Strecken betrachten. Die Definition beschränkt sich auf diesen Fall, weil es sich um eine Eigenschaft handelt, die sich in vielen Fällen leicht nachprüfen läßt; tatsächlich gilt aber eine viel allgemeinere Aussage. Um sie auch für den Fall der strikten Konvexität zu formulieren, brauchen wir zunächst eine weitere Definition:

Definition: *a)* Eine Teilmenge $A \subset \mathbb{R}^n$ heißt r-dimensionaler affiner Unterraum von $\mathbb{R}^n$, wenn es einen Punkt $P_0 \in \mathbb{R}^n$ gibt, so daß die Verbindungsvektoren $\overrightarrow{P_0P}$ für die sämtlichen Punkte $P \in A$ einen r-dimensionalen Untervektorraum von $\mathbb{R}^n$ bilden.
b) m Punkte $P_1, \ldots, P_m \in \mathbb{R}^n$ sind *in allgemeiner Lage,* wenn es keinen $(m-2)$-dimensionalen affinen Unterraum $A \subseteq \mathbb{R}^n$ gibt, der alle diese Punkte enthält.

Zwei Punkte sind also genau dann in allgemeiner Lage, wenn sie verschieden sind; von dreien erwarten wir zusätzlich, daß sie nicht auf einer Geraden liegen, und von vieren, daß es keine Ebene gibt, die alle drei enthält.

Lemma: *a)* Eine Teilmenge $\Delta \subseteq \mathbb{R}^n$ ist genau dann konvex, wenn für jedes $m \in \mathbb{N}$ gilt: Sind $P_1, \ldots, P_m \in \Delta$ und ist $(\lambda_1, \ldots, \lambda_m) \in \Delta_m$, so liegt auch $\lambda_1 P_1 + \cdots + \lambda_m R_m$ in Δ.
b) Eine Funktion $f\colon \Delta \to \mathbb{R}$ auf einer konvexen Teilmenge $\Delta \subseteq \mathbb{R}^n$ ist genau dann konvex, wenn für je m Punkte $P_1, \ldots, P_m \in \Delta$ und jedes Tupel $(\lambda_1, \ldots, \lambda_m) \in \Delta_m$ gilt:

$$f(\lambda_1 P_1 + \cdots + \lambda_m P_m) \le \lambda_1 f(P_1) + \cdots + \lambda_m f(P_m)\,.$$

c) Eine Funktion $f\colon \Delta \to \mathbb{R}$ auf einer konvexen Teilmenge $\Delta \subseteq \mathbb{R}^n$ ist genau dann strikt konvex, wenn für je m Punkte $P_1, \ldots, P_m \in \Delta$ in

allgemeiner Lage und jedes Tupel $(\lambda_1, \ldots, \lambda_m) \in \Delta_m$ gilt:

$$f(\lambda_1 P_1 + \cdots + \lambda_m P_m) \leq \lambda_1 f(P_1) + \cdots + \lambda_m f(P_m)$$

mit Gleichheit genau dann, wenn ein $\lambda_i = 1$ ist und die übrigen λ_j verschwinden.

Beweis: Die Definition der Konvexität ist jeweils gerade der Fall $m = 2$ des Lemmas, die der strikten Konvexität die Fälle $m = 1$ und $m = 2$; zu zeigen ist also nur die Gegenrichtung. Der Fall $m = 1$ ist dabei in allen drei Fällen trivial; wirklich zu zeigen sind also nur die Fälle $m \geq 3$. Wir beweisen diese jeweils durch vollständige Induktion mit dem Fall $m = 2$ als Induktionsanfang.

a) Wir haben m Punkte $P_1, \ldots, P_m \in \Delta$ und ein Tupel $(\lambda_1, \ldots, \lambda_m)$ aus Δ_m. Für $\lambda_m = 1$ verschwinden alle übrigen λ_j, und die Behauptung ist trivial; wir können uns also beschränken auf den Fall $\lambda_m \neq 1$. Dann können wir durch $1 - \lambda_m$ dividieren und das $(m-1)$-tupel

$$(\lambda_1^*, \ldots, \lambda_{m-1}^*) = \left(\frac{\lambda_1}{1-\lambda_m}, \ldots, \frac{\lambda_{m-1}}{1-\lambda_m} \right) \in \Delta_{m-1}$$

betrachten. Nach Induktionsannahme liegt der Punkt

$$P^* = \lambda_1^* P_1 + \cdots + \lambda_{m-1}^* P_{m-1}$$

daher in Δ, und nach Definition der Konvexität gilt dasselbe für $(1-\lambda_m)P^* + \lambda_m P_m = \sum_{i=1}^m \lambda_i P_i$.

b) f sei konvex, $P_1, \ldots, P_m$ seien wieder Punkte aus Δ und $(\lambda_1, \ldots, \lambda_m)$ ein Tupel aus Δ_m, und P^* sei der oben definierte Punkt. Nach Induktionsannahme ist dann $f(P^*) \leq \lambda_1^* f(P_1) + \cdots + \lambda_{m-1}^* f(P_{m-1})$, und nach Definition der Konvexität von f ist außerdem

$$\begin{aligned} f\big((1-\lambda_m)P^* + \lambda_m P_m\big) &= f(\lambda_1 P_1 + \cdots + \lambda_m P_m) \\ \leq (1-\lambda_m) f(P^*) + \lambda_m f(P_m) &= \lambda_1 f(P_1) + \cdots + \lambda_m f(P_m)\,. \end{aligned}$$

c) Wir müssen nur noch zeigen, daß für Punkte in allgemeiner Lage Gleichheit nur gilt, wenn alle λ_i mit einer Ausnahme verschwinden und die Ausnahme damit gleich eins ist.

Falls $\lambda_m = 1$ ist, gibt es nichts mehr zu zeigen; wir können uns also auf den Fall $\lambda_m < 1$ beschränken. Dann können wir wie oben den Punkt P^* definieren.

Für Punkte $P_1, \ldots, P_m$ in allgemeiner Lage sind auch die beiden Punkte P^* und P_m in allgemeiner Lage, denn zwei Punkte sind genau dann in allgemeiner Lage, wenn sie verschieden sind, und wäre $P^* = P_m$, so wäre P_m eine Linearkombination von P_1 bis P_{m-1}, läge also im von diesen Punkten aufgespannten $(m-2)$-dimensionalen affinen Unterraum. Somit ist

$$\begin{aligned} f(\lambda_1 P_1 + \cdots + \lambda_m P_m) &= f\big((1-\lambda_m)P^* + \lambda_m P_m\big) \\ &= (1-\lambda_m) f(P^*) + \lambda_m f(P_n) \end{aligned}$$

genau dann, wenn $\lambda_m = 0$ oder $\lambda_m = 1$ ist. Den Fall $\lambda_m = 1$ haben wir bereits ausgeschlossen; also ist $\lambda_m = 0$. Dann aber folgt die Behauptung sofort aus der Induktionsannahme. ∎

Die Aussage unter *b)* wird auch als *Ungleichung von* JENSEN bezeichnet; er bewies sie in einem Vortrag vom 17. Januar 1905 vor der dänischen Mathematikergesellschaft, wobei er allerdings nur voraussetzte, daß die Funktion f auf einem reellen Intervall definiert ist und dort die Ungleichung $f(x) + f(y) \geq 2f\left(\frac{x+y}{2}\right)$ erfüllt, was auf den ersten Blick etwas schwächer aussieht als die hier betrachtete Definition der Konvexität, nach dem Resultat von JENSEN aber äquivalent dazu ist.

Der dänische Mathematiker Johan Ludwig William Valdemar Jensen (1859–1925) studierte ab 1876 an der Københavns Tekniske Skole unter anderem Mathematik, Physik und Chemie. Sein Interesse konzentrierte sich immer mehr auf die Mathematik; zwischen 1879 und 1925 veröffentlichte er rund vierzig wissenschaftliche Arbeiten. Er war allerdings nie an einer Universität tätig und war auch von der Ausbildung her im wesentlichen Autodidakt. Sein gesamtes Berufsleben arbeitete er als Telephoningenieur bei der dänischen Telephongesellschaft. Außer der heute nach ihm benannten Ungleichung bewies er unter anderem auch Sätze im Umkreis der RIEMANN-Vermutung.

Wenn wir die λ_i als Wahrscheinlichkeiten interpretieren, können wir *b)* und *c)* auch als Aussagen über Erwartungswerte interpretieren:

Lemma: Für eine diskrete Zufallsvariable X und eine $\left\{\begin{matrix}\text{konvexe}\\ \text{konkave}\end{matrix}\right\}$ Funktion f auf dem Wertebereich von X gilt: $\mathbb{E}\big(f(X)\big)\ \left\{\begin{matrix}\geq\\ \leq\end{matrix}\right\}\ f\big(\mathbb{E}(X)\big)$.
Im Falle einer strikt $\left\{\begin{matrix}\text{konvexen}\\ \text{konkaven}\end{matrix}\right\}$ Funktion gilt Gleichheit genau dann, wenn X einen seiner Werte mit Wahrscheinlichkeit eins annimmt. ∎

Für mindestens zweimal stetig differenzierbare Funktionen läßt sich die Konvexität leicht anhand der zweiten Ableitung überprüfen. Im Fall einer Variablen haben wir einfach das

Lemma: *a)* Eine mindestens zweimal stetig differenzierbare Funktion $f\colon I \to \mathbb{R}$ auf einem Intervall $I \subseteq \mathbb{R}$ ist genau dann konvex, wenn ihre zweite Ableitung auf I keine negativen Werte annimmt; sie ist genau dann konkav, wenn f'' auf I keine positiven Werte annimmt.
b) Falls f'' im Innern von I nur positive Werte annimmt, ist f strikt konvex auf I; falls f'' dort nur negative Werte annimmt, ist f strikt konkav.

Beweis: a) Wir zeigen zunächst, daß im Falle der Konvexität die zweite Ableitung in ganz $(a,\, b)$ größer oder gleich null sein muß: Andernfalls gäbe es ein $x_0 \in (a,\, b)$ mit $f''(x_0) < 0$. Wir betrachten die Funktion $g(x) \underset{\text{def}}{=} f(x) - f'(x_0)(x - x_0)$. Als Summe von f und einer linearen Funktion ist g zweimal differenzierbar mit

$$g'(x_0) = f'(x_0) - f'(x_0) = 0 \quad \text{und} \quad g''(x_0) = f''(x_0) < 0\,.$$

Die Funktion $g'(x)$ ist also in einem hinreichend kleinen Intervall $(x_0 - h,\, x_0 + h)$ streng monoton fallend; sie ist daher positiv für $x < x_0$ und negativ für $x > x_0$. Somit ist g streng monoton wachsend für $x < x_0$ und streng monoton fallend für $x > x_0$; die Funktion g hat also bei x_0 ein lokales Maximum. Für ein $\varepsilon < h$ ist daher $g(x_0 \pm \varepsilon) < g(x_0)$ und damit ist auch

$$f(x_0) = g(x_0) > \frac{1}{2}g(x_0 - \varepsilon) + \frac{1}{2}g(x_0 + \varepsilon) = \frac{1}{2}f(x_0 - \varepsilon) + \frac{1}{2}f(x_0 + \varepsilon)\,.$$

Dies widerspricht aber der Konvexitätsbedingung für $x_{1/2} = x_0 \pm \varepsilon$ und $\lambda = \frac{1}{2}$. Somit muß $f''(x)$ in ganz $(a,\, b)$ größer oder gleich null sein.

Umgekehrt sei $f''(x) \geq 0$ für alle $x \in (a,\, b)$; wir müssen zeigen, daß f dann konvex ist. Seien also $x_1 < x_2$ zwei beliebige Punkte aus $(a,\, b)$ und $x = (1 - \lambda)x_1 + \lambda x_2$ mit $\lambda \in (0,\, 1)$. Nach dem Mittelwertsatz gibt es Punkte $\xi_1 \in (x_1, x)$ und $\xi_2 \in (x,\, x_2)$, so daß

$$f'(\xi_1) = \frac{f(x) - f(x_1)}{x - x_1} = \frac{f(x) - f(x_1)}{\lambda(x_2 - x_1)} \quad \text{und}$$
$$f'(\xi_2) = \frac{f(x_2) - f(x)}{x_2 - x} = \frac{f(x_2) - f(x)}{(1 - \lambda)(x_2 - x_1)}$$

ist. Da f'' nirgends negativ wird, ist f' monoton wachsend und damit insbesondere $f'(\xi_1) \leq f'(\xi_2)$. Diese Ungleichung können wir auch schreiben als

$$\frac{f(x) - f(x_1)}{\lambda(x_2 - x_1)} \leq \frac{f(x_2) - f(x)}{(1 - \lambda)(x_2 - x_1)}\,,$$

und da $x_1 < x_2$ ist, folgt daraus

$$\frac{f(x) - f(x_1)}{\lambda} \leq \frac{f(x_2) - f(x)}{1 - \lambda}\,.$$

Für $\lambda \in (0,\, 1)$ ändert sich nichts an dieser Ungleichung, wenn wir mit $\lambda(1 - \lambda)$ multiplizieren; dies führt auf

$$(1 - \lambda)f(x) - (1 - \lambda)f(x_1) \leq \lambda f(x_2) - \lambda f(x)$$

und damit die gewünschte Ungleichung

$$f(x) \leq (1 - \lambda)f(x_1) + \lambda f(x_2)\,,$$

die die Konvexität von f ausdrückt. Damit ist die Behauptung für konvexe Funktionen bewiesen.

Für konkave Funktionen folgt sie einfach daraus, daß $-f$ für eine konkave Funktion f konvex ist. ∎

Korollar: Die Funktion $f(x) = -x \log x$ ist über dem Intervall $[0,\, 1]$ konkav.

Beweis: $f'(x) = -x \cdot \frac{1}{x} - \log x = -1 - \log x$ hat als Ableitung die Funktion $f''(x) = -1/x$, die im Intervallinnern überall negativ ist. ∎

Damit gilt insbesondere für zwei beliebige Zahlen $p_1, p_2 \in [0,\ 1]$, daß

$$-\frac{p_1+p_2}{2}\log_2\frac{p_1+p_2}{2} \geq \frac{1}{2}(-p_1\log_2 p_1) + \frac{1}{2}(-p_2\log_2 p_2)$$

oder

$$-2\cdot\frac{p_1+p_2}{2}\log_2\frac{p_1+p_2}{2} \geq -p_1\log_2 p_1 - p_2\log_2 p_2\,.$$

Ersetzt man also im Ausdruck

$$H(p_1,\ldots,p_n) = -\sum_{i=1}^{m} p_i\log_2 p_i$$

irgendwelche zwei *verschiedene* Wahrscheinlichkeiten p_i und p_j durch ihren gemeinsamen Mittelwert $\frac{1}{2}(p_i+p_j)$, so wird die Entropie größer. Damit folgt fast sofort

Satz: Für m Zahlen $p_1,\ldots,p_m \in [0,1]$ mit $\sum_{i=1}^{m} p_i = 1$ gilt stets

$$0 \leq H(p_1,\ldots,p_m) = -\sum_{i=1}^{m} p_i\log_a p_i \leq \log m\,;$$

dabei steht rechts genau dann ein Gleichheitszeichen, wenn alle p_i gleich $1/m$ sind und links steht genau dann eines, wenn alle p_i mit einer Ausnahme verschwinden.

Beweis: Da H eine stetige Funktion auf der kompakten Menge Δ_m ist, nimmt sie sowohl ihr Maximum als auch ihr Minimum an. Wie wir gerade gesehen haben, kann es im Maximum keine zwei echt verschiedenen p_i geben, also müssen alle $p_i = \frac{1}{m}$ sein und das Maximum ist

$$H\left(\tfrac{1}{m},\ldots,\tfrac{1}{m}\right) = m\cdot\left(-\tfrac{1}{m}\log_2\tfrac{1}{m}\right) = -\log_2\tfrac{1}{m} = \log_2 m\,.$$

Umgekehrt ist $-p_i\log_2 p_i \geq 0$ für alle $p_i \in [0,\ 1]$ mit Gleichheit genau dann, wenn $p_i = 0$ oder $p_i = 1$ ist. Eine Summe Null entsteht somit genau dann, wenn alle $p_i \in \{0,1\}$ sind, d.h. wenn genau ein $p_i = 1$ ist und der Rest verschwindet. ∎

§3: Ein Beispiel

Um ein Gefühl für den SHANNONschen Informationsbegriff zu bekommen, wollen wir ein bekanntes Ratespiel informationstheoretisch betrachtet: Gegeben sind zwölf gleich aussehende Kugeln, von denen mindestens elf dasselbe Gewicht haben, sowie eine Balkenwaage. Man finde mit höchstens dreimaligem Wiegen heraus, ob es eine Kugel mit abweichendem Gewicht gibt, welche dies ist und ob sie leichter oder schwerer als der Rest ist.

Auf diese Frage gibt es 25 mögliche Antworten, die wir auf Grund unseres Informationsstands als gleich wahrscheinlich betrachten müssen; die korrekte Antwort hat somit einen Informationsgehalt von $\log_2 25$ Bit. Beim Wiegen erhalten wir eines von drei möglichen Ergebnissen (linke Seite schwerer, rechte Seite schwerer, beide Seiten gleich schwer); falls es uns gelingt, die zu vergleichenden Kugeln so auszuwählen, daß alle drei Ergebnisse gleich wahrscheinlich sind, bekommen wir eine Information von $\log_2 3$ Bit pro Wiegen. Bei dreimaligem Wiegen wären das $3\cdot\log_2 3 = \log_2 3^3 = \log_2 27$ Bit, was mehr ist als $\log_2 25$ Bit. Von daher spricht also nichts gegen die Lösbarkeit der Aufgabe, allerdings haben wir auch nicht viel Spielraum und müssen daher bei jedem Wiegen unbedingt darauf achten, daß die drei möglichen Resultate mit zumindest ungefähr gleicher Wahrscheinlichkeit auftreten.

Damit verbietet sich insbesondere der naheliegende Ansatz, zunächst zwei Sechsergruppen von Kugeln miteinander zu vergleichen: Da die Waage nur im Fall, daß alle Kugeln das gleiche Gewicht haben. im Gleichgewicht ist, tritt hier einer der drei Fälle nur mit einer Wahrscheinlichkeit von $1/25$ auf, die beiden anderen jeweils mit $8/25$, so daß wir nur eine Information von

$$-\frac{1}{25}\log_2\frac{1}{25} - \frac{16}{25}\log_2\frac{8}{25} \approx 1{,}202$$

Bit bekommen, was zu weit unter $\log_2 3 \approx 1{,}585$ liegt.

Stattdessen sollten wir mit Vierergruppen arbeiten: Wir numerieren die Kugeln von 1 bis 12 und vergleichen die Kugeln 1 bis 4 mit 5 bis 8. In neun der 25 Fälle erhalten wir das Ergebnis *gleich schwer,* nämlich

genau dann, wenn die zu leichte oder zu schwere Kugel unter denen mit Nummer 9 bis 12 zu finden ist oder aber alle Kugeln gleich schwer sind. Die rechte Seite mit den Kugeln 1 bis 4 ist genau dann schwerer, wenn entweder eine dieser vier Kugeln schwerer ist als die anderen oder wenn eine der Kugeln 5 bis 9 leichter ist als die anderen, also in jeweils acht Fällen. In den verbleibenden acht Fällen ist linke Seite schwerer; wir haben also drei Ergebnisse mit Wahrscheinlichkeiten $9/25$ und zweimal $8/25$; unsere Information ist

$$-\frac{9}{25}\log_2\frac{9}{25}-\frac{16}{25}\log_2\frac{8}{25}\approx 1{,}583\,,$$

was nur sehr knapp unter der maximal möglichen Information von $\log_2 3$ Bit liegt, die wir hier natürlich nicht erreichen können, da 25 nicht durch drei teilbar ist.

Wenn beide Seiten gleich schwer waren, wissen wir nicht nur, daß die gesuchte Kugel, so sie existiert, eine Nummer zwischen neun und zwölf hat, sondern wir wissen auch, daß die Kugeln eins bis acht allesamt das „übliche“ Gewicht haben. Wir haben somit „Referenzkugeln“, mit denen wir entscheiden können, ob eine gegebene Kugel leichter oder schwerer ist als der Rest.

Insgesamt haben wir neun mögliche Fälle (alle Kugeln gleich schwer, eine der Kugeln neun bis zwölf leichter *bzw.* schwerer); wir sollten so wiegen, daß jedes der drei möglichen Ergebnisse in drei der neun Fälle eintritt.

Dazu können wir beispielsweise die zwölfte Kugel auszeichnen und als eine Gruppe von drei Fällen den nehmen, daß entweder alle Kugeln gleich schwer sind oder aber die zwölfte das falsche Gewicht hat. In diesen drei Fällen haben also die Kugeln neun bis elf das richtige Gewicht.

Dies können wir entscheiden, in dem wir sie mit drei Referenzkugeln vergleichen, etwa den Kugeln eins bis drei; in den drei betrachteten Fällen sind beide Seiten der Waage gleich schwer.

Falls die Kugeln eins bis drei schwerer sind als neun bis elf, ist eine der letzteren leichter als der Rest, wofür es drei Fälle gibt; in den verbleibenden drei Fällen, wenn eine der Kugeln neun bis elf schwerer ist als

der Rest, sind auch die drei Kugeln zusammen schwerer als eins bis drei. Hier erhalten wir also die maximal mögliche Information von $\log_2 3$ Bit.

Falls die Waage im Gleichgewicht war, ist klar, wie wir weiter wiegen: Wir vergleichen Kugel zwölf mit irgendeiner anderen Kugel und erfahren, ob sie schwerer, leichter oder gleich schwer wie die anderen Kugeln ist; in diesem Fall liefert uns also auch das dritte Wiegen eine Information von $\log_2 3$ Bit.

In den beiden anderen Fällen wissen wir entweder, daß eine der drei Kugeln neun bis elf leichter ist als der Rest oder daß sie schwerer ist; wir müssen nur noch herausfinden, um welche der drei Kugeln es sich handelt. Dazu können wir beispielsweise die Kugeln neun und zehn miteinander vergleichen: Sind sie gleich schwer, so hat elf das abweichende Gewicht, andernfalls ist es im ersten Fall die leichtere, im zweiten die schwerere der beiden Kugeln. Hier erhalten wir also beim Wiegen wieder die maximal mögliche Information von $\log_2 3$ Bit.

Damit sind alle Fälle abgehandelt, bei denen die Waage beim ersten Einsatz ausbalanciert war; bleiben noch die, daß eine der beiden Seiten schwerer war.

Angenommen, die Kugeln von eins bis vier sind schwerer als die von fünf bis acht. Dann ist entweder eine der Kugeln eins bis vier zu schwer ist oder eine der Kugeln fünf bis acht zu leicht. Da wir in diesem Fall acht gleich wahrscheinliche Möglichkeiten haben und acht nicht durch drei teilbar ist, können wir beim Wiegen keine Information von $\log_2 3$ Bit bekommen; am meisten Information erhalten wir, wenn zwei der möglichen Ergebnisse in jeweils drei Fällen auftreten und das dritte in zweien. Ein solches Experiment, falls wir es realisieren können, liefert eine Information von

$$\begin{aligned}-2\cdot\frac{3}{8}\log_2\frac{3}{8}-\frac{1}{4}\log_2\frac{1}{4} &= -\frac{3}{4}(\log_2 3-\log_2 8)+\frac{1}{4}\log_2 4\\ &= -\frac{3}{4}\log_2 3+\frac{9}{4}+\frac{2}{4}=\frac{11}{4}-\frac{3}{4}\log_2 3\approx 1{,}561\end{aligned}$$

Bit, was etwas kleiner ist als $\log_2 3\approx 1{,}585$.

Im Gegensatz zur obigen Situation gibt es hier für jede Kugel nur noch zwei Möglichkeiten: Wenn ihr Gewicht von dem der restlichen Kugeln

abweicht, ist es im Falle der Kugeln eins bis vier notwendigerweise zu schwer, bei den vier anderen notwendigerweise zu leicht. Wir können deshalb keine Gruppe aus zwei Fällen konstruieren, indem wir nur *eine* Kugel betrachten; wir brauchen mindestens zwei.

Versuchen wir also, die beiden Fälle *Kugel eins zu schwer* und *Kugel zwei zu schwer* zu einer Fallgruppe zusammenzufassen. Die restlichen sechs Fälle müssen wir zu zwei Dreiergruppen zusammenfassen; aus Symmetriegründen sollte jede von diesen einen der beiden Fälle *Kugel drei zu schwer* und *Kugel vier zu schwer* enthalten sowie zwei Fälle mit zu leichten Kugeln.

Damit ist klar, wie wir weiter vorgehen können: Wir legen beispielsweise die Kugeln drei, fünf, sechs in die linke und vier, sieben, acht in die rechte Waagschale. Die linke Waagschale geht nach unten, wenn drei schwerer oder fünf oder sechs leichter ist, die rechte, wenn vier schwerer oder sieben oder acht leichter ist. In den verbleibenden Fällen, daß eins oder zwei schwerer ist, halten sich beide Seiten die Waage.

In diesem Fall müssen wir nur noch eins und zwei vergleichen; die schwerere der beiden Kugeln ist die abweichende, und sie ist schwerer als der Rest. Der Informationsgehalts dieses Vergleichs ist somit nur ein Bit.

In den anderen Fällen vergleichen wir jeweils die beiden möglicherweise zu leichten Kugeln. Ist eine davon tatsächlich leichter als die andere, ist sie die Lösung; andernfalls haben beide dasselbe Gewicht und die potentiell schwerere Kugel ist wirklich schwerer als der Rest. Hier bekommen wir also wieder $\log_2 3$ Bit Information.

Bleibt noch der Fall, daß die Kugeln von eins bis vier *leichter* sind als die von fünf bis acht; hier können wir natürlich vorgehen wie eben, nur daß die Begriffe *leichter* und *schwerer* miteinander vertauscht werden müssen.

Beim ersten Wiegen erhalten wir somit eine Information von

$$-\frac{16}{25}\log_2\frac{8}{25}-\frac{9}{25}\log_2\frac{9}{25} = \frac{-16}{25}(\log_2 8-\log_2 25)-\frac{9}{25}(\log_2 9-\log_2 25)$$
$$= \log_2 25 - \frac{48}{25} - \frac{18}{25}\log_2 3$$

Bit. In den 9/25 aller Fälle, in denen die Waage im Gleichgewicht ist, konnten wir beim zweiten und dritten Wiegen jeweils die maximal mögliche Information von $\log_2 3$ Bit realisieren, insgesamt also $2\log_2 3$ Bit. In den übrigen Fällen erhalten wir beim zweiten Wiegen nur eine Information von $\frac{11}{4} - \frac{3}{4}\log_2 3$ Bit und beim dritten erhalten wir in einem Viertel der Fälle nur ein Bit, ansonsten $\log_2 3$ Bit. Im Mittel bekommen wir somit genau die benötigte Information von

$$\left(\log_2 25 - \frac{48}{25} - \frac{18}{25}\log_2 3\right)$$
$$+\frac{9}{25}\cdot 2\log_2 3 + \frac{16}{25}\left(\frac{11}{4} - \frac{3}{4}\log_2 3 + \frac{1}{4} + \frac{3}{4}\log_2 3\right)$$
$$= \log_2 25 - \frac{48}{25} + \frac{16}{25}\cdot\frac{12}{4} = \log_2 25 \text{ Bit}.$$

Bei n Kugeln, von denen genau eine entweder schwerer oder leichter als die übrigen ist, haben wir offensichtlich keine Chance, das Problem mit r-maligem Wiegen zu lösen, wenn $\log_2(2n+1) > r\log_2 3$ ist oder, äquivalent, $2n+1 > 3^r$; schließlich können wir beim Wiegen nie eine größere Information als $\log_2 3$ Bit erhalten, und zumindest in einigen Fällen erhalten wir zwangsläufig weniger Information. Man kann sich fragen, ob wir im Falle $\log_2(2n+1) \le r\log_2 3$ *immer* eine Strategie finden können, bei der wir mit r maligem Wiegen auskommen. In diesem Fall sollte es also insbesondere möglich sein, mit dreimaligem Wiegen nicht nur das Problem mit zwölf Kugeln zu lösen, sondern sogar das mit dreizehn.

Hier gibt es 27 Möglichkeiten; um beim ersten Wiegen die maximal mögliche Information zu bekommen, sollten wir ein Experiment durchführen, bei dem jede der drei Alternativen genau neun Mal eintritt. Beim ersten Wiegen gibt es aber im wesentlichen nur einen Parameter, den wir beeinflussen können: Wir wählen irgendeine Zahl $m \le 6$ und legen in beide Waagschalen jeweils m Kugeln. In je $2m$ Fällen geht dann die linke oder rechte Waagschale nach unten; in den verbleibenden $27-4m$ Fällen sind sie ausbalanciert. Da sich neun nicht in der Form $9 = 2m$ schreiben läßt, können wir somit schon beim ersten Wiegen nicht die erforderliche Information von $\log_2 3$ Bit erreichen.

§4: Asymptotische Gleichverteilung

Wie wir am Ende des zweiten Paragraphen gesehen haben, kann die Entropie einer Quelle gelegentlich interpretiert werden als die mittlere Anzahl von Bit, die wir zur Kodierung eines Buchstabens benötigen. Dies funktioniert aber nicht immer: Bei einer Quelle, die drei Buchstaben mit gleicher Wahrscheinlichkeit produziert, kann es natürlich keine Kodierung geben, bei der wir im Durchschnitt genau $\log_2 3$ Bit brauchen – der mittlere Aufwand läßt sich bei jeder Kodierung als Bruch mit Nenner drei darstellen. Unsere beste Wahl besteht darin, daß wir einen der Buchstaben etwa als die Null darstellen und die beiden anderen als 10 und 11; der mittlere Aufwand beträgt dann $5/3$ Bit.

Wenn wir je zwei Buchstaben zu einer Gruppe zusammenfassen, haben wir neun Paare, die jeweils mit Wahrscheinlichkeit $1/9$ auftreten – falls wir annehmen, daß unsere Quelle Buchstaben jeweils unabhängig vom Vorgänger produziert. Kodieren wir die ersten vier Paare durch $000, 001, 010$ und 011, so können wir die restlichen fünf beispielsweise darstellen als $1000, 1001, 1010, 1011$ und 1100; der mittlere Aufwand ist also gesunken auf

$$\frac{4}{9} \times 3 + \frac{5}{9} \times 4 = \frac{32}{9} \text{ Bit}$$

pro Paar oder $16/9$ Bit pro Buchstabe. Bei Blöcken von fünf Buchstaben haben wir $3^5 = 243$ verschiedene Blöcke; indem wir einfach die Zahlen von 0 bis 242 im Zweiersystem darstellen, kommen wir also mit acht Bit aus (tatsächlich sogar geringfügig weniger, da wir noch ein paar 7-Bit-Kodierungen vergeben können) und sind damit bei knapp 1,6 Bit pro Buchstabe angelangt, was bereits recht nahe bei $\log_2 3 \approx 1{,}585$ angelangt.

Wir erwarten, daß wir durch Übergang zu immer größeren Blöcken dem Wert $\log_2 3$ immer näher kommen, auch wenn wir ihn zumindest in diesem Beispiel nie erreichen können: Wie man zeigen kann, ist $\log_2 3$ eine transzendente, insbesondere also irrationale Zahl, und der Aufwand pro Buchstabe ist bei dieser Quelle unabhängig von der Kodierung stets eine rationale Zahl.

Wir müssen uns daher begnügen mit einer Näherungsaussage.

Betrachten wir als Beispiel eine Folge voneinander unabhängiger Zufallsvariablen $X_1, X_2, \ldots$ mit einem Alphabet $\{0, 1\}$ aus zwei Buchstaben. Jede Variable X_i möge mit Wahrscheinlichkeit p eine Eins liefern und dementsprechend mit Wahrscheinlichkeit $q = 1 - p$ eine Null. Die Entropie ist dann jeweils $H(X_i) = -p \log_2 p - q \log_2 q$.

Das n-tupel $(X_1, \ldots, X_n)$ produziert Blöcke aus n Binärziffern; Erwartungswert für die Anzahl der Einsen in so einem Block ist pn. Um eine grobe Näherung für die Wahrscheinlichkeit eines *typischen* Blocks zu bekommen, nehmen wir erstens an, pn sei eine ganze Zahl, und zweitens, daß in einem typischen Block genau pn Einsen auftreten. Die Wahrscheinlichkeit eines solchen typischen Blocks ist dann

$$p^{np} q^{n-np} = p^{np} q^{nq} = 2^{np \log_2 p + nq \log_2 q} = 2^{-H(X_i)} .$$

Natürlich gibt es (außer im Fall $p = \frac{1}{2}$) auch Blöcke, deren Wahrscheinlichkeit deutlich von diesem Wert abweicht, zum Beispiel die beiden, die aus lauter Nullen *bzw.* lauter Einsen bestehen, aber nach dem Gesetz der großen Zahl sollte für hinreichend große n die Wahrscheinlichkeit einer größeren Abweichung von diesem Wert sehr gering sein.

Dieser Philosophie entsprechend definieren wir nun für Zufallsvariablen mit beliebiger Verteilung eine *typische Menge* und untersuchen deren Eigenschaften:

Definition: $X_1, X_2, \ldots$ sei eine Folge voneinander unabhängiger Zufallsvariablen mit Werten in einem Alphabet A, die allesamt die gleiche Wahrscheinlichkeitsverteilung haben. Für ein Tupel $(a_1, \ldots, a_n) \in A^n$ bezeichne

$$p(a_1, \ldots, a_n) = \prod_{i=1}^{n} p(a_i)$$

die Wahrscheinlichkeit dafür, daß für $i = 1, \ldots, n$ die Zufallsvariable X_i den Wert a_i annehme. Bezeichnet

$$H = -\sum_{x \in A} p(x) \log_2 p(x)$$

die Entropie der Zufallsvariablen X_i, definieren wir jedes $\varepsilon > 0$ eine *typische Menge*

$$A_\varepsilon^n = \left\{(a_1, \ldots, a_n) \in A^n \;\middle|\; 2^{-n(H+\varepsilon)} \le p(x_1, \ldots, a_n) \le 2^{-n(H-\varepsilon)}\right\}.$$

Einige wichtige Eigenschaften dieser Menge sind im folgenden Satz zusammengefaßt:

Satz: $X_1, X_2, \ldots$ sei eine Folge voneinander unabhängiger identisch verteilter Zufallsvariablen mit Werten in einem Alphabet A; ihre Entropie sei H. Dann gilt:

a) Für alle $(a_1, \ldots, a_n) \in A_\varepsilon^n$ ist $\left|-\frac{1}{n}p(a_1, \ldots, a_n) - H\right| < \varepsilon$.

b) Für hinreichend große Werte von n ist die Wahrscheinlichkeit dafür, daß ein Element von A^n in A_ε^n liegt, größer als $1 - \varepsilon$.

c) Die Kardinalität $\#A_\varepsilon^n$ von A_ε^n ist höchstens gleich $2^{n(H+\varepsilon)}$.

d) Für hinreichend große n ist $\#A_\varepsilon^n \ge (1-\varepsilon)2^{n(H+\varepsilon)}$.

Beweis: a) folgt sofort aus der Definition der typischen Menge, wenn wir in der definierenden Ungleichung zu Logarithmen übergehen und durch n dividieren.

Für *b)* erinnern wir uns an das (schwache) Gesetz der großen Zahlen: Danach konvergieren die Mittelwerte $\frac{1}{n}(Y_1 + \cdots + Y_n)$ einer Folge voneinander unabhängiger aber identisch verteilter reellwertiger Zufallsvariablen für $n \to \infty$ stochastisch gegen den gemeinsamen Erwartungswert der Y_i. Die Zufallsvariablen Y_i definieren wir für diesen Beweis wie folgt: Wenn X_i den Wert $a \in A$ liefert, soll Y_i den Wert $-\log_2 p(a)$ liefern. Der gemeinsame Erwartungswert der Y_i ist dann die Entropie

$$H = -\sum_{a \in A} p(a) \log_2 p(a)$$

der X_i; es gibt daher zu jedem $\delta > 0$ ein $N \in \mathbb{N}$, so daß die Wahrscheinlichkeit des Ereignisses $\left|-\frac{1}{n}p(a_1, \ldots, a_n) - H\right| < \varepsilon$ größer ist als $1 - \delta$ für alle $n \ge N$. Speziell können wir auch für $\delta = \varepsilon$ so ein N finden, womit *b)* bewiesen wäre.

c) folgt durch eine einfache Abschätzung: Da

$$1 = \sum_{(a_1,\ldots,a_n)\in A^n} p(a_1, \ldots, a_n) \ge \sum_{(a_1,\ldots,a_n)\in A_\varepsilon^n} p(a_1, \ldots, a_n)$$
$$\ge \sum_{(a_1,\ldots,a_n)\in A_\varepsilon^n} 2^{-n(H+\varepsilon)} = 2^{-n(H+\varepsilon)} \#A_\varepsilon^n$$

ist, muß $\#A_\varepsilon^n \le 2^{n(H+\varepsilon)}$ sein.

Zum Beweis von *d)* schließlich schätzen wir in umgekehrter Richtung ab, ausgehend von Aussage *b):* Für alle hinreichend großen n ist

$$1 - \varepsilon < \sum_{(a_1,\ldots,a_n)\in A_\varepsilon^n} p(a_1, \ldots, a_n) \le \sum_{(a_1,\ldots,a_n)\in A_\varepsilon^n} 2^{-n(H-\varepsilon)} = 2^{-n(H-\varepsilon)} \#A_\varepsilon^n$$

und damit $\#A_\varepsilon^n \ge (1-\varepsilon)2^{n(H-\varepsilon)}$. ∎

Als erste Anwendung können wir abschätzen, wie viele Bit wir brauchen, wenn wir in der vorliegenden Situation blockweise kodieren: Wenn wir zu vorgegebenem $\varepsilon > 0$ die Blocklänge n so groß wählen, daß *b)* erfüllt ist, haben wir einerseits höchstens $2^{n(H+\varepsilon)}$ Elemente in A_ε^n und kommen daher mit $n(H+\varepsilon)$ Bit pro Block aus, wenn wir nur diese Blöcke kodieren. Die Wahrscheinlichkeit dafür, daß einer der restlichen Blöcke auftritt, ist kleiner als ε; auch wenn wir für die Kodierung solcher Blöcke erheblich längere Bitfolgen ansetzen müssen, beispielsweise solche der Länge $n \log_2 m$, wobei m die Elementanzahl des Alphabets bezeichnet, wird diese Länge mit ε multipliziert, so daß wir im Mittel nur $n(H+\varepsilon) + \varepsilon \cdot n \log_2 m = n\big(H + \varepsilon(1 + \log_2 m)\big)$ brauchen, also $H + \varepsilon(1 + \log_2 m)$ pro Zeichen. Mit hinreichend großen Blocklängen kommen wir somit beim mittleren Kodierungsaufwand in der Tat beliebig nahe an die Entropie heran.

§5: Die Entropierate stochastischer Prozesse

a) Stochastische Prozesse

Im vorigen Paragraphen waren wir ausgegangen von einer Folge unabhängiger Zufallsvariablen. Wenn wir eine Quelle modellieren wollen,

die beispielsweise deutsche Texte produziert, ist das sicherlich eine unrealistische Annahme: E ist der häufigste Buchstabe des Alphabets, aber hinter einem E kommt fast nie ein weiteres E, und auch hinter einem C kommt nur selten ein E; viel wahrscheinlicher sind hier die (ansonsten eher nicht so häufigen) Buchstaben H und K.

Um zu realistischeren Modellen zu kommen, müssen wir daher auch Abhängigkeiten zwischen den Wahrscheinlichkeitsverteilungen der einzelnen Zufallsvariablen zulassen und damit auch allgemeinere stochastische Prozesse zulassen.

Unter einem stochastischen Prozess wollen wir dabei einfach eine Folge $X_1, X_2, \ldots$ von Zufallsvariablen verstehen; wir beschränken uns also weiterhin auf diskrete Zeit, und wir setzen auch weiterhin voraus, daß alle X_i Werte aus einem festen endlichen Alphabet A annehmen.

Bei Prozessen, die natürliche Sprachen beschreiben, sollte die Wahrscheinlichkeit, mit der ein gegebenes Wort vorkommt, nicht davon abhängen, ob wir den Anfang oder das Ende des Textes betrachten; solche Prozesse bezeichnen wir als *stationär:*

Definition: Ein stochastischer Prozess $(X_k)_{k\in\mathbb{N}}$ heißt *stationär,* wenn für alle $n \in \mathbb{N}$, alle $(x_1, \ldots, x_n) \in A^n$ und alle $m \in \mathbb{N}$ gilt: Die Wahrscheinlichkeit des Ereignisses $X_{m+1} = x_1, X_{m+2} = x_2, \ldots, X_{m+n} = x_n$ ist gleich der des Ereignisses $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$.

Stochastische Prozesse, die reale Phänomene beschreiben, haben oft sehr komplizierte Wahrscheinlichkeitsverteilungen; insbesondere wird der Wert von X_n oftmals von vielen, wenn nicht gar allen Werten der Vorgänger abhängen. Als einfache Idealisierung, die immerhin noch etwas realistischer ist als eine Folge unabhängiger Zufallsvariablen, sind MARKOV-Ketten ein beliebtes Modell. Hierbei handelt es sich um stochastische Prozesse ohne Gedächtnis, d.h. die Wahrscheinlichkeit, mit der die Zufallsvariable X_n einen Wert produziert, hängt nur ab vom Wert des unmittelbaren Vorgängers X_{n-1}. Formal:

Definition: *a)* Ein stochastischer Prozess $X_1, X_2, \ldots$ heißt MARKOV-Prozess oder MARKOV-Kette, wenn für alle $n \in \mathbb{N}$ gilt

$$p(X_{n+1} = x_{n+1} \mid X_1 = x_1, \ldots, X_n = x_n) = p(X_{n+1} = x_{n+1} \mid X_n = x_n)\,.$$

b) Eine MARKOV-Kette heißt *zeitinvariant,* wenn die bedingte Wahrscheinlichkeit $p(X_{n+1} = y \mid X_n = x)$ nicht von n abhängt. Ist $A = \{a_1, \ldots, a_m\}$, so setzen wir $p_{ij} = p(X_{n+1} = a_j \mid X_n = a_i)$ und bezeichnen die $m \times m$-Matrix mit Einträgen p_{ij} als die *Übergangsmatrix* des Prozesses.
c) Eine MARKOV-Kette heißt *irreduzibel,* wenn es für je zwei Buchstaben $a, b \in A$ und jede natürliche Zahl n ein $r \in \mathbb{N}$ gibt, so daß $p(X_{n+r} = b \mid X_n = a) > 0$ ist.

Bei einer irreduziblen MARKOV-Kette gibt es also keinen Buchstaben, dessen Auftreten das künftige Auftreten irgendeines anderen Buchstaben verhindert.

Der russische Mathematiker ANDREĬ ANDREEVIČ MARKOV (Андрей Андреевич Марков, 1856–1922) studierte in Sankt Petersburg, wo er später auch Professor wurde. Er beschäftigte sich zunächst hauptsächlich mit Zahlentheorie und Analysis; erst später kommen die wahrscheinlichkeitstheoretischen Arbeiten, für die er heute vor allem bekannt ist. Der Name Марков wird in lateinischen Buchstaben verschieden transkribiert; MARKOVs französische Arbeiten erschienen mit der Schreibweise MARKOFF; nach den klassischen deutschen Transkriptionsregeln müßte man MARKOW schreiben. Die Schreibweise MARKOV entspricht den englischen Regeln und scheint sich mittlerweile in der Mathematik ziemlich durchgesetzt zu haben.

Wir werden im folgenden, soweit nicht explizit etwas anderes gesagt ist, stets annehmen, daß unsere MARKOV-Ketten zeitinvariant sind. In diesem Fall können wir die Wahrscheinlichkeitsverteilungen aller Zufallsvariablen aus der von X_1 und der Übergangsmatrix berechnen: Ist allgemein $p_i^{(n)}$ die Wahrscheinlichkeit, mit der X_n dem Wert a_i annimmt, so ist

$$\begin{aligned} &p(X_n = a_{i_0}, X_{n+1} = a_{i_1}, \ldots, X_{n+r} = a_{i_r}) \\ &= p(X_n = a_{i_0}) \prod_{\ell=1}^{r} p(X_{n+\ell} = a_{i_\ell} \mid X_{n+\ell-1} = a_{\ell-1})\,. \end{aligned}$$

Für $r = 1$ wird das zu

$$p(X_n = a_i, X_{b+1} = a_j) = p(X_n = a_i)p(X_{n+1} = a_j \mid X_n = a_i) = p_i^{(n)} p_{ij}\,,$$

was wir auch einfacher mit Matrizen und Vektoren formulieren können: Ist $\mathfrak{p}^{(n)} = (p_1^{(n)}, \ldots, p_m^{(n)})^T$ der Spaltenvektor der Wahrscheinlichkeitsverteilung zu X_n, so ist $\mathfrak{p}^{(n+1)} = A^T \mathfrak{p}^{(n)}$ und damit $\mathfrak{p}^{(n)} = \left(A^T\right)^{n-1} \mathfrak{p}^{(1)}$.

Somit bestimmen $\mathfrak{p}^{(1)}$ und die Übergangsmatrix A die Wahrscheinlichkeitsverteilungen aller X_n und erlauben damit auch die Berechnung der Wahrscheinlichkeiten aller Teiltupel, die von der MARKOV-Kette produziert werden.

MARKOV-Ketten sind noch kein realistisches Modell für eine natürliche Sprache: Im Deutschen folgen beispielsweise auf ein C vorzugsweise die Buchstaben H und K; falls vor dem C aber ein S steht, sinkt die Wahrscheinlichkeit für ein K dramatisch. Trotzdem liefern MARKOV-Ketten ein deutlich besseres Modell für die deutsche Sprache als unabhängige Zufallsvariablen.

Wenn es uns darum geht, das Ergebnis eines stochastischen Prozesses zu kodieren, interessiert für lange Folgen vor allem die *mittlere* Entropie oder *Entropierate*

$$H \underset{\text{def}}{=} \lim_{n\to\infty} \frac{H(X_1, \ldots, X_n)}{n}$$

– sofern dieser Grenzwert existiert. Es ist nicht schwer, Beispiele zu finden, in denen er nicht existiert; bei den Prozessen, die uns interessieren, werden wir aber keine Probleme haben.

b) Wechselseitige Information

Bevor wir die mittlere Entropie einer MARKOV-Kette berechnen können, brauchen wir noch einige Vorbereitungen über die Entropie voneinander abhängiger Zufallsvariablen.

Wir betrachten daher zwei Zufallsvariablen X, Y mit Werten in nicht notwendigerweise übereinstimmenden Alphabeten A, B. Die gemeinsame Entropie der beiden ist einfach die Entropie der Zufallsvariablen $X \times Y$ mit Werten in $A \times B$, also

$$H(X,Y) = -\sum_{a\in A}\sum_{b\in B} p(X=a, Y=b) \log_2 p(X=a, Y=b)\,.$$

Die Abhängigkeit zwischen X und Y wird beschrieben durch die bedingten Wahrscheinlichkeiten; entsprechend dazu definieren wir

Definition: Die *bedingte Entropie* zweier Zufallsvariablen X, Y ist

$$H(Y|X) = -\sum_{a\in A}\sum_{b\in B} p(X=a, Y=b) \log_2 p(Y=b|X=a)\,;$$

für n Zufallsvariablen ist entsprechend $H(X_n|X_{n-1}, \ldots, X_1) =$

$$-\sum_{a\in A_1\times\cdots\times A_n} p(X_1=a_1, \ldots, X_n=a_n) \log_2 p(X_n=a_n|X_{n-1}=a_{n-1}, \ldots, X_1=a_1)\,.$$

Um Platz zu sparen werden wir künftig meist kurz $p(a_1, \ldots, a_n)$ und $p(a_n|a_{n-1}, \ldots, a_1)$ schreiben.

Für die bedingte Entropie gilt die folgende

Kettenregel: $H(X,Y) = H(X) + H(Y|X)$ und allgemein

$$H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1)\,.$$

Beweis: Der Übersichtlichkeit halber sei zunächst der Fall $n = 2$ behandelt:

$$\begin{aligned}
H(X,Y) &= -\sum_{a\in A}\sum_{b\in B} p(X=a, Y=b) \log_2 p(X=a, Y=b) \\
&= -\sum_{a\in A}\sum_{b\in B} p(X=a, Y=b) \log_2 \big(p(X=a)p(Y=b|X=a)\big) \\
&= -\sum_{a\in A}\left(\sum_{b\in B} p(X=a, Y=b)\right) \log_2 p(X=a) \\
&\qquad -\sum_{a\in A}\sum_{b\in B} p(X=a, Y=b) \log_2 p(Y=b|X=a) \\
&= -\sum_{a\in A} p(X=a) \log_2 p(X=a) + H(Y|X) \\
&= H(X) + H(Y|X)\,.
\end{aligned}$$

Der allgemeine Fall geht genauso: $H(X_1,\dots,X_n)$ ist nach Definition gleich

$$\begin{aligned}
&-\sum_{a\in A_1\times\cdots\times A_n} p(a_1,\dots,a_n)\log_2 p(a_1,\dots,a_n)\\
&= -\sum_{a\in A_1\times\cdots\times A_n} p(a_1,\dots,a_n)\log_2 \prod_{i=1}^{n} p(a_i|a_{i-1},\dots,a_1)\\
&= -\sum_{a\in A_1\times\cdots\times A_n}\ \sum_{i=1}^{n} p(a_1,\dots,a_n)\log_2 p(a_n|a_{i-1},\dots,a_1)\\
&= -\sum_{i=1}^{n}\ \sum_{a\in A_1\times\cdots\times A_n} p(a_1,\dots,a_n)\log_2 p(a_n|a_{i-1},\dots,a_1)\\
&= -\sum_{i=1}^{n}\ \sum_{a\in A_1\times\cdots\times A_i} p(a_1,\dots,a_n)\log_2 p(a_n|a_{i-1},\dots,a_1)\\
&= \sum_{i=1}^{n} H(X_i|X_{i-1},\dots,X_1)\,,
\end{aligned}$$

denn $\displaystyle\sum_{(a_{i+1},\dots,a_n)\in A_{i+1}\times\cdots\times A_n} p(a_1,\dots,a_n) = p(a_1,\dots,a_i).$ ∎

Die bedingte Entropie ist nicht das einzige Maß für die Information, die eine Zufallsvariable über eine andere gibt; um noch ein anderes zu definieren, betrachten wir zunächst den Fall zweier Zufallsvariablen X, Y mit Werten im gleichen Alphabet A. Diese unterscheiden sich nur in den Wahrscheinlichkeitsverteilungen: X nehme den Wert a an mit Wahrscheinlichkeit $p(a)$, Y mit Wahrscheinlichkeit $q(a)$.

Definition: Die KULLBACK-LEIBLER-Distanz zwischen X und Y oder p und q ist

$$D(X||Y) = D(p||q) = \sum_{a\in A} p(a)\log_2\frac{p(a)}{q(a)}\,.$$

Man beachte, daß die KULLBACK-LEIBLER-Distanz trotz des Namens *Distanz* keine Metrik ist: Im allgemeinen ist $D(p||q) \neq D(q||p)$. Andere

Namen für $D(p||q)$ sind *relative Entropie* oder KULLBACK-LEIBLER-*Divergenz*. $D(p||q)$ ist allerdings, wie es sich für eine Distanz gehört, nie negativ, denn wegen der Konkavität des Logarithmus ist

$$\begin{aligned}
\sum_{a\in A} p(a)\log_2\frac{p(a)}{q(a)} &= -\sum_{a\in A} p(a)\log_2\frac{q(a)}{p(a)}\\
&\le -\log_2\left(\sum_{a\in A} p(a)\frac{q(a)}{p(a)}\right) = -\log_2\sum_{a\in A} q(a) = -\log_2 1 = 0\,.
\end{aligned}$$

SOLOMON KULLBACK (1907–1994) studierte Mathematik am City College of New York und arbeitete zunächst als Lehrer. Schon 1930 wechselte er zum Signals Intelligence Service in Washington, D.C., wo er bei WILLIAM FRIEDMAN Kryptologie lernte. Daneben promovierte er 1934 an der George Washington University mit einer Arbeit aus dem Gebiet der Statistik. Während des zweiten Weltkriegs beschäftigte er sich mit dem Knacken deutscher und japanischer Codes; nach Gründung der *National Security Agency* im Jahr 1952 leitete er dort die Forschung und Entwicklung. Nach seiner Pensionierung 1962 arbeitete er als Professor für Statistik an der George Washington University.

DR. RICHARD A. LEIBLER

RICHARD ARTHUR LEIBLER (1914–2003) studierte Mathematik an der Northwestern University; nachdem er dort seinen Master erhalten hatte, promovierte er an der University of Illinois mit einer Arbeit über nichtlineare Differentialgleichungen. Nach einer kurzen Tätigkeit als Lehrer wechselte er zur Navy, die ihn im zweiten Weltkrieg im Pazifik einsetzte. 1953 kam er zur *National Security Agency*, wo er zunächst in der Forschungs- und Entwicklungsabteilung arbeitete; 1957 wurde er Leiter der Mathematischen Forschungsabteilung. 1958 wechselte er zur Kommunikationsabteilung des Instituts für Verteidigungsuntersuchungen in Princeton, deren Direktor er 1962 wurde. Von 1977 bis zu seiner Pensionierung 1980 arbeitete er wieder bei der NSA.

Um aus der KULLBACK-LEIBLER-Distanz ein Maß für die Abhängigkeit zweier beliebiger Zufallsvariablen X mit Werten in A und Y mit Werten in B zu machen, betrachten wir zwei Wahrscheinlichkeitsverteilungen

auf $A \times B$, nämlich einmal die gemeinsame Verteilung von X und Y, zum anderen die Verteilung, die wir hätten, wenn X und Y unabhängig wären, wenn also das Ereignis $(X = a, Y = b)$ die Wahrscheinlichkeit $p(X = a)p(Y = b)$ hätte. Die Distanz zwischen diesen beiden Verteilungen gibt uns ein Maß für die Abhängigkeit der beiden Zufallsvariablen:

Definition: Die *wechselseitige Information* zweier Zufallsvariablen X, Y ist

$$I(X;Y) \underset{\text{def}}{=} \sum_{a\in A}\sum_{b\in B} p(X=a, Y=b)\log_2 \frac{p(X=a, Y=b)}{p(X=a)p(Y=b)}\,.$$

Als spezieller Fall einer KULLBACK-LEIBLER-Distanz ist sie natürlich immer größer oder gleich Null.

Da $p(X = a, Y = b) = p(Y = b)p(X = a|Y = b)$ ist, können wir sie auch schreiben als

$$\begin{aligned}
I(X;Y) &= \sum_{a\in A}\sum_{b\in B} p(X=a, Y=b)\frac{\log_2 p(Y=b)p(X=a|Y=b)}{p(X=a)p(Y=b)}\\
&= \sum_{a\in A}\sum_{b\in B} p(X=a, Y=b)\frac{\log_2 p(X=a|Y=b)}{p(X=a}\\
&= \sum_{a\in A}\sum_{b\in B} p(X=a, Y=b)\log_2 p(X=a|Y=b)\\
&\qquad - \sum_{a\in A}\left(\sum_{b\in B} p(X=a, Y=b)\right)\log_2(X=a)\\
&= \sum_{a\in A}\sum_{b\in B} p(X=a, Y=b)\log_2 p(X=a|Y=b)\\
&\qquad - \sum_{a\in A} p(X=a)\log_2(X=a)\\
&= H(X) - H(X|Y)\,.
\end{aligned}$$

Somit ist $I(X;Y) = H(X) - H(X|Y)$ gerade die Differenz zwischen der Entropie von X und der bedingten Entropie von X bei Kenntnis von Y, was den Begriff *wechselseitige Information* besser erklärt als die

Formel aus der Definition. Die Nichtnegativität von $I(X;Y)$ beschreibt die Tatsache, daß die Entropie einer Zufallsvariable durch Zusatzinformation höchstens kleiner werden kann.

Auch die durch das Wort *wechselseitig* implizierte Symmetrie wird nun klar, denn da wir in obiger Rechnung die Rollen von X und Y vertauschen können, gilt auch die Formel

$$\begin{aligned}
I(X;Y) &= H(Y) - H(Y|X)\\
&= H(X) - \big(H(X,Y) - H(Y)\big)\\
&= H(X) + H(Y) - H(X,Y)\,,
\end{aligned}$$

aus der insbesondere folgt, daß $I(X;Y) = I(Y;X)$ ist.

Wenn wir bedingte Wahrscheinlichkeiten betrachten wollen, müssen wir auch hier wieder die Begriffe leicht abwandeln:

Definition: *a)* $p(x,y)$ und $q(x,y)$ seien zwei Wahrscheinlichkeitsverteilungen auf der Menge $A \times B$. Die *bedingte* KULLBACK-LEIBLER-*Distanz* zwischen p und q ist

$$D\big(p(y|x)\|q(y|x)\big) = \sum_{x\in A} p(x)\sum_{y\in B} p(y|x)\log_2 \frac{p(y|x)}{q(y|x)}\,.$$

b) Für drei Zufallsvariablen X, Y, Z ist die *bedingte wechselseitige Information von X und Y bei Kenntnis von Z*

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z)\,.$$

Für beide Größen gelten ähnliche Kettenregeln wie für die Entropie:

Lemma: *a)* $D\big(p(x,y)\|q(x,y)\big) = D\big(p(x)\|q(x)\big) + D\big(p(y|x)\|q(y|x)\big)$
b) Für $n+1$ Zufallsvariablen $X_1, \ldots, X_n$ und Y ist

$$I(X_1, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_1, \ldots, X_{i-1})\,.$$

Zum *Beweis* müssen wir in beiden Fällen einfach nachrechnen:

$$a)\ D\big(p(x,y)\|q(x,y)\big) = \sum_{x\in A}\sum_{y\in B} p(x,y)\log_2\frac{p(x,y)}{q(x,y)}$$

$$= \sum_{x\in A}\sum_{y\in B} p(x,y)\log_2\frac{p(x)p(y|x)}{q(x)q(y|x)}$$

$$= \sum_{x\in A}\sum_{y\in B} p(x,y)\log_2\frac{p(x)}{q(x)} + \sum_{x\in A}\sum_{y\in B} p(x,y)\log_2\frac{p(y|x)}{q(y|x)}$$

$$= \sum_{x\in A} p(x)\log_2\frac{p(x)}{q(x)} + \sum_{x\in A}\sum_{y\in B} p(x,y)\log_2\frac{p(y|x)}{q(y|x)}$$

$$= D\big(p(x)\|q(x)\big) + D\big(p(y|x)\|q(y|x)\big)$$

$$b)\ I(X_1,\dots,X_n;Y) = H(X_1,\dots,X_n) - H(X_1,\dots,X_n|Y)$$

$$= \sum_{i=1}^{n} H(X_i|X_1,\dots,X_{i-1}) - \sum_{i=1}^{n} H(X_i|X_1,\dots,X_{i-1},Y)$$

$$= \sum_{i=1}^{n} I(X_i;Y|X_1,\dots,X_{i-1})\,.$$

c) Berechnung der mittleren Entropie

In Abschnitt *a)* hatten wir die mittlere Entropie

$$H \underset{\text{def}}{=} \lim_{n\to\infty}\frac{H(X_1,\dots,X_n)}{n}$$

eines stochastischen Prozesses $\mathcal{X} = (X_n)_{n\in\mathbb{N}}$ definiert; nach den Vorbereitungen aus Abschnitt *b)* können wir jetzt untersuchen, unter welchen Bedingungen sie existiert, und wie sie auch anders berechnet werden kann.

Dazu betrachten wir die Information, die uns die n-te Zufallsvariable X_n *neu* liefert, wenn wir ihre Vorgänger $X_1,\dots,X_{n-1}$ bereits kennen, und lassen auch hier n gegen Unendlich gehen; falls der Grenzwert existiert, schreiben wir

$$H'(\mathcal{X}) \underset{\text{def}}{=} \lim_{n\to\infty} H(X_n|X_1,\dots,X_{n-1})\,.$$

Satz: Für einen stationären stochastischen Prozess $\mathcal{X} = (X_n)_{n\in\mathbb{N}}$ existieren die Grenzwerte $H(\mathcal{X})$ und $H'(\mathcal{X})$ und sind gleich.

Der *Beweis* besteht aus drei Schritten:

1. Schritt: $H'(\mathcal{X})$ existiert

Da die Entropie durch Zusatzinformation höchstens kleiner werden kann, ist

$$H(X_{n+1}|X_1,X_2,\dots,X_n) \le H(X_{n+1}|X_2,\dots,X_n)\,.$$

Wegen der Stationarität des Prozesses können wir auf der rechten Seite alle Indizes um eins erniedrigen, ohne daß sich der Wert der bedingten Entropie ändert; daher ist auch

$$H(X_{n+1}|X_1,X_2,\dots,X_n) \le H(X_n|X_1,X_2,\dots,X_{n-1})$$

für alle $n\in\mathbb{N}$. Somit ist die Folge der $H(X_n|X_1,\dots,X_{n-1})$ monoton fallend, und da auch bedingte Entropien nie negativ sind, ist sie nach unten beschränkt. Beides zusammen impliziert die Konvergenz.

2. Schritt: Konvergiert eine Folge $(x_n)_{n\in\mathbb{N}}$ von reellen Zahlen gegen einen Grenzwert a, so konvergiert auch die Folge der arithmetischen Mittel $y_n = \frac{1}{n}(x_1+\dots+x_n)$ gegen a (CESÀRO-Mittel).

Dazu müssen wir zeigen, daß es zu jedem $\varepsilon>0$ ein $N\in\mathbb{N}$ gibt derart, daß $|a-y_n|<\varepsilon$ ist für alle $n\ge N$.

Da die Folge der x_n gegen a konvergiert, gibt es jedenfalls ein $M\in\mathbb{N}$, so daß $|a-x_n|<\frac{1}{2}\varepsilon$ ist für alle $n\ge M$. Für solche n ist dann

$$|a-y_n| = \left|\frac{1}{n}\sum_{i=1}^{n}(a-x_i)\right| \le \frac{1}{n}\sum_{i=1}^{n}|a-x_i|$$

$$= \frac{1}{n}\sum_{i=1}^{M-1}|a-x_i| + \frac{1}{n}\sum_{i=M}^{n}|a-x_i|$$

$$< \frac{1}{n}\sum_{i=1}^{M-1}|a-x_i| + \frac{(n-M+1)\varepsilon}{n} \le \frac{1}{n}\sum_{i=1}^{M-1}|a-x_i| + \frac{\varepsilon}{2}\,.$$

Die Summe S über die ersten $M-1$ Abweichungen hängt nicht von n ab; wir können daher leicht ein $M'\in\mathbb{N}$ finden, so daß $S/n<\varepsilon/2$ ist

für alle $n \geq M'$. Ist n mindestens gleich dem Maximum N von M und M', so muß daher $|a - y_n| < \varepsilon$ sein, wie verlangt.

3. Schritt: Der Grenzwert $H(\mathcal{X})$ existiert und ist gleich $H'(\mathcal{X})$

Nach der Kettenregel für die Entropie ist

$$H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i | X_1, \ldots, X_{i-1}) ;$$

dividieren wir beide Seiten durch n, sind wir genau in der Situation des zweiten Schritts, wobei links die Glieder jener Folge stehen, als deren Grenzwert $H(\mathcal{X})$ definiert ist, und rechts das arithmetische Mittel der ersten n Terme der Folge, deren Grenzwert nach dem ersten Schritt existiert und mit $H'(\mathcal{X})$ bezeichnet wurde.

Damit ist der Satz vollständig bewiesen. ∎

ERNESTO CESÀRO (1859–1906) wurde in Neapel geboren und wuchs auf in der nahe gelegenen Kleinstadt Torre Annunziata, wo sei Vater einen landwirtschaftlichen Betrieb mit Hofladen führte. Nach seiner Schulausbildung in Neapel studierte er ab 1873 in Liège Mathematik. Nach dem Tod seines Vaters kehrte er 1879 nach Torre Annunziata zurück um den Betrieb weiterzuführen. Dank eines Stipendiums konnte er ab 1882 sein Studium in Liège fortführen; teilweise studierte er auch in Paris und ab 1884 schließlich an der Universität Rom. Obwohl er bereits zahlreiche Arbeiten veröffentlicht hatte, wurde er dort erst 1887 promoviert und bekam dann gleich einen Lehrstuhl an der Universität von Palermo. 1891 folgte er einem Ruf an die Universität Neapel, wo er bis zu seinem Tod lehrte. Der Großteil seiner Arbeiten befaßt sich mit Differentialgeometrie; er leistete aber auch Beiträge zur Zahlentheorie, unter anderem etwa zur Primzahlverteilung.

Speziell für (zeitinvariante) MARKOV-Ketten läßt sich $H'(\mathcal{X})$ leicht berechnen: Wegen der MARKOV-Eigenschaft ist $H(X_n | X_1, \ldots, X_{n-1}) = H(X_n | X_{n-1})$, und wegen der Zeitinvarianz ist das für alle n gleich $H(X_2 | X_1)$. Somit ist hier die Entropierate einfach die bedingte Entropie einer jeden Zufallsvariablen bei Kenntnis ihres Vorgängers.

§ 6: Datenkompression

Wir haben schon mehrfach gesehen, daß die Entropie einer Zufallsvariablen in einfachen Fällen interpretiert werden kann als die mittlere Bitanzahl für eine Kodierung ihres Alphabets. Diesen Zusammenhang und seine Anwendung auf die Komprimierung verschiedener Arten von Daten soll in diesem Paragraphen untersucht werden.

Zur Speicherung oder Übermittlung der von einer Zufallsvariablen oder einem stochastischen Prozess produzierten Daten müssen wir die Elemente des Alphabets A in geeigneter Weise kodieren mit Zeichenfolgen, die durch die verwendete Technik vorgegeben sind; heutzutage sind dies meist Bitfolgen. Wir gehen allgemein aus von einem Zeichensatz, der zwar meist, aber nicht immer, gleich der Menge $\{0, 1\}$ sein wird. Historisch bedeutsame Beispiele von Mengen C mit mehr als zwei Elementen sind etwa die Morsezeichen mit $C = \{\text{kurz, lang, Pause}\}$ oder die in der Seefahrt gebräuchlichen Flaggenalphabete.

Ziel der Datenkompression ist es, die Anzahl der Zeichen für die betrachteten Daten möglichst klein zu halten; bei einem verlustfreien Komprimierungsverfahren sollen aber die ursprünglichen Daten trotzdem noch exakt rekonstruierbar sein.

Nicht alle in der Praxis verwendeten Verfahren sind verlustfrei; beim mp3-Verfahren etwa wird ausgenutzt, daß manche Frequenzen unser Ohr für andere Frequenzen blockieren, so daß man diese aus dem Signal herausfiltern kann. Auch das JPEG-Verfahren, mit dem sich der letzte Abschnitt dieses Paragraphen beschäftigt, ist oft nicht verlustfrei; hier gibt es einen Qualitätsfaktor als Parameter, der die tolerierbaren Verluste quantifiziert.

Es ist klar, daß es keinen universellen Algorithmus zur Datenkompression geben kann: Gäbe es nämlich ein Verfahren, das für beliebige Dateien einen Kompressionsfaktor $\alpha < 1$ garantieren würde, so könnte man dieses Verfahren iterativ anwenden und nach n Anwendungen eine Kompressionsrate von α^n erreichen. Bei hinreichend großem n könnte man daher jede Datei auf weniger als ein Bit komprimieren, was natürlich absurd ist.

Ein Kompressionsverfahren kann also nur auf Dateien mit spezieller Struktur erfolgreich angewandt werden. Wir werden in diesem Paragraphen zwei Ansätze diskutieren: Die Entropiekodierung, bei der die Unterschiede zwischen den Häufigkeiten der einzelnen Buchstaben ausgenutzt wird, und die Datenkomprimierung durch lineare Transformationen, die bei einem stochastischen Prozess eine weitestgehende Dekorrelation der beteiligten Zufallsvariablen erreichen wollen.

a) Quellenkodierung

Wir gehen zunächst aus vom einfachsten Fall einer einzigen Zufallsvariablen X. Sie nehme Werte an in einem Alphabet $A = \{a_1, \ldots, a_m\}$, wobei der Buchstabe a_i mit Wahrscheinlichkeit p_i auftrete.

Zur Speicherung oder Übermittlung der von X produzierten Daten müssen wir die Elemente von A in geeigneter Weise kodieren mit Zeichenfolgen, die durch die verwendete Technik vorgegeben sind; heutzutage sind dies meist Bitfolgen. Wir gehen allgemein aus von einem Zeichensatz $\mathcal{D}$, der zwar meist, aber nicht immer, gleich der Menge $\{0, 1\}$ sein wird. Die Elementanzahl vonÂfl$\mathcal{D}$ bezeichnen wir mit D.

Das Wort *Kodierung* hat in der Informationstheorie mehrere deutlich verschiedene Bedeutungen: Wir haben einmal Codes, die dazu dienen allfällige Lese- und Übertragungsfehler zu erkennen und teilweise auch zu korrigieren; diese fehlererkennenden und fehlererkorrigierenden Codes bilden den Inhalt mathematischer Vorlesungen über *Kodierungstheorie;* Informationstechniker reden hier von *Kanalkodierung.* Dann gibt es schon seit mindestens zweieinhalb Jahrtausenden *Geheimcodes,* mit denen Text so verschlüsselt werden soll, daß ihn ein Unbefugter nicht rekonstruieren kann; diese werden in Vorlesungen über Kryptologie (oder, wenn die reine Verschlüsselung im Vordergrund steht, auch Kryptographie) behandelt. Die Kodierung, mit der wir uns hier beschäftigen, greift bereits eine Stufe vor diesen beiden und heißt *Quellenkodierung;* sie wird oft zusammen mit Kanalkodierung und/oder Verschlüsselung eingesetzt.

Definition: Ein *Quellencode* für eine Zufallsvariable X mit Werten in einem Alphabet A ist eine Abbildung C, die jedem Element $x \in A$

eine Folge von Elementen aus $\mathcal{D}$ zuordnet. Das Codewort zu $x \in A$ bezeichnen wir mit $C(x)$, seine Länge mit $\ell(x)$. Die *mittlere Länge* $L(C)$ ist der Erwartungswert

$$\mathbb{E}\big(\ell(X)\big) = \sum_{x \in A} p(x)\ell(x)\,.$$

Die wohl bekanntesten Beispiele von Quellencodes sind der ASCII-Code *(American Standard Code for Information Interchange),* der ein Alphabet aus 128 Zeichen durch Folgen aus je sieben Bit darstellt, denen zwecks Fehlererkennung praktisch immer ein Paritätsbit angehängt wird, sowie seine Erweiterungen wie ISO Latin-1, die ASCII zu einem echten Achtbitcode erweitern, in dem auch Umlaute und Ähnliches zum Alphabet gehören. Vor allem im *World Wide Web* wird oft auch der sogenannte *Unicode* verwendet, der mit 17 Ebenen zu je 16 Bit jedem irgendwo auf der Welt benutzten Schriftzeichen eine digitale Entsprechung zuordnen will.

Laut obiger Definition ist ein Quellencode einfach irgendeine Abbildung; wenn diese nicht injektiv ist, reden wir von einem *singulären* Code. Da solche Codes nur selten nützlich sind, werden wir uns im Folgenden auf nichtsinguläre Codes beschränken.

Auch bei diesen kann es aber noch Probleme mit der eindeutigen Rekonstruierbarkeit geben wenn wir uns bei der Kodierung nicht auf einzelne Buchstaben beschränken, sondern – wie dies wohl meist der Fall sein wird – Buchstaben*folgen* betrachten: Kodieren wir etwa die 26 Buchstaben des Alphabets durch die im Zweiersystem geschriebenen Zahlen von 0 bis 25, so ist diese Abbildung sicherlich injektiv; setzen wir aber die Codes $C(D) = 11$, $C(A) = 0$ und $C(S) = 10010$ einfach hintereinander, können wir das Ergebnis auch beispielsweise als GEC lesen statt als DAS.

Um dieses Problem zu vermeiden, könnten wir uns auf Codes beschränken, bei denen alle Codewörter $C(x)$ dieselbe Länge haben, wie es beispielsweise bei ASCII der Fall ist oder auch bei den heute kaum noch benutzten Fernschreibern. Wenn allerdings die Zeichen des Alphabets (wie etwa im Fall der Buchstaben eines deutschen Texts) deutlich verschiedene Wahrscheinlichkeiten haben, können wir die mittlere Länge

des Codes oft drastisch reduzieren, wenn wir den häufigen Buchstaben kurze Codewörter zuordnen. Die geschieht beispielsweise beim Morse-Code, der mit den drei Zeichen *kurz, lang* und *Pause* arbeitet; hier wird das E durch einmal kurz kodiert, das N durch einmal lang, das Y aber durch lang, kurz, lang, lang. Zum Trennen der einzelnen Buchstaben dient das dritte Zeichen, die Pause.

Wir interessieren uns für Codes variabler Länge, bei denen die eindeutige Dekodierbarkeit ohne spezielles Trennzeichen gewährleistet ist:

Definition: *a)* Ein Code heißt *eindeutig dekodierbar,* wenn es keine zwei Folgen von Buchstaben aus dem Alphabet A gibt, die auf dieselbe Zeichenfolge abgebildet werden.
b) Ein Code heißt *Praefixcode,* wenn es keine zwei Buchstaben $x, y \in A$ gibt, für die $C(x)$ mit den ersten $\ell(x)$ Zeichen von $C(y)$ übereinstimmt.

Offensichtlich ist jeder Praefixcode eindeutig dekodierbar; umgekehrt ist jedoch nicht jeder eindeutig dekodierbare Code ein Praefixcode: Kodieren wir etwa ein dreielementiges Alphabet durch die Vorschrift $C(a) = 0, C(b) = 001$ und $C(c) = 11$ und sehen beim Dekodieren als erstes Zeichen eine Eins, so kann der nächste Buchstabe nur ein b sein. Sehen wir eine Folge von n Nullen, gefolgt von einer geraden Anzahl von Einsen, so müssen wir n mal den Buchstaben a haben, gefolgt von einem c; folgt aber eine ungerade Anzahl von Einsen, so haben wir $n-2$ mal den Buchstaben a, gefolgt von einem b. Somit ist C ein Praefixcode, obwohl $C(a)$ ein Praefix von $C(b)$ ist.

Die eindeutige Dekodierbarkeit schränkt die Anzahl möglicher Codewörter einer vorgegebenen Länge ein; insbesondere gilt die folgende

Ungleichung von Kraft und McMillan: Ist C ein eindeutig dekodierbarer Code für das Alphabet A und bezeichnet D die Anzahl der Codezeichen, so ist

$$\sum_{x\in A} D^{-\ell(x)} \le 1\,.$$

Beweis: Für jede natürliche Zahl k läßt sich C fortsetzen zu einem Code für das Alphabet A^k, indem wir einem k-tupel $(x_1,\dots,x_n)$ von

Buchstaben die hintereinander gesetzten Codewörter $C(x_1),\dots,C(x_k)$ zuordnen, aufgefaßt als ein Codewort der Länge $\ell(x_1)+\dots+\ell(x_k)$. Wegen der eindeutigen Dekodierbarkeit von C ist auch dieser erweiterte Code nichtsingulär.

Nach dem Distributivgesetz ist

$$\left(\sum_{x\in A} D^{-\ell(x)}\right)^k = \sum_{(x_1,\dots,x_k)\in A^k} D^{-\ell(x_1)}\cdots D^{-\ell(x_k)} = \sum_{\mathbf{x}\in A^k} D^{-\ell(\mathbf{x})}\,.$$

Bezeichnet $a(n)$ die Anzahl der k-tupel $\mathbf{x}\in A^k$, denen ein Codewort der Länge n zugeordnet wird, können wir dies auch schreiben als

$$\sum_{n=1}^{k\ell_{\max}} a(n) D^{-n}\,,$$

wobei $\ell_{\max}$ das Maximum aller $\ell(x)$ für $x\in A$ bezeichnet. Da der Code nichtsingulär ist, kann $a(n)$ nicht größer sein als die Anzahl D^n aller möglicher Codewörter der Länge n; somit ist

$$\left(\sum_{x\in A} D^{-\ell(x)}\right)^k \le \sum_{n=1}^{k\ell_{\max}} D^n D^{-n} = k\ell_{\max}$$

und damit

$$\sum_{x\in A} D^{-\ell(x)} \le \sqrt[k]{k\ell_{\max}}\,.$$

Dies gilt für alle k, und da

$$\lim_{k\to\infty} \sqrt[k]{k\ell_{\max}} = 1$$

ist, folgt hieraus die zu beweisende Ungleichung. ∎

Dieser Satz wurde 1949 von LEON G. KRAFT im Rahmen seiner Master Thesis im Fach Elektrotechnik am MIT für Praefixcodes bewiesen. Für beliebige eindeutig dekodierbare Codes bewies sie BROCKWAY MCMILLAN von den Bell Telephone Labs unabhängig von Kraft 1956 in seiner Arbeit *Two Inequalities Implied by Unique Decipherability* in den IEEE Transaction on Information Theory, Band 2(4), S. 115-116.

Umgekehrt gilt

Satz: Ist $\ell\colon A \to \mathbb{N}$ eine Abbildung des Alphabets A in die natürlichen Zahlen mit

$$\sum_{x\in A} D^{-\ell(x)} \leq 1\,,$$

so gibt es einen Praefixcode C mit D-elementigem Zeichensatz, für den das Codewort $C(x)$ die Länge $\ell(x)$ hat.

Beweis: Die D Zeichen, aus denen die Codewörter gebildet werden, seien $z_1, \ldots, z_D$, und $\ell_{\max}$ sei das Maximum der Längen $\ell(x)$; das zu kodierende Alphabet sei $A = \{a_1, \ldots, a_m\}$.

Wir zeichnen einen Baum, indem wir ausgehend von einem festen Punkt, der Wurzel, D gerichtete Strecken zeichnen, die wir mit z_1 bis z_D markieren. Vom Endpunkt einer jeden dieser Strecken zeichnen wir wieder D solche Strecken und so weiter, bis wir Streckenzüge der Länge $\ell_{\max}$ haben. Punkte, die vom Anfangspunkt aus über einen Streckenzug der Länge n erreichbar sind, bezeichnen wir als Knoten der Tiefe n. Jeder Knoten ist eindeutig beschreibbar durch die Folge der Markierungen $z_{i_1} \ldots z_{i_n}$ der Strecken, die zu ihm führen; Knoten der Tiefe n entsprechen also potentiellen Codewörtern der Länge n. Diese ordnen wir lexikographisch durch die Übereinkunft, daß z_i vor z_j kommen soll, wenn $i < j$ ist.

Zur Konstruktion des gewünschten Codes ordnen wir als erstes dem Buchstaben a_1 das Codewort zu, das aus $\ell(a_1)$ Zeichen z_1 besteht. Danach entfernen wir den zu diesem Codewort gehörenden Knoten der Tiefe n aus dem Baum sowie alle Knoten, die über diesen Punkt erreichbar sind – sie entsprechen schließlich Codewörtern, die das gerade vergebene Codewort als Anfang hätten.

Wenn wir Codewörter für a_1 bis a_{r-1} vergeben haben, konstruieren wir das Codewort $C(a_r)$ wie folgt: Unter allen noch verbleibenden Knoten der Tiefe $\ell(a_r)$ nehmen wir den lexikographisch ersten und definieren $C(a_r)$ als das dazugehörige Codewort; sodann streichen wir wieder sowohl diesen Knoten als auch alle über ihn erreichbaren Knoten größerer Tiefe.

Das kann natürlich nur dann funktionieren, wenn es noch einen Knoten der Tiefe $\ell(a_r)$ gibt. Dazu genügt es, wenn wir zeigen, daß es noch

mindestens einen Knoten der Tiefe $\ell_{\max}$ gibt, denn der Weg zu einem solchen Knoten führt durch Knoten aller kleineren Tiefen.

Wenn wir ein Codewort der Länge n vergeben, streichen wir mit dem zugehörigen Knoten der Tiefe n auch alle über diesen Knoten erreichbare tiefere; in Tiefe $\ell_{\max}$ sind das $D^{\ell_{\max}-n}$ Stück. Durch die Vergabe von Codewörtern für a_1 bis a_{r-1} wurden somit

$$\sum_{i=1}^{r-1} D^{\ell_{\max}-\ell(a_i)} = D^{\ell_{\max}} \sum_{i=1}^{r-1} D^{-\ell(a_i)}$$

Knoten gelöscht. Anfangs gab es zu jedem n nach Konstruktion D^n Knoten der Tiefe n, also $D^{\ell_{\max}}$ Knoten der maximalen Tiefe. Nach Voraussetzung ist für $r \leq n$

$$\sum_{i=1}^{r-1} D^{-\ell(a_i)} < \sum_{x\in A} D^{-\ell(x)} \leq 1\,,$$

die Anzahl bereits gestrichener Knoten maximaler Tiefe ist also echt kleiner als die ursprünglich vorhandene Anzahl. ∎

Zusammen zeigen die beiden gerade bewiesenen Sätze, daß es zu jedem eindeutig dekodierbaren Code einen Praefixcode gibt, dessen Codewörter exakt dieselbe Länge haben. Auf der Suche nach möglichst guten Codes können und werden wir uns daher auf Praefixcodes beschränken.

b) Optimale Codes

Da sowohl die Speicherung als auch die Übertragung von Daten oftmals knappe Ressourcen in Anspruch nehmen wie beispielsweise den Speicher einer kleinen Digitalkamera oder die Kapazität eines zumindest zeitweise sehr stark belasteten Mobilfunknetzes, ist es für viele Anwendungen wichtig, den benötigten Aufwand auf das unbedingt notwendige Minimum zu beschränken. Auf dem Niveau der Quellenkodierung bedeutet dies insbesondere, daß die mittlere Länge des verwendeten Codes möglichst klein sein soll.

Sei also X eine Zufallsvariable mit Werten in einem m-elementigen Alphabet A, dessen i-tes Element mit Wahrscheinlichkeit p_i angenommen werde. Wir suchen einen Code, dessen mittlere Länge

$$L = \sum_{i=1}^{m} p_i \ell_i$$

minimal ist; dabei steht ℓ_i für die Länge des Codeworts zum i-ten Element des Alphabets.

Wie wir gerade gesehen haben, müssen die Längen der Codewörter eines eindeutig dekodierbaren Quellencodes die Ungleichung von KRAFT und MCMILLAN erfüllen; umgekehrt gibt es auch zu jeder Längenverteilung, die diese Ungleichung erfüllt, einen Praefixcode. Somit müssen wir L minimieren unter der Nebenbedingung

$$\sum_{i=1}^{m} D^{-\ell_i} \leq 1 \,.$$

Wenn wir für den Augenblick vergessen, daß die ℓ_i natürliche Zahlen sein müssen, haben wir hier ein klassisches Extremwertproblem mit Nebenbedingung, das wir mit Hilfe eins LAGRANGE-Multiplikators lösen können: Da eine lineare Funktion keine lokalen Extrema hat, muß die Nebenbedingung für alle Extrema eine Gleichung sein, und somit müssen

$$\operatorname{grad} L = \begin{pmatrix} p_1 \\ \vdots \\ p_m \end{pmatrix} \quad \text{und} \quad \operatorname{grad} \sum_{i=1}^{m} D^{-\ell_i} = D^{-\ell_i} \ln D$$

proportional sein, d.h. es gibt ein $\lambda \in \mathbb{R}$, so daß

$$p_i = \lambda D^{-\ell_i} \ln D \quad \text{für alle } i \,.$$

Da sowohl die Summe aller p_i als auch die Summe der $D^{-\ell_i}$ gleich eins ist, muß $\lambda = 1/\ln D$ sein, d.h.

$$p_i = D^{-\ell_i} \quad \text{und} \quad \ell_i = -\log_D p_i \,.$$

Der Wert der Zielfunktion in diesem Punkt ist

$$L = \sum_{i=1}^{m} p_i \ell_i = -\sum_{i=1}^{m} p_i \log_D p_i \,,$$

in Falle $D = 2$ also die Entropie $H(X)$, ansonsten eine dazu proportionale Größe, die wir der Kürze halber mit $H_D(X)$ bezeichnen wollen.

Bislang haben wir nur eine *notwendige* Bedingung für ein Extremum gefunden; das folgende Lemma zeigt, daß wir tatsächlich das globale Minimum gefunden haben. (Der Leser sollte sich davon überzeugen, daß der Beweis auch funktioniert, wenn die ℓ_i beliebige reelle Zahlen sind, die der Ungleichung von KRAFT und MCMILLAN genügen.)

Lemma: Die mittlere Länge L eines jeden eindeutig dekodierbaren Quellencodes mit D-elementigem Zeichensatz ist mindestens gleich $H_D(X)$ mit Gleichheit genau dann, wenn für alle Wahrscheinlichkeiten p_i gilt: $p_i = D^{-\ell_i}$ mit einem $\ell_i \in \mathbb{N}_0$.

Beweis: Die Ungleichung von KRAFT und MCMILLAN sagt uns, daß $c = \sum_{i=1}^{m} D^{-\ell_i} \leq 1$ ist; setzen wir $q_i = D^{-\ell_i}/c$, so definieren auch die q_i eine Wahrscheinlichkeitsverteilung q und

$$\begin{aligned} L - H_D(X) &= \sum_{i=1}^{m} p_i \ell_i + \sum_{i=1}^{m} p_i \log_D p_i \\ &= -\sum_{i=1}^{m} p_i \log_D D^{-\ell_i} + \sum_{i=1}^{m} p_i \log_D p_i \\ &= \sum_{i=1}^{m} p_i \log_D \frac{p_i}{c r_i} = \sum_{i=1}^{m} p_i \log_D \frac{p_i}{r_i} - \sum_{i=1}^{m} p_i \log_D c \\ &= \frac{D(p\|q)}{\log_2 D} - \log_D c \geq 0 \,, \end{aligned}$$

da die KULLBACK-LEIBLER-Distanz zweier Wahrscheinlichkeitsverteilungen nicht negativ ist und $\log_D c \leq 0$, da $c \leq 1$ ist. Falls die Differenz gleich Null ist, muß $c = 1$ und $D(p\|q) = 0$ sein, d.h. für jedes i ist $p_i = q_i = D^{-\ell_i}$.

■

Damit haben wir eine untere Grenze für L gefunden, die allerdings nur in recht speziellen Situationen wirklich angenommen wird. Wie der folgende Satz zeigt, können wir aber immer einen Code finden, für den sie um weniger als eins überschritten wird:

Satz: X sei eine Zufallsvariable mit Werten in einem m-elementigen Alphabet A, dessen i-tes Element mit Wahrscheinlichkeit p_i angenommen werde. Dann gibt es einen Quellencode für A mit einem D-elementigen Zeichensatz, dessen mittlere Länge L die Ungleichung

$$H_D(X) \le L < H_D(X) + 1$$

erfüllt.

Beweis: Wir suchen natürliche Zahlen $\ell_1, \dots, \ell_m$, für die $\sum p_i \ell_i$ möglichst klein wird, die aber die Ungleichung von KRAFT und MCMILLAN erfüllen. Im Reellen wird für $\ell_i = -\log_D p_i$ das Minimum angenommen; wir gehen zur nächstgrößeren ganzen Zahl, definieren ℓ_i also als diejenige natürliche Zahl, für die gilt

$$-\log_D p_i \le \ell_i < -\log_D p_i + 1 \,.$$

Dann ist

$$\sum_{i=1}^{m} D^{-\ell_i} \le \sum_{i=1}^{m} D^{\log_D p_i} = \sum_{i=1}^{m} p_i = 1 \,,$$

die Ungleichung von KRAFT und MCMILLAN ist somit erfüllt, so daß es einen Praefixcode C gibt, der dem i-ten Buchstaben des Alphabets A ein Codewort der Länge ℓ_i zuordnet. Für dessen mittlere Länge gilt

$$H_D(X) = \sum_{i=1}^{m} p_i(-\log_D p_i) \le L = \sum_{i=1}^{m} p_i \ell_i$$
$$< \sum_{i=1}^{m} p_i(-\log_D p_i + 1) = H_D(X) + 1 \,.$$

In den ersten Paragraphen dieses Kapitels haben wir mehrere Beispiele betrachtet, bei denen sich die Annäherung der mittleren Bitzahl pro Buchstabe an die Entropie verbessern ließ, wenn wir statt einzelner Buchstaben Blöcke von Buchstaben betrachten. Der gerade bewiesene Satz erklärt auch das:

Korollar: X sei eine Zufallsvariable mit Werten in einem m-elementigen Alphabet A, dessen i-tes Element mit Wahrscheinlichkeit p_i angenommen werde. Dann gibt es einen Quellencode für A^n mit einem D-elementigen Zeichensatz, dessen mittlere Länge L pro Buchstaben aus A die Ungleichung

$$H_D(X) \le L < H_D(X) + \frac{1}{n}$$

erfüllt.

Beweis: Wir betrachten ein n-tupel aus unabhängigen Zufallsvariablen $X_1, \dots, X_n$, die allesamt Werte im Alphabet A annehmen und alle die gleiche Wahrscheinlichkeitsverteilung haben wie X. Die Entropie dieses n-tupels (zur Basis D) mit Werten in A^n ist $nH_D(X)$; nach dem gerade bewiesenen Satz gibt es also einen Quellencode, dessen mittlere Länge zwischen $nH_D(X)$ und $nH_D(X) + 1$ liegt. Die Länge L bezogen auf die Buchstaben aus A ist ein n-tel davon, erfüllt also die behauptete Ungleichung. ∎

Dieses Korollar zeigt insbesondere, daß wir den mittleren Aufwand pro Buchstabe mit hinreichend großen Blocklängen beliebig nahe an die Entropie annähern können.

c) Huffman-Codes

DAVID HUFFMAN stellte 1951 ein Verfahren vor, wie man zu einer gegebenen Häufigkeitsverteilung einen Praefixcode mit minimaler mittlerer Länge konstruieren kann. In seiner Arbeit

> DAVID A. HUFFMAN: A Method for the Construction of Minimum-Redundancy Codes, *Proc. I.R.E., Sept. 1952, 1098–1101*

geht er aus von einer endlichen Menge von Nachrichten, also dem, was wir hier immer als das Alphabet $A = \{a_1, \dots, a_m\}$ bezeichnen, und ordnet sie so, daß für die Wahrscheinlichkeit p_i von a_i gilt: $p_i \ge p_j$ falls $i < j$.

Für einen Code C mit minimaler mittlerer Länge kann man dann ohne Beschränkung der Allgemeinheit davon ausgehen, daß für die Länge ℓ_i

des Codeworts $C(a_i)$ gilt: $\ell_i \leq \ell_j$ fälls $i < j$. Wäre nämlich $\ell_j > \ell_i$, so wäre $p_j\ell_j + p_i\ell_i > p_i\ell_j + p_j\ell_i$, d.h. die mittlere Länge des Codes würde echt kleiner durch Vertauschen der Codewörter $C(a_i)$ und $C(a_j)$, im Widerspruch zur vorausgesetzten Minimalität.

Außerdem muß $\ell_m = \ell_{m-1}$ sein, denn die ersten ℓ_{m-1} Zeichen von $C(A_m)$ können wegen der Praefixbedingung kein Codewort für einen anderen Buchstaben sein; wenn wir etwaige folgende Zeichen von $C(a_m)$ streichen, erhalten wir einen neuen Praefixcode, dessen mittlere Länge im Fall $\ell_m > \ell_{m-1}$ kleiner wäre als die von C.

Somit gibt es mindestens zwei Buchstaben, denen ein Codewort maximaler Länge zugeordnet wird. Wir können aber noch mehr sagen: Es gibt unter den Codewörtern maximaler Länge mindestens zwei, die sich nur in ihrem letzten Zeichen unterscheiden: Wäre dies nicht der Fall, könnten wir bei allen Codewörtern maximaler Länge das letzte Zeichen streichen ohne die Praefixbedingung zu verletzen.

Bislang gilt alles für Codes mit einer beliebigen Anzahl D von Zeichen; für die Konstruktion eines Codes anhand der aufgestellten Prinzipien wollen wir uns aber – genau wie HUFFMAN – zunächst auf den Fall $D = 2$ beschränken, und die Modifikationen für $D > 2$ anschließend kurz diskutieren.

Angenommen, wir haben einen optimalen binären Code für ein Alphabet aus $m > 2$ Buchstaben. Dann wissen wir, daß es unter den Codewörtern maximaler Länge zwei gibt, die sich nur im letzten Bit unterscheiden; indem wir die Codes der Buchstaben mit maximaler Codelänge nötigenfalls permutieren, können wir annehmen, daß es sich dabei im die beiden Buchstaben a_m und a_{m-1} handelt. (Man beachte, daß HUFFMAN die Buchstaben nach ihrer Häufigkeit anordnet.)

Für einen beliebigen binären Code, bei dem die beiden Buchstaben geringster Wahrscheinlichkeit Codewörter maximaler Länge haben, die sich nur im letzten Bit unterscheiden, können wir die HUFFMAN-Reduktion bilden wie folgt:

Wir betrachten ein neues Alphabet A^* aus $m-1$ Buchstaben; es enthält die Buchstaben a_1 bis a_{m-2} sowie einen neuen Buchstaben a^*, der

mit Wahrscheinlichkeit $p_{m-1} + p_m$ auftreten soll. Zu diesem Alphabet betrachten wir den Code C^*, der den a_i mit $i \leq m-2$ das Codewort $C(a_i)$ zuordnet; $C^*(a^*)$ sei $C(a_m)$ ohne das letzte Bit. Umgekehrt läßt sich C aus C^* fast eindeutig rekonstruieren: Wir setzen $C(a_{m-1})$ auf $C^*(a^*)$ gefolgt von einer Null und $C(a_m)$ auf $C^*(a^*)$ gefolgt von einer Eins (oder umgekehrt). Die mittlere Länge L^* von C^* läßt sich leicht durch die mittlere Länge L von C ausdrücken:

$$L^* = \sum_{i=1}^{m-1} p_i\ell_i + (p_{m-1} + p_m)(\ell_m - 1) = \sum_{i=1}^{m} p_i\ell_i - p_{m-1} - p_m$$
$$= L - p_{m-1} - p_m\,,$$

denn $\ell_{m-1} = \ell_m$.

HUFFMANs Konstruktion beruht wesentlich auf der folgenden Beobachtung: *C ist genau dann optimal, wenn C^* optimal ist.*

Ist nämlich C^* nicht optimal, so gibt es einen Code D^* mit kleinerer mittlerer Länge $L^{**} < L^*$. Daraus läßt sich ein Code D für A konstruieren, bei dem $D(a_{m-1})$ und $D(a_m)$ aus $D^*(a^*)$ entstehen durch Anhängen einer Null *bzw.* einer Eins; die mittlere Länge dieses Codes ist $L^{**} + p_{m-1} + p_m < L$, so daß auch C nicht optimal ist. Entsprechend folgt auch die andere Richtung.

Damit ist die rekursive Struktur des Algorithmus von HUFFMAN klar: Wir wollen ein Alphabet A kodieren, das m Buchstaben enthält.

Im Fall $m = 2$ können wir einfach jedem der beiden Buchstaben eines der beiden Codezeichen zuordnen; die mittlere Länge des Codes ist dann eins, und kürzer kann sie bei keinem Code sein.

Ist $m > 2$, so führen wir eine HUFFMAN-Reduktion durch, d.h. wir fassen die beiden am wenigsten wahrscheinlichen Zeichen zusammen zu einem Zeichen a^*, dem wir die Summe von deren Wahrscheinlichkeiten zuordnen. Dank unserer rekursiven Vorgehensweise können wir für dieses Alphabet aus $m-1$ Elementen einen optimalen Code konstruieren; um auch einen für das gegebene Alphabet zu erhalten, kodieren wir die beiden seltensten Zeichen so, daß wir an das Codewort für a^* eine Null

bzw. eine Eins anhängen. Wie wir uns gerade überlegt haben, ist dann auch der neue Code optimal.

Als Beispiel betrachten wir eine Zufallsvariable X mit Werten in einem sechselementigen Alphabet $A = \{a, b, c, d, e, f\}$; die Wahrscheinlichkeiten der sechs Buchstaben seien (in alphabetischer Reihenfolge) $\frac{1}{3}, \frac{1}{4}, \frac{1}{6}, \frac{1}{8}, \frac{1}{12}$ und $\frac{1}{24}$.

Die beiden seltensten Zeichen sind e und f; wir fassen sie also zusammen zu einem Zeichen ef mit Wahrscheinlichkeit $\frac{1}{12} + \frac{1}{24} = \frac{1}{8}$. Im neuen Alphabet $\{a, b, c, d, ef\}$ sind d und ef die beiden seltensten Buchstaben; wir fassen sie also zusammen zu einem neuen Buchstaben $d(ef)$ mit Wahrscheinlichkeit $\frac{1}{8} + \frac{1}{8} = \frac{1}{4}$. Im Alphabet $\{a, b, c, d(ef)\}$ ist c der seltenste Buchstabe; für den zweitseltensten haben wir die Auswahl zwischen b und $d(ef)$, die jeweils mit Wahrscheinlichkeit $\frac{1}{4}$ auftreten und müssen uns für einen der beiden entscheiden. Um Klammern zu sparen fassen wir b und c zusammen zu einem neuen Buchstaben bc mit Wahrscheinlichkeit $\frac{1}{6} + \frac{1}{4} = \frac{5}{12}$. Dies führt auf das Alphabet $\{a, bc, d(ef)\}$, in dem $d(ef)$ und a die beiden seltensten Buchstaben sind; wir fassen sie zusammen zum „Buchstaben“ $a\big(d(ef)\big)$. Damit sind wir bei einem zweibuchstabigen Alphabet angelangt; einer der beiden optimalen Codes besteht darin, daß wir $a\big(d(ef)\big)$ mit Null und bc mit Eins kodieren.

Nun müssen wir nacheinander die HUFFMAN-Reduktionen rückgängig machen. Als erstes wird bc aufgespalten in b mit dem Code 10 und c mit dem Code 11; das Umgekehrte wäre natürlich genauso gut möglich. Sodann wir $a\big(d(ef)\big)$ aufgespalten in a und $d(ef)$ durch die Kodierung a = 00 und $d(ef)$ = 01. Als nächstes wird $d(ef)$ aufgespalten durch d = 010 und ef = 011, und im letzten Schritt setzen wir e = 0110 und f = 0111.

Damit haben wir einen optimalen Code gefunden; seine mittlere Länge ist

$$\frac{1}{3}\cdot 2 + \frac{1}{4}\cdot 2 + \frac{1}{6}\cdot 2 + \frac{1}{8}\cdot 3 + \frac{1}{12}\cdot 4 + \frac{1}{24}\cdot 4 = 2\frac{3}{8} = 2{,}375\,,$$

also nur wenig größer als die Entropie

$$\begin{aligned} H &= \frac{1}{3}\log_2 3 + \frac{1}{4}\log_2 4 + \frac{1}{6}\log_2 6 + \frac{1}{12}\log_2 12 + \frac{1}{24}\log_2 24 \\ &= \frac{4}{3} + \frac{5}{8}\log_2 3 \approx 2{,}324\,. \end{aligned}$$

Falls wir optimale Codes mit einem Zeichensatz von $D > 2$ Elementen suchen, können wir im wesentlichen genauso vorgehen; allerdings können wir nun für jedes Alphabet mit höchstens D Elementen einen Code mit mittlerer Länge eins finden. Bei jeder HUFFMAN-Reduktion außer der ersten fassen wir die D am wenigsten wahrscheinlichen Buchstabe zu einem neuen Buchstaben a^* zusammen; dadurch verringert sich die Elementanzahl dies Alphabets um $D - 1$. Falls $m - 1$ durch $D - 1$ teilbar ist, können wir dies auch im ersten Schritt tun; andernfalls fassen wir dort nur m_0 Elemente zusammen, wobei $2 \le m_0 < D$ so gewählt wird, daß $D - m_0$ durch $D - 1$ teilbar ist.

DAVID ALBERT HUFFMAN (1925-â^1999) studierte Elektrotechnik an der Ohio State University; nachdem er 1944 im Alter von 18 Jahren seinen Bachelor-Abschluß bekommen hatte, verbrachte er den Rest der Kriegszeit als Radaroffizier auf einem Schiff der US Navy. Nach dem Krieg kehrte er an die Ohio State University zurück; nach seinem Master 1949 wechselte er zur Promotion ans MIT. wo er danach von 1953 bis 1957 auch lehrte. Den HUFFMAN-Code entwickelte er dort als Studienarbeit während seines Promotionsstudiums. 1957 gründete er das Cumputer Science Department der Universität von Kalifornien in Santa Cruz, wo er 1994 emeritiert wurde. Die meisten seiner Arbeiten beschäftigen sich mit Informations- und Kodierungstheorie; er befaßte sich beispielsweise mit GAUSSschen Flächen der Krümmung null.

d) Komprimierung durch Dekorrelation

HUFFMAN-Codes sind optimal, wenn wir eine einzige Zufallsvariable betrachten oder, was auf dasselbe hinausläuft, eine Folge von unabhängigen identisch verteilten Zufallsvariablen. Eines der Haupteinsatzgebiete der Datenkompression sind aber Bild- und Audiodaten, bei denen diese Annahme sicherlich nicht erfüllt ist: Bei einer digitalen Tonaufnahme etwa wird der Schalldruck 44 100-mal pro Sekunde gemessen und auf einen Wert zwischen 0 und $2^{24} - 1$ = 16 777 215 oder 0 und $2^{16} - 1$ = 65 535 skaliert. Eine Aufnahme, bei der die Lautstärke 44 100-mal pro Sekunde zufällig wechselt, werden nur wenige Hörer als Lieblingsmusik wählen: Dramatische Wechsel in der Lautstärke sind zwar

ein wichtiges kompositorisches Element, aber es muß sparsam eingesetzt werden. Von den 44 100 Werten, die pro Sekunde aufgezeichnet werden, unterscheidet sich die überwiegende Mehrzahl nur wenig von ihrem Vorgänger.

Ähnlich ist es bei Bilddaten: Selbstverständlich sind abrupte Übergänge auch hier ein oft eingesetztes Stilmittel, aber verglichen mit der Anzahl der Pixel, mit denen Bilder typischerweise digitalisiert werden, unterscheidet sich auch hier der Großteil aller Farb- oder Grauwerte nur wenig von den entsprechenden Werten der Nachbarpixel. Hier werden Grau- oder Farbwerte typischerweise nur mit Werten zwischen 0 und $2^8 - 1 = 256$ kodiert, da unser Auge bei gedruckten oder auf eine Leinwand projizierten Bildern selbst bei nur 64 verschiedenen Werten keine Artefakte mehr erkennen kann, wohingegen unser Gehör noch auf sehr viel feinere Unterschiede reagiert.

In beiden Fällen steckt also ein zumindest im Mittel wesentlicher Teil der Information bereits im Vorgänger; wir können dies dadurch quantifizieren, daß wir die typische Korrelation eines Schalldruck oder Farbwerts mit seinem Vorgänger (oder sonstigen Nachbar) berechnen. Diese Korrelation bezeichnet man als die *Autokorrelation* des stochastischen Prozesses, durch den wir ein typisches Musikstück oder Bild beschreiben.

Auf der folgenden Doppelseite sind einige in Lehrbüchern der Bildverarbeitung beliebte Grauwerttestbilder zu sehen zusammen mit den Daten über minimale, maximale und mittlere Helligkeit $x_{\min}$, $x_{\max}$ und μ, Varianz σ^2, Standardabweichung σ sowie der Autokorrelation ρ. Alle diese aus

> P.M. FARELLE: Recursive Block Coding for Image Data Compression, *Springer,* 1990

entnommenen Daten beziehen sich natürlich auf die Originalbilder und nicht auf das, was Ihr Bildschirm oder Drucker daraus macht. Trotzdem sollte der Vergleich von Bildern und Daten einen einigermaßen korrekten Eindruck zumindest der relativen Situation vermitteln, da hoffentlich alle hier abgedruckten Bilder in derselben Weise verunstaltet sind.

Wie die Daten zeigen, können wir bei der Komprimierung von Bilddaten zumindest ungefähr von einer Autokorrelation um die 95% ausgehen; jeder Wert ist somit im Mittel bereits zu 95% durch seinen Vorgänger bestimmt. Trotzdem übertragen oder speichern wir zumindest mir den uns bislang bekannten Kompressionsverfahren immer wieder 100% der Information einer jeden Zufallsvariablen.

Ein möglicher Ansatz zur Datenkompression wäre daher, daß wir immer nur die Differenz zum Vorgänger übertragen oder speichern: Haben X und Y beide Erwartungswert μ und Varianz σ^2, so hat $Z = Y - X$ den deutlich kleineren Erwartungswert $(1-\rho)\mu$ und nur noch Varianz $2(1-\rho)\sigma^2$. Die Kovarianz zwischen X und Z ist

$$\begin{aligned}\operatorname{Cov}(X,Z) &= \operatorname{Cov}(X, Y-\rho X) = \operatorname{Cov}(X,Y) - \rho\operatorname{Cov}(X,X)\\ &= \rho\sigma^2 - \rho\sigma^2 = 0\,;\end{aligned}$$

X und Z sind also unabhängig voneinander.

Verfahren, die dies ausnutzen, gibt es in der Tat; sie müssen aber mit dem Problem fertig werden, daß die Differenz zwar *meistens* klein ist, daß aber gerade die Ausnahmen, bei denen sie groß ist, sehr wesentlich für die Rekonstruktion des Original sind.

Der am häufigsten verwendete Ansatz geht daher anders vor: Die „Nachricht“ wird in Blöcke aufgeteilt; bei der Schallaufzeichung auf CD sind dies bei den derzeit auf dem Massenmarkt erhältlichen Geräten Blöcke zu je acht Werten, bei Bildern sind es Teilbilder von jeweils 8×8 Pixel.

Anstelle der einzelnen Zufallsvariablen kodieren wir die Blöcke; wie wir bereits mehrfach gesehen haben, läßt sich allein dadurch der mittlere Aufwand meistens reduzieren, wenn auch nicht so dramatisch wie bei den hier betrachteten Komprimierungsverfahren.

Der Einfachheit halber beschränken wir uns auf ein eindimensionales Modell, mit dem wir es beispielsweise bei Audiodaten zu tun haben. Wir betrachten Blöcke aus einer gewissen Anzahl n von Zufallsvariablen, die allesamt dasselbe in $\mathbb{R}$ enthaltene Alphabet und dieselbe Wahrscheinlichkeitsverteilung haben. Dabei nehmen wir an, daß jede der Zufallsvariable mit ihrem Nachbarn die Korrelation ρ habe.

Peppers

μ = 115,6
σ^2 = 5632
σ = 75,0
ρ = 0,98
$x_{\min}$ = 0
$x_{\max}$ = 237

Lenna

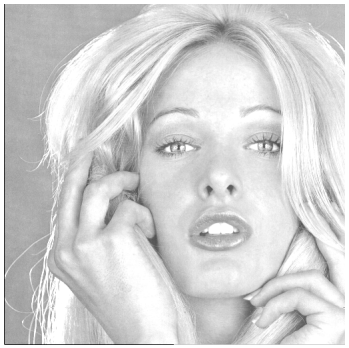
μ = 99,1
σ^2 = 2796
σ = 52,9
ρ = 0,97
$x_{\min}$ = 3
$x_{\max}$ = 248

Sailboat

μ = 124,3
σ^2 = 6027
σ = 77,6
ρ = 0,97
$x_{\min}$ = 0
$x_{\max}$ = 249

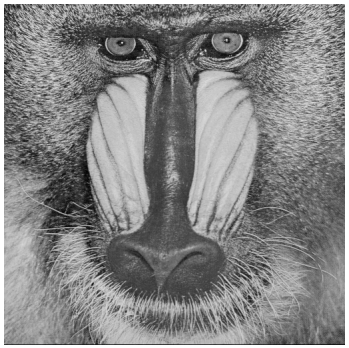
Stream

μ = 113,8
σ^2 = 2996
σ = 54,7
ρ = 0,94
$x_{\min}$ = 0
$x_{\max}$ = 255



Tiffany

μ = 208,6
σ^2 = 1126
σ = 33,6
ρ = 0,87
$x_{\min}$ = 3
$x_{\max}$ = 255



Baboon

μ = 128,9
σ^2 = 2282
σ = 47,8
ρ = 0,86
$x_{\min}$ = 0
$x_{\max}$ = 236

Als erstes sollten wir uns fragen, wie es mit der Korrelation zwischen weiter entfernten Variablen aussieht. Dazu können wir *a priori* nichts sagen; wenn wir im Extremfall nur zwei Zufallsvariablen mit Korrelation ρ haben, so daß alle Folgenglieder mit geradem Index gleich der einen und alle übrigen gleich der anderen sind, ist die Korrelation gleich ρ, wenn sich die Indizes um eine ungerade Zahl unterscheiden, und eins sonst. Dieser Fall wird freilich bei Bild- und Audiodaten kaum vorkommen.

Dort geht man üblicherweise aus vom sogenannten ar(1)-Modell, wonach – in Analogie zu MARKOV-Ketten – alle Abhängigkeiten ausschließlich auf die zwischen unmittelbaren Nachbarn zurückzuführen sind, so daß die Korrelation zwischen zwei Zufallsvariablen gleich ρ hoch Betrag der Indexdifferenz ist.

Beim blockweisen Kodieren nach diesem Modell betrachten wir somit einen Vektor $(X_1, \dots, X_n)$ von Zufallsvariablen; alle X_i haben denselben Erwartungswert μ und dieselbe Varianz σ^2; die Korrelation zwischen X_i und X_j ist $\rho^{|i-j|}$. Die Korrelationsmatrix, bei der an der Stelle ij dieser Wert steht, ist somit die symmetrische Matrix

$$\begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{pmatrix};$$

die Kovarianzmatrix ist das σ^2-fache davon.

Wir betrachten nun anstelle der Zufallsvariablen X_i neue Variablen

$$Y_i = \sum_{j=1}^{n} a_{ij} X_j$$

mit zunächst irgendwelchen reellen Zahlen a_{ij}. Dann ist

$$\begin{aligned} \operatorname{Cov}(Y_i, Y_k) &= \operatorname{Cov}\left(\sum_{j=1}^{n} a_{ij} X_j, \sum_{\ell=1}^{n} a_{k\ell} X_\ell \right) \\ &= \sum_{j=1}^{n} \sum_{\ell=1}^{n} a_{ij} \operatorname{Cov}(X_j, X_\ell) a_{k\ell} \,. \end{aligned}$$

Stünde in der letzten Summe ganz rechts $a_{\ell k}$ anstelle von $a_{k\ell}$, so wäre die Summe gerade der ik-Eintrag des Produkts A mal Kovarianzmatrix der X_i mal A, wobei A die Matrix mit Einträgen a_{ij} bezeichnet. Da tatsächlich $a_{k\ell}$ dasteht, müssen wir als letzten Faktor des Matrixprodukts die transponierte Matrix A^T anstelle von A nehmen; die Kovarianzmatrix der neuen Zufallsvariablen ist also

$$\operatorname{Cov}(Y_1, \dots, Y_n) = A \operatorname{Cov}(X_1, \dots, X_n) A^T \,.$$

Als nächstes beachten wir, daß die Kovarianzmatrix der X_i symmetrisch ist; nach dem (im Anhang zu diesem Paragraphen bewiesenen) Spektralsatz gibt es daher eine Orthonormalbasis des $\mathbb{R}^n$, bezüglich derer sie Diagonalgestalt hat. Es gibt also eine Matrix A, so daß $A \operatorname{Cov}(X_1, \dots, X_n) A^{-1}$ eine Diagonalmatrix ist, und da die Matrix A zu einem Basiswechsel zwischen zwei Orthonormalbasen gehört, ist AA^T die Einheitsmatrix, d.h. $A^{-1} = A^T$.

Definieren wir daher $Y_i = \sum a_{ij} X_j$ mit den Einträgen dieser Matrix A, so ist $\operatorname{Cov}(Y_1, \dots, Y_n)$ eine Diagonalmatrix; die verschiedenen Y_i sind also voneinander unabhängige Zufallsvariablen.

Unter den Annahmen des ar(1)-Modells können wir somit jede Folge von Zufallsvariablen durch eine lineare Transformation in eine Folge unkorrelierter Zufallsvariablen überführen. Diese Transformation bezeichnet man, obwohl sie zuerst von HOTELLING vorgeschlagen wurde, als KARHUNEN-LOÈVE-Transformation.

HAROLD HOTELLING (1895–1973) war ein amerikanischer Statistiker und Ökonom; er lehrte an der Columbia University und der University of North Carolina. In einer 1933 veröffentlichten Arbeit im *Journal of Educational Psychology* schlug er erstmalig diese Transformation vor, die von Statistikern heute in Anlehnung an den Titel seiner Arbeit meist als *Hauptkomponentenanalyse* bezeichnet wird. In Europa erschien die Transformation fast gleichzeitig um 1947 *bzw.* 1948 in wahrscheinlichkeitstheoretischen Arbeiten des Finnen KARI KARHUNEN (1915–1992) und des Franzosen MICHEL LOÈVE (1907–1979), nach denen sie in der technischen Literatur benannt wird.

Die Matrix A der linearen Transformation hängt nur von ρ ab und kann daher für gängige Werte von ρ vorberechnet werden; die KARHUNEN-LOÈVE-Transformation besteht somit einfach in der Multiplikation mit einer bekannten Matrix. Da die Abhängigkeit von ρ im hier relevanten Bereich relativ schwach ist, verschlechtern sich die Ergebnisse kaum, wenn man sich dabei auf *ein* typisches ρ beschränkt, etwa auf $\rho = 0{,}95$.

Trotzdem wird die KARHUNEN-LOÈVE-Transformation praktisch nie verwendet, denn durch einen anderen Ansatz kommt man mit deutlich geringerem Aufwand auf fast dasselbe Ergebnis. Um zu sehen, wie dieser Funtkioniert, betrachten wir die für Komprimierungsverfahren in der Unterhaltungselektronik typischen Werte $n = 8$ und $\rho = 0{,}95$. Die Eigenvektoren der Kovarianzmatrix

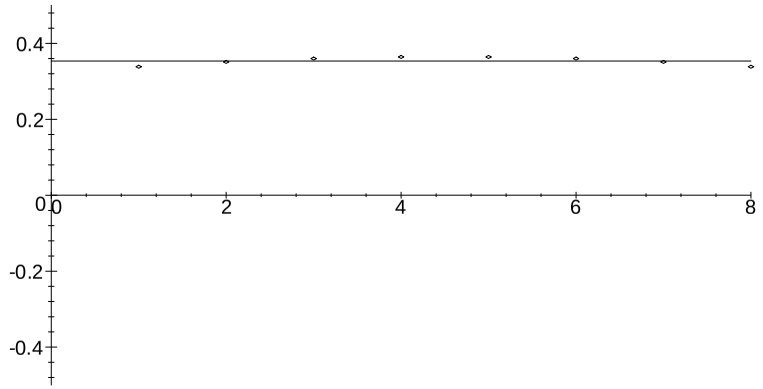
$$\mathrm{Cov}(X_1, \ldots, X_8) = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 & \rho^5 & \rho^6 & \rho^7 \\ \rho & 1 & \rho & \rho^2 & \rho^3 & \rho^4 & \rho^5 & \rho^6 \\ \rho^2 & \rho & 1 & \rho & \rho^2 & \rho^3 & \rho^4 & \rho^5 \\ \rho^3 & \rho^2 & \rho & 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^5 & \rho^4 & \rho^3 & \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^6 & \rho^5 & \rho^4 & \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^7 & \rho^6 & \rho^5 & \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}$$

können dann zumindest numerisch leicht bestimmt werden; da man acht Vektoren mit jeweils acht Einträgen wenig Struktur ansehen kann, habe ich die Ergebnisse in den folgenden acht Zeichnungen graphisch dargestellt: Die eingezeichneten Punkte sind (i, x_i) für $i = 1, \ldots, 8$, wobei x_i jeweils die i-te Komponente des auf Länge eins normierten Eigenvektors bezeichnet. Zusätzlich ist in der j-ten Zeichnung noch die Kurve

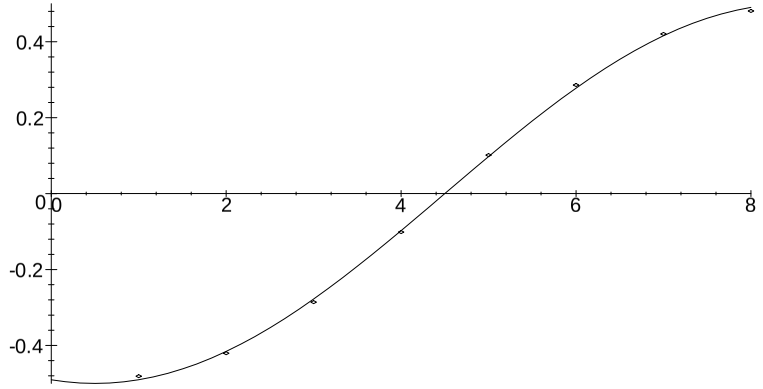
$$y = \cos\left(\frac{(2x-1)(j-1)\pi}{16}\right)$$

eingezeichnet; wie man sieht, liegen die Punkte (i, x_i) zwar nicht exakt auf diesen Kurven, aber doch sehr in deren Nähe. Entsprechendes gilt auch für andere Werte von n und andere Korrelationskoeffizienten ρ nahe eins.

Der Grund dafür, daß man in der Praxis lieber mit den so durch Kosinuswerte angenäherten Basisvektoren arbeitet, liegt nicht darin, daß

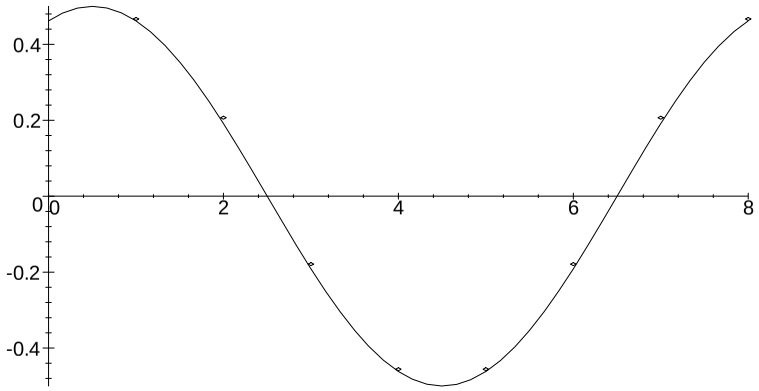


Der erste Eigenvektor der Korrelationsmatrix

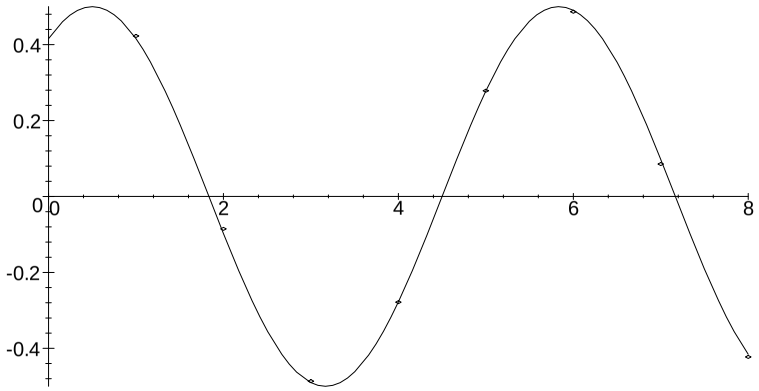


Der zweite Eigenvektor der Korrelationsmatrix

diese einfacher zu berechnen sind – die Eigenvektoren sind schließlich Konstanten des Komprimierungsverfahren. Für die Transformation eines Blocks in die neue Basis brauchen wir aber eine Matrix-Vektor-Multiplikation, d.h. n^2 Multiplikationen sowie $n(n-1)$ Additionen reeller Zahlen. Wenn wir stattdessen mit den durch Kosinuswerte gegebenen Vektoren arbeiten, können wir eine sogenannte schnelle FOURIER-



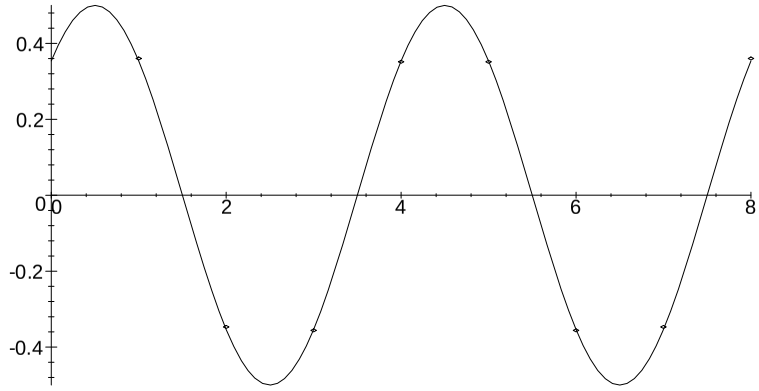
Der dritte Eigenvektor der Korrelationsmatrix



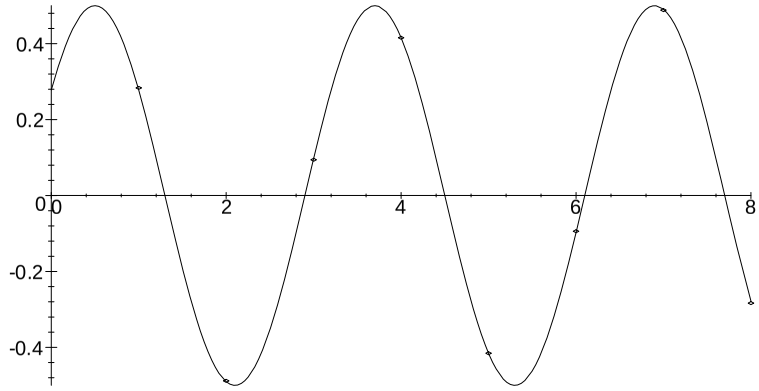
Der vierte Eigenvektor der Korrelationsmatrix

bzw. Kosinustransformation anwenden, die nur ungefähr $n \log_2 n$ Multiplikationen benötigt. Ihre Dekorrelationseffiziens liegt im Fall $n = 8$ und $\rho \approx 0{,}95$ bei etwa 98%; wir verlieren also fast nichts im Vergleich zur aufwendigeren KARHUNEN-LOÈVE-Transformation.

Schnelle FOURIER- und Kosinustransformationen gibt es auch in höheren Dimensionen; insbesondere können wir im Zweidimensionalen Blöcke



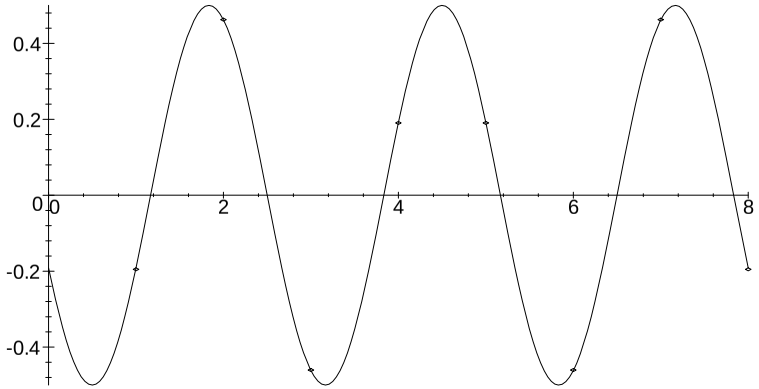
Der fünfte Eigenvektor der Korrelationsmatrix



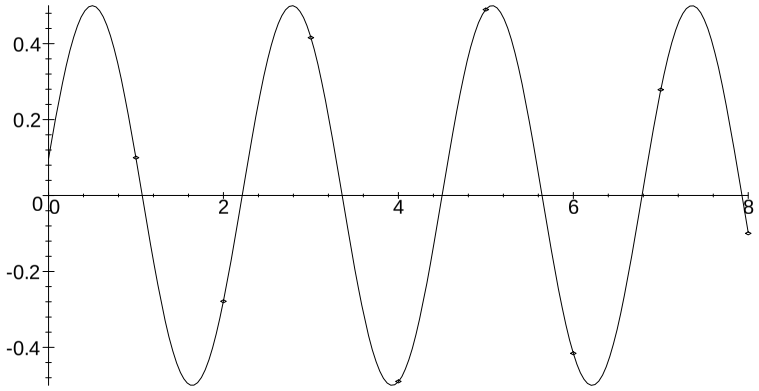
Der sechste Eigenvektor der Korrelationsmatrix

von 8×8 Pixel schnell so in eine neue Basis des 64-dimensionalen Raums transformieren, daß sie praktisch unabhängig voneinander sind.

Die diskrete Kosinustransformation ist Teil fast aller gängiger Normen zur Bildkomprimierung: Sowohl der JPEG-Standard für Photographien, die Standards MPEG 1 und 2 für digitale (Unterhaltungs-)Videos als auch der Standard CCITT H.261 für Videokonferenzen enthalten (ne-



Der siebte Eigenvektor der Korrelationsmatrix



Der achte Eigenvektor der Korrelationsmatrix

ben anderen Bestandteilen) jeweils eine diskrete Kosinustransformation; auch im mp3-Standard ist sie ein Teil der Codierung.

Die Transformation allein ist natürlich noch keine Komprimierung: Schließlich haben wir nur einen Vektor in einer anderen Basis hingeschrieben, und die Anzahl der reellen Zahlen, die man zur Beschreibung eines solchen Vektors benötigt, ist unabhängig von der Basis. Der

wesentliche Vorteil der neuen Basis ist, daß man statistisch recht gute Aussagen über die Größe der Komponenten machen kann. Hier wollen wir auf exakte statistische Berechnungen verzichten und stattdessen informell diskutieren, warum dies der Fall sein könnte.

Wie die Abbildungen der Basisvektoren zur KARHUNEN-LOÈVE-Transformation und die Formeln für die Basisvektoren zur diskreten Kosinustransformation zeigen, werden die Basisvektoren, wenn man sie in der hier angegebenen Reihenfolge betrachtet, immer hochfrequenter. Von einem hinreichend fein abgetasteten Bild- oder Audiosignal erwarten wir, daß hochfrequente Schwankungen keine große Rolle spielen und somit die entsprechenden Basisvektoren nur kleine Koeffizienten haben oder in vielen Fällen sogar gleich gar nicht auftreten. Dementsprechend genügt es, für die Übertragung dieser Koeffizienten nur wenige Bits bereitzustellen; bei nur geringen Abstrichen an die Qualität kann man auf gewisse Koeffizienten sogar ganz verzichten.

Ein Kompressionsverfahren wird daher, je nach Anspruch an die Qualität, entweder alle Koeffizienten des Signals in der neuen Basis übertragen und durch eine geeignete Darstellung der Daten dafür sorgen, daß Folgen von Nullen nur wenig Platz benötigen, oder aber es wird nur eine Auswahl der Koeffizienten übertragen und auch für diese jeweils festlegen, wie viele Bit dafür in Anspruch genommen werden. Diese Anzahl wird umso geringer sein, je höher die Frequenz des jeweiligen Basisvektors ist; bei einigen Verfahren wie etwa JPEG können die Anzahlen auch variabel in Abhängigkeit von einer Qualitätszahl gewählt werden.

Zum Schluß sei noch ganz kurz erwähnt, daß die KARHUNEN-LOÈVE-Transformation und damit (mit ganz geringen Abstrichen) auch die diskrete Kosinustransformation zwar die Korrelationsmatrix in optimaler Weise diagonalisieren, daß aber daraus nicht folgt, daß sie auch optimale Kompressionsverfahren liefern: Ausßer der Kovarianz gibt es noch weitere Quellen für Redundanz eines Bildes.

Ein gewisser Nachteil der Kosinustransformation ist außerdem, daß man für abrupte Übergänge, wie sie etwa bei Kanten immer wieder einmal auftauchen, die hochfrequenten Basisvektoren braucht, die dann aber

nicht nur die Kante selbst beeinflussen, sondern das gesamte Quadrat, auf das die Transformation angewandt wird.

Eine bessere Möglichkeit wäre es daher, wenn man anstelle von Kosinusfunktionen Funktionen verwenden könnte, die sowohl im Zeit- als auch im Frequenzbereich lokalisiert sind. Solche Funktionen gibt es in der Tat, etwa die sogenannten *Wavelets*. Hierbei handelt es sich um schnell abklingende Wellen, und neuere Arbeiten deuten darauf hin, daß diese für gewisse Bildmodelle (die im Gegensatz zum hier betrachteten nicht mit Wahrscheinlichkeiten arbeiten) nicht zu weit vom Optimum entfernt sein sollten. Im Rahmen dieser Vorlesung ist es jedoch zeitlich weder möglich, auf diese Modelle einzugehen, noch ist an eine genauere Behandlung von Wavelets zu denken.

Einen allgemein verständlichen Überblick über Wavelets findet man etwa bei

BARBARA BURKE HUBBARD: Wavelets: Die Mathematik der kleinen Wellen, *Birkhäuser* 1997;

das zitierte Optimalitätsresultat ist beschrieben im Vortrag

STÉPHANE MALLAT: Applied Mathematics meets signal processing

auf dem Internationalen Mathematikerkongress 1998 in Berlin, nachzulesen in Band I der Proceedings, S. 319–338, oder unter http://www.mathematik.uni-bielefeld.de/documenta/xvol-icm/00/Mallat.MAN.html.

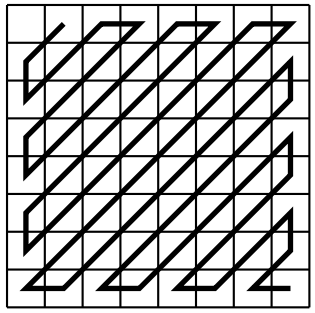
e) Datenkomprimierung mit JPEG

Als praktisches Beispiel eines Komprimierungsalgorithmus wollen wir zumindest kurz des JPEG-Standard der *Joint Photographers Expert Group* betrachten, der in vielen Digitalkameras verwendet wird.

Er beginnt damit, das Bild in Blöcke von 8×8 Pixel aufzuteilen und jeden wie im vorigen Abschnitt beschrieben einer Kosinustransformation zu unterziehen. Das Ergebnis ist eine neue 8×8-Matrix reeller Zahlen, die zunächst quantisiert, d.h. auf Werte aus einer diskreten Menge gerundet werden. Größe und Aufbau dieser Menge hängen sowohl von der Position in der Matrix als von einem wählbaren Qualitätsfaktor ab.

Der Matrixeintrag links oben entspricht dem Basisvektor, dessen sämtliche Einträge gleich sind; dort steht also der Mittelwert der Einträge der ursprünglichen 8×8-Matrix. Er wird (abgesehen natürlich vom ersten Block) gespeichert als die Differenz vom Mittelwert des Vorgängerblocks; Vorgänger bezieht sich dabei auf die zeilenweise Anordnung.

Die übrigen 63 Einträge eines jeden Blocks werden mäanderförmig nach dem Schema im nächsten Bild durchlaufen:



Die so erhaltene Folge von 63 Werten wird meist viele Nullen enthalten; gespeichert werden daher nur die von Null verschiedenen Werte zusammen mit der Anzahl der Nullfelder vor so einem Wert.

Im letzten Schritt schließlich wird die gesamte so erhaltene Zeichenfolge HUFFMAN-kodiert.

f) Anhang: Der Spektralsatz

Für diejenigen Leser, die nicht mit Eigenwerten und Eigenvektoren symmetrischer Matrizen vertraut sind, seien hier die entsprechenden Sätze bewiesen. Um die Notation festzulegen, beginne ich mit der Definition von Eigenwerten und Eigenvektoren:

a) Eigenwerte und Eigenvektoren: Die Matrix einer linearen Abbildung $\varphi\colon V \to V$ bezüglich einer Basis $\mathcal{B} = (b_1, \ldots, b_n)$ ist genau dann eine Diagonalmatrix, wenn jeder der Basisvektoren b_i von φ auf ein Vielfaches $\lambda_i b_i$ von sich selbst abgebildet wird; alsdann ist die Abbildungsmatrix gleich der Diagonalmatrix mit Einträgen $\lambda_1, \ldots, \lambda_n$. Somit

hängt es sehr von der Basis ab, ob die Abbildungsmatrix Diagonalgestalt hat oder nicht.

Die Matrix

$$A = \begin{pmatrix} 1 & 2 & 1 & -2 \\ 2 & 1 & -2 & 1 \\ 1 & -2 & 1 & 2 \\ -2 & 1 & 2 & 1 \end{pmatrix} \in \mathbb{R}^{4\times 4}$$

etwa ist ganz sicher keine Diagonalmatrix. Betrachten wir die lineare Abbildung

$$\varphi\colon \mathbb{R}^4 \to \mathbb{R}^4; \quad v \mapsto Av$$

aber bezüglich der Basis $\mathcal{B} = (b_1, b_2, b_3, b_4)$ mit

$$b_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \quad b_2 = \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix} \quad b_3 = \begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \end{pmatrix} \quad \text{und} \quad b_4 = \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix},$$

so rechnet man leicht nach, daß

$$\varphi(b_1) = Ab_1 = 2b_1, \quad \varphi(b_2) = 2b_2, \quad \varphi(b_3) = -4b_3 \quad \text{und} \quad \varphi(b_4) = 4b_4$$

ist, bezüglich $\mathcal{B}$ hat φ also die Diagonalmatrix

$$D = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & -4 & 0 \\ 0 & 0 & 0 & 4 \end{pmatrix}$$

als Abbildungsmatrix. Wenn keine schwerwiegenden anderen Gründe dagegen sprechen, wird es bei umfangreichen Rechnungen mit der Matrix A meist eine gute Idee sein, statt mit der Standardbasis von $\mathbb{R}^4$ mit der Basis $\mathcal{B}$ zu rechnen. So ist beispielsweise die Berechnung der inversen Matrix von D eine einfache Kopfrechenaufgabe, wohingegen die Berechnung von

$$A^{-1} = \frac{1}{8}\begin{pmatrix} 2 & 1 & 2 & -1 \\ 1 & 2 & -1 & 2 \\ 2 & -1 & 2 & 1 \\ -1 & 2 & 1 & 2 \end{pmatrix}$$

doch einiges an Aufwand erfordert. Genauso ist die Berechnung von

$$A^{10} = \begin{pmatrix} 524800 & 0 & -523776 & 0 \\ 0 & 524800 & 0 & -523776 \\ -523776 & 0 & 524800 & 0 \\ 0 & -523776 & 0 & 524800 \end{pmatrix}$$

mit erheblicher Arbeit verbunden, während man für die von D^{10} nur wissen muß, daß $2^{10} = 1\,024$ und $2^{20} = 1\,048\,576$ ist.

Definition: Ein Vektor $v \in k^n \setminus \{0\}$ heißt *Eigenvektor* der Matrix $A \in k^{n\times n}$, wenn es eine Zahl $\lambda \in k$ gibt, so daß $Av = \lambda v$ ist. Dieses λ bezeichnen wir als einen *Eigenwert* von A.

Der seltsame Name *Eigenwert* läßt sich vielleicht am besten verstehen, wenn man seine Anwendung in der Quantenmechanik betrachtet: Dort werden *Observable*, d.h. physikalische Meßgrößen, durch Matrizen beschrieben und Zustände durch Vektoren. Die möglichen Ergebnisse einer Messung sind die Eigenwerte der zugehörigen Matrix, und nach der Messung ist der Zustand des Systems ein Eigenvektor zum gemessenen Eigenwert. Die in der Quantenmechanik auftretenden Matrizen sind allesamt so, daß es eine Basis aus Eigenvektoren gibt.

Falls es also zu einer Matrix A eine Basis aus Eigenvektoren gibt, können wir sie also bezüglich dieser Basis als Diagonalmatrix darstellen.

Offensichtlich ist mit einem Vektor v auch jedes Vielfache (außer dem nach Definition ausgeschlossenen Nullvektor) ein Eigenvektor zum selben Eigenwert; allgemeiner ist sogar jede Linearkombination (außer 0) von Eigenvektoren zum Eigenwert λ wieder ein Eigenvektor zum Eigenwert λ, d.h. die Eigenvektoren zu einem festen Eigenwert λ bilden zusammen mit dem Nullvektor einen Untervektorraum von V, den sogenannten *Eigenraum* von λ.

Definition: Die Dimension des Eigenraums von λ heißt *geometrische Vielfachheit* des Eigenwerts λ.

Lemma: Sind $v_1, \ldots, v_r \in V$ Eigenvektoren der Matrix A zu verschiedenen Eigenwerten $\lambda_1, \ldots, \lambda_r$, so sind diese Vektoren linear unabhängig.

Beweis: Angenommen, $v_1, \ldots, v_r$ seien linear abhängig. Dann können wir eine Zahl $2 \leq s \leq r$ finden, so daß zwar $v_1, \ldots, v_s$ linear abhängig sind, nicht aber $v_1, \ldots, v_{s-1}$. Es gibt daher Skalare $\alpha_i \in k$, so daß

$$\alpha_1 v_1 + \cdots + \alpha_s v_s = 0$$

ist. Wenden wir beide Seiten dieser Gleichung mit A multiplizieren und beachten, daß $Av_i = \lambda_i v_i$ ist, folgt, daß auch

$$\alpha_1 \lambda_1 v_1 + \cdots + \alpha_s \lambda_s v_s = 0$$

ist. Andererseits können wir obige Gleichung auch einfach mit λ_s multiplizieren mit dem Ergebnis

$$\lambda_s \alpha_1 v_1 + \cdots + \lambda_s \alpha_s v_s = 0\,.$$

Durch Subtraktion der letzten beiden Gleichungen voneinander erhalten wir eine lineare Abhängigkeit

$$\alpha_1(\lambda_s - \lambda_1)v_1 + \cdots + \alpha_{s-1}(\lambda_s - \lambda_{s-1})v_{s-1} = 0$$

zwischen $v_1, \ldots, v_{s-1}$. Da diese Vektoren linear unabhängig sind, müssen alle Koeffizienten verschwinden. Da die Eigenwerte $\lambda_1, \cdots, \lambda_s$ aber allesamt verschieden sind, ist dies nur möglich, wenn α_1 bis α_{s-1} verschwinden. Wegen $v_s \neq 0$ muß dann aber auch α_s verschwinden, im Widerspruch zur angenommenen linearen Unabhängigkeit von $v_1, \ldots, v_s$.

■

b) Vielfachheiten von Eigenwerten: Ist x eine Nullstelle eines Polynoms $f(X)$, so kann $f(X)$ bekanntlich durch $(X - x)$ geteilt werden, und x heißt *r-fache Nullstelle* von $f(X)$, wenn $f(X)$ durch $(X - x)^r$ teilbar ist, nicht aber durch $(X - x)^{r+1}$.

Definition: Wir sagen, der Eigenwert λ von A habe die *algebraische Vielfachheit* r, wenn λ eine r-fache Nullstelle des charakteristischen Polynoms $\det(A - \lambda E)$ ist.

Im obigen Beispiel hatte also der Eigenwert Null die algebraische Vielfachheit zwei, die anderen beiden hatten algebraische Vielfachheit eins.

Die Dimension des jeweiligen Eigenraums, die geometrische Vielfachheit also, war genauso groß, jedoch muß dies im allgemeinen nicht der Fall sein: Für die Matrix

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

etwa hat das charakteristische Polynom

$$\det(A - \lambda E) = \begin{vmatrix} 1-\lambda & 1 \\ 0 & 1-\lambda \end{vmatrix} = (1-\lambda)^2$$

die doppelte Nullstelle eins, $\lambda = 1$ ist also ein Eigenwert mit algebraischer Vielfachheit zwei. Der zugehörige Eigenraum ist die Lösungsmenge des linearen Gleichungssystems

$$\begin{aligned} 0x_1 + 1x_2 &= 0 \\ 0x_1 + 0x_2 &= 0\,, \end{aligned}$$

also gerade die Menge aller Vektoren der Form $\binom{x}{0}$ und somit eindimensional. Die geometrische Vielfachheit des Eigenwerts eins ist daher nur eins.

Das Beispiel der Abbildung

$$\varphi\colon \mathbb{R}^2 \to \mathbb{R}^2; \quad \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} x\cos\vartheta - y\sin\vartheta \\ y\cos\vartheta + x\sin\vartheta \end{pmatrix}$$

mit Abbildungsmatrix

$$A = \begin{pmatrix} \cos\vartheta & -\sin\vartheta \\ \sin\vartheta & \cos\vartheta \end{pmatrix} \in \mathbb{R}^{2\times 2}$$

zeigt, daß es überhaupt keine Eigenwerte geben muß, denn hier ist das charakteristische Polynom gleich

$$\begin{vmatrix} \cos\vartheta - \lambda & -\sin\vartheta \\ \sin\vartheta & \cos\vartheta - \lambda \end{vmatrix} = (\cos\vartheta - \lambda)^2 + \sin^2\vartheta\,.$$

Abgesehen vom Fall $\sin\vartheta = 0$, wenn A gleich der positiven oder negativen Einheitsmatrix ist, hat dieses Polynom keine reelle Nullstelle, da es nur positive Werte annimmt. Es hat aber natürlich die beiden komplexen Nullstellen

$$\lambda_{1/2} = \cos\vartheta \pm i\sin\vartheta = e^{\pm i\vartheta}\,;$$

fassen wir φ als Abbildung von $\mathbb{C}^2$ nach $\mathbb{C}^2$ auf, gibt es also zwei Eigenwerte. Beide haben die algebraische und geometrische Vielfachheit eins; zugehörige Eigenvektoren sind etwa $\binom{1}{i}$ und $\binom{1}{-i}$. Wählen wir diese beiden Vektoren als Basis, so wird die Abbildungsmatrix von φ bezüglich dieser neuen Basis zur Diagonalmatrix

$$\begin{pmatrix} e^{i\vartheta} & 0 \\ 0 & e^{-i\vartheta} \end{pmatrix} .$$

Allgemein gilt für die algebraischen und geometrischen Vielfachheiten von Eigenvektoren

Satz: *a)* Die geometrische Vielfachheit eines Eigenwerts ist stets kleiner oder gleich der algebraischen Vielfachheit.
b) Die Summe der algebraischen Vielfachheiten der verschiedenen Eigenwerte einer linearen Abbildung ist kleiner oder gleich der Dimension des Vektorraums.

Beweis: a) Der Eigenwert λ der $n \times n$-Matrix A habe die geometrische Vielfachheit r, d.h. der zugehörige Eigenraum habe die Dimension r. Wir wählen eine Basis $b_1, \ldots, b_r$ dieses Eigenraums und ergänzen sie zu einer Basis des gesamten Vektorraums; bezüglich dieser Basis sei C die Abbildungsmatrix der linearen Abbildung

$$\varphi: \begin{cases} k^n \to k^n \\ v \mapsto Av \end{cases} .$$

Da $b_1, \ldots, b_r$ Eigenvektoren zum Eigenwert λ sind, ist $\varphi(b_i) = \lambda b_i$. In den ersten r Spalten von C steht also jeweils in der Diagonalen das Element λ und ansonsten überall die Null. A hat somit die Form

$$A = \begin{pmatrix} \lambda & 0 & \ldots & 0 & * & \ldots & * \\ 0 & \lambda & \ldots & 0 & * & \ldots & * \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda & * & \ldots & * \\ & & & & & & \\ 0 & 0 & \ldots & 0 & & & \\ \vdots & \vdots & \ddots & \vdots & & M & \\ 0 & 0 & \ldots & 0 & & & \end{pmatrix} ,$$

wobei uns weder die mit $*$ bezeichneten Körperelemente noch die $(n-r) \times (n-r)$-Matrix M weiter zu interessieren brauchen.

Für $C - xE$ gilt dasselbe, nur daß jetzt $\lambda - x$ in der Diagonalen steht, d.h. diese Matrix hat die Form

$$\begin{pmatrix} \lambda - x & 0 & \ldots & 0 & * & \ldots & * \\ 0 & \lambda - x & \ldots & 0 & * & \ldots & * \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda - x & * & \ldots & * \\ & & & & & & \\ 0 & 0 & \ldots & 0 & & & \\ \vdots & \vdots & \ddots & \vdots & & M - xE_{n-r} & \\ 0 & 0 & \ldots & 0 & & & \end{pmatrix} ,$$

wobei E_{n-r} die $(n-r) \times (n-r)$-Einheitsmatrix bezeichnet.

Zur Berechnung ihrer Determinanten verwenden wir den LAPLACEschen Entwicklungssatz: Da in der ersten Zeile (oder Spalte) nur an der ersten Stelle ein von Null verschiedener Eintrag steht, ist diese Determinante gleich $(\lambda - x)$ mal der Determinante jener Matrix, die durch Streichen der ersten Zeile und Spalte entsteht. Falls $r > 1$ ist, hat diese neue Matrix dieselbe Form, wir können den LAPLACEschen Entwicklungssatz also noch einmal anwenden *usw.;* wir erhalten schließlich

$$\det(C - xE) = (\lambda - x)^r \det(M - xE_{n-r}) .$$

Somit ist $\det(C - xE)$ durch $(x - \lambda)^r$ teilbar.

Was uns wirklich interessiert, ist aber nicht $\det(C - xE)$, sondern $\det(A - xE)$. Ist B die Matrix des Basiswechsels von der Standardbasis des k^n auf die Basis $\{b_1, \ldots, b_n\}$, jene Matrix also, deren Spaltenvektoren die b_i sind, so ist $C = B^{-1}AB$ und

$$\begin{aligned} \det(C - xE) &= \det(B^{-1}AB - xE) = \det(B^{-1}AB - xB^{-1}EB) \\ &= \det\big(B(A - xE)B^{-1}\big) = \det B \det(A - xE)(\det B)^{-1} \\ &= \det(A - xE) . \end{aligned}$$

A und C haben also dasselbe charakteristische Polynom, und somit ist auch das charakteristische Polynom von A durch $(x - \lambda)^r$ teilbar. Die algebraische Vielfachheit von λ ist daher mindestens r.

Unabhängig von diesem Ergebnis wollen wir noch festhalten, daß nach der gerade durchgeführten Rechnung für eine beliebige Matrix A und eine invertierbare Matrix B die beiden Matrizen A und BAB^{-1} dasselbe charakteristische Polynom haben; insbesondere haben also die Abbildungsmatrizen einer linearen Abbildung zu verschiedenen Basen dasselbe charakteristische Polynom.

b) Sind $\lambda_1, \ldots, \lambda_\ell$ die verschiedenen Eigenwerte von φ und sind $r_1, \ldots, r_\ell$ ihre algebraischen Vielfachheiten, so ist das charakteristische Polynom $\det(A - xE)$ teilbar durch

$$(x - \lambda_1)^{r_1} \cdots (x - \lambda_\ell)^{r_\ell} .$$

Dies ist ein Polynom vom Grad $r_1 + \cdots + r_\ell$, wohingegen das charakteristische Polynom Grad n hat; daher ist

$$r_1 + \cdots + r_\ell \leq n ,$$

denn der Grad eines Teilers kann nicht größer sein als der des Polynoms selbst.

■

c) Eigenwerte symmetrischer und Hermitescher Matrizen: Wie wir im letzten Paragraphen gesehen haben, kann die geometrische Vielfachheit eines Eigenwerts kleiner sein als die algebraische, und im Falle einer reellen Matrix müssen nicht auch die Eigenwerte reell sein. In diesem Abschnitt wollen wir sehen, daß solche Dinge bei symmetrischen (und auch den noch zu definierenden HERMITEschen) Matrizen nicht möglich sind.

a) Für eine Matrix $A = (a_{ij}) \in \mathbb{C}^{n\times n}$ bezeichnen wir die Matrix $\overline{A} = \left(\overline{a_{ij}}\right)$ als die zu A *konjugiert komplexe Matrix.*
b) $A \in \mathbb{C}^{n\times n}$ heißt HERMITEsch, falls $A^T = \overline{A}$ ist.
c) Zu einem Vektor $v = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$ heißt $\overline{v} = \begin{pmatrix} \overline{v_1} \\ \vdots \\ \overline{v_n} \end{pmatrix}$ der *konjugiert komplexe* Vektor.

Schließlich wollen wir Vektoren hier mit $1 \times n$-Matrizen identifizieren; insbesondere rechnen wir mit dem „transponierten Vektor“

$$v^T = (v_1, \ldots, v_n) .$$

Mit dieser Bezeichnung kann das Standardskalarprodukt zweier Vektoren $v, w \in \mathbb{R}^n$ als Matrixprodukt $v^T w$ geschrieben werden; das Standard-HERMITEsche Produkt in $\mathbb{C}^n$ ist entsprechend $v^T\, \overline{w}$.

Da die komplexe Konjugation auf $\mathbb{R}$ keine Wirkung hat, ist eine HERMITEsche Matrix mit reellen Einträgen einfach eine symmetrische Matrix; wir können uns im folgenden bei den Beweisen daher auf HERMITEsche Matrizen beschränken und erhalten trotzdem Ergebnisse, die auch für reelle symmetrische Matrizen gelten.

Das Hauptziel dieses Abschnitts ist

Satz: A sei eine symmetrische reelle oder HERMITEsche (komplexe) Matrix.
a) Dann sind alle Eigenwerte von A reell.
b) Eigenvektoren zu verschiedenen Eigenwerten sind orthogonal bezüglich des Standard- *bzw.* HERMITEschen Skalarprodukts.
c) Für jeden Eigenwert von A ist die geometrische Vielfachheit gleich der algebraischen Vielfachheit.
d) $\mathbb{R}^n$ *bzw.* $\mathbb{C}^n$ hat eine Orthonormalbasis aus Eigenvektoren von A.

Beweis: a) Ist $\lambda \in \mathbb{C}$ ein Eigenwert von A, so gibt es nach Definition einen Vektor $v \neq 0$, so daß $Av = \lambda v$ ist. Da die komplexe Konjugation mit sämtlichen Grundrechenarten vertauschbar ist, folgt, daß

$$\overline{A}\,\overline{v} = \overline{\lambda}\,\overline{v} , \quad \text{d.h.} \quad v^T \overline{A}\,\overline{v} = v^T \overline{\lambda}\,\overline{v} = \overline{\lambda}\, v^T\, \overline{v} .$$

Bislang gilt alles noch für beliebige $n \times n$-Matrizen; um die Symmetrie *bzw.* HERMITE-Eigenschaft von A ins Spiel zu bringen, betrachten wir den Vektor $({}^T Av) = v^T\, A^T$. Da nach Voraussetzung $A^T = \overline{A}$ ist, können wir die rechte Seite der Gleichung auch als $v^T\, \overline{A}$ schreiben, und die linke Seite als $(\lambda v)^T = \lambda v^T$, da v Eigenvektor von A ist. Somit können wir die Zahl $v^T \overline{A}\,\overline{v}$ auch schreiben als

$$v^T \overline{A}\,\overline{v} = \left(v^T\, \overline{A}\right)\overline{v} = \lambda\, v^T\, \overline{v} .$$

Somit haben wir die beiden Darstellungen

$$v^T \overline{A}\,\overline{v} = \lambda\, v^T \overline{v} \quad \text{und} \quad v^T \overline{A}\,\overline{v} = \overline{\lambda}\, v^T \overline{v} ,$$

die nur dann beide richtig sein können, wenn $\lambda = \overline{\lambda}$ und somit reell ist; denn $v^T\,\overline{v}$ kann wegen der Definitheit HERMITEscher Skalarprodukte für einen Vektor $v \neq 0$ nicht verschwinden.

b) v sei Eigenvektor zum Eigenwert λ, und w sei Eigenvektor zum davon verschiedenen Eigenwert μ, d.h.

$$Av = \lambda v \quad \text{und} \quad Aw = \mu w \quad \text{und} \quad \lambda \neq \mu\,.$$

Dann ist

$$\begin{aligned}\lambda\, v^T\overline{w} &= (\lambda v)^T\overline{w} = (Av)^T\overline{w} = v^T\,A^T\overline{w}\\ &= v^T\overline{A}\,\overline{w} = v^T\overline{Aw} = v^T\overline{\mu w} = \overline{\mu}\,v^T\overline{w}\,.\end{aligned}$$

Wie wir schon wissen, sind alle Eigenwerte reell, d.h. $\overline{\mu} = \mu \neq \lambda$. Die obige Gleichungskette kann daher nur richtig sein, wenn $v^T\,\overline{w}$ verschwindet, d.h. wenn v und w orthogonal sind.

Beim Beweis von *c)* gehen wir im wesentlichen genauso vor wie im vorigen Abschnitt, als wir zeigten, daß die geometrische Vielfachheit eines Eigenwerts stets kleiner oder gleich der algebraischen ist; die zusätzliche Annahme über die Matrix A wird zeigen, daß hier die beiden Vielfachheiten sogar gleich sind.

λ sei also ein Eigenwert von A mit geometrischer Vielfachheit r, d.h. der zugehörige Eigenraum habe die Dimension r. Wir wählen eine Basis $\{b_1, \ldots, b_r\}$ davon und ergänzen sie zu einer Basis $\mathcal{B} = \{b_1, \ldots, b_n\}$ des gesamten Vektorraums $V = \mathbb{R}^n$ oder $\mathbb{C}^n$. Indem wir nötigenfalls das GRAM-SCHMIDTsche Orthogonalisierungsverfahren anwenden und anschließend die Längen aller Vektoren auf eins normieren, können wir annehmen, daß es sich dabei um eine Orthonormalbasis handelt.

Nun betrachten wir die lineare Abbildung

$$\varphi\colon V \to V; \quad v \mapsto Av\,.$$

Bezüglich der Standardbasis hat sie A als Abbildungsmatrix; für uns interessanter ist aber die Abbildungsmatrix C bezüglich der neuen Basis $\mathcal{B}$. Dazu sei B die Matrix mit Spaltenvektoren b_i; da der Eintrag an der Stelle (i,j) eines Matrixprodukts das (Standard-)Skalarprodukt des i-ten Zeilenvektors des ersten Faktors mit dem j-ten Spaltenvektor des zweiten Faktors ist, steht an der Stelle (i,j) der Matrix $B^T\overline{B}$

das (Standard) HERMITEsche Produkt der Vektoren b_i und b_j. Da $\mathcal{B}$ als Orthonormalbasis gewählt wurde, ist daher

$$B^T\overline{B} = E \quad \text{und damit} \quad B^T = \overline{B}^{\,-1} = \overline{B^{-1}}\,.$$

Aus dieser Formel folgt, daß mit A auch C eine HERMITEsche Matrix ist, denn

$$C^T = (B^{-1}AB)^T = B^T\,A^T\,i\big(B^T\big)^{-1} = \overline{B^{-1}}\,\overline{A}\,\overline{B} = \overline{B^{-1}AB} = \overline{C}\,.$$

Die ersten r Basisvektoren b_i sind Eigenvektoren von A zum Eigenwert λ; für $i \leq r$ ist daher $\varphi(b_i) = \lambda b_i$, d.h. in der i-ten Spalte von C steht an der i-ten Stelle die reelle Zahl λ und ansonsten überall die Null, genau wie auch im vorigen Abschnitt. Im Gegensatz zu dort haben wir nun aber eine HERMITEsche Matrix; da in der i-ten Spalte abgesehen von λ auf der Hauptdiagonalen nur Nullen stehen, muß daher dasselbe auch für die i-te Zeile gelten; die Matrix C hat also die Form

$$C = \begin{pmatrix} \lambda & 0 & \ldots & 0 & 0 & \ldots & 0\\ 0 & \lambda & \ldots & 0 & 0 & \ldots & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \ldots & \lambda & 0 & \ldots & 0\\ 0 & 0 & \ldots & 0 & & & \\ \vdots & \vdots & \ddots & \vdots & & \mathrm{M} & \\ 0 & 0 & \ldots & 0 & & & \end{pmatrix},$$

wobei M eine $(n-r) \times (n-r)$-Matrix ist, die uns nicht weiter zu interessieren braucht. Damit hat $C - xE$ die Form

$$\begin{pmatrix} \lambda - x & 0 & \ldots & 0 & 0 & \ldots & 0\\ 0 & \lambda - x & \ldots & 0 & 0 & \ldots & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \ldots & \lambda - x & 0 & \ldots & 0\\ 0 & 0 & \ldots & 0 & & & \\ \vdots & \vdots & \ddots & \vdots & & \mathrm{M} - \mathrm{x}\mathrm{E}_{n-r} & \\ 0 & 0 & \ldots & 0 & & & \end{pmatrix},$$

wobei E_{n-r} die $(n-r) \times (n-r)$-Einheitsmatrix bezeichnet.

Wie wir uns schon im vorigen Abschnitt überlegten beim Beweis, daß die geometrische Vielfachheit eines Eigenwerts immer kleiner oder gleich der algebraischen ist, haben A und C dasselbe charakterische Polynom; da wir die Matrix C besser kennen, rechnen wir mit ihr.

Wie in Abschnitt *d)* folgt auf Grund der obigen Form der Matrix $C - xE$ aus dem LAPLACEschen Entwicklungssatz, daß

$$\det(A - xE) = \det(C - xE) = (\lambda - x)^r \det(M - xE_{n-r})$$

ist, wobei E_{n-r} die $(n-r) \times (n-r)$-Einheitsmatrix bezeichnet. Wir müssen zeigen, daß die algebraische Vielfachheit von λ *genau* gleich r ist, daß also λ keine Nullstelle von $\det(M - xE_{n-r})$ sein kann.

Wäre λ Nullstelle von $\det(M - xE_{n-r})$, so hätte M den Eigenwert λ, es gäbe also einen $(n-r)$-dimensionalen Eigenvektor w von M. Wegen der speziellen Form der Matrix C ist für jeden Eigenvektor

$$w = \begin{pmatrix} w_{r-1} \\ \vdots \\ w_n \end{pmatrix} \quad \text{von } M \text{ der Vektor} \quad v = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ w_{r-1} \\ \vdots \\ w_n \end{pmatrix}$$

ein Eigenvektor von C und damit von A – Eigenvektoren hängen schließlich nur von der linearen Abbildung ab, nicht von einer speziellen Abbildungsmatrix. Dies widerspricht aber der Voraussetzung, daß der Eigenraum zum Eigenwert λ von $b_1, \ldots, b_r$ erzeugt wird, denn v ist linear unabhängig von diesen b_i.

Also hat λ die algebraische Vielfachheit r, und *c)* ist gezeigt.

d) ist nun eine einfache Folgerung aus den übrigen Aussagen und dem sogenannten *Fundamentalsatz der Algebra,* wonach jedes reelle oder komplexe Polynom über den komplexen Zahlen in Linearfaktoren zerfällt:

Wir wissen, daß die Summe der algebraischen Vielfachheiten aller Eigenwerte gleich der Dimension n des Vektorraums ist und daß alle Eigenwerte reell sind; da die algebraischen gleich den geometrischen

Vielfachheiten sind, gibt es also n Eigenvektoren, die eine Basis von V bilden.

Für jeden einzelnen Eigenraum können wir die Eigenvektoren nach GRAM-SCHMIDT so wählen, daß sie eine Orthonormalbasis bilden; da Eigenvektoren zu verschiedenen Eigenwerten stets orthogonal sind, ist die Vereinigungsmenge dieser Basen Orthonormalbasis von V. ∎