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.

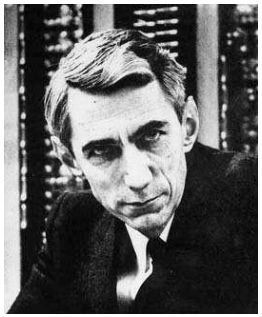
Der Wert solcher Informationen kann allerdings nicht objektiv beziffert werden: Die Nachricht, daß in einer südafrikanischen Goldmine eine neue stark erzführende Ader entdeckt wurde, ist wertlos für jemanden, der kein Geld für Aktienkäufe hat oder grundsätzlich nur in Europa investiert; für einen südafrikanischen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

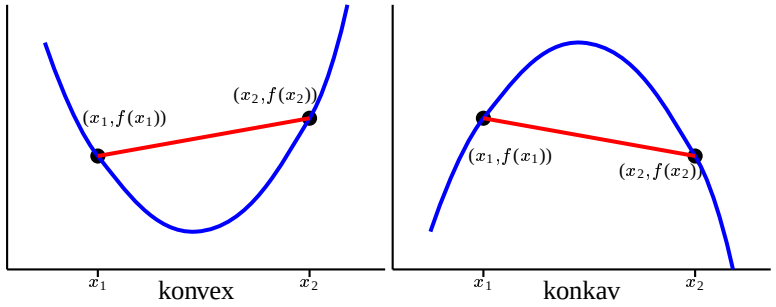
§2: Konvexität

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:

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

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

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

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:

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

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

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

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

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

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.

Fassen wir also, die beiden Fälle *Kugel eins zu schwer* und *Kugel zwei zu schwer* zu einer Fallgruppe zusammen. Die restlichen sechs Fälle sollen zwei Dreiergruppen bilden; 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 Informationsgehalt 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; indem wir die Begriffe *leichter* und *schwerer* miteinander vertauschen, können wir diese genauso behandeln.

Beim ersten Wiegen erhalten wir somit eine Information von

$$\begin{aligned} &-\frac{16}{25}\log_2\frac{8}{25}-\frac{9}{25}\log_2\frac{9}{25}\\ =\ &\frac{-16}{25}(\log_2 8-\log_2 25)-\frac{9}{25}(\log_2 9-\log_2 25)\\ =\ &\log_2 25-\frac{48}{25}-\frac{18}{25}\log_2 3\text{ Bit};\end{aligned}$$

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

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

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.