Einführung in die Informationstheorie
1–semestrige Vorlesung für Studenten der Physik

BERND POMPE

Institut für Physik der Universität Greifswald

Studentenfassung vom 8. April 2005

Inhaltsverzeichnis

Kapitel 1

Einleitung

Unsere Zeit wird häufig als Informationszeitalter charakterisiert, ist doch der Begriff der *Information* aus unserem Alltag nicht mehr wegzudenken. Täglich dringen über die modernen Kommunikationsmedien Presse, Radio Fernsehen und Computernetze eine Fülle an Nachrichten auf uns ein, die ein einzelner nicht verarbeiten kann. Digitale Rechenanlagen verarbeiten sekundenschnell Millionen und Abermillionen von Informationseinheiten. Mit der Geschwindigkeit des Lichtes werden Informationen rund um die Erde gesandt und durchlaufen die Weiten des Universums.

Der Menschheit größte technische Anlage bildet das Telefonnetz — eine Einrichtung, die ausschließlich zur Übertragung von Informationen erdacht wurde. Mit ihr sind nahezu zehn Prozent aller Menschen miteinander verbunden, heute noch vorrangig durch die Laut– und Schriftsprache, aber mit der weltweiten Entwicklung des dienstintegrierten Digitalnetzes (ISDN) in absehbarer Zeit auch zunehmend über Bilder. Tragen weltweite Computernetze auch teilweise noch recht anarchische Züge, so sind sie doch bereits fester Bestandteil der heutigen Kommunikationstechnik, die noch vor einem Jahrhundert allzu phantastisch schien. Der technische Daten– bzw. Informationsaustausch macht heute einen qualitativen Wandel durch, indem er weniger zentralistisch wird, dafür aber immer stärker sich selbst organisierende, disperse Strukturen zeigt. Die Unidirektionalität des Austausches etwa beim klassischen Radioempfang wird zunehmend abgelöst durch die Bi- oder gar Multidirektionalität im World Wide Web (WWW) bzw. Internet. Die klassischen Kommunikationsmedien werden hier eingebettet. Die Entwicklung des WWW zeigt erstaunliche Ähnlichkeiten etwa zur Vernetzung natürlicher neuronaler Netze. Pespektivisch werden die rund 10^{10} Menschen via WWW in Verbindung stehen, dies ist

von der gleichen Größenordnung wie die rund 10^{10} Neuronen im Nervensystem eines einzelnen Menschen. Hier bilden sie ein komplexes neuronales System, eine neue Qualität, die sich über die Daseinsweise eines einzelnen Neurons erhebt. Im WWW vernetzen sich Menschen zu eine Art „Überorganismus“, dessen Konturen heute schon zu erahnen sind. Auch im Mikrokosmos der Elementarteilchen lassen sich solch neue Entitäten finden, die sich aus der Wechselwirkung (dem Informationsaustausch) elementarer Einheiten ergeben, etwa dann, wenn unter 1K abgekühlte Elektronen in einem starken Magnetfeld zusammenwirken und gewisse Quasiteilchen bilden.

Das Wesen des Zusammenwirkens elementarer Objekte in komplexen Systemen besteht weniger im Energie– als vielmehr im Informationsaustausch, wenngleich sich beide wechselseitig bedingen. Dieser Standpunkt wird in immer stärkeren Maße von jenen eingenommen, die komplexe Systeme erforschen Einen allgemeinverständlichen und aktuellen Überblick hierzu vermittelt z. B. Mainzer [11]. Die sich zu erwartende Renaissance der Informationstheorie drückte Wheeler [20] wie folgt aus:

> „Tomorrow we will have learned to understand and express all of physics in the language of information.“[1]

Der Begriff der Information ist Bestandteil unserer Umgangssprache. Wird er etwa im Sinne von Nachricht, Auskunft oder Belehrung gebraucht, so findet man, daß Informationen seit jeher unter Menschen ausgetauscht wurden – mittels Gesten, Lauten, Rauch– und Lichtzeichen u.v.a.m. Der immer effektivere, schnellere und weiterreichende Informationsaustausch stellt ein wesentliches Moment der menschlichen Evolution dar. So hatte beispielsweise die Erfindung der Schrift zur Speicherung oder Übermittlung von Sprachinformationen weitreichende Konsequenzen. Aber auch die Werke der bildenden Kunst oder Musik haben eine ihnen eigene Sprache, ihr spezifisches Material zur Speicherung und Übermittlung von Informationen, wie überhaupt jedes gegenständliche Wirken des Menschen charakteristische Spuren und somit Informationen für die Nachkommenden hinterläßt.

Der Begriff der Information bzw. des Informationsaustausches ist nicht an zivilisatorische Entwicklungen gebunden, wenngleich er uns gerade mit Hinblick hierauf in den folgenden Vorlesungen interessieren wird. Vielmehr werden Informationen in der gesamten Natur ausgetauscht

[1] In freier Übersetzung: „Eines Tages werden wir alle Physik aus der Sicht der Information(stheorie) verstehen und ausdrücken. “

und verarbeitet, in Flora, Fauna und der unbelebten Natur, allerdings auf sehr unterschiedlichem Niveau. Letztlich ist Informationsaustausch Ausdruck der Wechselwirkung der Materie. Über Wechselwirkungen hält sich die Materie in Bewegung, tauschen Bereiche der objektiven Realität Informationen aus und verändern dabei ihren Zustand, wodurch die erhaltene Information „konserviert“ wird. Sie kann dann von jenen abgerufen werden, die es verstehen, sie zu lesen und zu interpretieren.

So berichten uns Mondkrater vom Zusammenprall mit einst vagabundierenden Himmelskörpern. Lebewesen wachsen nach den in der DNS–Doppelhelix gespeicherten Informationen. Das wohl beeindruckenste Beispiel hierfür hat die Natur im Menschen hervorgebracht, der die ihm über die Sinnesorgane zugeführten Informationen durch komplexe bio– und physikochemische Operationen codiert (verschlüsselt), speichert, abruft, decodiert — kurzum: *verarbeitet*. Dabei darf jedoch nicht übersehen werden, daß die Wirkung einer Information auf den Empfänger wesentlich von seinem Systemzustand abhängt, von seiner Fähigkeit zur Interpretation (Decodierung) der Informationen und seinen Möglichkeiten der Reaktion in Form der eigenen Zustandsänderung sowie der Aussendung von Informationen an die Umwelt.

Die erwähnten Beispiele machen deutlich, daß die Begriffe der Information und des Informationsaustausches objektive Kategorien sind, also des Menschen und seiner Reflexionen nicht bedürfen. Sie sind der Daseinsweise der Natur inhärent. Insbesondere existieren sie unabhängig von der modernen Kommunikationstechnik. Allerdings wurde der Begriff der Information erst mit der Entwicklung dieser Kommunikationstechnik in unserem Jahrhundert zur wissenschaftlichen Kategorie, als es dem amerikanischen Mathematiker und Nachrichtentechniker Claude E. Shannon Ende der vierziger Jahre gelang, ein Maß für die Information zu begründen. Dadurch wurde etwas meßbar, was zuvor als unmeßbar galt. Shannon stellte seine Ergebnisse zunächst auf einer Tagung in New York City im März 1948 vor. Im gleichem Jahr erschien seine berühmte Arbeit *A mathematical theory of communication* im Bell System Technical Journal. Eine weitere, in diesem Zusammenhang wichtige Erfindung fällt in den Juli desselben Jahres — die öffentliche Ankündigung der technischen Realisierung eines Festkörper–Verstärkers, des sogenannten *Transistors*. Seine Erfinder um den Amerikaner Shockley entwickelten diesen Transistor in den Bell Telephone Laboratorien, wo auch Shannon angestellt war. Im Jahre 1948 erschien aber auch das Buch des Amerikaners Norbert Wiener mit dem Titel *Cybernetics or control and communication in the animal and the machine*. Damit begründete Wie-

NER die Kybernetik als die Wissenschaft von den abstrakten dynamischen Systemen, die über Informationsverarbeitung, Selbstregulation und –organisation verfügen. Mit diesen Theorien, Entdeckungen und Erfindungen gilt das Jahr 1948 heute als das Geburtsjahr der Informationstheorie und der modernen Kommunikationstechnik, als der Beginn des technischen Informationszeitalters. Selbstverständlich hatten die erwähnten Fortschritte ihre Vorläufer und Begleiter. So sind bezüglich der SHANNONschen Informationstheorie beispielsweise Namen und Zeiten wie GABOR (1946), KOTELNIKOV, NYQUIST (1924/28) und HARTLEY (1928) zu nennen.

Nimmt ein System aus seiner Umwelt Informationen auf, so führt dies infolge der Informationsspeicherung zur Erhöhung des Ordnungszustandes in diesem System. Damit verringert sich seine thermodynamische Entropie. Folglich erwarten wir eine grundsätzliche Beziehung zwischen dem nachrichtentechnischen Informationsbegriff und der thermodynamischen Entropie, welche CLAUSIUS im Jahre 1865 einführte. Solche Zusammenhänge konnten tatsächlich aufgezeigt werden, vor allem durch die Untersuchungen von BOLTZMANN (1872/77), SZILARD (1929), WIENER (1949) und BRILLOUIN (1956).

Charakteristisch für SHANNONs Theorie ist, daß sie keine Theorie des Nachrichtenmittels ist, sondern die Theorie der Nachricht selbst. Damit ist sie allgemein genug, um in den unterschiedlichen Wissenschaften, von der Physik über die Biologie bis hin zur Psychologie, den Erziehungs– und Kunstwissenschaften u.a. Bedeutung erlangen zu können. Die *Information* ist damit eine der allgemeinsten wissenschaftlichen Begriffe überhaupt.

Versucht man eine Klassifizierung der großen Aufgabenstellungen, welchen sich die Physik und die darauf aufbauenden technischen Disziplinen vorrangig widmen, so findet man Komplexe, welche sich um die Begriffe *Stoff*, *Energie* und *Information* ranken. Zum ersten Begriff gehören die Materialwissenschaften. Hier geht es darum, Materialien mit gewissen ausgezeichneten Eigenschaften zu erzeugen, die der Oberflächenveredlung dienen, der effektiven Leitung elektrischer Ströme z. B. mit Supraleitern u.v.a.m. Die zweite Kategorie ist mit den Bemühungen verknüpft, umweltverträgliche und kostengünstige Energieträger großtechnisch zu nutzen. Hier gehören Begriffsfelder her wie die Kernfusion und die Photovoltaik. Der Informationsbegriff eröffnet eine in seiner Bedeutung ebenbürtige dritte Sparte, wenngleich generell anzumerken ist, daß sich diese Gebiete gegenseitig bedingen und durchdringen.

Im Unterschied zu HARTLEY (1928) gebrauchte SHANNON (1948) den Begriff Information nicht. Vielmehr sprach er von der *Kommunikation* (communication). Nach allgemeiner Auffas-

sung geht der Gebrauch des Begriffes Information im hier interessierenden Zusammenhang auf WIENER zurück. In seinem bereits erwähnten Buch aus dem Jahre 1948 schreibt er

> „Information is information not matter or energy.“ [2]

Die Dreiheit von *Stoff, Energie und Information* haben wir bereits hervorgehoben. Haben wir ein stoffliches Gebilde mit einem bestimmten Energieinhalt, so beschreibt die Information (Entropie) gewissermaßen seinen Ordnungszustand.

In den 50er und 60er Jahren gab es wesentliche Weiterentwicklungen der Informationstheorie mit breiten Anwendungen in den verschiedensten Wissenschaftsdisziplinen. Jedoch wurde die Bedeutung dieser Theorie dabei teilweise auch überschätzt. Aber bis in die Gegenwart hinein spielen Begriffe aus der Informationstheorie eine teilweise zentrale Rolle in sich neu entwickelnden Theorien, wie jener dynamischer Systeme (Chaostheorie).

Gegenstand der Informationstheorie

Die Informationstheorie beschäftigt sich mit theoretischen Problemen bei der Aufbewahrung (Speicherung), Verschlüsselung (Codierung) und Übertragung von Informationen. SHANNONs Betrachtungen setzen ein allgemeines Übertragungsschema voraus, wie es in der Abbildung 1.1 zu sehen ist. Jede Schnittstelle in diesem Schema kann ihrerseits als ein Paar aus Informationsquelle und –senke aufgefaßt werden. Quelle und Senke der Information heißen auch *Sender* bzw. *Empfänger*.

Träger der Information sind Signale, die in der Nachrichtentechnik häufig aus einer Folge verschiedener Pegel eines elektrischen Potentials bestehen. Diese Signale werden zunächst geeignet codiert, was man *Quellencodierung* nennt. Die Informationstheorie macht Aussagen darüber, was man unter einer optimalen Codierung einer Informationsquelle zu verstehen hat und wie man dieses Optimum erreicht. Eine Codierung ist optimal, wenn sie zu einer möglichst kompakten (kurzen) Darstellung führt. Denken wir z. B. an Dateien in einem Personalcomputer. Um hier mit den Speicherressourcen möglichst sparsam zu sein, werden sogenannte Packer verwendet, welche die Dateilängen verkürzen. Dabei dürfen selbstverständlich keine Informationen verloren gehen — mit den Entpackern müssen wir jederzeit den Originalfile eindeutig rekonstruieren können, ebenso wie unsere Hand einen Schwamm zunächst auf kleinem Raum zusammendrückt und dann wieder frei gibt, worauf sich die usprünglichen Konturen zeigen.

[2] In freier Übersetzung: „Information ist etwas Eigenständiges, das weder Stoff noch Energie ist.“

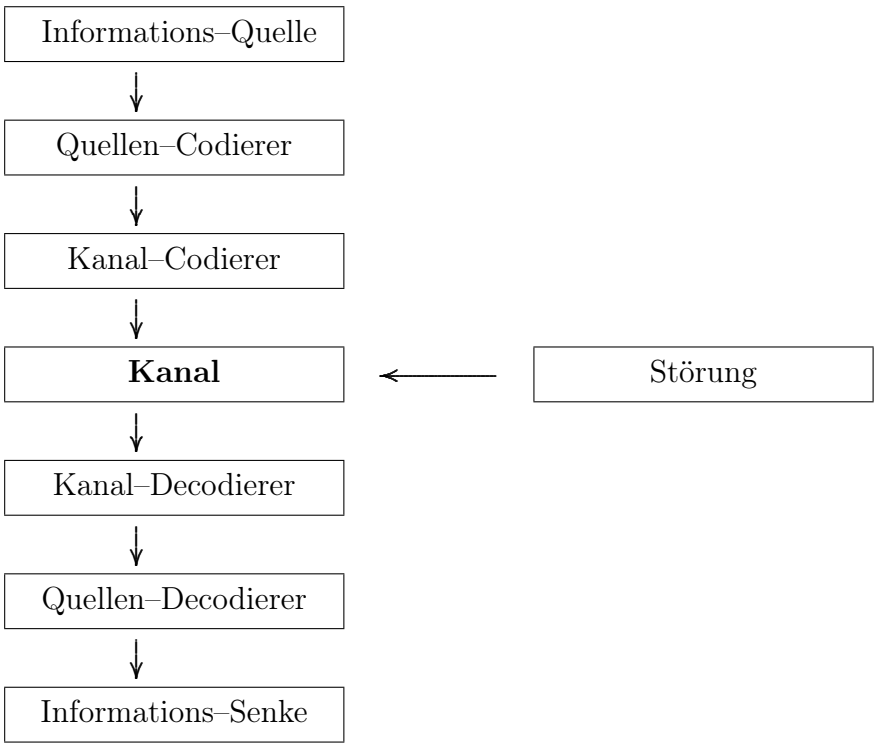

Abb. 1.1: Allgemeines Schema zur Informationsübertragung

Der Quellencodierung folgt die sogenannte *Kanalcodierung*. Sie ist notwendig, weil in realen Systemen die Übertragung immer mehr oder weniger stark gestört wird, besonders dann, wenn hochfrequente Signale übertragen werden. Deshalb kann in der Zeiteinheit nicht beliebig viel Information durch einen realen Nachrichtenkanal gelangen, ebenso, wie durch ein Rohr in der Sekunde nicht beliebig viel Wasser fließen kann. Es ist eine Grundaussage der Informationstheorie, daß trotz gewisser Störungen prinzipiell immer eine bestimmte Informationsmenge *fehlerfrei* übertragen werden kann. Der maximal mögliche (fehlerfreie) Informationsfluß ist die *Kanalkapazität*. Soll sie gut ausgenutzt werden, so muß durch eine geeignete Codierung das Signal an den Kanal angepaßt werden. Die Informationstheorie macht Aussagen darüber, wie dies zu erfolgen hat.

Generell sei jedoch angemerkt, daß die Informationstheorie kaum konstruktiv ist. Die grundlegenden Algorithmen zur optimalen Codierung können nur in besonders einfachen Fällen in der Praxis umgesetzt werden. In ihren wesentlichen ursprünglichen Aussagen gibt sie eher prinzipielle Grenzen für das praktisch Machbare an. Der Ingenieur muß dann seine ganze Findigkeit einsetzen, um mit der aufgebauten Apparatur diesen Grenzen möglichst nahe zu kommen. An diesen theoretisch abschätzbaren Grenzen mißt sich dann der Wert der Apparatur.

Aus der Informations– ist die Codierungstheorie hervorgegangen, welche wegen ihrer praktischen Bedeutung zuweilen als eigenständige Disziplin aufgefaßt wird. Hier geht es u. a. darum, Codes mit „kontrollierter Redundanz“ zu konstruieren. Denken wir dazu beispielsweise an unsere natürliche Schriftsprache, mit der dieses Buch *geschriben* ist. Der Sinn des obigen Satzes ist offenbar nicht verfälscht, obwohl im vorletzten Wort ein Buchstabe fehlt. Dieser Buchstabe trägt also zu einer scheinbar unnötigen Weitschweifigkeit (Redundanz) des geschriebenen Textes bei. Tatsächlich kann man viele weitere Beispiele für eine derartige Weitschweifigkeit finden, wenngleich in einigen Fällen durchaus schon ein einzelner fehlender oder falscher Buchstabe sinnentstellend sein kann: *Das Kind ißt eine Möhre. — Das Kind ist eine Göre.* Entsprechende Beispiele finden sich auch in der mündlichen Sprache. Tatsächlich erfüllt die Weitschweifigkeit aber eine wichtige Funktion: durch sie wird die Informationsübertragung bei zufälligen Störungen sicherer. Wählt man geschickte Codierungen, so kann man mit einer möglichst geringen Weitschweifigkeit einen möglichst großen Gewinn an Übertragungssicherheit erzielen.

Die mathematischen Grundlagen der Informationstheorie bilden Wahrscheinlichkeitrechnung und Statistik. Sie werden in dieser Vorlesung zumeist auf einem elementaren Niveau benötigt.

Etymologie des Informationsbegriffes

Die Worte *informieren* und *Information* wurden im 15. bzw. 16. Jahrhundert dem lateinischen Verb *in–formare* bzw. dem Substantiv *informatio* entlehnt. Das Verb wurde in dessen übertragener Bedeutung „durch Unterweisung bilden, unterrichten“ gebraucht. Eigentlich bedeutet in–formare „eine Gestalt geben“, „formen“, „bilden“ oder auch „in eine Form geben“. Das Substantiv Information wurde zunächst im Sinne von „Nachricht, Auskunft, Belehrung“ gebraucht. In der Zeit des Humanismus und der Renaissance trat dann die Bedeutung „Bildung (Unterweisung) durch den Dorfschullehrer (Informator)“ in den Vordergrund. Bis weit in das 19. Jahrhundert war die Bezeichnung „Informator“ für den Hauslehrer geläufig. Dann ging dieser Sinn mehr und mehr verloren. Schließlich blieb das Substantiv bis in unsere Tage im Sinne von „Auskunft, Mitteilung, Nachricht, Neuigkeit“ in der Alltagssprache erhalten. Der Gebrauch der Informationsbegriffes in den Wissenschaften setzt allerdings seine strenge formalisierte Definition voraus. Dabei sollte sich unser intuitiver, in der natürlichen Sprache entwickelter Informationsbegriff in der Formalisierung wiederfinden. Dies wird zumindest in einigen Aspekten auch tatsächlich

erreicht. Nur deshalb konnten wir bislang nahezu unmißverständlich den Informationsbegriff gebrauchen.

Nach genauer Analyse unseres intuitiven Informationsbegriffes stellt sich heraus, daß man ihn unter drei verschiedenen Gesichtspunkten sehen kann:

1. syntaktisch–statistisch,
2. semantisch und
3. pragmatischer Aspekt.

Der erste Gesichtspunkt berücksichtigt nur die Wahrscheinlichkeiten, mit denen die Zeichen aus einer Informationsquelle gesandt werden. Hingegen sagt der zweite Aspekt etwas über die Bedeutung aus, welche die Zeichen für den Empfänger haben. Schließlich berücksichtigt der dritte Aspekt die Wirkung der Zeichen auf den Empfänger.

Wegen der Komplexität unseres intuitiven Informationsbegriffes kann es nicht verwundern, daß ein formalisierter Informationsbegriff immer nur gewissen Teilaspekten des intuitiven Begriffes Rechnung tragen kann. So trägt der SHANNONsche Informationsbegriff nur dem syntaktisch–statistischen Aspekt Rechnung. Semantische Aspekte werden nicht erfaßt, sie sind für den klassischen Nachtrichteningenieur unwichtig. Allerdings gab es Bemühungen, allgemeinere Informationsmaße zu begründen. Wir werden in einem späteren Abschnitt auf derartige Verallgemeinerungen eingehen.

Kapitel 2

Klassische Informationsmaße

2.1 Hartley–Information

Wir betrachten eine Informationsquelle, welche Buchstaben a_m aus dem endlichen *Alphabet*

$$A = \{a_1, a_2, \ldots, a_m, \ldots, a_{|A|}\} \tag{2.1}$$

sendet. Darin bezeichnet $|A|$ die Mächtigkeit von A, welche z. B. beim Zeichensatz von Tab. 2.1 den Wert 112 hat. In diesem Kapitel sei $|A|$ immer endlich. Beobachten wir eine Informationsquelle über die Zeitdauer T, so registrieren wir eine Buchstabenfolge

$$a_{m_1}, a_{m_2}, \ldots, a_{m_t}, \ldots, a_{m_T} \quad . \tag{2.2}$$

Darin ist $m_t \in \{1, 2, \ldots, |A|\}$ der Buchstabenindex zur Zeit t. Wenn wir die Buchstaben des Alphabetes (2.1) durchnumerieren, dann können wir anstelle von (2.2) auch

$$m_1, m_2, \ldots, m_t, \ldots, m_T \tag{2.3}$$

schreiben. Hinsichtlich des Informationsgehaltes ist dies bei bekannter Numerierung vollkommen gleichwertig aber in der Schreibweise einfacher. Aus kombinatorischer Sicht gibt es genau $|A|^T$ viele verschiedene Folgen (2.3), das sind alle möglichen Variationen mit Wiederholung. Allerdings ist eine bestimmte reale Informationsquelle nicht immer dazu fähig, eine jede der kombinatorisch möglichen Buchstabenfolgen auch wirklich zu senden, d.h., es kann sein, dass nicht eine jede dieser Möglichkeiten eine von Null verschiedene Realisierungs–Wahrscheinlichkeit hat.

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	ɐ	ɑ	ɒ	ɔ	ʌ	ɓ	ƀ	ɸ	ɓ	β	ɞ	ɕ	ȼ	đ	ɗ	ɠ
1	ɗ	ɖ	ʤ	ð	ɒ	ə	ɚ	ɘ	ɛ	ɜ	ɝ	ɞ	ɡ	ɠ	ɢ	ɣ
2	ɣ	ɤ	ɦ	ħ	ɦ	ɧ	ɥ	ɨ	ɨ	ɩ	ɪ	ɨ	ɟ	ɬ	ɬ	ɬ
3	ɭ	ɮ	ʎ	ɤ	ɱ	ɯ	ɯ	ɲ	ŋ	ɳ	ɴ	⊙	ɵ	ɔ	ω	ω
4	ɷ	ɸ	ƥ	ɸ	ɾ	ɾ	ɽ	ɹ	ɻ	ɻ	ʀ	ʁ	ʂ	ʃ	ʃ	σ
5	ʈ	ʧ	ʇ	θ	ʉ	ʊ	ʊ	ʊ	ʊ	ʊ	ʍ	χ	ʎ	ʏ	ƶ	ʐ
6	ʒ	ʒ	ʔ	ʕ	ƅ	ƿ	ʕ	˩	ˈ	ˌ	ˌ	˥	˧	˦	˨	˩
7	ˈ	ː	ˑ	ˈ	˄	˅	˂	˃	˚	˳	˷	~	~	ˬ	˭	ˮ

Tab. 2.1: Alphabet der internationalen Lautschrift mit diakritischen Zeichen (TEX–Zeichensatz `wsuipa`)

Wir wollen nun die Information messen, welche wir durch Beobachtung der Folge (2.3) erhalten. Ein plausibler Ansatz ist offenbar, der Folge die Information

$$I = T \tag{2.4}$$

zuzuordnen. Eine doppelt solange Folge würde dann auch die zweifache Informationsmenge beinhalten. Haben wir andererseits gar kein Symbol beobachtet, so ist die Information null.

Denken wir z. B. an den Kauf eines Sachbuches, wo wir die Wahl zwischen zwei Büchern gleichen Preises aber unterschiedlichen Umfangs haben mögen. Dann entscheiden wir uns zunächst für das „dickere“, versprechen wir uns doch von diesem mehr Information.

Die Vorgehensweise entspricht im Grunde auch jener bei der Berechnung von Telefongebühren, wo die Dauer des Telefongesprächs zur Berechnungsgrundlage dient. Den Buchstaben unseres abstrakten Alphabetes (2.1) entsprechen hier die Laute, welche im Alphabet der internationalen Lautschrift zusammengefaßt sind (Tab. 2.1).

Die Anzahl der übermittelten Laute beträgt bei mittlerer Sprechgeschwindigkeit ca. 10 pro Sekunde. Die Dauer des Gespräches ist also im wesentlichen der Anzahl der übertragenen Laute proportional, und die Kosten des Gespräches entsprechen der übertragenen Informationsmenge im Sinne von Gl. (2.4).

Wenn wir die Buchstaben des Alphabets (2.1) eindeutig umkehrbar auf ein anderes Alphabet

$$B_r = \{b_1, b_2, \ldots, b_r\} \tag{2.5}$$

abbilden, so ist auch (2.2) ein–eindeutig auf eine Folge

$$b_{m_1}, b_{m_2}, \ldots, b_{m_t}, \ldots, b_{m_T} \tag{2.6}$$

abbildbar, welche dieselbe Informationsmenge liefert. Der Übergang von der Buchstabenfolge (2.2) zur Indexfolge (2.3) gibt hierfür bereits ein Beispiel.

Wenn im neuen Alphabet B_r weniger Buchstaben sind als in A, also falls

$$r = |B_r| < |A|,$$

gilt, so kann die Abbildung zwischen den Alphabeten nicht eindeutig umkehrbar sein. Wir betrachten dann anstelle von B_r ein l–faches cartesisches Mengenprodukt davon, also

$$\boxtimes_l B_r \equiv \underbrace{B_r \times B_r \ldots \times B_r}_{l\text{–mal}}.$$

Dann müssen wir l nur hinreichend groß wählen, um genügend viele verschiedene Buchstaben zu bekommen, vorausgesetzt, daß $r \geq 2$ gilt, denn die Mächtigkeit von $\boxtimes_l B_r$ ist $|\boxtimes_l B_r| = r^l$.

Die Buchstaben von $\boxtimes_l B_r$ sind die sogenannten *Codeworte*. Im Falle $r = 2$ spricht man vom *Binärcode*, und wir schreiben $b_1 \equiv 0$ und $b_2 \equiv 1$.

Als simples Beispiel betrachten wir einen zufälligen Versuch, in welchem aus einem französischen Skatspiel eine Karte gezogen und wieder zurückgesteckt wird. Führen wir den Versuch T–mal durch, so erhalten wir möglicherweise die Buchstabenfolge

$$\underbrace{\heartsuit\,8 \quad \spadesuit\,\mathrm{K} \quad \diamondsuit\,7 \quad \clubsuit\,9 \quad \ldots \quad \heartsuit\,8 \quad \heartsuit\,\mathrm{B}}_{T\text{ viele Buchstaben (Karten)}} \tag{2.7}$$

Die Abstände zwischen den Karten, die Leerstellen, bilden keinen zusätzlichen Buchstaben, sie dienen nur der besseren Übersicht. Insbesondere würde ihr Fortlassen den Informationsgehalt nicht ändern, vorausgesetzt, wir halten uns immer an die Konvention, einen Buchstaben (das ist eine Karte) durch die Angabe der Farbe gefolgt vom Kartenwert darzustellen.

Nach unserem intuitiven Verständnis sollten wir mit der Serie (2.7) die gleiche Informationsmenge erhalten, wie mit einer entsprechenden Folge von Binärworten, welche die Karten eindeutig umkehrbar codieren. Das muß zumindest ein 5–stelliger Binärcode sein, stehen uns doch dann gerade $2^5 = 32$ verschiedene Binärworte zur Verfügung. Wie wir diese Binärworte den 32 Karten zuordnen, ist zunächst gleichgültig, wir fordern nur, daß sie eindeutig umkehrbar

ist. In der Regel verwendet man eine gewisse, auf natürliche Weise gegebene Systematik der Zuordnung, wie beispielsweise die folgende:

$$\begin{aligned} \diamondsuit\ 7 &\leftrightarrow 00\,0\,00 \\ &\dots \\ \heartsuit\ 8 &\leftrightarrow 01\,0\,01 \\ &\dots \\ \heartsuit\ \mathrm{B} &\leftrightarrow 01\,1\,00 \\ &\dots \\ \spadesuit\ \mathrm{K} &\leftrightarrow 10\,1\,10 \\ &\dots \\ \clubsuit\ 9 &\leftrightarrow 11\,0\,10 \end{aligned}$$

Die ersten beiden Binärstellen codieren die Farbe, die dritte Position unterscheidet zwischen Zahl oder Kopf bzw. As. Damit haben wir die 32 Karten schon in 8 Klassen zu je 4 Mitglieder aufgeteilt, und die letzten beiden Positionen differenzieren innerhalb einer jeden dieser Klassen. Wir benötigen dann $5\,T$ Binärzeichen zur Darstellung der Originalfolge (2.7),

$$\underbrace{01001\,10110\,00000\,11010\,\dots\,01001\,01100}_{5T\text{ viele Nullen und Einsen}} \tag{2.8}$$

Mit unserem Informationsmaß (2.4) würden wir dieser Folge die Information $5\,T$ zuordnen, wohingen für (2.7) T erhalten wird. Das widerspricht aber unserer Intuition, sind doch die Folgen eindeutig umkehrbar aufeinander abbildbar, so daß bei der Codierung weder zusätzliche Information entstanden noch verlorengegangen sein kann. Unser bisheriges Informationmaß versagt also. Darüber hinaus hätten wir eine eindeutige Decodierbarkeit auch mit Binärworten der Länge 6, 7 oder noch längeren erreichen können, so daß unser Maß noch größere Informationsgehalte vortäuschen würde.

Das Problem kann dadurch gelöst werden, daß wir die Information einer beliebigen Buchstabenfolge grundsätzlich wie folgt messen: Wir bilden die Buchstabenfolge *eindeutig umkehrbar* auf eine *möglichst kurze Binärfolge* ab. Die Anzahl der Nullen und Einsen in dieser Folge definieren wir dann als Information der Buchstabenfolge.

Die Situation ist mit der Messung beliebiger physikalischer Größen vergleichbar. Messen wir z. B. eine Länge, so verwenden wir das Meter und geben jede andere Länge in möglicherweise gebrochenen Vielfachen dieser genormten Längeneinheit an. Ebenso müssen wir bei der Messung

der Information eine Informationseinheit definieren. Das geschieht hier dadurch, daß immer mit Binärworten und möglichst sparsam aber noch eindeutig umkehrbar codiert wird. Die Maßeinheit der Information ist dann das *bit*.[1] In unserem Beispiel der Kartenfolge würden wir also einer Folge der Länge T die Information $5T$ bit zuordnen.

Selbstverständlich könnte man anstelle des binären auch andere Standardalphabete zur Grundlage nehmen. Es gibt jedoch gute Gründe für diese Wahl. Einer davon besteht darin, daß das binäre Alphabet das kleinste ist, welches als Ausgangspunkt für eine Codierung taugt — mit nur einem Symbol kann man nicht mehr als einen Buchstaben eindeutig umkehrbar codieren und Folgen aus nur einem Buchstaben haben keinen Informationsgehalt. Tatsächlich ändert sich am Grundgehalt informationstheoretischer Aussagen nichts, wenn sogenannte *Radix–r–Codes* mit $r \geq 2$ verwendet werden. Weiter unten in diesem Abschnitt werden wir noch einige praktische Gründe für die Wahl des binären Alphabetes anführen. Aber auch Oktal– und Hexadezimal–Codes ($r = 8$ bzw. $r = 16$) sind besonders in der „Welt der Computer“ weit verbreitet. Der Dezimalcode ($r = 10$) ist den Menschen auf natürliche Weise durch seine zehn Finger gegeben.

Besteht ein Alphabet A aus $|A| = 2^l$ vielen Buchstaben, so können diese mit Binärworten der Länge $\operatorname{ld} 2^l = l$ codiert werden.[2] Einer Buchstabenfolge der Länge T ordnen wir somit die Gesamtinformation $I_{\text{ges}} = lT$ bit zu. Pro Buchstaben der Folge erhalten wir den Informationsgehalt $I_{\text{ges}}/T = \operatorname{ld}|A| = l$ bit. Wir definieren deshalb: Bekommen wir Kenntnis über einen konkreten Buchstaben aus einem vorweg bekannten Alphabet A der Mächtigkeit $|A| > 0$, so erhalten wir die Information

$$\boxed{I_H(A) \equiv \operatorname{ld}|A|} \quad \text{Hartley–Information} \tag{2.9}$$

Dieses Informationsmaß hat Hartley im Jahre 1928 eingeführt [6]. Es stellt den Prototyp aller weiteren Maße dar, wie wir weiter unten noch ausführen werden. Zunächst bedarf es jedoch noch einiger Rechtfertigungen.

Rechtfertigung der Hartley–Information

Bei der Einführung der Hartley–Information (2.9) sind wir davon ausgegangen, daß die Buchstabenanzahl $|A|$ als ganzzahlige Potenz von 2 darstellbar ist. Dies ist eine allzu einschränkende Annahme, die den praktischen Wert dieses Maßes fraglich erscheinen läßt. Wir könnten nun

[1] Im Englischen bedeutet „bit“ soviel wie „ein kleines Stück“ oder „eine kleinste Quantität“.

[2] Wir werden im folgenden zumeist den *Logarithmus dualis*, $\operatorname{ld} \equiv \log_2$, verwenden.

das Alphabet A mit ungenutzten Buchstaben auffüllen, bis die Anzahl als ganzzahlige Potenz von 2 darstellbar ist. Dieser Weg wird bei praktischen Codierungen häufig begangen, hier führt er jedoch in die Irre.

Tatsächlich lösen wir das Problem, indem wir anstelle einzelner Buchstaben von A Worte der Länge d binär codieren — Worte, die aus Buchstaben von A bestehen und somit aus dem Alphabet

$$\boxtimes_d A \equiv \underbrace{A \times A \times \ldots \times A}_{d\text{–mal}} \tag{2.10}$$

stammen. Das sind genau $|A|^d$ viele Worte. Nehmen wir nun an, daß l_d–stellige Binärwörter zur eindeutig umkehrbaren Codierung aller Worte aus (2.10) ausreichen, aber $(l_d - 1)$–stelligen Binärwörter noch nicht. Es soll also

$$2^{l_d - 1} < |A|^d \leq 2^{l_d} \tag{2.11}$$

gelten. Die Ungleichung ändert sich nicht, wenn wir von einem jeden Term den Logarithmus dualis bilden und durch d dividieren, was zur Darstellung

$$\frac{l_d - 1}{d} < \operatorname{ld} |A| \leq \frac{l_d}{d} \tag{2.12}$$

führt. Der Quotient l_d/d kann gebrochen sein. Er gibt die mittlere Anzahl von Binärstellen an, die bei der Codierung pro Buchstabe des Ausgangsalphabetes A benötigt wird. Mit d wird auch l_d wachsen. Lassen wir nun die Wortlänge d gegen unendlich laufen, so finden wir aus (2.12) schließlich

$$\lim_{d \to \infty} \frac{l_d}{d} = \operatorname{ld} |A|. \tag{2.13}$$

Diese Beziehung sagt aus, daß wir zur eindeutig umkehrbaren binären Codierung einer langen $(T \to \infty)$ Buchstabenfolge (2.2) in der Tat mit $\operatorname{ld} |A|$ Binärstellen *pro Buchstabe* auskommen. Ist $|A|$ keine ganzzahlige Potenz von 2, so muß man nur hinreichend lange $(d \to \infty)$ Worte binär codieren. Dies rechtfertigt die allgemeine Anwendung des Hartleyschen Informationsmaßes (2.9) auf Alphabete, deren Mächtigkeit nicht notwendig eine ganzzahlige Potenz von 2 ist.

Beispiel: Codierung beim Würfeln

Betrachten wir als Beispiel das Würfeln, wobei einer der möglichen Werte 1, 2, 3, 4, 5 oder 6 erhalten wird. Diese sechs Ziffern bilden unser Originalalphabet A. Das Informationsmaß

d	l_d	l_d/d	$\frac{l_d/d}{\mathrm{ld}\,6}$
1	3	3	1,1605...
2	6	3	1,1605...
3	8	2,6666...	1,0316...
4	11	2,75	1,0638...
5	13	2,6	1,0058...
6	16	2,6666...	1,0316...
7	19	2,7142...	1,0500...
8	21	2,6265...	1,0154...
9	24	2,6666...	1,0316...
10	26	2,6	1,0058...
100	259	2,59	1,0019...
1000	2585	2,585	1,00001...

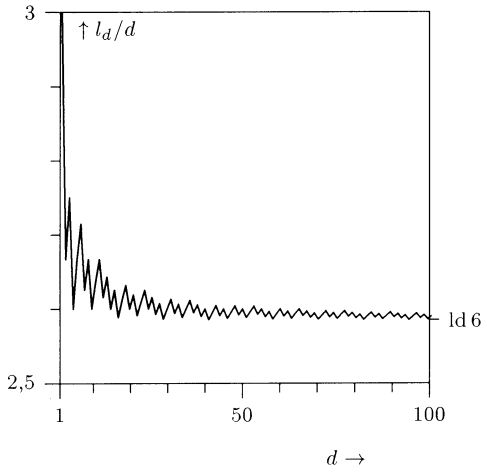

Abb. 2.1: Konvergenz der Bitdichte l_d/d (= kleinste mittlere Anzahl von Binärstellen zur eindeutig umkehrbaren binären Codierung von Worten aus d Würfelergebnissen) gegen die Hartleysche Information $I_H = \mathrm{ld}\,6$ mit zunehmender Wortlänge d beim Würfelversuch

(2.9) sagt dann, daß wir mit jedem Würfelergebnis die Information $I_H = \mathrm{ld}\,6 = 2,5849\ldots$ bit erhalten. Die codierungstheoretische Rechtfertigung dieser Aussage lautet folgendermaßen: Eine eindeutig umkehrbare Codierung der sechs Ziffern gelingt offenbar mit Binärworten der Länge $l_1 = 3$. Hingegen können mit zweistelligen Binärworten höchstens $2^{l_1-1} = 2^2 = 4$ Ziffern (Würfelergebnisse) codiert werden. Andererseits haben vier– oder höherstellige Binärwörter zumindest eine überflüsse Binärstelle. Wollen wir aber eine (lange) Folge von Würfelergebnissen codieren, so bilden wir aus der Ziffernfolge eine Wortfolge, indem wir immer d aufeinander folgende Ziffern zu einem Wort zusammenfassen. Für wachsendes d erhalten wir somit pro codierter Ziffer Binärwortlängen l_d/d, wie sie in der Abbildung 2.1 aufgelistet sind. Die Werte machen deutlich, daß es nicht immer vorteilhaft ist, die Wortlänge zu vergrößern. Allerdings ist die generelle Tendenz gemäß der Beziehung (2.13) nicht zu verkennen. Wenn wir immer fünf Würfelergebnisse gemeinsam binär codieren, so ist die benötigte Binärstellenzahl pro codiertem Würfelergebnis mit 2,6 schon dichter als 1% beim theoretisch *best* möglichen Wert von $\mathrm{ld}\,6$. Daß dieser Wert tatsächlich nicht unterboten werden kann, folgt unmittelbar aus (2.12).

Rechtfertigung der Verwendung von Binärcodes

Alle bisherigen Überlegungen lassen sich auf den Fall übertragen, daß anstelle der Binär– nun Trinärcodes oder allgemeiner Radix–r–Codes mit $r \geq 2$ verwendet werden. In der Hartleyschen

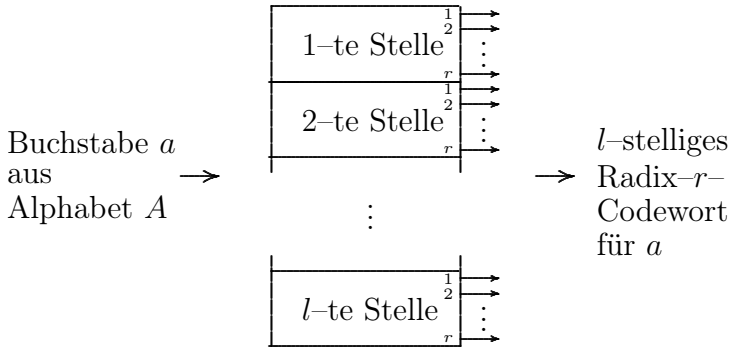

Abb. 2.2: Schematischer Radix–r–Coder

Formel (2.9) würde dann anstelle des Logarithmus dualis der Logarithmus zur Basis r, $\log_r$, zu verwenden sein. Die Einheit der Information wäre dann allerdings nicht „bit“, sondern eine Einheit, die wir mit „ bit_r“ bezeichnen könnten. Im Spezialfall gilt dann $\mathrm{bit}_2 \equiv \mathrm{bit}$. Unter Beachtung der Beziehung

$$\log_r x = \frac{\mathrm{ld}\, x}{\mathrm{ld}\, r}$$

finden wir aber $1\,\mathrm{bit}_r = \mathrm{ld}\, r\,\mathrm{bit}$. Die verschiedenen Informationseinheiten unterscheiden sich also nur in einem Umrechnungsfaktor, so daß sich kein wesentlicher Unterschied zur Verwendung des dualen Logarithmus ergibt. Tatsächlich ist die Verwendung der Einheit „ bit_r“ für $r \neq 2$ nicht gebräuchlich.[3]

Wir geben nun ein praktisches Argument für die Verwendung von Binärcodes: Wenn wir einen Buchstaben a aus einem Alphabet A codieren wollen, so brauchen wir dazu ein Gerät, das wir Coder nennen und in Abb. 2.2 schematisch dargestellt haben. Ein gewisses, wenn auch recht grobes Maß für die Komplexität des Coders liefert die Kostenfunktion

$$K = l\, r \quad .$$

Darin ist l die Anzahl der Stellen des Codewortes und r die Anzahl der möglichen Ausgangswerte einer Coderstelle. Mit dem Coder lassen sich r^l viele Buchstaben eindeutig umkehrbar codieren. Wir suchen nun das Minimum der Kosten K unter der Nebenbedingung, daß die Anzahl der codierten Buchstaben konstant ist, also $r^l = \text{const.} \equiv c$. Ersetzen wir in der Kostenfunktion

[3]Wenn wir im folgenden einmal den Informationsgehalt nur mit einem Zahlenwert angeben, so versteht es sich von selbst, das die Einheit bit gemeint ist. Einige Autoren gebrauchen auch den *logarithmus naturalis* ln. Die Informationseinheit heißt dann „nit“ statt „bit“.

mit Hilfe der Nebenbedingung die Variable l, so folgt

$$K(r) = r \log_r c \ .$$

Wir interessieren uns nun für den Wert für r, der die geringsten Kosten verursacht. Dazu ermitteln wir zunächst die Ableitung von K nach r,

$$\frac{\mathrm{d}K}{\mathrm{d}r} = \frac{\mathrm{d}}{\mathrm{d}r}\left(\ln c \frac{r}{\ln r}\right) = \ln c \frac{\ln r - 1}{(\ln r)^2}$$

Die Ableitung verschwindet für $\ln r - 1 = 0$ also für

$$r = \mathrm{e} = 2,718\ldots$$

Die zweite Ableitung $\mathrm{d}^2K/\,\mathrm{d}r^2$ an dieser Stelle ist $\mathrm{e}^{-1} > 0$, so daß die Kosten für $r = \mathrm{e}$ tatsächlich *minimal* sind. Da wir keine Radix–r–Codes mit gebrochenem oder gar irrationalem Wert für r konstruieren können, wählen wir für r die am nächsten gelegenen ganzen Zahlen, also 2 oder 3. Würde man nun zu einer Kostenfunktion übergehen, die der Realität näher kommt, so findet man leicht Argumente für die Wahl $r = 2$. Das sind u.a. Gesichtspunkte aus der digitalen Schaltungstechnik, was hier nicht weiter vertieft werden soll.

Eigenschaften des Hartleyschen Informationsmaßes

Seien

$$A = \{a_m\}_{m=1}^{|A|} \quad \text{und} \quad B = \{b_n\}_{n=1}^{|B|}$$

zwei Alphabete mit den Mächtigkeiten $|A|$ bzw. $|B|$. Das Hartleysche Informationsmaß (2.9) hat dann die folgenden Eigenschaften:

1. (Additivität)

$$I_H(A \times B) = I_H(A) + I_H(B),$$

2. (Monotonie)

$$I_H(A) < I_H(B), \quad \text{falls} \quad |A| < |B|,$$

3. (Festlegung der Einheit)

$$I_H(A) = 1, \quad \text{falls} \quad |A| = 2.$$

Die erste Eigenschaft folgt aus der Verwendung der Logarithmus–Funktion, die ein Produkt in eine Summe überführt,

$$I_H(A \times B) = \operatorname{ld}(|A|\,|B|) = \operatorname{ld}|A| + \operatorname{ld}|B| = I_H(A) + I_H(B).$$

Das entspricht offenbar unserem intuitiven Verständnis vom Informationsbegriff, besagt es doch folgendes: Wenn wir erfahren, welcher Buchstabe (a_m, b_n) aus einem Alphabet

$$A \times B = \{(a_m, b_n)\}_{m,n=1}^{|A|,|B|}$$

eingetreten ist, dann sollte das dieselbe Informationsmenge sein, die im Wissen um die einzelnen Buchstaben (Komponenten) a_m und b_n steckt.

Auch die zweite Eigenschaft entspricht offenbar unserer Intuition. Denken wir zum Beispiel an die beiden Versuche Münzwurf und Würfeln, wo es zum einen zwei und zum anderen sechs mögliche Versuchsausgänge gibt. Folglich ist der Überraschungsgehalt beim Münzwurf geringer, was die Relation $\operatorname{ld} 2 < \operatorname{ld} 6$ ausdrückt.

Schließlich legt die dritte Eigenschaft die Einheit der Information fest. Sie bestimmt die Basis r des in (2.9) verwendeten Logarithmus zu $r = 2$.

Das Hartleysche Informationsmaß berücksichtigt nur die Anzahl einer möglichen Menge von Versuchsausgängen, das sind die Buchstaben eines Alphabetes. Dabei ist es vollkommen gleich, wie diese Buchstaben konkret heißen. Auch dies ist typisch für eine Vielzahl weiterer Informationsmaße, denen wir uns später noch zuwenden.

Grenzen des Hartleyschen Informationsmaßes

Die Ausführungen machen deutlich, daß das Hartleysche Maß durchaus einige „wünschenswerte“ Eigenschaften und Interpretationen besitzt, welche wir intuitiv mit einem Informationsmaß verknüpfen. Nichtsdestotrotz weist es aber zumindest einen wesentlichen Mangel auf, den das folgende Beispiel illustriert: Wir betrachten zwei unterschiedliche Versuche, die jeweils nur zwei unterschiedliche Ausgänge haben. Nennen wir sie a_1 und a_2 für den ersten und b_1 und b_2 für den zweiten Versuch. Sei p_m die Wahrscheinlichkeit, a_m zu beobachten, und die für b_n sei q_n. Es gelten immer $p_1 + p_2 = 1$ sowie $q_1 + q_2 = 1$. Nehmen wir nun an, daß die Ausgänge a_m gleichwahrscheinlich sind, also $p_1 = p_2 = 1/2$. Hingegen sei b_2 extrem unwahrscheinlich. Als konkretes

Beispiel kann für den ersten Versuch der Wünzwurf stehen und für den zweiten der Volltreffer beim Lottospiel „6 aus 49“, wofür die Wahrscheinlichkeit bei nur $\binom{49}{6} = q_2 \lesssim 10^{-7}$ liegt. Beobachten wir nun eine Versuchserie zum Münzwurf, so werden wir mit hoher Wahrscheinlichkeit eine recht abwechslungsreiche Folge von Buchstaben a_1 und a_2 erhalten. Hingegen wird beim Lottospiel aller Wahrscheinlichkeit nach zumeist allein das wahrscheinlichere Ereignis b_1 — „kein Volltreffer“ — registriert werden. Wegen dieser Einförmigkeit werden wir rasch ermüden und vielleicht vom Lottospiel ablassen. Wenden wir nun auf die Versuche das Hartleysche Informationsmaß an, so sagt es, daß wir bei einem jeden Versuch pro Versuchsergebnis die Information von einem bit erhalten. Dies widerspricht aber offenbar unserer Intuition, sollte doch der Informationsgehalt der allzu monotonen Lottoserie geringer sein. Allerdings steckt in der einzelnen Nachricht vom Lottogewinn sehr viel Überraschung also auch sehr viel Information. Diese Nachricht ist aber so selten, daß *im Mittel* die Serie zum Lottoversuch wenig Information liefert.

Wir sehen also, daß das Hartleysche Maß in Situationen, wo ungleich wahrscheinliche Ereignisse (Buchstaben) auftreten, unserer intuitiven Vorstellung allzu krass widersprechen kann. Deshalb wenden wir uns nun einem weiteren Informationsmaß zu, welches die unterschiedlichen Wahrscheinlichkeiten berücksichtigt.

2.2 Shannon–Information

Das bereits eingeführte Hartleysche Informationsmaß (2.9) hat einige erhaltenswerte Eigenschaften. Es taugt vor allem dann, wenn die Informationsquelle die Buchstaben des Alphabetes mit der gleichen Wahrscheinlichkeit sendet. Das nun neu zu entwickelnde Maß I sollte deshalb im Spezialfall gleichverteilter Buchstaben in das Hartleysche übergehen.

Zur Entwicklung dieses neuen Maßes gehen wir zunächst von einem Alphabet (2.1) aus, dessen Buchstaben *gleichwahrscheinlich* sind. Wenn wir über die Realisierung eines bestimmten Buchstaben a aus A Kenntnis erhalten, so bekommen wir nach (2.9) die Information $\mathrm{ld}\,|A|$. Dieselbe Informationsmenge können wir aber auch in zwei Schritten erhalten, nämlich indem wir zunächst erfahren, daß a einer bestimmten Teilmenge A_m von A angehört und im zweiten Schritt noch Kenntnis darüber bekommen, um welchen Buchstaben aus A_m es sich handelt.

Denken wir z. B. an unser Kartenspiel vom vorigen Abschnitt. Teilen wir seine 32 Karten

die zwei Gruppen, in der einen die Damen in der anderen der Rest,

$$\begin{aligned} A_1 &= \{\diamondsuit D, \quad \heartsuit D, \quad \spadesuit D, \quad \clubsuit D\} \\ A_2 &= A \setminus A_1. \end{aligned}$$

Nehmen wir nun an, daß $\heartsuit D$ gezogen wurde. Damit erhalten wir die Gesamtinformation $\mathrm{ld}\,|A| = \mathrm{ld}\,32$. Das sollte aber gleich der Informationmenge sein, die wir damit erhalten, daß es sich erstens um eine Dame handelt und daß es zweitens die Farbe $\heartsuit$ ist. Alle Karten sind gleichwahrscheinlich. Deshalb konnten wir die Gesamtinformation schon angeben. Ebenso einfach ist es mit der zweiten Teilinformation $\mathrm{ld}\,|A_1| = \mathrm{ld}\,4$, denn wenn wir schon wissen, daß die Karte aus A_1 stammt, dann gibt es nur noch vier Möglichkeiten, die jeweils die gleiche Wahrscheinlichkeit von $1/4$ haben. Anders verhält es sich jedoch mit der ersten Teilinformation. Eine Dame zu ziehen, hat offenbar die Wahrscheinlichkeit $4/32 = 1/8$ und demzufolge ist die Wahrscheinlichkeit für eine Karte aus A_2 gleich $1 - 1/8 = 7/8$. Das bedeutet aber, die Information darüber, aus welchem der Teilalphabete A_1 und A_2 die Karte stammt, sollten wir wegen der ungleichen Wahrscheinlichkeiten nicht auch noch mit dem Hartleyschen Informationsmaß messen. Benennen wir diese erste Teilinformation deshalb zunächst allgemein mit I_1. Die Forderung, daß die Gesamtinformation gleich der Summe der beiden Teilinformationen ist, lautet dann

$$\mathrm{ld}\,|A| = I_1 + \mathrm{ld}\,|A_1| \ .$$

Lösen wir nun nach der Unbekannten I_1 auf, so erhalten wir

$$I_1 = \mathrm{ld}\,|A| - \mathrm{ld}\,|A_1| = -\,\mathrm{ld}\,\frac{|A_1|}{|A|} \equiv -\,\mathrm{ld}\,p_1 \ .$$

Der Quotient $p_1 \equiv |A_1|/|A|$ ist gerade die Wahrscheinlichkeit, eine Dame zu ziehen. Wir haben somit einen Ausdruck erhalten, der die Information mißt, welche in der Nachricht über ein Ereignis steckt, das die Wahrscheinlichkeit p_1 hat. Vollkommen analog erhalten wir $I_2 = -\,\mathrm{ld}\,\frac{|A_2|}{|A|} \equiv -\,\mathrm{ld}\,p_2$ als die Information darüber, daß die Karte keine Dame ist, also zum Teilalphabet A_2 gehört. Wiederholen wir nun unseren Versuch viele Male, dann wird die erste Teilinformation (‚Dame' oder ‚nicht Dame') im Mittel die Information

$$I = p_1 I_1 + p_2 I_2 = -\sum_{m=1}^{2} p_m \,\mathrm{ld}\,p_m$$

liefern.

Wir verallgemeinern nun unsere Überlegungen auf ein abstraktes Alphabet A mit gleichwahrscheinlichen Buchstaben. Dazu zerlegen wir zunächst A in *paarweise disjunkte* Teilalphabete,

$$A_1, A_2, \ldots, A_m, \ldots, A_M \ ,$$

mit

$$A = \cup_{m=1}^{M} A_m \quad \text{und} \quad A_{m_1} \cap A_{m_2} \quad \text{falls} \quad m_1 \neq m_2 \ . \tag{2.14}$$

Wir erhalten dann für das Teilalphabet A_m die Wahrscheinlichkeit

$$p_m \equiv |A_m|/|A| \ ,$$

und die *mittlere* Information über die Zugehörigkeit einer gezogenen Karte zu einem dieser Teilalphabete ist das gesuchte Informationsmaß :[4]

$$\boxed{I = -\sum_{m=1}^{M} p_m \operatorname{ld} p_m} \quad \text{Shannon–Information} \tag{2.15}$$

Sind alle Ereignisse (Buchstaben) gleich wahrscheinlich, also $p_m = 1/M$, so folgt aus (2.15) $I = \operatorname{ld} M$. Die Shannon–Information geht also in die Hartley–Information über.

Funktionale Begründung der Shannon–Information

Unsere Entwicklung der Formel (2.15) läßt nur rationale Werte für die Wahrscheinlichkeiten zu. Es stellt sich somit die Frage nach ihrem Sinn für beliebige reelle Werte. Weiter unten werden wir die Anwendbarkeit der Formel auch für diesen Fall codierungstheoretisch begründen. Hier soll zunächst eine axiomatische (funktionale) Charakterisierung der Shannon–Information gegeben werden, die auf CHINTSCHIN (1953) und FADDEJEW (1956) zurückgeht [4, 5]:

Satz 2.1 *Seien x_m, $m = 1, 2, \ldots, M$, zufällige Ereignisse, die mit den Wahrscheinlichkeiten p_m eintreffen. Das Informationsmaß I hänge nur von den Wahrscheinlichkeiten ab und erfülle die folgenden Axiome:*

1. *Die Funktion $I(p, 1-p)$ ist für $0 \leq p \leq 1$, stetig und in mindestens einem Punkt positiv. (Wir haben hier $p \equiv p_1$ und $p_2 \equiv 1 - p$ gesetzt.)*

[4] Neben SHANNON hat nahezu gleichzeitig auch NORBERT WIENER in der zweiten Hälfte der 1940er Jahre die Bedeutung dieser Formel erkannt. Sie ist ähnlich fundamental wie z. B. EINSTEINs Energie–Masse–Äquivalenz $E = mc^2$ in der speziellen Relativitätstheorie oder PLANCKs Energie–Frequenz–Beziehung $E = h\nu$ in der Quantenmechanik.

2. $I(p_1, p_2, \ldots, p_M)$ *ist symmetrisch in allen Argumenten,*

$$I(p_1, \ldots p_{m_1}, \ldots p_{m_2}, \ldots, p_M) = I(p_1, \ldots p_{m_2}, \ldots p_{m_1}, \ldots, p_M)$$

für alle $m_1, m_2 = 1, 2, \ldots, M$.

3. Für $M \geq 2$ *und* $p \equiv p_1 + p_2$ *gilt,*

$$I(p_1, p_2, p_3, \ldots, p_M) = I(p, p_3, \ldots, p_M) + p\, I\left(\frac{p_1}{p}, \frac{p_2}{p}\right) \quad . \tag{2.16}$$

4. $I\left(\frac{1}{2}, \frac{1}{2}\right) = 1$.

Die Funktion (2.15) ist die einzige*, welche die obigen Axiome erfüllt.*

Beweis: (s. Anhang)

□

Das erste Axiom impliziert, daß die gesuchte Funktion im *gesamten* Definitionsbereich nicht negativ ist ($I \geq 0$) ist. Dies ist offenbar eine sinnvolle Forderung bzw. Eigenschaft, denn mit der Kenntnis eines Versuchsausganges sollte sich unsere Ungewißheit nicht vergrößern können.

Das zweite Axiom verlangt, daß es gleich ist, in welcher Reihenfolge die Buchstaben numeriert sind.

Das dritte Axiom impliziert die Additivität der Information und ihre statistische Wichtung. Ziehen wir z. B. aus unserem Kartenspiel eine Karte, so erhalten wir die Information 5 bit. Das ist dieselbe Information, die wir bekommen, wenn wir zunächst zwischen den beiden schwarzen Buben nicht unterscheiden und somit nur 31 mögliche Versuchsausgänge haben. Wird ein schwarzer Buben gezogen, so benötigen wir zur vollständigen Charakterisierung des Versuchsausgangs zusätzlich die Information von einem bit, welche zwischen ♠B und ♣B zu unterscheiden gestattet. Dies wird aber nur mit der Wahrscheinlichkeit $p_1 + p_2 = 1/16$ erforderlich sein, mit der ein schwarzer Bube auftritt.

Schließlich bestimmt das vierte Axiom die Maßeinheit, indem in (2.15) die Basis des Logarithmus auf 2 festgelegt wird. Der Satz 2.1 gibt eine funktionale Rechtfertigung der Shannon–Information für *beliebige* diskrete Wahrscheinlichkeitsverteilungen.

Die Abbildung 2.3 zeigt den Verlauf der Shannon–Information I für zwei Ereignisse. Für $p = 1/2$ wird $I(p, 1-p)$ maximal. Haben die Ereignisse sehr unterschiedliche Wahrscheinlichkeiten

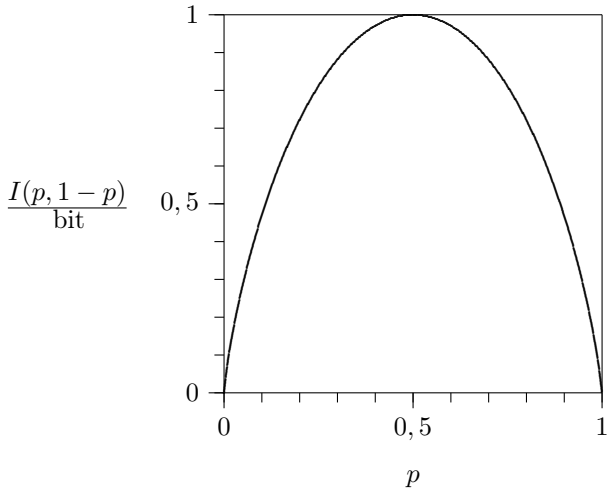

Abb. 2.3: Shannon–Information für zwei Ereignisse mit den Wahrscheinlichkeiten p und $1-p$; $I(p,1-p) = -p\,\mathrm{ld}\,p - (1-p)\,\mathrm{ld}\,(1-p)$

($p \to 0$ oder $p \to 1$), so verschwindet I.[5] Für unseren Versuch „Volltreffer im Lotto“ vom Ende des vorigen Abschnittes erhalten wir z. B.

$$I \approx 10^{-7}\,\mathrm{ld}\,10^{7} + (1-10^{-7})\,\mathrm{ld}\,(1-10^{-7})^{-1} < 25 \cdot 10^{-7}\,\mathrm{bit}\ .$$

Die Information $\mathrm{ld}\,10^{7} \approx 23\,\mathrm{bit}$ bekommen wir, wenn das Ereignis „Volltreffer“ eingetreten ist. Diese Information ist für sich genommen sehr groß. Allerding wird sie bei der Mittelung mit der äußerst geringen Wahrscheinlichkeit dieses Ereignisses multipliziert, so daß das Ereignis „Volltreffer“ *im Mittel* wenig Information liefert. Umgekehrt ist das Ereignis „kein Volltreffer“ zwar sehr wahrscheinlich, allerdings es liefert mit $\mathrm{ld}\,(1-10^{-7})^{-1} \approx 10^{-7}\,\mathrm{bit}$ äußerst wenig Information.

Die Shannon–Information ist nach oben und unten beschränkt, denn es gilt der

Satz 2.2 *Für eine beliebige diskrete Wahrscheinlichkeitsverteilung $\mathcal{P} \equiv \{p_1, \ldots, p_m, \ldots, p_M\}$ gilt für die die Shannon–Information (2.15):*

$$0 \le I(\mathcal{P}) \le \mathrm{ld}\,M\ . \tag{2.17}$$

Die untere Grenze wird genau dann angenommen, wenn alle Wahrscheinlichkeiten p_m bis auf eine null sind. Die obere Grenze nimmt I nur für die Gleichverteilung an.

Beweis: (Übungsaufgabe)

□

[5] Die l'Hospitalsche Regel liefert $\lim_{p\to 0} p \log p = 0$.

Es ist üblich, anstelle von $I(\mathcal{P})$ auch $I(\xi)$ zu schreiben, wobei ξ eine diskrete Zufallsgröße ist, welche gewisse Werte $x_1, x_2, \ldots, x_M$ mit der Wahrscheinlichkeitsverteilung $\mathcal{P} = p_1, p_2, \ldots, p_M$ annimmt. Bei der Schreibweise $I(\xi)$ sollte aber nicht vergessen werden, daß unser Informationsmaß I nur von der Verteilung $\mathcal{P}$ abhängt, nicht aber von den konkreten Werten x_m, welche ξ annimmt. Wir können uns ξ auf irgend einem Alphabet A in natürlicher Weise definiert denken, $\xi : a_m \to x_m \in \mathbb{R}$.

Darüber hinaus sei erwähnt, daß I nicht nur *Information* genannt wird, sondern auch *Unsicherheit* oder *Entropie*. Die zu wählende Sprechweise ergibt sich aus dem Kontext: Erhalten wir über ein zufälliges Ereignis eine Nachricht, so bekommen wir die Information I. Vor dem Eintreffen der Nachricht gibt dieselbe Zahl I unsere Unsicherheit bezüglich der durch $\mathcal{P}$ charakterisierten Ereignisvielfalt an. Dabei dürfen wir nicht vergessen, daß I als *Mittelwert* eingeführt wurde, so daß wir also streng genommen von einer *mittleren* Information bzw. *mittleren* Unsicherheit sprechen müßten, dies aber zur Vereinfachung der Sprechweise häufig unterlassen.

2.3 Bedingte und relative Information

Wir wenden uns im folgenden der Frage zu, wieviel Information wir über eine zufällige Größe η erhalten, wenn wir die Realisierung einer im allgemeinen anderen zufällige Größe ξ kennen. Dabei wird sich herausstellen, daß solch eine Information im Mittel genau dann vorhanden ist, wenn ein gewisser statistischer Zusammenhang besteht. Die einzuführende *relative Information* ist ein quantitatives Maß für diesen Zusammenhang.

Denken wir z. B. an zwei Versuchpersonen, von denen die eine würfelt und die andere eine Münze wirft. Gibt es zwischen ihnen keinerlei Absprachen, also keinen statistischen Zusammenhang, so werden wir aus der Kenntnis eines konkreten Ergebnisses zum Münzwurf nichts über das Ergebnis zum Würfeln erfahren und umgekehrt ebenso. Die relative Information ist hier null. Besteht aber der zweite Versuch anstelle des Münzwurfes darin, daß uns gesagt wird, ob eine gerade Augenzahl gewürfelt wurde, so hat sich offenbar damit unsere Ungewißheit über die gewürfelte Augenzahl verringert, gibt es doch dann anstelle der ürsprünglichen sechs nur noch drei gleich wahrscheinliche Möglichkeiten. Die relative Information ist dann 1 bit und die verbleibende bedingte Unsicherheit $\operatorname{ld} 3$. Addiert ergeben beide $1\,\text{bit} + \operatorname{ld} 3 = \operatorname{ld} 6$, dies ist die Information, welche das Würfelergebnis vollständig charakterisiert. Wir präzisieren nun die

entsprechende Begriffsbildung auf der Grundlage des Shannonschen Informationsmaßes (2.15).

Erfährt man vom Ausgang eines zufälligen Versuches ξ, so bekommt man die Information

$$I(\xi) = -\sum_{m=1}^{M} p_m \operatorname{ld} p_m \ ,$$

falls ξ die Werte $x_1, x_2, \ldots, x_M$ mit den Wahrscheinlichkeiten $p_1, p_2, \ldots, p_M$ annimmt. Wir betrachten einen zweiten Versuch η, der die möglichen Resultate $y_1, y_2, \ldots, y_N$ mit den Wahrscheinlichkeiten $q_1, q_2, \ldots, q_N$ annehme. Der Ausgang von η liefert die Information

$$I(\eta) = -\sum_{n=1}^{N} q_n \operatorname{ld} q_n \ .$$

Hängen die Versuchsergebnisse von ξ und η nicht statistisch ab, dann erhalten wir insgesamt die Information

$$I\big((\xi,\eta)\big) = I(\xi) + I(\eta) \ . \tag{2.18}$$

Darin ist (ξ,η) der aus den einzelnen Zufallsgrößen ξ und η gebildete Zufallsvektor. Hängen aber beide voneinander ab, so erhält man mit dem Ausgang des einen Versuches schon eine gewisse Information über den Ausgang des anderen, so daß im allgemeinen

$$I\big((\xi,\eta)\big) \leq I(\xi) + I(\eta) \tag{2.19}$$

gelten sollte. Diese Gleichung besagt, daß die Information über den Ausgang des Verbundversuches, der durch den Zufallsvektor (ξ,η) beschrieben ist, nicht größer sein kann, als die Summe der Information über die Einzelversuche.

Um den Zusammenhang zwischen $I\big((\xi,\eta)\big)$ und $I(\xi)$ sowie $I(\eta)$ aufzudecken, betrachten wir die Verbundwahrscheinlichkeiten

$$\{s_{mn}\}_{m,n=1}^{M,N} \ ,$$

die zum Zufallsvektor (ξ,η) gehören. Der Ausgang von (ξ,η) liefert die Information

$$I\big((\xi,\eta)\big) = -\sum_{m,n=1}^{M,N} s_{mn} \operatorname{ld} s_{mn} \ . \tag{2.20}$$

Nimmt nun η den Wert y_n an, so ist der Ausgang von ξ durch die bedingte Wahrscheinlichkeitsverteilung $\{p_{m|n}\}_{m=1}^{M}$ charakterisiert, mit

$$p_{m|n} = \begin{cases} s_{mn}/q_n & : \quad q_n > 0 \\ 0 & : \quad q_n = 0 \ . \end{cases} \tag{2.21}$$

Damit haben wir unter der Bedingung $\eta = y_n$ über ξ die Ungewißheit

$$I(\xi|\eta = y_n) = -\sum_{m=1}^{M} p_{m|n} \operatorname{ld} p_{m|n} \ . \tag{2.22}$$

η nimmt den Wert y_n mit der Wahrscheinlichkeit q_n an, so daß wir im Mittel über ξ die Ungewißheit

$$I(\xi|\eta) \equiv \sum_{n=1}^{N} q_n I(\xi|\eta = y_n) \tag{2.23}$$

haben. Dies ist die (mittlere) *bedingte Entropie über ξ, unter der Bedingung, daß η bekannt ist.* Setzt man (2.22) in (2.23) ein, so folgt

$$\begin{aligned} I(\xi|\eta) &= -\sum_{m,n=1}^{M,N} q_n p_{m|n} \operatorname{ld} p_{m|n} \\ &= -\sum_{m,n=1}^{M,N} s_{mn} \operatorname{ld} \frac{s_{mn}}{q_n} \\ &= -\sum_{m,n=1}^{M,N} s_{mn} (\operatorname{ld} s_{mn} - \operatorname{ld} q_n) \\ &= -\sum_{m,n=1}^{M,N} s_{mn} \operatorname{ld} s_{mn} + \sum_{n=1}^{N} q_n \operatorname{ld} q_n . \end{aligned}$$

Es gilt also

$$\boxed{I(\xi|\eta) = I\big((\xi,\eta)\big) - I(\eta)} \tag{2.24}$$

Die (mittlere) Entropie von ξ unter der Bedingung, daß η bekannt ist, ist gleich der Differenz aus der Verbundentropie (Gesamtunsicherheit), $I\big((\xi,\eta)\big)$, und der Einzelentropie, $I(\eta)$. Wichtige Eigenschaften der bedingten Entropie beschreibt der

Satz 2.3 *Sind ξ und η zwei diskrete Zufallsgrößen, so erfüllt die durch (2.24) gegebene bedingte Entropie $I(\xi|\eta)$ die Relationen*

$$0 \le I(\xi|\eta) \le I(\xi) \ . \tag{2.25}$$

Die untere Grenze wird genau dann angenommen, wenn ξ eine Funktion von η ist. Hingegen gilt $I(\xi|\eta) = I(\xi)$ dann und nur dann, wenn ξ und η statistisch unabhängig sind.

Beweis: (s. Anhang)

□

Man kann auch umgekehrt nach der Unsicherheit über η fragen, unter der Bedingung, daß ξ bekannt ist. Vollkommen analog erhält man

$$I(\eta|\xi) = I\big((\xi,\eta)\big) - I(\xi) \ . \tag{2.26}$$

$I(\xi)$ ist nach unserer bisherigen Diskussion die Unsicherheit über ξ. Ziehen wir von dieser die Unsicherheit (2.24) ab, erhalten wir die sogenannte *relative Information* [6]

$$\boxed{I_T(\xi,\eta) \equiv I(\xi) - I(\xi|\eta)} \tag{2.27}$$

Sie beschreibt die (mittlere) *Information*, welche in η über ξ enthalten ist. Die Umkehrung gilt auch, d.h., die Transinformation ist symmetrisch in den beiden Argumenten, denn

$$\begin{aligned} I_T(\xi,\eta) &= I(\xi) - I(\xi|\eta) \\ &= I(\xi) - \Big[I\big((\xi,\eta)\big) - I(\eta)\Big] \\ &= I(\eta) - \Big[I\big((\eta,\xi)\big) - I(\xi)\Big] \\ &= I_T(\eta,\xi) \ . \end{aligned}$$

Es gilt der

Satz 2.4 *Sind ξ und η zwei diskrete Zufallsgrößen, so erfüllt die durch (2.27) gegebene Transinformation $I_T(\xi,\eta)$ die Relationen*

$$0 \le I_T(\xi,\eta) \le \min\{I(\xi), I(\eta)\} \ . \tag{2.28}$$

Die untere Grenze wird genau dann angenommen, wenn ξ und η unabhängig sind. Hingegen gilt $I_T(\xi,\eta) = I(\xi)$ dann und nur dann, wenn ξ eine Funktion von η ist.

Beweis: (Die Aussagen dieses Satzes folgen unmittelbar aus Satz 2.3.)

□

Nach diesem Satz ist $I(\eta) \ge I(\xi)$ eine notwendige Bedingung dafür, daß ξ eine Funktion von η ist, sie ist aber nicht hinreichend. Darüber hinaus folgt wegen der Symmetrie (2.28) auch, daß $I_T(\xi,\eta) = I(\eta)$ dann und nur dann gilt, wenn η eine Funktion von ξ ist.

[6] Sie wird auch *Transinformation* oder *Synentropie* genannt. Im Englischen ist die Bezeichnung *mutual information* üblich.

Beispiel: Kartenspiel

Wir betrachten zwei Versuche, welche vom Ziehen einer Karte aus einem Skatspiel abgeleitet sind:

1. Versuch, ξ:

Ereignis	Wahrscheinlichkeit
E_1: *Ziehe rote Karte.*	$p_1 = 1/2$
E_2: *Ziehe schwarze Karte.*	$p_2 = 1/2$

2. Versuch, η:

Ereignis	Wahrscheinlichkeit
F_1: *Ziehe Karte* ♢.	$q_1 = 1/4$
F_2: *Ziehe Karte* ♡.	$q_2 = 1/4$
F_3: *Ziehe Karte* ♠.	$q_3 = 1/4$
F_4: *Ziehe Karte* ♣.	$q_4 = 1/4$

Somit beträgt unsere Ungewißheit über den Ausgang des ersten Versuches $I(\xi) = 1\,\text{bit}$ und über den zweiten Versuch $I(\eta) = 2\,\text{bit}$. Der Verbundversuch (ξ, η) hat die Verteilung

$$\begin{array}{llll} s_{11} = 1/4 & s_{12} = 1/4 & s_{13} = 0 & s_{14} = 0 \\ s_{21} = 0 & s_{22} = 0 & s_{23} = 1/4 & s_{24} = 1/4 \ . \end{array}$$

Im Verbundversuch steckt also die Ungewißheit $I\big((\xi, \eta)\big) = 2\,\text{bit}$. Für die bedingten Entropien folgt dann nach (2.24) und (2.26)

$$I(\eta|\xi) = 1\,\text{bit} \quad \text{bzw.} \quad I(\xi|\eta) = 0 \ .$$

Dies ist ein plausibles Resultat, denn kennen wir die Farbe (Ergebnis zum Versuch ξ), so haben wir über η noch die Ungewißheit von 1 bit, um zu entscheiden, ob „♢ oder ♡“ bzw. „♠ oder ♣“ gezogen wurde. Umgekehrt haben wir mit der Kenntnis des Versuchsausganges η keine Ungewißheit mehr über ξ. Für die Transinformation ergibt sich also $I_T(\xi, \eta) = 1\,\text{bit}$.

Bergersches Diagramm

Aus der Sicht der Informationsübertragung über einen gestörten Nachrichtenkanal kann man sich die verschiedenen Informationsbegriffe im *Bergerschen Diagramm* veranschaulichen (Abb 2.4).

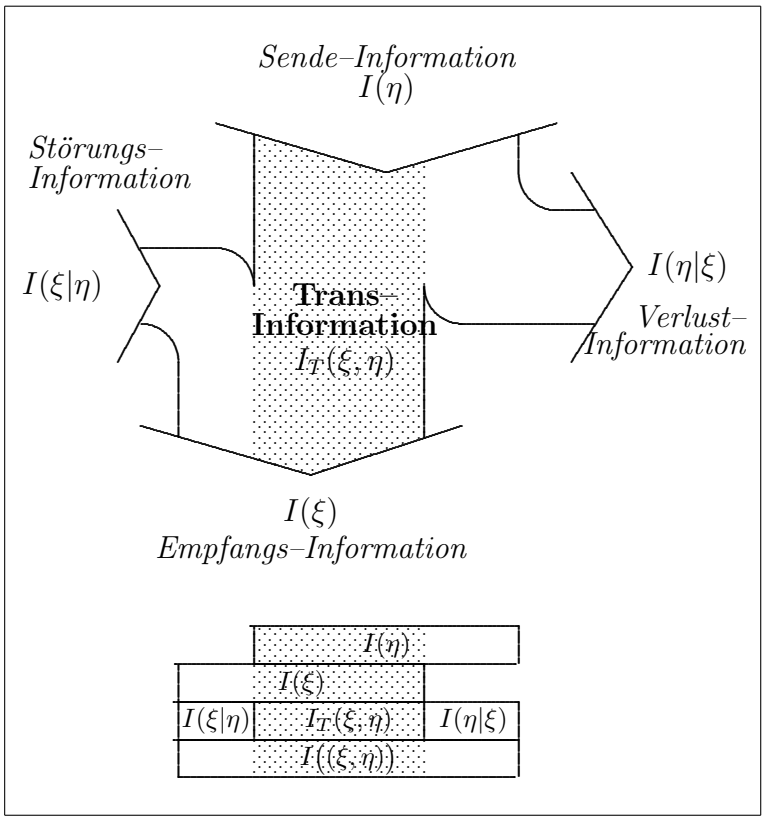

Abb. 2.4: Bergersches Diagramm zum Informationsfluß bei der Nachrichtenübertragung über einen gestörten Kanal

Auf den Kanal wird eine gewisse *Sende–Information* $I(\eta)$ gegeben. Im allgemeinen kommt jedoch aus den verschiedensten technischen Gründen (infolge von Störungen) nicht die gesamte gesandte Information beim Empfänger an. Hier erhalten wir die *Empfangs–Information* $I(\xi)$. Sie enthält zum einen die Transinformation $I_T(\xi,\eta)$ (punktiert gezeichnet) und darüber hinaus die Information von den Störungen $I(\xi|\eta)$, welche auch *Irrelevanz* heißt.[7] Andererseits kommt beim Empfänger ein Teil der gesandten Information nicht an. Das ist die sogenannte *Verlust–Information*[8], $I(\eta|\xi)$. Wir werden weiter unten im Abschnitt zur Nachrichtenübertragung über gestörte Kanäle auf dieses Schema noch genauer eingehen.

[7]Weitere Bezeichnungen für die *Störungs–Information* sind *Fehl–*, *Streu–* oder *Kontext–Information* sowie *Dissipation.*

[8]Die *Verlust–Information* heißt auch *Rückschluß–Entropie* und *Äquivokation.*

Entropie diskreter Nachrichtenquellen

Wir verwenden nun den Begriff der bedingten Entropie zur Charakterisierung einer Nachrichtenquelle, die wir als diskreten stationären stochastischen Prozeß auffassen,

$$\ldots \xi_{t-2}\, \xi_{t-1}\, \xi_t\, \xi_{t+1} \ldots \tag{2.29}$$

Die Zufallsgröße ξ_t nehme die Werte $x_1, x_2 \ldots x_M$ mit den Wahrscheinlichkeiten $p_1, p_2 \ldots p_M$ an. Wegen der geforderten Stationarität hängen die Wahrscheinlichkeiten nicht vom Zeitpunkt t ab. Die Verbundwahrscheinlichkeiten zum d–dimensionalen Zufallsvektor

$$\boldsymbol{\xi}_{t,d} \equiv (\xi_{t-d}, \ldots, \xi_{t-1})$$

bezeichnen wir mit

$$p_{m_d \ldots m_1} \quad .$$

Wir fragen nun nach der Unsicherheit über den zukünftigen Zustand ξ_t, unter der Bedingung, daß d vergangene $\boldsymbol{\xi}_{t,d}$ bekannt sind. Nach (2.24) ist diese durch die folgenden Gleichungen gegeben (beachte: $(\boldsymbol{\xi}_{t,d}, \xi_t) = \boldsymbol{\xi}_{t+1,d+1}$):

$$\begin{aligned}
& I(\xi_t | \boldsymbol{\xi}_{t,d}) \\
&= I(\boldsymbol{\xi}_{t+1,d+1}) - I(\boldsymbol{\xi}_{t,d}) \\
&= -\sum_{m_d,\ldots,m_1,m_0} p_{m_d \ldots m_1 m_0}\ \mathrm{ld}\, p_{m_d \ldots m_1 m_0} + \ldots \\
&\quad \ldots + \sum_{m_d,\ldots,m_1} p_{m_d \ldots m_1}\ \mathrm{ld}\, p_{m_d \ldots m_1} \\
&= -\sum_{m_d,\ldots,m_1,m_0} p_{m_d \ldots m_1 m_0}\ \mathrm{ld}\, \frac{p_{m_d \ldots m_1 m_0}}{p_{m_d \ldots m_1}} \\
&= \sum_{m_d,\ldots,m_1} p_{m_d \ldots m_1} \left(-\sum_{m_0} p_{m_0 | m_d \ldots m_1}\ \mathrm{ld}\, p_{m_0 | m_d \ldots m_1} \right) .
\end{aligned} \tag{2.30}$$

Hierin ist

$$p_{m_0 | m_d \ldots m_1} \equiv \begin{cases} \frac{p_{m_d \ldots m_1 m_0}}{p_{m_d \ldots m_1}} & : \quad p_{m_d \ldots m_1} > 0 \\ 0 & : \quad \text{sonst} \end{cases}$$

die Wahrscheinlichkeit für $\xi_t = x_{m_0}$, unter der Bedingung $\boldsymbol{\xi}_{t,d} = (x_{m_d,\ldots,m_1})$. Der Zeitpunkt t ist wegen der vorausgesetzten Stationarität beliebig.

Wenn wir nun die Anzahl d der als bekannt vorausgesetzten vergangenen Werte erhöhen, so verringert sich im allgemeinen die Unsicherheit über den zukünftigen Wert ξ_t oder bleibt

konstant, erhöhen kann sie sich dadurch jedoch *nicht*. Andererseits ist $I(\xi_t|\boldsymbol{\xi}_{t,d})$ nach unten durch null beschränkt. Folglich existiert der Grenzwert

$$\boxed{h \equiv \lim_{d\to\infty} I(\xi_t|\boldsymbol{\xi}_{t,d})} \tag{2.31}$$

Er ist die (mittlere) Unsicherheit über den zukünftigen Buchstaben x_{m_0}, wenn alle vergangenen Buchstaben $\ldots, x_{m_{-2}}, x_{m_{-1}}$ bekannt sind. Damit charakterisiert h eine Nachrichtenquelle bzw. einen Prozeß (2.29) auf fundamentale Weise. h heißt *Entropie der Nachrichtenquelle* (2.29). Die Bezeichnung *Quellentropie* für h ist auch üblich. Alternativ zu (2.31) können wir auch wie folgt schreiben:

$$\begin{aligned} h &= \lim_{d\to\infty} \frac{I(\boldsymbol{\xi}_{t,d})}{d} \\ &= \lim_{d\to\infty} I(\boldsymbol{\xi}_{t+1,d+1}) - I(\boldsymbol{\xi}_{t,d}) \ . \end{aligned}$$

Wir betrachten nun zwei Spezialfälle:

1. Beispiel: Vollkommen unabhängiger Prozeß Sind die ξ_t vollkommen statistisch unabhängig[9], so lassen sich die Verbundwahrscheinlichkeiten als Produkt der entsprechenden Einzelwahrscheinlichkeiten berechnen,

$$p_{m_d \ldots m_1 m_0} = p_{m_d} \ldots p_{m_1} p_{m_0} \ .$$

Wir erhalten dann nach kurzer Rechnung

$$I(\xi_t|\boldsymbol{\xi}_{t,d}) = I(\xi_t) \ .$$

Hierin können wir wegen der vorausgesetzten Stationarität auch $I(\xi_t) = I(\xi)$ schreiben. Für die Quellentropie (2.31) folgt somit

$$h = I(\xi) = -\sum_{m=1}^{M} p_m \ \mathrm{ld}\, p_m .$$

Beim fairen Würfel gilt z. B. $h = \mathrm{ld}\, 6$, beim Münzwurf $h = 1\,\mathrm{bit}$, und beim Kartenziehen (mit Zurücklegen und Mischen!) erhalten wir $h = 5\,\mathrm{bit}$.

[9] $\{\xi_t\}$ sei also ein sogenannter i.i.d. (independent identically distributed) Prozeß.

2. Beispiel: Markoff–Ketten Ist unsere Nachrichtenquelle eine Markoff–Kette der Ordnung D, so gilt

$$p_{m_0|m_d \ldots m_1} = p_{m_0|m_D \ldots m_1} \text{ , für alle } \quad d \geq D \ ,$$

und wir erhalten

$$h = I(\xi_t | \boldsymbol{\xi}_{t,D}) \ .$$

Hier liefern also schon D vergangene Werte alle verfügbare Information über den zukünftigen Buchstaben. Mit anderen Worten, jeder weitere vergangene Zustand ξ_{t-d}, $d > D$, liefert keine zu $\boldsymbol{\xi}_{t,D}$ *zusätzliche* Information über ξ_t. (Daraus darf aber *nicht* gefolgert werden, daß ξ_t und ξ_{t-d} für $d > D$ statistisch unabhängig wären und damit $I(\xi_t|\xi_{t-d})$ verschwinden würde.)

Redundanz einer Nachrichtenquelle

Allgemein liegt die Quellentropie im Bereich

$$0 \leq h \leq \operatorname{ld} M \ ,$$

wenn M die Anzahl der Werte ist, welche ξ_t annehmen kann. Dies motiviert die Einführung einer Größe, welche die Quellentropie ins Verhältnis zum maximal möglichen Wert $\operatorname{ld} M$ setzt,

$$\boxed{R \equiv \frac{\operatorname{ld} M - h}{\operatorname{ld} M}} \tag{2.32}$$

R heißt *Redundanz der Nachrichtenquelle* oder auch *Weitschweifigkeit.*

Empfangen wir Buchstaben von einer Nachrichtenquelle mit der Redundanz $R = 1$ (also, $h = 0$), dann bekommen wir im allgemeinen mit jedem später empfangenen Buchstaben immer weniger Information. Der Überraschungsgehalt pro Buchstabe wird also immer kleiner, und asymptotisch können wir jeden zukünftigen Buchstaben aus den vergangenen mit an Sicherheit grenzender Wahrscheinlichkeit exakt vorhersagen.

Hingegen bedeutet $R = 0$ (also, $h = \operatorname{ld} M$), daß mit jedem Buchstaben ein Maximum an möglicher Information ausgesandt wird, wobei die Kenntnis der Vergangenheit den Überraschungsgehalt zukünftiger Werte nicht verringert. Hier bringen also die vergangenen Beobachtungen (Buchstaben) keinen Vorhersagenutzen. Andererseits ist aber bei solchen Quellen der

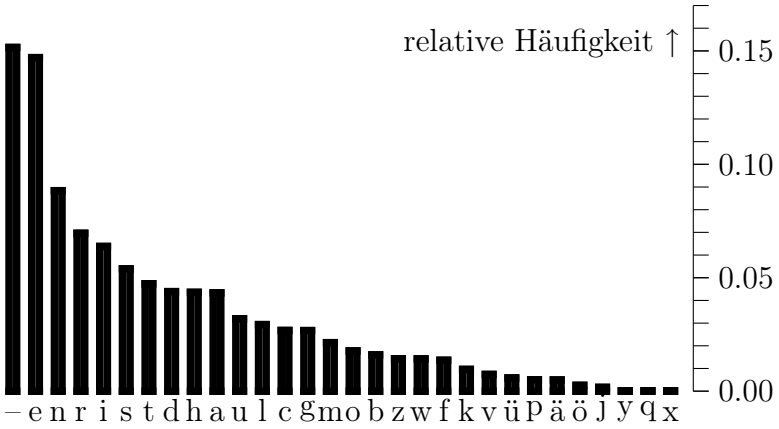

Abb. 2.5: Relative Häufigkeiten von Buchstaben deutscher Texte (nach [19], S. 218)

Datenstrom pro Zeiteinheit (die Bitrate) hinsichtlich der Informationsübertragungsrate am effektivsten genutzt. Weiter unten werden wir sehen, daß es ein Hauptziel einer Codierung von Buchstabenfolgen (Signalen) ist, die Redundanz möglichst gering zu halten, um die Kapazität des Nachrichtenkanals, Information zu übertragen, optimal auszunutzen. Man kann aber auch ein gewisses Maß an „kontrollierter Redundanz“ in den Code einbauen, um Übertragungsfehler erkennen und möglicherweise auch korrigieren zu können.

Beispiel:

Quellentropie natürlicher Schriftsprachen

Für einige natürliche Schriftsprachen wurde die Quellentropie von verschiedenen Autoren geschätzt. Probleme entstehen dabei durch die allzu große Anzahl der Buchstaben bzw. der aus ihnen bildbaren Worte. Geht man z. B. von 32 Buchstaben aus[10], so kann man allein zur Länge 5 schon $32^5 = 2^{25} \gtrsim 3 \times 10^7$ viele verschiedene Worte bilden, zu denen die Wahrscheinlichkeiten aus möglichst langen repräsentativen Stichproben (Texten) geschätzt werden müssen. Tatsächlich tritt aber nur ein geringer Bruchteil aller kombinatorischen Möglichkeiten (Variationen mit Wiederholung) mit Wahrscheinlichkeiten auf, die signifikant größer als null sind. So kennt man im Englischen wie im Deutschen jeweils nur etwa 500 000 verschiedenen Worte, Fachbegriffe sind hier ausgeschlossen. Die Verbundwahrscheinlichkeiten niedriger Ordnung können noch durch Auswertung langer Texte ausgezählt werden. Im Deutschen erhält man für die Einzelwahrscheinlichkeiten die Verteilung in Abb. 2.5. Am häufigsten ist hier das Zeichen

[10]Es werden hierbei nur die Buchstaben und einige Sonderzeichen unterschieden, zwischen Groß– und Kleinschreibung wird nicht differenziert.

Wortlänge d	Deutsch	Englisch	Französisch
0	4,70 oder 4,75	4,70 oder 4,75	4,70 oder 4,75
1	4,10	4,03 .. 4,14	3,95
2		3,32 .. 3,56	3,17
3		3,1 .. 3,3	2,83
4			
5		2,1 .. 2,6	
8		2,3	
9		1,9	
∞	**1,3**	**0,6 .. 1,3**	**1,22 .. 1,40**
	Russisch	Spanisch	Samoa
0	5 oder 5,13		4 oder 4,09
1	4,35 .. 4,55	3,98	3,40
2	3,44 .. 3,52		2,68
3	3,01		
4			
5			
8			
9			
∞	**0,83 .. 1,40**		

Tab. 2.2: Bedingte Entropien $I(\xi_t|\boldsymbol{\xi}_{t,d})$ verschiedener Schriftsprachen in Einheiten von bit/Buchstabe. d: Anzahl vergangener Buchstaben

„–“, welches für den Wortzwischenraum steht. Im Englischen sind die häufigsten Buchstaben `etanori` und im Französischen `esiantur`.

Für größere Wortlängen ist eine direkte Schätzung von Wahrscheinlichkeiten durch Auszählen recht ungenau. Hier bedient man sich anderer Verfahren, etwa der Ratemethode von Shannon und Kolmogoroff (s. [7], S. 197 ff.). Hierbei wird das Wissen von Versuchspersonen ausgenutzt wird, welche die Sprache gut kennen und deshalb angefangene Wortteile, Wörter, Satzteile oder gar Sätze durch Raten mehr oder weniger „sinnvoll“ fortsetzen können.

Einen Überblick zu den bedingten Entropien (2.30) bzw. zur Quellentropie (2.31) verschie-

dener Schriftsprachen gibt Tab. 2.2. Für die verschiedenen Sprachen sind in der Zeile $d = 0$ die Werte $\operatorname{ld} M$ eingetragen, wobei M die Anzahl der unterschiedenen Buchstaben ist. Im Deutschen, Englischen und Französischen wurden einmal $M = 26$ ($\operatorname{ld} 26 \approx 4,70\,\text{bit}$) und ein andermal $M = 27$ ($\operatorname{ld} 27 \approx 4,75\,\text{bit}$) Buchstaben unterschieden, je nachdem ob der Wortzwischenraum als zusätzliches Zeichen interpretiert wurde oder nicht. Im Russischen wurde einmal das bis zum Jahre 1917 gebräuchliche Alphabet aus $M = 35$ ($\operatorname{ld} 35 \approx 5,13\,\text{bit}$) Buchstaben verwendet und ein andermal das modernere aus $M = 32$ ($\operatorname{ld} 32 = 5\,\text{bit}$) vielen Buchstaben. Die polynesische Sprache Samoa unterscheidet nur 16 Buchstaben. Die mittlere Wortlänge beträgt hier nur ca. 3,2 Buchstaben. In den anderen Sprachen ist sie deutlich größer, z. B. im Englischen ca. 4,1 Buchstaben/Wort und im Russischen ca. 5,3 Buchstaben/Wort. Folglich hat der Wortzwischenraum in Samoa eine relativ große Auftrittswahrscheinlichkeit.

Neben der Wahl der Alphabete spielt noch die Art des untersuchten Textes eine wesentliche Rolle. So haben z. B. Geschäftstexte, die wiederholt gewisse Floskeln enthalten, einen geringeren Informationgehalt (sprich: kleinere Entropien) als etwa literarische Texte. Für einige Wortlängen d lassen sich Entropien nur recht unzuverlässig schätzen, weshalb in der Tabelle einige Zeilen recht unvollständig sind.

Wie unterschiedlich die Alphabete der verschiedenen Sprachen auch sein mögen, asymptotisch, d.h. für Gedächtnistiefen $d \to \infty$, stellt sich ein Grenzwert h ein (in der Tabelle fett geschrieben), der im Rahmen der Schätzgenauigkeiten für alle untersuchten Sprachen gleich groß ist. Dafür gibt es eine einfache Erklärung: Die Schriftsprachen haben sich aus natürlichen, zunächst nur gesprochenen Sprachen entwickelt. Eine Buchstabenfolge kann umkehrbar eindeutig durch eine Folge von Lauten ausgedrückt werden. Bei den in der Tabelle berücksichtigten Sprachen entspricht ein Laut etwa einem Buchstaben, in anderen Sprachen, so z. B. bei den altägyptischen Hieroglyphen oder dem Chinesischen, kann ein Buchstabe (Zeichen) allerdings auch mehreren Lauten entsprechen. Nun werden bei normaler Sprechgeschwindigkeit etwa 10 Laute pro Sekunde artikuliert. Folglich vermittelt der Mensch mit der Sprache etwa 10bit/s. Dies entspricht ungefähr der Information, die der Mensch pro Sekunde ins Langzeitgedächtnis aufnehmen kann. Gehen wir also davon aus, daß alle Menschen hinsichtlich dieser Aufnahmerate biologisch gleich veranlagt sind, so verwundert es nicht, daß die Informationsrate verschiedener natürlicher Sprachen gleich groß ist.

2.4 Information kontinuierlicher Quellen

Wir wollen nun die bisherige Begriffsbildung auf Informationsquellen anwenden, deren Alphabet (Wertevorrat) unendlich groß ist. Unser Shannonsches Informationsmaß (2.15) ist in diesem Falle offenbar wenig sinnvoll. Betrachten wir z. B. ein Alphabet A mit $|A| = M$ gleich wahrscheinlichen Buchstaben, die statistisch unabhängig gesendet werden. Dann erhalten wir pro Buchstabe die Information $I = \mathrm{ld}\, M$. Im Grenzfall unendlich vieler Buchstaben wird also die Information pro Buchstabe unendlich groß, $\lim_{M\to\infty} \mathrm{ld}\, M = \infty$.

Aber auch für andere Wahrscheinlichkeitsverteilungen auf abzählbar unendlich großen Alphabeten kann die Reihe (2.15) divergieren, so z. B. für Verteilungen der Form

$$p_m = \begin{cases} p & : \quad m = 1, \quad 0 \le p < 1 \\ \dfrac{1-p}{M-1} & : \quad m = 2, 3, \ldots, M \ . \end{cases} \tag{2.33}$$

Das Informationsmaß (2.15) liefert hier

$$I = -p\,\mathrm{ld}\, p - (1-p)\,\mathrm{ld}\,(1-p) + (1-p)\,\mathrm{ld}\,(M-1) \ , \tag{2.34}$$

so daß

$$\lim_{M\to\infty} I = \infty \quad \text{für } 0 \le p < 1$$

gilt. Wir können also eine Folge statistisch unabhängiger Buchstaben erzeugen, wo ein einzelner Buchstabe, z. B. a_1, beliebig häufig auftritt. Halten wir p_1 beliebig nahe bei 1 fest, so kann dennoch der mittlere Überraschungsgehalt pro Buchstabe beliebig groß werden, wenn nur hinreichend viele andere Buchstaben a_m, $m = 2, \ldots, M$, mit positiven (dicht bei 0 liegenden) Wahrscheinlichkeiten $(1-p)/(M-1)$ auftreten. Die Abbildung 2.6 verdeutlicht dies für $p = 0,99$ und $M = 2^{16}$. Würden wir alle Parameter beibehalten und nur M vergrößern, so erhielten wir zur gleichen Zeitreihenlänge ähnliche Bilder, weil unser Auge die feinen Unterschiede in der Lage der Punkte nicht ausmachen könnte. Der Informationsgehalt I würde aber beliebig anwachsen, wenn auch recht langsam: für $M \to M^2$ auf etwa das Doppelte.

Bevor wir unsere Betrachtungen zu unendlich großen Alphabeten fortsetzen, wollen wir anmerken, daß keine praktischen Situationen denkbar sind, wo unendlich viele Buchstaben tatsächlich auftreten würden. Denken wir z. B. an eine elektrische Meßspannung U, die über einem Kondensator der Kapazität C anliegt. Der kleinste mögliche Spannungssprung ist dann

$$\Delta U_{\min} = e/C \ ,$$

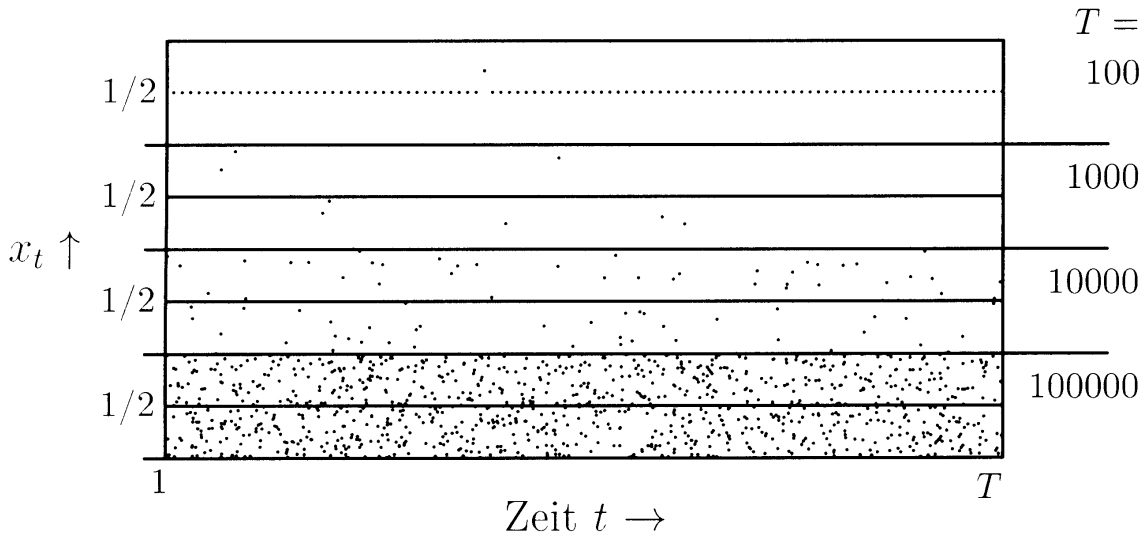

Abb. 2.6: Realisierungen statistisch unabhängiger Zufallsfolgen (Zeitreihen) $\{x_t\}_{t=1}^{T}$, mit dem Wertevorrat $\{x_m = (m-1)/M\}_{m=1}^{M}$, $M = 2^{16}$. Die Verteilung ist im wesentlichen durch (2.33) gegeben, nur daß hier die Einzelwahrscheinlichkeit $p_{1+M/2} = p = 0,99$ für das Ereignis $x_t = 1/2$ gesetzt wurde. Folglich tritt in der Zeitreihe der Wert $1/2$ recht häufig auf, während ein jeder der anderen Werte die geringe Wahrscheinlichkeit von $(1-p)/(M-1)$ hat. Erst für recht lange Reihen erhält man einen repräsentativen Eindruck von der zugrunde liegenden Verteilung. Nach (2.34) ist der Informationsgehalt $I \approx 0,24\,\mathrm{bit/Buchstabe}$. Für größere Werte von M würden die Bilder ebenso aussehen, der Informationsgehalt I würde aber mit M gegen unendlich streben

mit der Elementarladung $e \approx 1,602 \times 10^{-19}$ As. Eine jede reale elektrische Kapazität C ist endlich, und das elektrische Ladungsquant e kann nicht unterschritten werden. Folglich ist der kleinste beobachtbare Spannungssprung auch größer als null, $\Delta U_{\min} > 0$. Damit werden wir in einem endlich großen Spannungsbereich, etwa der Breite $\Delta U = 10\mathrm{V}$, immer nur *endlich* viele verschiedene Werte, das sind die Buchstaben a_m unseres abstrakten Alphabetes A, messen können. Allerding kann diese endliche Anzahl $|A| \equiv M$ recht groß werden. In unserem Beispiel erhalten wir für die Kapazität $C = 10^{-3}\,\mathrm{F}$,

$$M = \frac{\Delta U}{\Delta U_{\min}} = \frac{\Delta U \times C}{e} \approx 6 \times 10^{16} \quad .$$

Im Falle der Gleichverteilung würden wir also mit jeder Messung die Information

$$\mathrm{ld}\, M \approx 96 \times \mathrm{ld}\, 10 \approx 320\,\mathrm{bit}$$

erhalten (wenn die Meßwerte statistisch unabhängig sind). Dies ist eine recht große Informationsmenge, und sie wächst mit der Kapazität C, was uns eine gewisse Motivation gibt, auch stetige Zufallsgrößen zu betrachten. Letztlich gründet aber unser Interesse an kontinuierlichen

Zufallsgößen, bzw. an Alphabeten unendlicher Mächtigkeit, vor allem darin, daß sie in verschiedenen Theorien eine wichtige Rolle spielen, etwa jener dynamischer Systeme (Ergodentheorie). Im Abschnitt **??** werden wir dies noch genauer ausführen.

Sei ξ eine Zufallsgröße, die durch eine stetige Verteilungsfunktion $P(x)$ charakterisiert ist. Wir nähern nun ξ durch eine diskrete Zufallsgröße

$$\xi_M \equiv \frac{[M\xi]}{M} \tag{2.35}$$

an, worin $[x]$ den ganzen Teil von x bezeichnet. ξ_M wird die Verteilung

$$p_{M,m} = \text{prob}\left(\xi_M = \frac{m}{M}\right) = \text{prob}\left(\frac{m}{M} \leq \xi < \frac{m+1}{M}\right)$$

zugeordnet, wobei m alle ganzen Zahlen durchläuft. Wir setzen voraus, daß $I(\xi_1)$ endlich ist.[11] Dann ist auch $I(\xi_M)$ für jedes $M < \infty$ endlich, denn aus der Eigenschaft (2.16) folgt

$$I(\xi_M) \leq I(\xi_1) + \text{ld}\, M \ . \tag{2.36}$$

Ist $P(x)$ stetig, so strebt $I(\xi_M)$ mit M gegen ∞.

Informationsdimension

Das Anwachsen von M in (2.35) bedeutet, daß die kontinuierliche Zufallsgröße ξ immer genauer durch ξ_M approximiert wird, folglich erhalten wir mit der Kenntnis darüber, welchen Wert ξ_M annimmt, mit wachsender Genauigkeit M auch immer mehr Information über ξ, $\lim_{M\to\infty} I(\xi_M) = I(\xi) = \infty$. Allerdings zeigt es sich, daß $I(\xi_M)$ mit recht unterschiedlichen Geschwindigkeiten wachsen kann, was für große Werte von M durch den folgenden Quotienten beschrieben wird,

$$\boxed{D_{\text{I}}(\xi) \equiv \lim_{M\to\infty} \frac{I(\xi_M)}{\text{ld}\, M}} \tag{2.37}$$

Existiert dieser Grenzwert, so heißt er *Informationsdimension der Zufallsgröße* ξ.[12] Für $M \to \infty$ können wir also

$$I(\xi_M) \simeq D_{\text{I}}(\xi) \times \text{ld}\, M$$

[11] $I(\xi_1)$ ist z.B. dann endlich, wenn ξ eine Wahrscheinlichkeitsdichte $\text{d}P(x)/\,\text{d}x \equiv p(x)$ mit einem beschränkten Träger besitzt. Wir können uns dann ξ geeignet normiert denken, so daß der Träger das Intervall $[0,1]$ ist. Für das folgende ist diese Normierung aber unwesentlich.

[12] Existiert dieser Grenzwert nicht, so können anstelle von lim die Limites $\overline{\lim}$ (limes superior) und $\underline{\lim}$ (limes inferior) betrachtet werden. Wir wollen hier jedoch immer die Existenz des Grenzwertes annehmen.

schreiben.

Aus (2.36) und $I(\xi_1) \leq \operatorname{ld} M$ folgt, daß die Informationsdimension einer eindimensionalen Zufallsgröße ξ nicht größer als 1 sein kann. Andererseits kann aber $D_{\mathrm{I}}(\xi)$ auch nicht negativ sein, denn mit $I(\xi_M)$ ist auch der Quotient $I(\xi_M)/\operatorname{ld} M$ immer größer oder gleich null. Allgemein gilt also

$$0 \leq D_{\mathrm{I}}(\xi) \leq 1 \tag{2.38}$$

Ist ξ absolut stetig verteilt, so gilt $D_{\mathrm{I}} = 1$.[13]

Zur Illustration betrachten wir die Zeitreihen in Abb. 2.6. Ihnen liegt eine Verteilungsfunktion

$$P_M(x) \equiv \sum_{m:x_m<x} p_m$$

zugrunde, die für $M \to \infty$ gegen die folgende kontinuierliche Verteilung strebt (Abb. 2.7):

$$P_\infty(x) = \begin{cases} 0 & : \quad x \leq 0 \\ (1-p)x & : \quad 0 < x \leq 1/2 \\ p + (1-p)x & : \quad 1/2 < x \leq 1 \\ 1 & : \quad 1 < x \ . \end{cases} \tag{2.39}$$

Die Information zur entsprechend vergröberten Zufallsgröße ξ_M ist hier durch (2.34) gegeben. Setzen wir dies in (2.37) ein, so erhalten wir die Informationsdimension

$$D_{\mathrm{I}}(\xi) = 1 - p \ .$$

Der Parameter $p \in [0, 1/2]$, also die Einzelwahrscheinlichkeit dafür, daß ein ganz bestimmter Wert, $x = 1/2$, angenommem wird, steuert den Wert der Informationsdimension zwischen 0 und 1. Je wahrscheinlicher dieser eine Wert ist, desto geringer ist die Informationsdimension und desto langsamer wächst der Informationsgehalt bei Vergrößerung der Auflösung.[14]

[13] Eine Zufallsgröße ξ heißt *absolut stetig verteilt bez. des Maßes* λ, wenn für jede λ–meßbare Menge *Nullmenge* $B \subseteq \mathbb{R}$, für die $\lambda(B) = 0$ gilt, auch $\mathrm{prob}(\xi \in B) = 0$ erfüllt ist. Wird zur Vereinfachung der Sprechweise das Maß λ nicht genannt, so versteht es sich von selbst, daß λ das Lebesgue–Maß auf der σ–Algebra der Borelmengen in $\mathbb{R}$ ist. Ist ξ absolut stetig verteilt, dann gibt es eine Funktion $p(x)$, so daß die Verteilungsfunktion $P(x)$ von ξ wie folgt darstellbar ist

$$P(x) = \int_{-\infty}^{x} p(x^*)\,\mathrm{d}x^* \ .$$

In der Integrationstheorie schreibt man hier genauer $\mathrm{d}\lambda$ statt $\mathrm{d}x^*$.

[14] Für $p > 0$ hat die Verteilung (2.39) eine Sprungstelle an der Stelle $x = 1/2$. Damit ist die Verteilung, bzw. eine so verteilte Zufallsgröße ξ, nicht absolut stetig, denn für das Lebesgue–Maß erhalten wir $\lambda(1/2) = 0$, wohingegen $\mathrm{prob}(\xi = 1/2) = p > 0$ gilt.

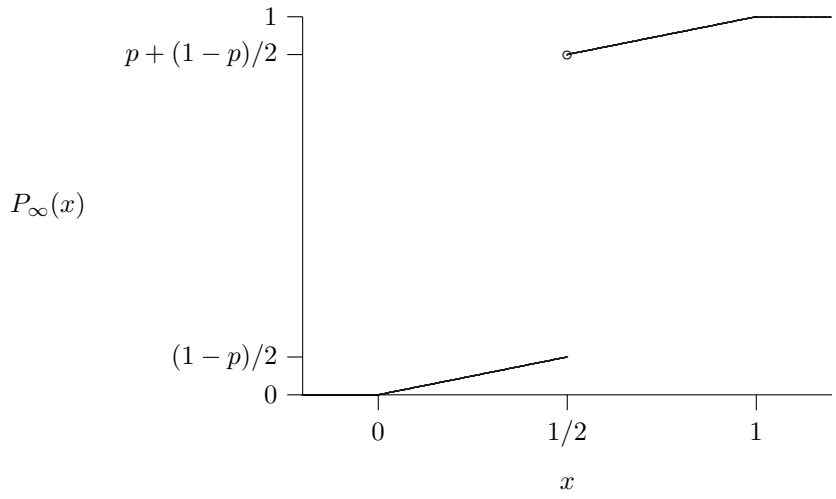

Abb. 2.7: Wahrscheinlichkeitsverteilung Gl. (2.39). Sie hat die Informationsdimension $1 - p$

Unser Beispiel läßt sich leicht wie folgt verallgemeinern: Sei ξ eine Zufallsgröße, deren Verteilungsfunktion $P(x)$ stetig ist, bis auf genau K Sprungstellen

$$p_1, \ldots, p_K, \quad \text{mit } S \equiv \sum_{k=1}^{K} p_k \leq 1 \ .$$

Dann gilt

$$I(\xi_M) \lesssim -\sum_{k=1}^{K} p_k \operatorname{ld} p_k - (1-S) \operatorname{ld} (1-S) + (1-S) \operatorname{ld} (M-K) \ .$$

Mit (2.37) erhalten wir die Informationsdimension

$$D_{\mathrm{I}}(\xi) = 1 - \sum_{k=1}^{K} p_k \ .$$

Für jede diskrete Zufallsgröße mit endlichem Wertevorrat gilt $1 = \sum_{k=1}^{K} p_k$. Ihre Informationsdimension ist also null.

Entropieintegral

Sei ξ wiederum absolut stetig verteilt, mit der Dichte $p(x)$, und ξ_M bezeichne die durch Vergröberung abgeleitete diskrete Zufallsgröße (2.35). Dann gilt[15]

$$\boxed{\lim_{M\to\infty}\big(I(\xi_M) - \operatorname{ld} M\big) = -\int_{\mathrm{R}} p(x)\operatorname{ld} p(x)\,\mathrm{d}x} \tag{2.40}$$

Hier setzen wir voraus, daß das Integral tatsächlich existiert. Wir nennen es dann *Entropieintegral* .

Das Entropieintegral hat eine augenfällige Ähnlichkeit mit dem in (2.15) definierten Informationsmaß I. Tatsächlich muß es aber vollkommen anders interpretiert werden. Das wird unter anderem dadurch deutlich, daß das Entropieintegral im Unterschied zu I auch negativ werden kann. Betrachten wir z. B. die Gleichverteilung auf dem Intervall $[a, b]$, mit der Dichte

$$p(x) = \begin{cases} 1/(b-a) & : \quad a \le x < b \\ 0 & : \quad \text{sonst.} \end{cases} \tag{2.41}$$

Hier erhalten wir

$$-\int_{\mathrm{R}} p(x)\operatorname{ld} p(x)\,\mathrm{d}x = \operatorname{ld}(b-a) \ .$$

Für $b - a < 1$ ist das Entropieintegral der Gleichverteilung also negativ.

Den Unterschied zwischen dem Entropieintegral und der Shannon–Information I können wir auch wie folgt verdeutlichen: Denken wir uns den Träger der Wahrscheinlichkeitsdichte $p(x)$ in paarweise disjunkte Intervalle (Boxen) B_m der Länge Δx unterteilt. Jeder Box ordnen wir eine Wahrscheinlichkeit

$$p_m \equiv \int_{B_m} p(x)\,\mathrm{d}x = p(x_m) \times \Delta x \tag{2.42}$$

zu, mit einer geeigneten Zwischenstelle $x_m \in B_m$. Wir setzen voraus, daß die zugehörige Information

$$\begin{aligned} I\big(\{p_m\}\big) &= -\sum_m p_m \operatorname{ld} p_m \\ &= -\sum_m p(x_m)\Delta x \operatorname{ld}\big(p(x_m)\Delta x\big) \end{aligned}$$

[15] Ein Beweis für (2.40), unter der (eigentlich überflüssigen) Bedingung, daß $p(x)$ beschränkt ist, findet sich z. B. in [15], S. 476 f.

für $\Delta x > 0$ endlich ist. Diesen Ausdruck können wir nun unter Beachtung von (2.42) umformen,

$$\begin{aligned} I\big(\{p_m\}\big) &= -\sum_m \big(p(x_m)\,\mathrm{ld}\,p(x_m)\big)\Delta x - \sum_m \big(p(x_m)\,\mathrm{ld}\,\Delta x\big)\Delta x \\ &= -\sum_m \big(p(x_m)\,\mathrm{ld}\,p(x_m)\big)\Delta x - \sum_m p_m\,\mathrm{ld}\,\Delta x \\ &= -\sum_m \big(p(x_m)\,\mathrm{ld}\,p(x_m)\big)\Delta x - \mathrm{ld}\,\Delta x \end{aligned} \tag{2.43}$$

Strebt nun Δx gegen null, d.h., wird die Auflösung immer feiner, so strebt in der unteren Zeile die Summe gegen das Entropieintegral (2.40) und $-\,\mathrm{ld}\,\Delta x$ divergiert. Es gilt also analog zu (2.40)

$$\lim_{\Delta x\to 0}\left(I\big(\{p_m\}\big) - \mathrm{ld}\,\frac{1}{\Delta x}\right) = -\int_{\mathrm{R}} p(x)\,\mathrm{ld}\,p(x)\,\mathrm{d}x \quad .$$

Ist der Träger der Dichte $p(x)$ das Einheitsintervall $[0,1]$, so liefert $1/\Delta x = M$ gerade die Anzahl der Boxen, d.h., die Mächtigkeit unseres Alphabetes.

Die tiefere Bedeutung des Entropieintegrals (2.40) wird aber erst auf der Grundlage eines weiteren zentralen Begriffes klar, den wir im folgenden Abschnitt einführen.

2.5 Informationsgewinn

Wir führen den Begriff des Informationsgewinns zunächst heuristisch ein, ebenso, wie wir im Abschnitt 2.2 die Shannon–Information einführten. Dazu betrachten wir also wieder ein Alphabet

$$A = a_1, a_2, \ldots, a_i, \ldots, a_{|A|}$$

aus $|A|$ gleich wahrscheinlichen Buchstaben. A unterteilen wir in paarweise disjunkte Teilalphabete,

$$A_1, A_2, \ldots, A_m, \ldots, A_M \quad ,$$

mit der Eigenschaft (2.14). Die Wahrscheinlichkeit, daß ein Buchstabe aus A in A_m liegt, ist dann

$$p_m = |A_m|/|A| \quad . \tag{2.44}$$

Die entsprechende Verteilung bezeichnen wir mit

$$\mathcal{P} \equiv \{p_m\}_{m=1}^{M} \quad .$$

Wir betrachten wiederum den zufälligen Versuch, der darin besteht, daß die Zugehörigkeit eines gezogenen Buchstaben zu einem der Teilalphabete bestimmt wird. Die entsprechende Zufallsgröße sei ξ mit der Verteilung $\mathcal{P}$. Soweit ist unsere Begriffsbildung also dieselbe, wie im Abschnitt 2.2.

Wir führen nun ein zweites Alphabet $A^* \subseteq A$ ein und fragen danach, wieviel Information wir aus der

$$\text{Bedingung}: \quad a_i \in A^* \tag{2.45}$$

über die Zufallsgröße ξ im Mittel gewinnen können.

Unter der Bedingung (2.45) haben wir es mit einer neuen Verteilung zu tun:

$$\mathcal{Q} \equiv \{q_m\}_{m=1}^{M} \ , \quad q_m = |A_m^*|/|A^*| \ , \quad A_m^* \equiv A^* \cap A_m \ . \tag{2.46}$$

Um nun den Informationsgewinn zu bestimmen, der entsteht, wenn die Verteilung $\mathcal{P}$ durch die Verteilung $\mathcal{Q}$ ersetzt wird, gehen wir wie folgt vor: $\mathrm{ld}\,|A|$ ist die (a priori) Gesamtungewißheit über den gezogenen Buchstaben. $\mathrm{ld}\,|A^*|$ ist die (a posteriori) Gesamtungewißheit, also die Ungewißheit, die verbleibt falls wir schon wissen, daß (2.45) gilt. Damit enthält die Aussage (2.45) also insgesamt die *Information*

$$\mathrm{ld}\,|A| - \mathrm{ld}\,|A^*| = \mathrm{ld}\,(|A|/|A^*|) \ .$$

Diese Information können wir uns aber auch in zwei Teile zerlegt denken,

Erste Teilinformation: Information darüber, in welchem Teilalphabet A_m der gezogene Buchstabe a_i liegt, wenn (2.45) bekannt ist. Das ist der gesuchte partielle Informationsgewinn über ξ. Wir bezeichnen ihn mit $I(q_m \| p_m)$ und werden ihn später noch konkret angeben.

Zweite Teilinformation: Information darüber, welcher Buchstabe im Alphabet A_m gezogen wurde, wenn (2.45) schon bekannt ist. Diese Teilinformation können wir sofort angeben, denn die m–te (a priori) Teilungewißheit ist $\mathrm{ld}\,|A_m|$ und die entsprechende (a posteriori) Teilungewißheit ist $\mathrm{ld}\,|A_m^*|$, so daß also die Aussage (2.45) die Teilinformation

$$\mathrm{ld}\,|A_m| - \mathrm{ld}\,|A_m^*| = \mathrm{ld}\,(|A_m|/|A_m^*|)$$

birgt.

Damit gilt also

$$\mathrm{ld}\,(|A|/|A^*|) = I(q_m\|p_m) + \mathrm{ld}\,(|A_m|/|A_m^*|) \ ,$$

und nach I aufgelöst erhalten wir unter Beachtung von (2.44) und (2.46),

$$I(q_m\|p_m) = \mathrm{ld}\,(q_m/p_m) \ . \tag{2.47}$$

Dieser m–te partielle Informationsgewinn kann sowohl negative wie auch positive Werte annehmen. Allerdings interessieren wir uns im folgenden für seinen Mittelwert. Unter der Bedingung (2.45) tritt $a \in A_m$ mit der Wahrscheinlichkeit q_m auf. Der mittlere Informationsgewinn ist somit

$$I(\mathcal{Q}\|\mathcal{P}) \equiv \sum_{m=1}^{M} q_m I(q_m\|p_m) \ .$$

Mit (2.47) können wir dafür wie folgt schreiben:

$$\boxed{I(\mathcal{Q}\|\mathcal{P}) = \sum_{m=1}^{M} q_m \,\mathrm{ld}\,\frac{q_m}{p_m}} \tag{2.48}$$

Die Größe (2.48) heißt *Informationsgewinn, der entsteht, wenn $\mathcal{P}$ durch die Verteilung $\mathcal{Q}$ ersetzt wird.* [16] [17] Sie ist definiert, falls eine genaue Zuordnung der Wahrscheinlichkeiten beider Verteilungen $\mathcal{Q}$ und $\mathcal{P}$ besteht, und falls $p_m > 0$ für alle $m = 1, \ldots, M$ gilt. Der Informationsgewinn ist einer der zentralen Begriffe der Informationstheorie. Kann der m–te partielle Informationsgewinn auch negativ werden, im Mittel ist dies nicht der Fall, denn allgemein gilt der

Satz 2.5 *Sind $\mathcal{P} = \{p_m\}_{m=1}^M$, mit $p_m > 0$, und $\mathcal{Q} = \{q_m\}_{m=1}^M$ zwei diskrete Wahrscheinlichkeitsverteilungen, so ist der Informationsgewinn (2.48) beidseitig wie folgt beschränkt,*

$$0 \le I(\mathcal{Q}\|\mathcal{P}) \le \max_{m=1,2,\ldots,M}\{-\mathrm{ld}\,p_m\} \ . \tag{2.49}$$

$I(\mathcal{Q}\|\mathcal{P}) = 0$ gilt genau dann, wenn die Verteilungen übereinstimmen, also wenn $p_m = q_m$ für $m = 1, 2, \ldots, M$.

[16] Sind ξ und η zwei Zufallsgrößen mit den Wahrscheinlichkeitsverteilungen $\mathcal{P}$ bzw. $\mathcal{Q}$, so werden wir für den Informationsgewinn $I(\mathcal{Q}\|\mathcal{P})$ auch $I(\eta\|\xi)$ schreiben, obwohl der Informationsgewinn nur von den Verteilungen abhängt.

[17] Der Informationsgewinn heißt auch *Kullback–Leibler–Entropie* bzw. *Kullback–Leibler–Information* [9].

Beweis: (Der Beweis für die linke Relation in (2.49) folgt aus der Jensenschen Ungleichung, s. Anhang.)

□

Transinformation als Informationsgewinn

Im Abschnitt 2.3 haben wir die Transinformation (2.27) zweier diskreter Zufallsgrößen ξ und η eingeführt. Mit den dortigen Bezeichnungen können wir wie folgt schreiben,

$$\begin{aligned} I_T(\xi,\eta) &= I(\xi) + I(\eta) - I\big((\xi,\eta)\big) \\ &= \sum_{m,n=1}^{M,N} s_{mn} \operatorname{ld} \frac{s_{mn}}{p_m q_n} \ . \end{aligned}$$

Damit ist also die Transinformation gleich dem Informationsgewinn, der entsteht, wenn die Verteilung

$$S_{\mathrm{u}} \equiv \{p_m q_n\}_{m,n=1}^{M,N}$$

durch die Verteilung

$$S_{\mathrm{r}} \equiv \{s_{mn}\}_{m,n=1}^{M,N}$$

ersetzt wird. S_{u} kann als Hypothese über die statistische Unabhängigkeit von ξ und η angesehen werden, und S_{r} ist die tatsächliche (reale) Verbundverteilung. Damit mißt also die Transinformation den Abstand zwischen der hypothetischen und der realen Verteilung.

Informationsgewinn als Differenz von subjektiver und objektiver Information

Wir können den Informationsgewinn (2.48) wie folgt umschreiben:

$$I(\mathcal{Q}\|\mathcal{P}) = \sum_{m=1}^{M} q_m \operatorname{ld} \frac{1}{p_m} - \sum_{m=1}^{M} q_m \operatorname{ld} \frac{1}{q_m} \ . \tag{2.50}$$

Um die beiden Summen in dieser Darstellung des Informationsgewinns zu interpretieren, werden wir zunächst eine codierungstheoretische Interpretation unseres Informationsmaßes (2.15) geben, womit wir den ausführlicheren Darstellungen im Kapitel 3 etwas vorgreifen.

Nehmen wir an, wir wollen eine statistisch unabhängige Buchstabenfolge, in der die einzelne Buchstaben mit der Verteilung $\mathcal{Q} = \{q_m\}_m$ auftreten, durch möglichst wenige *Entscheidungs-*

fragen[18] aufklären. Unter der Voraussetzung, daß die Verteilung $\mathcal{Q}$ bekannt ist, können wir eine *optimale Fragestrategie* entwickeln, so daß wir pro Buchstabe im Mittel mit möglichst wenigen Fragen auskommen. Es zeigt sich nun, daß die Shannon–Information

$$I_{\mathrm{obj}}(\mathcal{Q}) = -\sum_m q_m \,\mathrm{ld}\, q_m$$

gerade die mittlere Anzahl von Fragen pro Buchstabe ist, die bei einer optimalen Fragestrategie mit an Sicherheit grenzender Wahrscheinlichkeit benötigt wird. Mit weniger Fragen kommt man „höchstwahrscheinlich“ nicht aus.

In der Praxis wird man die objektiven Auftrittswahrscheinlichkeiten $\mathcal{Q}$ jedoch nicht genau kennen, sondern nur eine gewisse Schätzung $\mathcal{P} = \{p_m\}_m$, die subjektiven Wahrscheinlichkeiten. Wird nun auf deren Grundlage eine optimale Fragestrategie entwickelt, so benötigt man im „objektiven“ Mittel

$$I_{\mathrm{sub}}(\mathcal{Q}, \mathcal{P}) = -\sum_m q_m \,\mathrm{ld}\, p_m$$

Fragen pro Buchstabe.[19] Immer dann, wenn die subjektive Verteilung von der objektiven abweicht, werden mehr Fragen benötigt. Der Informationsgewinn (2.50) beschreibt gerade diese Abweichung,

$$I(\mathcal{Q}\|\mathcal{P}) = I_{\mathrm{sub}}(\mathcal{Q}, \mathcal{P}) - I_{\mathrm{obj}}(\mathcal{Q}) \ .$$

Informationsgewinn kontinuierlicher Zufallsgrößen

Im Abschnitt 2.4 haben wir gesehen, daß eine Anwendung der zunächst für endliche Alphabete entwickelten Begriffe auf unendlich mächtige Alphabete bzw. auf kontinuierliche Zufallsgrößen nicht ohne weiteres möglich ist. Mit dem Informationsgewinn ist es jedoch anders. Um dies zu verdeutlichen, gehen wir wieder wie auf S. 45 f. vor:
Seien ξ und η zwei absolut stetige Zufallsgrößen, mit den Verteilungsdichten $p(x)$ und $q(x)$. Der Träger von q sei in dem von p enthalten.[20] $\{B_m\}$ bezeichne eine Zerlegung des Trägers von $p(x)$ in paarweise disjunkte Intervalle der Länge Δx. Dann können wir analog zu (2.42) wiederum vergröberte (diskrete) Verteilungen $\mathcal{P}_{\Delta x} \equiv \{p_m\}$ und $\mathcal{Q}_{\Delta x} \equiv \{q_m\}$ ein-

[18] Entscheidungsfragen werden immer entweder mit „ja“ oder „nein“ beantwortet.
[19] Die Größe $I_{\mathrm{sub}}(\mathcal{Q}, \mathcal{P})$ heißt auch *Bongard–Information* oder *Bongard–Entropie* [1].
[20] Bezeichnet $\mathrm{supp}(p) \equiv \{x \in \mathbb{R} : p(x) > 0\}$ den Träger der Dichte p, so gelte also $\mathrm{supp}(q) \subseteq \mathrm{supp}(p)$.

führen. Möge der aus diesen Verteilungen gebildete Informationsgewinn (2.48) existieren.[21] Dann können wir wie folgt schreiben: [22]

$$I(\mathcal{Q}_{\Delta x}\|\mathcal{P}_{\Delta x}) = \sum_m q(y_m)\Delta x \,\mathrm{ld}\,\frac{q(y_m)\Delta x}{p(x_m)\Delta x} \quad . \tag{2.51}$$

In (2.51) erkennen wir nun, daß sich Δx im Argument von ld kürzt und somit die Divergenzprobleme wie in (2.43) nicht auftreten. Tatsächlich erhalten wir

$$\lim_{\Delta x\to 0} I(\mathcal{Q}_{\Delta x}\|\mathcal{P}_{\Delta x}) = \int_{\mathrm{R}} q(x)\,\mathrm{ld}\,\frac{q(x)}{p(x)}\,\mathrm{d}x \quad .$$

Existiert dieses Integral, so können wir es also vollkommen analog zum diskreten Fall (2.48) interpretieren als *Informationsgewinn, der entsteht, wenn die Verteilungsdichte $p(x)$ durch $q(x)$ ersetzt wird.* Wir schreiben dann

$$\boxed{I(\xi\|\eta) = \int\limits_{\mathrm{R}} q(x)\,\mathrm{ld}\,\frac{q(x)}{p(x)}\,\mathrm{d}x} \tag{2.52}$$

Insbesondere können wir also auch für den Fall kontinuierlicher Zufallsgrößen eine Transinformation betrachten: Bezeichnet $s(x,y)$ die Verteilungsdichte des Zufallsvektors (ξ,η), und sind

$$p(x) = \int_{\mathrm{R}} s(x,y)\,\mathrm{d}y \quad \text{und} \quad q(y) = \int_{\mathrm{R}} s(x,y)\,\mathrm{d}x$$

die zugehörigen Randdichten, so gilt

$$\boxed{I_{\mathrm{T}}(\xi,\eta) = \int\int s(x,y)\,\mathrm{ld}\,\frac{s(x,y)}{p(x)q(y)}\,\mathrm{d}x\,\mathrm{d}y} \tag{2.53}$$

Informationsgewinn als Abstandsmaß

Allgemein könnten wir den Informationsgewinn (2.48) bzw. (2.52) als ein Abstandsmaß zwischen Verteilungen $\mathcal{Q}$ und $\mathcal{P}$ ansehen. Allerdings wäre dieser Abstand unsymmetrisch,

$$I(\xi\|\eta) \neq I(\eta\|\xi) \quad .$$

[21] Ist der Träger von $p(x)$ beschränkt, so gibt es für $\Delta x > 0$ nur endlich viele Summanden in (2.48), und folglich existiert dieser Informationsgewinn.

[22] Bei der Diskretisierung müssen wir im allgemeinen für die Dichten $p(x)$ und $q(x)$ unterschiedliche Stützstellensysteme $\{x_m\}$ bzw. $\{y_m\}$ verwenden, um die Normierung $\sum_m p_m = 1$ und $\sum_m q_m = 1$ zu gewährleisten. Für den folgenden Grenzübergang ist dies aber belanglos.

Ein symmetrisches Abstandsmaß liefert aber z. B. die Größe

$$I_{\text{dist}}(\xi,\eta) \equiv I(\xi\|\eta) + I(\eta\|\xi) \ . \tag{2.54}$$

Offenbar hat es die folgenden Eigenschaften:

Symmetrie:

$$I_{\text{dist}}(\xi,\eta) = I_{\text{dist}}(\eta,\xi) \ .$$

Positive Definitheit:

$$I_{\text{dist}}(\xi,\eta) \geq 0 \ .$$

Die Gleichheit gilt genau dann, wenn ξ und η die gleiche Verteilung haben, also für $\mathcal{P} = \mathcal{Q}$.

Beispiel:

Transinformation Normalverteilter Zufallsgrößen

Seien ξ und η im Paar normalverteilt, mit der Verbunddichte

$$s(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right] \ . \tag{2.55}$$

Der Korrelationskoeffizient ist dann gegeben durch

$$\rho \equiv \int\limits_{\mathrm{R}^2} xy\, s(x,y)\ \mathrm{d}x\ \mathrm{d}y \ , \tag{2.56}$$

und die Randverteilungsdichten sind

$$\begin{aligned} p(x) &= \int\limits_{\mathrm{R}} s(x,y)\,\mathrm{d}y = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{x^2}{2}\right] \\ &\text{bzw.} \\ q(y) &= \int\limits_{\mathrm{R}} s(x,y)\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{y^2}{2}\right] \ . \end{aligned} \tag{2.57}$$

Setzen wir diese Dichten in (2.53) ein, so erhalten wir

$$I_{\mathrm{T}}(\xi,\eta) = -\frac{1}{2}\,\mathrm{ld}\,(1-\rho^2) \ . \tag{2.58}$$

Die Transinformation unterscheidet also nicht zwischen positiver und negativer Korrelation. Das Verschwinden der Korrelation zwischen ξ und η ist hier gleichbedeutend damit, daß die Transinformation verschwindet, daß es also keine statistischen Abhängigkeiten zwischen den Zufallsgrößen gibt. Insofern ist die Normalverteilung ein Sonderfall.[23] Für $\rho \to 1$ divergiert die

[23] Im allgemeinen folgt aus der Unkorreliertheit zweier Zufallsgrößen nicht auch deren statistische Unabhängigkeit. Die Umkehrung gilt jedoch immer.

Transinformation (2.58). Dies kann nicht verwundern, denn der Zusammenhang zwischen ξ und η ist für $\rho = 1$ deterministisch, das heißt, die eine Zufallsgröße ist hier eine Funktion der anderen. Die Information der einen über die andere muß dann unendlich sein, denn die vollständige Charakterisierung einer kontinuierlichen Zufallsgröße verlangt unendlich viel Information.

Entropieintegral als Informationsgewinn

Seien ξ und η im Intervall $[a, b]$ stetig verteilte Zufallsgrößen mit den Verteilungsdichten $p(x)$ bzw. $q(x)$, und der Träger von q sein in dem von p enthalten, $\mathrm{supp}(q) \subseteq \mathrm{supp}(p)$. Ist ξ gleichverteilt in $[a, b]$, so daß also $p(x)$ durch (2.41) gegeben ist, so erhalten wir für den Informationsgewinn, der entsteht, wenn p durch q ersetzt wird,

$$I(\eta\|\xi) = \int q(x) \operatorname{ld} q(x)\, \mathrm{d}x + \operatorname{ld}(b-a) \ .$$

Auf der rechten Seite taucht gerade das Entropieintegral (2.40) auf, angewandt auf die Dichte q und mit negativen Vorzeichen. Damit läßt sich also das Entropieintegral auf einen Informationsgewinn zurückführen. Für $b - a = 1$ ist es gerade gleich dem negativen Wert des Informationsgewinns $I(\eta\|\xi)$.

Hat die Verteilungsdichte $q(x)$ von η einen beschränkten Träger, mit a und b als größte untere bzw. kleinste obere Schranke, dann liefert die Transformation (Streckung bzw. Stauchung)

$$\eta \to \eta^* = \eta/(b-a) \tag{2.59}$$

gerade eine Zufallsgröße η^*, deren Dichte

$$q^*(x) = (b-a) \times q\big(x/(b-a)\big)$$

ihren Träger im Intervall $[a/(b-a), b/(b-a)]$ der Länge 1 hat. Ist dann ξ in diesem Intervall gleichverteilt, so können wir wie folgt schreiben:

$$I(\eta^*\|\xi) = \int q^*(x) \operatorname{ld} q^*(x)\, \mathrm{d}x \ .$$

Nach der Transformation (2.59) kann das Entropieintegral also als negativer Informationsgewinn aufgefaßt werden, der entsteht, wenn die Gleichverteilung durch die Verteilung $q^*(x)$ ersetzt wird.

Kapitel 3

Elemente der Codierungstheorie

Wir legen hier einige Grundzüge der Codierungstheorie dar. Damit wird letztendlich eine tiefer gehende Deutung unseres Shannonschen Informationsmaßes (2.15) ermöglicht und seine Bedeutung für die Praxis der redundanzarmen Codierung von Nachrichten (Signalen) aufgezeigt.

3.1 Kraftsche Ungleichung

Gegeben sei eine endliche Menge A, unser *Alphabet*, das aus den Buchstaben a_m, $m = 1, 2, \ldots, M$, besteht. Sei nun ein zweites Alphabet $B_r = \{b_1, \ldots, b_r\}$ gegeben, das aus r verschiedenen Zeichen (Buchstaben) besteht, mit $r \geq 2$. Die Elemente des kartesischen Produktes

$$\boxtimes_l B_r \equiv \underbrace{B_r \times \ldots \times B_r}_{l\text{–mal}}$$

nennen wir *Codewort* der Länge l, das aus dem erzeugenden Alphabet $B_r \equiv \boxtimes_1 B_r$ gebildet wird. Ist dies z.B. das binäre Alphabet $B_2 = \{0, 1\}$, dann gilt

$$\begin{aligned} \boxtimes_2 B_2 &= \{(0,0),\ (0,1),\ (1,0),\ (1,1)\} \\ \boxtimes_3 B_2 &= \{(0,0,0),\ (0,0,1),\ (0,1,0), \ldots, (1,1,1)\} \quad \text{usw.} \end{aligned}$$

Offenbar hat $\boxtimes_l B_r$ gerade r^l Elemente, die sogenannten *Codeworte*. Unter einem *Radix–r–Code* $\mathcal{C}_r$ des Alphabetes A verstehen wir eine ein–ein–deutige Abbildung

$$\mathcal{C}_r : A \longrightarrow \bigcup_{l^*=1}^{l} \boxtimes_{l^*} B_r \ .$$

Für den Radix $r = 2$ ist $\mathcal{C}_r$ ein *Binärcode*.

Nicht jeder Code ist eindeutig decodierbar. Kann aus einer beliebigen Folge von Codeworten eindeutig auf eine Folge von Buchstaben aus A geschlossen werden, so heißt der Code *eindeutig decodierbar*. Eindeutig decodierbar sind z.B. sogenannte *Präfixcodes*, die sich dadurch auszeichnen, dass kein Codewort der Anfang eines anderen Codewortes ist.[1] So ist z.B.

$$\begin{aligned} a_1 &\rightarrow 1 \\ a_2 &\rightarrow 01 \\ a_3 &\rightarrow 00 \end{aligned} \tag{3.1}$$

ein binärer Präfixcode für die Buchstaben a_1, a_2, a_3, wohingegen

$$\begin{aligned} a_1 &\rightarrow 1 \\ a_2 &\rightarrow 0 \\ a_3 &\rightarrow 00 \end{aligned} \tag{3.2}$$

kein solcher Code ist.[2] Für den Code (3.1) erhalten wir beispielsweise

$$\begin{array}{rcccccc} 10001101 & = & 1 & 00 & 01 & 1 & 01 \\ & \rightarrow & a_1 & a_3 & a_2 & a_1 & a_2 \end{array}$$

Hingegen würde im Fall (3.2)

$$\begin{array}{rcccccccc} 10001101 & = & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ & \rightarrow & a_1 & a_2 & a_2 & a_2 & a_1 & a_1 & a_2 & a_1 \end{array}$$

oder

$$\begin{array}{rcccccccc} 10001101 & = & 1 & 00 & 0 & 1 & 1 & 0 & 1 \\ & \rightarrow & a_1 & a_3 & a_2 & a_2 & a_1 & a_1 & a_2 \end{array}$$

decodiert werden. Die Umkehrung des Codes, die *Decodierung*, ist also nicht eindeutig.

Wird der Buchstabe a_m durch das Codewort

$$B \equiv b_{n_1} b_{n_2} \ldots b_{n_{l_m}}$$

[1] Spiegelt man die Codeworte eines Präfixcodes, d.h., liest man die Codeworte von rechts nach links, so erhält man einen *Suffixcode*.

[2] Man beachte hierbei, dass in einer Codewortfolge kein Trennzeichen zwischen den Codeworten gestattet ist, würde dies doch einem zusätzlichen Zeichen entsprechen, so dass der Radix des Codes nicht r, sondern $r+1$ wäre.

codiert, dann ist l_m die zugehörige *Codewortlänge*. Unter der *Länge* $\mathcal{L}$ *des Codes* $\mathcal{C}_r$ *eines Alphabetes* $A = \{a_m\}_{m=1}^{M}$ verstehen wir die maximale Codewortlänge,

$$\mathcal{L}(\mathcal{C}_r) \equiv \max_{m=1,\ldots,M} \{l_m\} \ . \tag{3.3}$$

Sind alle Codewortlängen gleich, so spricht man von einem *Blockcode*. Ein Blockcode ist immer auch Präfix– sowie Suffixcode und somit eindeutig decodierbar.

Es gilt nun der folgende Satz (**(Kraftsche Ungleichung)**):

1. Sei ein Alphabet $A = \{a_m\}_{m=1}^{M}$ gegeben, sowie ein System von Codewortlängen $\{l_m\}_{m=1}^{M}$. Ist dann die Ungleichung

$$\boxed{\sum_{m=1}^{M} r^{-l_m} \leq 1} \tag{3.4}$$

   erfüllt, so existiert ein eindeutig decodierbarer Radix–r–Code des Alphabetes A mit eben diesen Codewortlängen $\{l_m\}$.

2. Jeder eindeutig decodierbare Code mit den Codewortlängen $\{l_m\}$ erfüllt die Ungleichung (3.4).

Beweis: Jeder eindeutig decodierbare Code kann durch einen Päfixcode mit gleichen Wortlängen $\{l_m\}$ ersetzt werden. (Davon machen wir hier ohne Beweis Gebrauch.) Deshalb können wir im Folgenden ohne Beschränkung der Allgemeinheit einen Präfixcode voraussetzen. Seien nun die Codewortlängen der Größe nach geordnet,

$$l_1 \leq \ldots \leq l_m \leq \ldots \leq l_M \ .$$

Mit dem Code sind offenbar höchstens r^{l_M} Codeworte möglich.[3]

Der Buchstabe a_1 habe die Codewortlänge l_1. Da wir nur Präfixcodes betrachten, darf das Codewort von a_1 nicht Präfix eines anderen Codewortes sein. Für $l_1 = l_M$ sind somit nur noch $r^{l_M} - 1$ viele Codeworte möglich. Für $l_1 = l_M - 1$ sind aber nur noch $r^{l_M} - r^1$ andere Codeworte möglich usw. Allgemein sind durch die Existenz eines Codewortes der Länge l_1 nur noch $r^{l_M} - r^{l_M - l_1}$ weitere Codeworte möglich.

Der Buchstabe a_2 habe die Codewortlänge l_2. Folglich sind nun ingesamt nur noch

$$r^{l_M} - r^{l_M - l_1} - r^{l_M - l_2}$$

Codeworte möglich usw. Die Anzahl der noch nicht belegten aber noch möglichen Codeworte reduziert sich also nach M Schritten auf

$$r^{l_M} - \sum_{m=1}^{M} r^{l_M - l_m} \ .$$

[3]All diese Codeworte treten bei voll genutzten Blockcodes auf, wo $l_1 = l_M$ gilt, also alle Codeworte gleich lang sind.

Wird diese Anzahl negativ, so hätten wir nicht genügend viele Codeworte, um einen jeden der M Buchstaben eindeutig umkehrbar zu codieren. Wir müssen demzufolge fordern,

$$0 \leq r^{l_M} - \sum_{m=1}^{M} r^{l_M - l_m} \ .$$

Nach Division beider Seiten durch r^{l_M} folgt die Behauptung des Satzes.

□

Die Relation (3.4) heißt *Kraftsche Ungleichung*. Für den Code in (3.1) gilt,

$$\sum_{m=1}^{3} 2^{-l_m} = 2^{-1} + 2^{-2} + 2^{-2} = 1 \ .$$

Er erfüllt also die Kraftsche Ungleichung. Nach der zweiten Aussage des obigen Satzes kann dies nicht verwundern, haben wir ihn doch schon als eindeutig decodierbar erkannt.

Die Gleichheit in (3.4) zeigt an, dass alle zum System $\{l_m\}$ bildbaren und eindeutig decodierbaren Radix–r–Codeworte tatsächlich verwendet werden. Betrachten wir z. B. anstelle von (3.1) den binären Präfixcode

$$\begin{aligned} a_1 &\rightarrow 11 \\ a_2 &\rightarrow 01 \\ a_3 &\rightarrow 00 \ , \end{aligned} \tag{3.5}$$

so erhalten wir

$$\sum_{m=1}^{3} 2^{-l_m} = \frac{3}{4} < 1 \ .$$

Daraus lesen wir ab, dass zumindest ein Codewort mit einer Codewortlänge, die nicht größer als $\mathcal{L}(\mathcal{C}_r) = 2$, noch ungenutzt ist. In unserem Beispiel (3.5) ist dies gerade das Codewort 10.

Für den Code (3.2) erhalten wir

$$\sum_{m=1}^{3} 2^{-l_m} = 2^{-1} + 2^{-1} + 2^{-2} = 1,25 > 1 \ .$$

Die Kraftsche Ungleichung ist also verletzt, woraus sofort geschlossen werden kann, dass der Code nicht eindeutig decodierbar ist. Auch das hatten wir bereits erkannt.

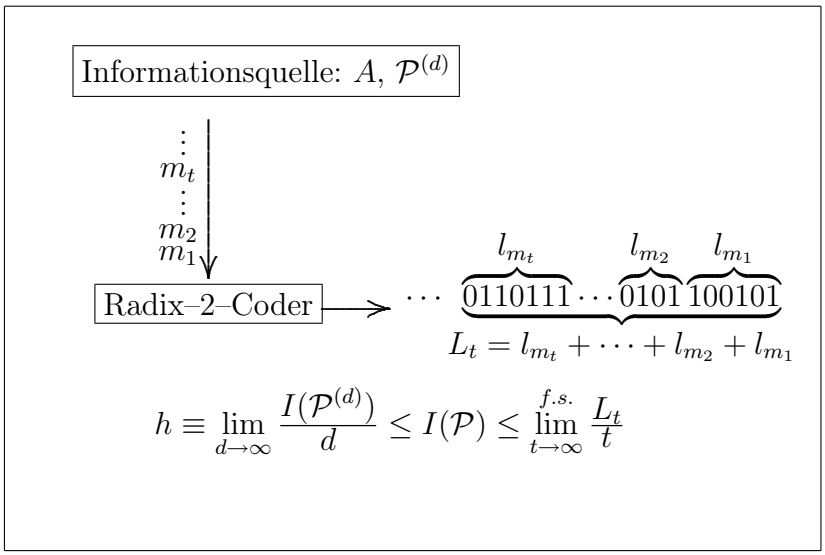

Abb. 3.1: Binäres Codierschema einer zeitlichen Buchstabenfolge aus einem Alphabet A mit der Sendeverteilungen $\mathcal{P}^{(d)}$ für Worte der Länge d aus $\boxtimes_d A$. Der zur Zeit t gesandte Buchstabe $a_{m_t} \equiv m_t$ wird mit einem binären Codewort der Länge l_{m_t} codiert. Werden die Buchstaben vollkommen statistisch unabhängig gesendet, dann ist die Shannon–Entropie I der Verteilung $\mathcal{P} \equiv \mathcal{P}^{(1)}$ die kleinst mögliche mittlere Codewortlänge, die bei einer optimalen Codierung fast sicher (f.s.) erreicht werden kann. Im allgemeinen Fall, in dem auch statistische Abhängigkeiten zwischen gesandten Buchstaben auftreten können, ist die Quellentropie h die kleinst mögliche mittlere Codewortlänge pro Buchstabe. Werden die Buchstaben jedoch vollständig unabhängig gesandt, so gilt $h = I(\mathcal{P})$

3.2 Fundamentalsatz der Codierung

Gegeben sei eine stationäre Informationsquelle, die Buchstaben a_m aus einem endlichen Alphabet A mit den Wahrscheinlichkeiten $\mathcal{P} \equiv \{p_m\}_{m=1}^{M}$ sendet. Die Buchstabenfolge

$$m_1, m_2, \ldots, m_t, \ldots \,^{4} \tag{3.6}$$

werde durch einen Radix–2–Code in eine eindeutig decodierbare binäre Sequenz von Nullen und Einsen überführt (Abb. 3.1).

Eine solche Codierung geschieht häufig für einzelne Buchstaben. Sie kann aber auch wortweise erfolgen, indem z. B. immer zwei aufeinander folgende Buchstaben

$$m_1 m_2, \quad m_3 m_4, \quad m_5 m_6, \ldots$$

[4]Zur Vereinfachung der Schreibweise notieren wir anstelle des Buchstabens $a_{m_t} \in \{a_1, a_2, \ldots, a_m, \ldots, a_M\}$ nur dessen Index $m_t \in \{1, 2, \ldots, m, \ldots, M\}$.

binär codiert werden. Es gibt M^2 viele Worte der Länge 2. Wir können also im Allgemeinen Worte der Länge $d \geq 1$ aus

$$\boxtimes_d A \equiv \underbrace{A \times A \times \ldots \times A}_{d\text{–mal}}$$

binär codieren. Offenbar gibt es M^d viele Worte der Länge d, die selbst wieder ein Alphabet $\boxtimes_d A$ bilden. Bei einer wortweisen, eindeutig umkehrbaren Codierung einer Buchstabenfolge kann die Anzahl $L^{(d)}$ der Nullen und Einsen, die pro Buchstabe nötig sind, mit wachsender Wortlänge d fallen oder auch wachsen. Im Zusammenhang mit der Abbildung 2.1 haben wir dies bereits ausführlich diskutiert. Allerdings wird $L^{(d)}$ im Trend immer fallen, d.h., es gibt zu jeder Wortlänge d eine noch größere Wortlänge $d^* > d$, so daß $L^{(d^*)} \leq L^{(d)}$ gilt.

Werden nun die ersten t Buchstaben der Folge (3.6) durch L_t Nullen und Einsen (einzeln oder wortweise) codiert, so hat der Code pro Buchstabe die *mittlere Codewortlänge*

$$L \equiv \overset{f.s.}{\lim_{t\to\infty}} \frac{L_t}{t} \quad . \tag{3.7}$$

Wenn immer Worte $(a_{m_1}, \ldots, a_{m_d}) \in \boxtimes_d A$ mit einem bestimmten binären Codewort der Länge $l_{m_1 \ldots m_d}$ codiert werden, so kann man bei vorausgesetzter Ergodizität das obige Zeitmittel auch als Scharmittel über die Wahrscheinlichkeitsverteilung der Worte erhalten,

$$L^{(d)} = \frac{1}{d} \sum_{m_1,\ldots,m_d=1}^{M} p_{m_1 \ldots m_d} l_{m_1 \ldots m_d} \quad . \tag{3.8}$$

Um den technischen Codierungsaufwand möglichst klein zu halten, wird man versuchen, diese mittlere Codewortlänge pro Buchstabe möglichst klein zu halten. Der folgende Satz macht eine Aussage, inwieweit dies grundsätzlich möglich:

(Fundamentalsatz der Codierung) Wird eine stationäre ergodische Buchstabenfolge (3.6) der Länge t mit L_t Binärzeichen (Nullen und Einsen) eindeutig umkehrbar codiert, so werden im Grenzfall einer unendlich langen Buchstabenfolge fast sicher pro Buchstabe nicht weniger Binärzeichen benötigt, als die Quellentropie

$$h \equiv \lim_{d\to\infty} \frac{1}{d} \sum_{m_1,\ldots,m_d=1}^{M} p_{m_1 \ldots m_d} \operatorname{ld} p_{m_1 \ldots m_d}$$

angibt. Es gilt also

$$h \leq \overset{f.s.}{\lim_{t\to\infty}} \frac{L_t}{t} \equiv L \quad .$$

Es gibt eine eindeutig umkehrbare Codierung, so daß L beliebig nahe bei der Quellentropie h liegt. Hingegen gibt es keine eindeutig umkehrbare Codierung, bei der L fast sicher kleiner als

h ist. **Beweis:** Sei $l_{m_1\ldots m_d}$ die Länge des Binärwortes, mit dem das Wort $(a_{m_1},\ldots,a_{m_d})$ codiert ist, welches aus Buchstaben des Alphabetes $A=\{a_1,\ldots,a_m,\ldots,a_M\}$ gebildet wird. $\mathcal{P}^{(d)}\equiv\{p_{m_1\ldots m_d}\}$ bezeichne die Wahrscheinlichkeiten, mit der die entsprechenden Worte auftreten.

Wir können nun die Längen $l_{m_1\ldots m_d}$ gerade so bilden, daß

$$-\operatorname{ld} p_{m_1\ldots m_d}\le l_{m_1\ldots m_d}<1-\operatorname{ld} p_{m_1\ldots m_d} \tag{3.9}$$

gilt. Nach Mittelung dieser Ungleichung über die Wortverteilung $\mathcal{P}^{(d)}$ (Multiplikation mit $p_{m_1\ldots m_d}$ und Summation über alle Indizes) und anschließender Division durch die Wortlänge d erhalten wir

$$\frac{I(\mathcal{P}^{(d)})}{d}\le L^{(d)}<\frac{I(\mathcal{P}^{(d)})}{d}+\frac{1}{d}\ . \tag{3.10}$$

Hieraus lesen wir ab, daß im Grenzfall unendlich langer Codeworte die mittlere Codewortlänge pro Buchstabe gegen die Quellentropie strebt,

$$\lim_{d\to\infty}L^{(d)}=\lim_{d\to\infty}\frac{I(\mathcal{P}^{(d)})}{d}=h\ .$$

Wir zeigen nun, daß die untere Grenze in (3.10) nicht unterschritten werden kann: Um einen eindeutig decodierbaren Code zu gewährleisten, muß das System der Codewortlängen notwendig die Kraftsche Ungleichung (3.4) erfüllen,

$$\sum_{m_1,\ldots,m_d=1}^{M}2^{-l_{m_1\ldots m_d}}=1\ .$$

(Hier setzen wir voraus, daß alle möglichen Codeworte tatsächlich verwendet werden.) Folglich kann das System $\{2^{-l_{m_1\ldots m_d}}\}_{m_1\ldots m_d}$ als Wahrscheinlichkeitsverteilung interpretiert werden, und nach Satz 2.5, Gl. (2.49), gilt

$$0\le\sum p_{m_1\ldots m_d}\operatorname{ld}\frac{p_{m_1\ldots m_d}}{2^{-l_{m_1\ldots m_d}}}\ .$$

Eine einfache Umformung dieser Ungleichung liefert

$$-\frac{1}{d}\sum p_{m_1\ldots m_d}\operatorname{ld}p_{m_1\ldots m_d}\le\frac{1}{d}\sum p_{m_1\ldots m_d}l_{m_1\ldots m_d}\ ,$$

also

$$\frac{I(\mathcal{P}^{(d)})}{d}\le L^{(d)}\ .$$

Im Grenzfall $\mathrm{d}\to\infty$ gilt somit

$$h\le L^{(\infty)}=L\ .$$

Daraus folgt die Behauptung des Satzes. [5]

□

Nach diesem Satz können wir also eine konkrete eindeutig umkehrbare Codierung bewerten, indem wir die mittlere Codewortlänge $L^{(d)}$ nach (3.7) bzw. (3.8) bestimmen und mit der Quellentropie h vergleichen. Je dichter L bei h liegt, desto „besser“ (kürzer) ist der Code.

[5] Einen strengeren Beweis, der insbesondere auch die fast sichere Konvergenz zeigt, gibt z.B. Rényi [15], S. 455 ff.

Überschreite beispielsweise bei einer konkreten praktischen Codierung L die untere Grenze h um nur $1\% = (L-h)/h \times 100\%$. Daraus lesen wir ab, daß eine weitere Reduzierung der Codelänge um 1% der Quellentropie wohl möglich ist, indem anders codiert wird. Aus dem Beweis des Satzes bekommen wir sogar einen Hinweis darauf, wie dies zu bewerkstelligen wäre, wenn wir Kenntnis über die Verbundwahrscheinlichkeiten hinreichend langer Worte hätten. Diese Kenntnis haben wir jedoch in der Praxis zumeist nicht, obgleich wir die Quellentropie h, die eigentlich auch nur aus Verbundwahrscheinlichkeiten berechnet wird, wenigstens näherungsweise kennen. In der Praxis muß dann beurteilt werden, ob sich die Aufwendungen für die Entwicklung eines neuen Codes mit dann zumeist größerem (Online–) Codier– bzw. Decodieraufwand lohnen, bei der Aussicht, die mittlere Codelänge um nicht mehr als gerade einmal 1% zu kürzen.

Fragestrategien

Für eine anschauliche Interpretation des Fundamentalsatzes der Codierung stellen wir uns einen Lehrer, L, und einen Schüler, S, vor. L möge den Ausgang eines wiederholt durchgeführten zufälligen Versuches registrieren, in dessen Ergebnis eines von M Ereignissen $\{a_1, \ldots, a_M\}$ mit den Wahrscheinlichkeiten $\{p_1, \ldots, p_M\}$ eintritt. Die Versuchsergebnisse seien voneinander statistisch unabhängig. Man denke z. B. an das Ziehen einer Karte aus einem Kartenspiel von 32 Karten, wobei gezogene Karten zurückgelegt werden und vor dem erneuten Ziehen „gut“ gemischt wird. S soll nicht sehen, was L gezogen hat, aber er kann L nach der gezogenen Karte fragen. Dabei seien nur *Entscheidungsfragen* zugelassen, die L entweder mit „Ja“ oder „Nein“ beantwortet muß. Nach dem obigen Satz ist es dann prinzipiell möglich, daß S bei Kenntnis der Wahrscheinlichkeiten eine *optimale Fragestrategie* entwickelt, so daß er bei wiederholter Durchführung des Versuches mit an Sicherheit grenzender Wahrscheinlichkeit für eine jede zu erfragende Karte im Mittel nur beliebig wenig mehr als h Fragen benötigt. Dabei gilt hier wegen der statistischen Unabhängigkeit der Versuche $h = I(\mathcal{P})$. Werden nur wenige Versuche durchgeführt, so könnte S im Mittel pro Versuch auch weniger als h Fragen benötigen. Steigt aber die Anzahl der Versuche, so wird dies immer unwahrscheinlicher und letztlich praktisch unmöglich.

Stellen wir uns nun vor, S würde immer in einer bestimmten Reihenfolge nur nach einer

bestimmten Karte fragen, etwa

Fragestrategie (FS):

1. Frage: „Ist es ♢ 7?“
2. Frage: „Ist es ♢ 8?“

   ⋮

31. Frage: „Ist es ♣ K?“

Sollte S bei der 31. Frage anlangen und wird diese verneint, so braucht er offenbar nicht mehr nach der letzten Karte ♣ As fragen, denn dann hat L offenbar ♣ As gezogen. Hat S aber viel Glück, so lautet gleich die erste Antwort „Ja“, und er braucht nicht weiter zu fragen. Die Wahrscheinlichkeit hierfür fällt jedoch mit $p = 1/32$ recht gering aus. Für den zweiten Versuch möge L die zuerst gezogene Karte zurücklegen, gut mischen und erneut eine Karte ziehen. Fragt S wiederum entsprechend der Strategie FS, so hat er erneut eine Wahrscheinlichkeit von $1/32$ für die Antwort „Ja“ auf die erste Frage. Allerdings liegt die Wahrscheinlichkeit, in zwei Versuchen mit je einer Frage auszukommen, bei nur $p^2 = 1/32^2 \approx 10^{-3}$. Sollte er in t Versuchen hintereinander mit nur einer Frage auskommen wollen, so wird dies mit einer wachsenden Versuchszahl t rasch immer unwahrscheinlicher. Die Wahrscheinlichkeit hierfür ist p^t, sie fällt also *exponentiell* mit t. Höchst wahrscheinlich müßte S also pro Versuch mehrmals fragen. Im ungünstigsten Fall sind dies gerade 31 Fragen. Aber wieviele Fragen benötigt er im Mittel pro Versuch?

Um dies zu beantworten, führen wir die Wahrscheinlichkeit $q^{(n)}$ ein, in einem Versuch genau n Entscheidungsfragen zu verwenden. Dann folgt für die mittlere Anzahl von Fragen

$$L(\mathrm{FS}) = \sum_{n=1}^{31} q^{(n)} n \ . \tag{3.11}$$

Hat nun L die Karte ♢ 7 gezogen, so würde S genau eine Frage benötigen. Dies tritt mit der Wahrscheinlichkeit $q^{(1)} = 1/32$ auf. Wurde ♢ 8 gezogen, so benötigt S genau $n = 2$ Fragen. Dies tritt ebenfalls mit der Wahrscheinlichkeit $q^{(2)} = 1/32$ auf, usw. Allgemein würde S offenbar genau n Fragen mit der Wahrscheinlichkeit $q^{(n)} = 1/32$ benötigen. Eine Ausnahme bilden hier lediglich die beiden letzten Fälle, in denen L ♣ K oder ♣ As gezogen hat. Hier benötigt S genau 31 Fragen, so daß also $q^{(31)} = 2/32$ gilt. Setzen wir diese Wahrscheinlichkeiten in (3.11) ein, so

erhalten wir die mittlere Anzahl von Fragen pro Versuch,

$$\begin{aligned} L(\mathrm{FS}) &= \left(\frac{1}{32} \times \sum_{n=1}^{30} n\right) + \frac{2}{32} \times 31 \\ &= \frac{15 \times 31 + 2 \times 31}{32} = 16\,{}^{15}/_{32} \approx 16,47 \ . \end{aligned}$$

Man überlegt sich leicht, daß sich hieran nichts ändert, wenn S die Reihenfolge der Fragen in der Strategie FS von Versuch zu Versuch zufällig permutiert.

Nach unserem Satz 3.2 müßte S aber bei einer optimalen Fragestrategie $\mathrm{FS}_{\mathrm{opt}}$ mit nur

$$L(\mathrm{FS}_{\mathrm{opt}}) = -\sum_{m=1}^{32} p_m \,\mathrm{ld}\, p_m = \mathrm{ld}\, 32 = 5$$

Fragen auskommen. Aus der Tatsache

$$L(\mathrm{FS}) > L(\mathrm{FS}_{\mathrm{opt}})$$

lesen wir ab, daß die Fragestrategie FS von S nicht optimal ist.

Tatsächlich findet S aber recht schnell eine optimale Strategie, indem er einen Fragebaum konstruiert, so daß er auf jedem Zweig des Baumes nach gewissen Untergruppen *möglichst gleicher* Wahrscheinlichkeit fragt. So könnte er z. B. wie folgt fragen:

Fragestrategie ($\mathrm{FS}_{\mathrm{opt}}$):

1. Frage: „Ist die Karte rot?“
2*a*. Frage (falls 1. bejaht): „Ist die Karte ♢?“
2*b*. Frage (falls 1. verneint): „Ist die Karte ♠?“
3. Frage: „Ist die Karte eine Zahl?“
4*a*. Frage (falls 3. bejaht): „Ist die Karte kleiner als 9?“
4*b*. Frage (falls 3. verneint): „Ist die Karte B oder D?“
5*a*. Frage (falls 4a. bejaht): „Ist die Karte 7?“
5*b*. Frage (falls 4a. verneint): „Ist die Karte 9?“
5*c*. Frage (falls 4b. bejaht): „Ist die Karte B?“
5*d*. Frage (falls 4b. verneint): „Ist die Karte K?“

Damit hat S schon nach 2 Fragen vollständige Kenntnis darüber, ob die von L gezogene Karte zur Gruppe ♢, ♡, ♠ oder ♣ gehört. Nach der vierten Antwort hat S nur noch die Ungewißheit über zwei Karten, so daß er also nur noch eine letzte 5. Frage zu stellen braucht. Mit dieser

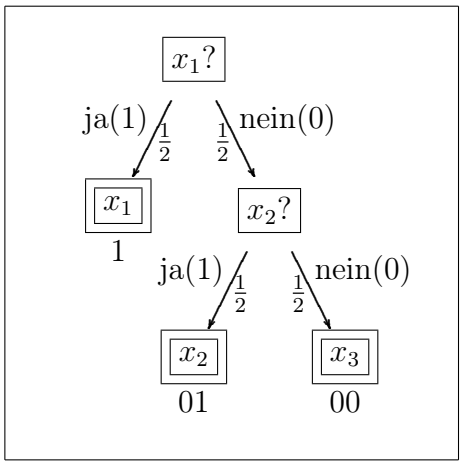

Abb. 3.2: Optimale Fragestrategie gemäß dem Code (3.1) und den Wahrscheinlichkeiten $p_1 = 1/2$ **sowie** $p_2 = p_3 = 1/4$

Strategie benötigt S also in einem jeden Versuch genau 5 Fragen, die mittlere Anzahl von Fragen pro Versuch ist dann ebenfalls 5. Mit dieser Strategie $\mathrm{FS}_{\mathrm{opt}}$ benötigt S also im Mittel die kleinst mögliche Anzahl von Fragen, was sie als optimal ausweist. Im Allgemeinen gibt es nicht nur eine optimale Strategie.

Entscheidend für eine optimale Fragestrategie ist es, daß auf einer beliebigen Stufe des Fragebaumes mit einer Entscheidungsfrage ein Maximum an Information erlangt wird. Dies tritt genau dann ein, wenn wir mit einer jeden unserer Fragen nach der Zugehörigkeit zu einer von zwei *gleich wahrscheinlichen* Gruppen fragen, in denen alle Ereignisse (Karten) verbleiben, die mit den vorherigen Antworten noch nicht ausgeschlossen worden sind. Genau dann liefert eine jede Antwort ein bit an Information. Haben die zwei Gruppen und somit die beiden möglichen Antworten ungleiche Wahrscheinlichkeiten, so ist die Information pro Antwort immer geringer als ein bit. Dies kann deshalb keine optimale Fragestrategie liefern.

Aus der Fragestrategie kann man unmittelbar einen Binärcode ableiten, wenn man die Antworten „Ja“ mit 1 und „Nein“ 0 codiert, oder umgekehrt. Sind z. B. nur drei Versuchsergebnisse (Buchstaben) mit den Wahrscheinlichkeiten $p_1 = 0,5$ sowie $p_2 = p_3 = 0,25$ möglich, so erreicht man das Optimum schon mit der Codierung einzelner Buchstaben, etwa mit dem Code in (3.1). Eine mögliche Fragestrategie ist als Fragebaum in Abbildung 3.2 angegeben. Mit der Wahrscheinlichkeit $1/2$ bekommt man schon mit der ersten Frage Kenntnis über den eingetretenen Buchstaben. Mit der gleichen Wahrscheinlichkeit benötigt man jedoch eine weitere Frage. Im Mittel sind also $1/2 \times 1 + 1/2 \times 2 = 1,5$ Fragen zur vollständigen Charakterisierung des

m	p_m	Guppen	Code
1	0,4	} 1	1
2	0,2	} 1	01
3	0,2	} 1	001
4	0,1	} 1	0001
5	0,05	} 1	00001
6	0,05	} 0	00000

Tab. 3.1: Beispiel zur Shannon–Fano–Codierung (1. Variante)

Versuchsergebnisses erforderlich. Für die Entropie (Shannon–Information) erhält man ebenfalls $h = I(\mathcal{P}) = 1,5\,\text{bit}$. Aus dem Satz 3.2 können wir also für dieses Beispiel schließen, daß unsere Fragestrategie (bzw. unser Code) bereits optimal ist — mit weniger als $1,5$ Fragen werden wir im Langzeitmittel nicht auskommen können.

Shannon–Fano–Codierung

Es stellt sich nun die Frage nach einer Systematik, wie wir zu einem beliebigen (endlichen) Alphabet mit der gegebenen Wahrscheinlichkeitsverteilung $\mathcal{P} = \{p_1, \ldots, p_M\}$ eine optimale Fragestrategie (einen optimalen Code) entwickeln können . Dabei hat uns die Beweisskizze zum Satz 3.2 bereits einige Hinweise gegeben. Danach sollte man z. B. die Codewortlänge l_m des $m$ten Buchstabens a_m möglichst nahe bei (aber nicht kleiner als) $-\operatorname{ld} p_m$ wählen. Generell sollte man versuchen, mit jeder Entscheidungsfrage (jeder Stelle des Codewortes) möglichst 1 bit Information zu bekommen. Dazu müssen die Antworten möglichst unabhängig von den vorherigen Antworten sein und mit gleicher Wahrscheinlichkeit „Ja“ bzw. „Nein“ lauten. SHANNON und FANO haben dazu in den Jahren 1948/49 einen Vorschlag gemacht, der durch das Beispiel in Tabelle 3.1 illustriert wird. Man verfährt dabei in den folgenden Schritten:

1. Ordne die Buchstaben nach der Größe ihrer Wahrscheinlichkeiten.

2. Bilde zwei Gruppen von Buchstaben mit möglichst gleich großen Wahrscheinlichkeiten. In der ersten Gruppe seien die wahrscheinlicheren und in der zweiten die weniger wahrscheinlichen Buchstaben zusammengefaßt, entsprechend der im ersten Schritt entstandenen Ordnung.

3. Wiederhole den vorherigen Schritt für eine jede Buchstabengruppe, solange, bis jede Gruppe aus nur noch einem Buchstaben besteht.

m	p_m	Guppen	Code
1	0,4	1 | 1	11
2	0,2	1 | 0	10
3	0,2	0 | 1	01
4	0,1	0 | 0 | 1	001
5	0,05	0 | 0 | 0 | 1	0001
6	0,05	0 | 0 | 0 | 0	0000

Tab. 3.2: Beispiel zur Shannon–Fano–Codierung (2. Variante)

Ordnet man bei jeder Aufteilung der einen Gruppe die Null und der anderen die Eins zu, so erhält man schließlich für einen jeden Buchstaben ein binäres Codewort. Man überlegt sich leicht, daß dies ein Präfixcode ist.

Hinsichtlich des resultierenden Codes ist der Algorithmus nicht immer eindeutig. So könnte man für unser obiges Beispiel auch wie in Tabelle 3.2 verfahren. Allerdings erhält man bei aller Willkür immer dieselbe mittlere Codewortlänge (3.8). In unserem Beispiel beträgt sie für beide Codes 2,3. Die Güte dieser Codierung mißt sich nach dem Satz 3.2 am Abstand zur Shannon–Entropie. Wir erhalten

$$I(\mathcal{P}) = -\sum_{m=1}^{6} p_m \operatorname{ld} p_m = \operatorname{ld} 5 - 0,1 \approx 2,22 < 2,3.$$

Damit liegt unsere mittlere Codewortlänge also um weniger als 4% über der kleinst möglichen Länge $h = I(\mathcal{P})$, so daß sich unsere Codes aus praktischer Sicht schon als nahezu optimal erweisen. Wollten wir uns dem theoretisch Möglichen noch stärker nähern, dann sollten wir nach der obigen Beweisführung zum Fundamentalsatz der Codierung (s. S. 61) anstelle einzelner Buchstaben Worte codieren.

Huffman–Codierung

Der Shannon–Fano–Algorithmus liefert im Allgemeinen nicht einen kürzesten Code. Bevor wir dies mit einem Beispiel illustrieren, wenden wir uns jedoch einem zweiten Codier–Algorithmus zu, den HUFFMAN 1952 angegeben hat. Seine Anwendung auf unser Beispiel zeigt Tabelle 3.3.

Allgemein lautet der Huffman–Algorithmus wie folgt:

1. Ordne die Buchstaben des Alphabetes

$$A = \{a_1, a_2, \ldots a_{M-2}, a_{M-1}, a_M\}$$

	A		$A^{(1)}$		$A^{(2)}$		$A^{(3)}$		$A^{(4)}$	
m	p_m	$\mathcal{C}$	$p_m^{(1)}$	$\mathcal{C}^{(1)}$	$p_m^{(2)}$	$\mathcal{C}^{(2)}$	$p_m^{(3)}$	$\mathcal{C}^{(3)}$	$p_m^{(4)}$	$\mathcal{C}^{(4)}$
1	0,4	0	0,4	0	0,4	0	0,4	0	0,6	1
2	0,2	10	0,2	10	0,2	10	0,4	11	0,4	0
3	0,2	111	0,2	111	0,2	111	0,2	10		
4	0,1	1101	0,1	1101	0,2	110				
5	0,05	11001	0,1	1100						
6	0,05	11000								

Tab. 3.3: Beispiel zur Huffman–Codierung. Angegeben sind jeweils die Wahrscheinlichkeiten sowie der Code des Ausgangsalphabetes A sowie der reduzierten Alphabete $A^{(k)}$, $k = 1, \ldots, 4$.

nach der Größe ihrer Wahrscheinlichkeiten. Sei also

$$p_1 \geq p_2 \geq \ldots \geq p_M \quad .$$

2. Fasse die beiden Buchstaben a_m und a_{m-1} mit den geringsten Wahrscheinlichkeiten zusammen, so daß ein reduziertes Alphabet

$$A^{(1)} = \{a_1, a_2, \ldots a_{M-2}, a_{M-1}^{(1)}\}$$

mit den Wahscheinlichkeiten

$$\{p_1, p_2, \ldots, p_{M-2}, p_{M-1}^*\}$$

entsteht, wobei $p_{M-1}^* = p_{M-1} + p_M$.

3. Wiederhole die beiden vorherigen Schritte für das gerade entwickelte reduzierte Alphabet.

Der Algorithmus bricht ab, wenn nach $M-2$ Reduzierungen ein Alphabet $A^{(M-2)}$ entstanden ist, das nur noch aus zwei Buchstaben besteht. Den gesuchten Code des Ausgangsalphabetes A erhält man, indem die beiden Buchstaben im letzten reduzierten Alphabet $A^{(M-2)}$ mit 0 und 1 codiert werden und dann die Entwicklung zurück verfolgt wird (also in der Tabelle 3.3 entgegen der Pfeilrichtung), wobei an jeder Verzweigung mit 0 und 1 unterschieden wird.

Der Algorithmus ist wiederum nicht eindeutig. So ist es z. B. willkürlich, welchem Teil man bei einer Verzweigung den Codebuchstaben 0 bzw. 1 zuordnet. Darüber hinaus ist die

Ordnung der Buchstaben des Alphabetes bzw. eines reduzierten Alphabetes nach der Größe der Wahrscheinlichkeiten nicht eindeutig, wenn einige Wahrscheinlichkeiten gleich groß sind. Allerdings kann man zeigen, daß diese Willkürlichkeiten keinen Einfluß auf die mittlere Länge L des Codes haben. In unserem Beispiel erhalten wir die mittlere Codewortlänge $L = 2,3$ — denselben Wert lieferte auch die Shannon–Fano–Codierung. Im Unterschied zu dieser gilt hier aber im Allgemeinen der folgende Satz:
Ist $A = \{a_1, \ldots, a_m, \ldots, a_M\}$ ein Alphabet, dessen Buchstaben a_m mit den Wahrscheinlichkeiten p_m auftreten. Dann liefert der Huffman–Algorithmus einen Binärcode mit der kleinst möglichen mittleren Codewortlänge, die man bei der binären Codierung einzelner Buchstaben erreichen kann.

Codiert man anstelle einzelner Buchstaben alle Worte

$$(a_{m_1}, a_{m_2}, \ldots, a_{m_d}),$$

die aus d Buchstaben von A gebildet sind und mit den Wort–Wahrscheinlichkeiten

$$p_{m_1 m_2 \ldots m_d}$$

auftreten, dann liefert der Huffman–Algorithmus einen Binärcode, dessen mittlere Codewortlänge pro Buchstabe des Ausgangsalphabetes A für $d \to \infty$ gegen die Quellentropie h aus Gl. (2.31) strebt. (Ein Beweis findet sich z. B. in [13], S. 15 f.)

Will man also eine (unendliche) stationäre Folge $m_1, m_2, \ldots, m_t, \ldots$ von Buchstaben codieren, wobei auch statistische Abhängigkeiten auftreten dürfen, so liefert der Huffman–Algorithmus optimale Codes, d.h., bei der Codierung immer längerer Worte würde asymptotisch die kleinst mögliche mittlere Codewortlänge pro Buchstaben erhalten werden, also die Quellentropie (2.31). Allerdings scheitert die Umsetzung dieses Verfahrens in der Regel an der Unkenntnis der Wortwahrscheinlichkeiten hinreichend großer Ordnung. Deshalb müssen in der Praxis andere Wege beschritten werden. Die hier diskutierten Grundideen finden sich aber in diesen praktischen Codieralgorithmen, mehr oder weniger versteckt, wieder. Im Spezialfall der Unabhängigkeit gilt allerdings

$$p_{m_1 m_2 \ldots m_d} = p_{m_1} p_{m_2} \ldots p_{m_d} \quad .$$

Hier können wir also von den Einzelwahrscheinlichkeiten sofort auf die Wortwahrscheinlichkeiten beliebig hoher Ordnung schließen, und es gilt $h = -\sum_m p_m \operatorname{ld} p_m$.

3.3 Fehlererkennung und –korrektur

Bei der Übertragung oder Speicherung von Codewörtern treten im Allgemeinen Fehler auf. Durch eine geschickte Codierung gelingt es, einen Teil dieser Fehler zu erkennen und eventuell auch zu korrigieren. Denken wir z. B. an einen geschriebenen Text. Solange die Anzahl der orthographischen Fehler nicht zu groß ist, wird der Inhalt des Textes noch richtig gedeutet, wie das folgende Beispiel zeigt:

1: Dies ist ein Text mit Schreibfehlern.
2: Dies ist ein Text mit Schreibfelern.
3: Dies ist Text mit Schreibfehlern.
4: Dies ist kein Text mit Schreibfehlern.

Der erste Satz ist syntaktisch fehlerfrei, wenngleich dies der Aussage, der Semantik, des Satzes widerspricht. Hier wird jedoch allein der syntaktische Aspekt, also die Beziehung der Zeichen untereinander berücksichtigt, nicht aber der semantische, der die Bedeutung der Zeichenfolgen widerspiegelt. Im letzten Wort des zweiten Satzes fehlt offenbar der Buchstabe „h“. Dieser Fehler kann erkannt werden, weil das Wort „Schreibfelern“ in der deutschen Schriftsprache nicht vorkommt. Aber selbst der dritte Satz enthält alle Information, obgleich hier ein ganzes Wort fehlt. Schließlich führt im vierten Satz ein zusätzlicher Buchstabe „k“ zu einem Satz mit korrekter Syntax. Folglich kann der Fehler hier nicht bemerkt werden, es sei denn, daß der Satz aus seinem Kontext, wie es diese erklärenden Worte sind, als fehlerhaft erkannt wird.[6]

Wir sehen also, daß einige Fehler erkannt werden können. Voraussetzung hierfür ist es, daß nicht alle Worte, die kombinatorisch möglich sind, auch tatsächlich vorkommen. Natürliche Sprachen verwenden nur einen äußerst geringen Bruchteil aller kombinatorisch möglichen Worte und Sätze. Geht man z. B. vom Alphabet { a, b, c, ..., x, y, z, ä, ö, ü } aus, so können daraus $\sum_{d=1}^{5} 29^d = 21\,243\,689$ Worte der Längen $d = 1, \ldots, 5$ gebildet werden. Die deutsche Sprache kennt aber nur etwa 500 000 Worte, wobei darunter Worte mit noch viel größeren Wortlängen als

[6] Die Möglichkeiten zur Fehlerkorrektur sind bei natürlichen Sprachen aber noch viel umfangreicher. Führen z. B. Schreibfehler innerhalb eines Wortes zu einem noch zulässigen Wort, so stimmt dann häufig nicht mehr die Grammatik, wie etwa beim richtigen Satz „*Dort steht die Bank.*“, der mit Schreibfehlern „*Dort steht der Bank.*“ lauten möge. Kommt es gar zur Sinnentstellung, wie in „*Dort geht die Bank.*“, so erkennen wir ebenso einen Fehler. Durch Raten könnten wir diese Sinnentstellung sogar mit hoher Wahrscheinlichkeit korrigieren, zumal dann, wenn uns der Satz mündlich erreicht und mit einer entsprechenden Geste unseres Gesprächspartners verbunden ist.

5 sind. Die Folge hiervon ist, daß geringfügige Schreibfehler in der Regel erkannt und korrigiert werden können.

Begriffsbildungen

Im Folgenden werden einige allgemeine Verfahren der Fehlererkennung vorgestellt. Sie basieren darauf, daß man Codes mit *kontrollierter Redundanz* verwendet. Unter der Redundanz $R(\mathcal{C}, \mathcal{P})$ eines Codes $\mathcal{C}$ wird hier die Differenz der mittleren Codewortlänge L von der Quellentropie h verstanden,

$$R(\mathcal{C}, \mathcal{P}) \equiv L - h \ . \tag{3.12}$$

Nach dem Fundamentalsatz der Codierung (s. S. 60) gilt immer

$$0 \leq R(\mathcal{C}, \mathcal{P}) \ . \tag{3.13}$$

Im Folgenden betrachten wir nur sogenannte *Blockcodes*. Bei einem solchen Code haben alle Codeworte die gleiche Länge, $l_m = L$ für $m = 1, \ldots, M$. Nach (3.3) ist dann die mittlere Codewortlänge L gleich der Länge des Codes, $L = \mathcal{L}(\mathcal{C}_r)$. Darüber hinaus wird hier ein Radix 2 Code, also $r = 2$, vorausgesetzt und deshalb kurz $\mathcal{C}$ statt $\mathcal{C}_2$ geschrieben.

Unter dem *Abstand zweier Codeworte* $B_1 = b_{11}b_{12}\ldots b_{1L}$ und $B_2 = b_{21}b_{22}\ldots b_{2L}$ eines Blockcodes versteht man die Anzahl der Buchstaben, in denen sich B_1 und B_2 unterscheiden,

$$\mathrm{dist}(B_1, B_2) \equiv \#\{l \ : \ b_{1l} \neq b_{2l}, \ l = 1, 2, \ldots, L\} \ . \tag{3.14}$$

So gelten z. B. $\mathrm{dist}(0000, 0000) = 0$ und $\mathrm{dist}(0010, 1101) = 4$. Allgemein gilt offenbar immer

$$0 \leq \mathrm{dist}(B_1, B_2) \leq L \ . \tag{3.15}$$

Binäre Blockcodes $\mathcal{C}$ der Länge L enthalten höchstens 2^L Codeworte. Den kleinsten Abstand zwischen allen (verschiedenen) Codeworten eines binären Blockcodes nennt man *Hamming–Distanz des Codes*,

$$\mathrm{dist}_{\mathrm{H}}(\mathcal{C}) \equiv \min_{m,n \,:\, m \neq n} \{\mathrm{dist}(B_m, B_n)\} \ . \tag{3.16}$$

Da alle Codeworte eines eindeutig decodierbaren Codes unterschiedlich sind, gilt hier

$$1 \leq \mathrm{dist}_{\mathrm{H}}(\mathcal{C}) \leq L \ . \tag{3.17}$$

Die Anzahl von Einsen eines Codewortes heißt *Gewicht des Codewortes*. Unter einem *gleichgewichtigen Binärcode* versteht man einen Binärcode, bei dem ein jedes Codewort die gleiche Anzahl von Einsen hat.

Korrekturradius

Werden mit einem binären Blockcode der Länge L genau 2^L Buchstaben codiert, so hat der Code die Hamming–Distanz 1, und jeder Übertragungsfehler, bei dem entweder „0“ statt „1“ empfangen wird oder umgekehrt, ist nicht korrigierbar. Um einen solchen Fehler korrigieren zu können, muß die Hamming–Distanz offenbar größer als 1 sein. Dann unterscheiden sich die verwendeten Codeworte in zumindest 2 Stellen, so daß ein einzelner Fehler notwendig zu einem nicht zulässigen Codewort führt. Treten jedoch in einem Codewort mehr als ein Fehler auf, so wird der Fehler im Allgemeinen nicht erkannt, nämlich genau dann nicht, wenn das durch den Doppelfehler entstandene Codewort ein zulässiges Codewort ist. Im Allgemeinen werden also n–fache Fehler in einem Codewort eines binären Blockcodes $\mathcal{C}$ mit Sicherheit erkannt, wenn $n < \mathrm{dist}_{\mathrm{H}}(\mathcal{C})$ gilt.

Zur Konstruktion eines fehlererkennenden Codes werden häufig sogenannte *Prüfbits* verwendet. Zur Erläuterung betrachten wir das Beispiel der Codierung von 8 Buchstaben mit einem binären Blockcode:

Buchstabe	Code 1	Code 2
a_1	000	000 **0**
a_2	001	001 **1**
a_3	010	010 **1**
a_4	011	011 **0**
a_5	100	100 **1**
a_6	101	101 **0**
a_7	110	110 **0**
a_8	111	111 **1**
	$\mathrm{dist}_{\mathrm{H}} = 1$	$\mathrm{dist}_{\mathrm{H}} = 2$
		Parität gerade

Der Code 1 hat die Hamming–Distanz 1. Folglich können hier Fehler, einzelne wie auch mehrfache, nicht erkannt werden. Beim Code 2 ist ein zusätzliches Prüfbit in der letzten Position eingefügt, so daß das Gewicht (die Anzahl der Einsen) eines jeden Codewortes gerade ist. Ebensogut hätte das Gewicht eines jeden Codewortes durch Invertierung des Prüfbits immer

ungerade gewählt werden können. Im ersten Fall spricht man von der *geraden* und im zweiten, nicht dargestellten Fall von einer *ungeraden Parität*. Die Länge des Codes 2 ist um eins größer als die von Code 1. Seine Hamming–Distanz ist 2 und folglich werden nun einzelne Bitfehler erkannt, aber ebenso werden Dreifachfehler erkannt. Offenbar wird im Allgemeinen bei Codes mit einer Hammingdistance von 2 eine ungerade Anzahl von Fehlern erkannt.

Es stellt sich die Frage, ob in unserem obigen Beispiel die erkennbaren Einzelfehler bei Verwendung von Code 2 auch korrigierbar sind. Dazu betrachten wir zunächst den Fall, daß 100 0 empfangen wird. Das Gewicht (die Parität) ist hier ungerade, so daß offenbar ein Fehler aufgetreten sein muß. Der Fehler ist aber nicht korrigierbar. Gehen wir z. B. davon aus, daß es sich um einen Einzelfehler handelt, dann kann sowohl 000 0 (Korrektur in der ersten Stelle) als auch 100 1 (Korrektur in der vierten Stelle) als korrigiertes Codewort auftreten. Eine eindeutige Korrektur ist also nicht möglich. In der Praxis der Nachrichtenübertragung bedeutet dies, daß der Empfänger eine erneute Übertragung des Codewortes beim Sender beantragen muß.

Unter welchen Bedingungen kann nun ein Fehler nicht nur erkannt, sondern auch korrigiert werden, so daß eine Wiederholung der Übertragung des Codewortes nicht nötig ist? Gehen wir vom Beispiel der Codierung von nur zwei Buchstaben aus:

Buchstabe	Code 1	Code 2
a_1	0	0 **00**
a_2	1	1 **11**
	$\mathrm{dist_H} = 1$	$\mathrm{dist_H} = 3$

Der Code 2 ist nun so gewählt, daß neben Einzelfehlern sogar Doppelfehler erkannt würden, denn seine Hamming–Distanz ist 3. Wir gehen aber davon aus, daß nur ein Einzelfehler erkannt wird, dafür aber auch korrigierbar sein soll. Tatsächlich ist eine solche Korrektur nun möglich, denn unter der Annahme, daß höchstens ein Einzelfehler aufgetreten ist, werden alle empfangenen Worte 1 00, 0 10 und 0 01 zu 0 00 korrigiert und die übrigen, also 0 11, 1 01 und 1 10, zu 1 11. Allgemein findet man, daß die Anzahl E_{kor} der korrigierbaren Fehler, der sogenannte *Korrekturradius* des Codes $\mathcal{C}$, mit der Hamming–Distanz $\mathrm{dist_H}$ des Codes wie folgt zusammenhängt,

$$E_{\mathrm{kor}}(\mathcal{C}) \leq \begin{cases} (\mathrm{dist_H}(\mathcal{C}) - 1)/2 & : \quad \mathrm{dist_H}(\mathcal{C}) \text{ ungerade} \\ (\mathrm{dist_H}(\mathcal{C}) - 2)/2 & : \quad \mathrm{dist_H}(\mathcal{C}) \text{ gerade.} \end{cases} \tag{3.18}$$

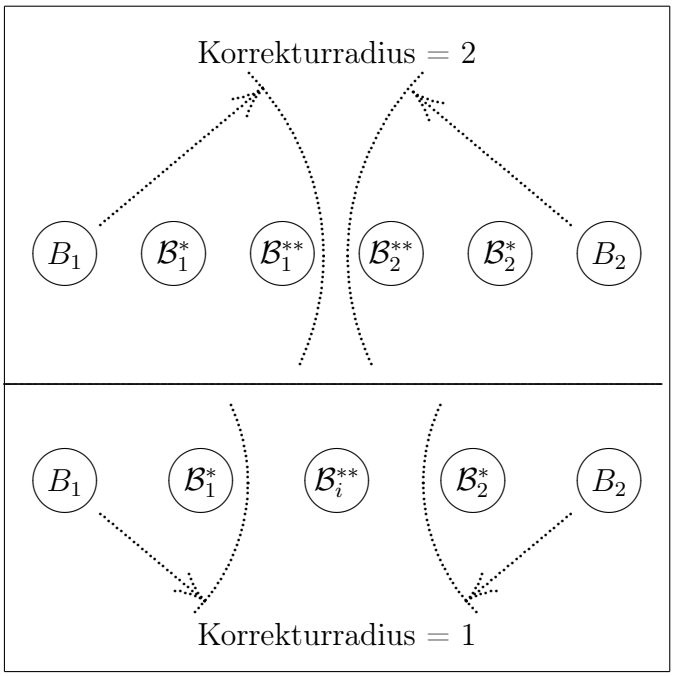

Abb. 3.3: Korrekturradii von Codes der Hamming–Distanz $\mathrm{dist_H} = 5$ **(oben) und** $\mathrm{dist_H} = 4$ **(unten). Im oberen Fall werden bis zu Vierfachfehler erkannt, während Einfach– und Doppelfehler auch korrigierbar sind. Im unteren Fall werden bis zu Dreifachfehler erkannt, während nur Einfachfehler auch korrigierbar sind**

Anschaulich stellen wir uns dazu zwei Codeworte der Hamming–Distanz $\mathrm{dist_H} > 1$ vor, wie es in der Abbildung 3.3 für die beiden Fälle $\mathrm{dist_H} = 5$ und $\mathrm{dist_H} = 4$ dargestellt ist. B_1 und B_2 seien tatsächlich auftretende Codeworte des Codes $\mathcal{C}$. Für ihren Abstand gilt folglich

$$\mathrm{dist}(B_1, B_2) \geq \mathrm{dist_H}(\mathcal{C}) \ .$$

Im Bild ist der geringste Abstand, also $\mathrm{dist}(B_1, B_2) = \mathrm{dist_H}(\mathcal{C})$ angenommen, dies ist offenbar der kritischste Fall. $\mathcal{B}_m^*$ enthalte alle Binärworte der Länge $l = L(\mathcal{C})$, die aus B_m durch Änderung einer Binärstelle entstehen, also

$$\mathcal{B}_m^* = \{B \ : \ \mathrm{dist}(B, B_m) = 1\} \quad \text{für } m = 1,\ 2 \ .$$

Entsprechend gelte

$$\mathcal{B}_m^{**} = \{B \ : \ \mathrm{dist}(B, B_m) = 2\} \quad \text{für } m = 1,\ 2 \ .$$

Offenbar enthalten weder $\mathcal{B}_m^*$ noch $\mathcal{B}_m^{**}$ zulässige Codeworte, weil ihr Abstand zu einem zulässigen Codewort (zu B_m) kleiner als die Hamming–Distanz $\mathrm{dist_H}$ des Codes $\mathcal{C}$ ist. Nehmen wir nun an, ein Codewort wird fehlerhaft übertragen, so daß statt B_m ein Binärwort B empfangen

wird. Im Falle $\mathrm{dist_H}(\mathcal{C}) = 5$ (oberer Teil der Abbildung 3.3) kann B offenbar nur aus genau einem zulässigen Codewort entstanden sein, solange die Anzahl der Bitfehler kleiner als 3 ist. Für $\mathrm{dist_H}(\mathcal{C}) = 4$ (unterer Teil der Abbildung 3.3) ist die eindeutige Rekonstruktion aber nur bei höchstens einem Bitfehler möglich, denn schon bei einem Doppelfehler kann das fehlerhafte Wort B aus B_1 oder B_2 entstanden sein. Man findet somit, daß im allgemeinen Fall der Korrekturradius die Bedingung (3.18) erfüllt.

Liegt ein jedes der kombinatorisch möglichen Binärworte der Länge L im Korrekturradius um ein zulässiges Codewort, so spricht man von einem *dichtgepackten Code*. Offenbar kann ein Code mit einer geraden Hamming–Distanz, wie z. B. in der Abbildung 3.3 unten, nicht dichtgepackt sein.

Kann ein Code einen Fehler korrigieren, so heißt er *Hamming–Code*. Nach den obigen Ausführungen ist seine Hamming–Distanz zumindest 3.

Prüfworte

Ein einfaches Verfahren zur Fehlererkennung und –korrektur ist durch das Einfügen eines *Prüfwortes* gegeben. Zur Illustration dieses Verfahrens gehen wir vom Code 2 in der Tabelle auf der Seite 72 aus. Denken wir uns nun eine Folge von möglichen Codeworten übereinander dargestellt, in der Reihenfolge, wie sie ausgesandt bzw. empfangen werden.

Zeitschritt	Codewort aus Code 2	
t	Code 1	Prüfbit
	$\pi_1\pi_2\pi_3$	π_4
1	011	**0**
2	111	**1**
3	000	**0**
4	100	**1**
$5 = T$	011	**0**
Prüfwort	**011**	**0**

Das Prüfwort $\pi_1\pi_2 \ldots \pi_L$ wird in der Weise gebildet, daß das Gewicht einer jeden Spalte gerade ist. Eine Ausnahme macht hier nur die letzte Stelle π_L, die das Paritätsbit des Prüfwortes

bildet. Allgemein gilt also

$$\pi_l = \begin{cases} 0 & : \quad \sum_{t=1}^{T} b_{tl} \text{ gerade} \\ 1 & : \quad \text{sonst} \end{cases} , \qquad l = 1, 2, \ldots, L-1 \ .$$

Hierbei ist T die Anzahl der Codeworte, die zwischen zwei Prüfworten übertragen wird, und b_{tl} ist die l–te Binärstelle im t–ten Codewort. Tritt nun in der Bitmatrix ein Einzelfehler auf, so kann die Fehlerstelle genau lokalisiert, die Zeile wird durch das Prüfbit und die Spalte durch das Prüfwort erkannt. (Einzelfehler im Prüfbit oder Prüfwort brauchen nicht korrigiert werden.) Fehler in mehreren Codeworten sind jedoch im Allgemeinen nicht mehr eindeutig lokalisierbar und folglich auch nicht korrigierbar. In der Praxis muß man deshalb hinreichend häufig Prüfworte senden, also die Anzahl T der Codeworte zwischen aufeinanderfolgenden Prüfworten hinreichend klein machen, so daß die Wahrscheinlichkeit für einen Mehrfachfehler gering ist und somit Korrekturen in der Regel möglich sind.

Prüfmuster

Wir betrachten nun ein weiteres Beispiel für einen fehlererkennenden und –korrigierenden Code (s. Tab.3.4). Seine Codewortlänge beträgt 7. $b_1 b_2 b_4$ sind Prüfbits (Paritätsbits) für jeweils eine Prüfgruppe von 4 Binärstellen. Die restlichen vier Stellen $b_3 b_5 b_6 b_7$ codieren die 16 Buchstaben. Tritt nun an einer der sieben möglichen Stellen eines Codewortes $b_1 b_2 b_3 b_4 b_5 b_6 b_7$ ein Einzelfehler auf, so wird dieser durch Überprüfung aller drei Prüfgruppen erkannt und lokalisiert und damit auch korrigierbar, denn wie die untere Tabelle zeigt, führt jede fehlerhafte Codewortstelle auf ein unterschiedliches Prüfmuster. Durch die drei Prüfstellen können offenbar genau $2^3 = 7 + 1$ Codewortstellen codiert werden. Die achte Stelle entspricht der fehlerfreien Übertragung, bei der alle Prüfbedingungen erfüllt sind.

Gehen wir im Allgemeinen von einem binären Blockcode der Länge L aus, der k Prüfstellen hat, so muß offenbar

$$2^k \geq 1 + L$$

gelten, damit ein Einzelfehler erkannt und korrigiert werden kann. Es stehen also $L - k$ Codewortstellen für die Codierung der eigentlichen Nachricht, also des Buchstabens, zur Verfügung. Für die notwendige Prüfstellenzahl k erhält man

$$k \geq \operatorname{ld}(1 + L) \ .$$

codierter Buchstabe	Codewort						
	$\boldsymbol{b_1}$	$\boldsymbol{b_2}$	b_3	$\boldsymbol{b_4}$	b_5	b_6	b_7
a_0	**0**	**0**	0	**0**	0	0	0
a_1	**1**	**1**	0	**1**	0	0	1
a_2	**0**	**1**	0	**1**	0	1	0
a_3	**1**	**0**	0	**0**	0	1	1
a_4	**1**	**0**	0	**1**	1	0	0
a_5	**0**	**1**	0	**0**	1	0	1
a_6	**1**	**1**	0	**0**	1	1	0
a_7	**0**	**0**	0	**1**	1	1	1
a_8	**1**	**1**	1	**0**	0	0	0
a_9	**0**	**0**	1	**1**	0	0	1
a_{10}	**1**	**0**	1	**1**	0	1	0
a_{11}	**0**	**1**	1	**0**	0	1	1
a_{12}	**0**	**1**	1	**1**	1	0	0
a_{13}	**1**	**0**	1	**0**	1	0	1
a_{14}	**0**	**0**	1	**0**	1	1	0
a_{15}	**1**	**1**	1	**1**	1	1	1

fehlerhafte Codewortstelle	nicht erfüllte Prüfbedingungen (Prüfmuster)		
	b_1	b_2	b_4
b_1	×	·	·
b_2	·	×	·
b_3	×	×	·
b_4	·	·	×
b_5	×	·	×
b_6	·	×	×
b_7	×	×	×
kein Fehler	·	·	·

Paritätsbit	Prüfgruppe						
b_1	b_1		b_3		b_5		b_7
b_2		b_2	b_3			b_6	b_7
b_4				b_4	b_5	b_6	b_7

Tab. 3.4: Fehlererkennender und –korrigierender binärer Blockcode der Länge $L = 7$, mit drei Paritätsbits. Das Prüfmuster $b_1b_2b_4$ codiert die Position eines Einzelfehlers im Codewort $b_1b_2b_3b_4b_5b_6b_7$. Dabei bedeutet $b_1b_2b_4 = \cdots$, daß kein Fehler aufgetreten ist

Offenbar werden die Paritätsbits des Codes „gut“ ausgenutzt, wenn für die Codewortlänge gerade $L = 2^k - 1$ gilt, worin k eine ganze (positive) Zahl ist, die Anzahl der Paritätsbit bzw. die Länge des Prüfmusters. Dies ist im Beispiel der Tabelle 3.4 gegeben, für das $k = 3$ gilt. Im Falle $k = 4$ wäre die Codewortlänge $L = 15$. Hier würden 4 Paritätsbits auftreten, mit denen Einzelfehler in den $2^4 - 1 = 15$ Binärstellen des Codes codiert werden können. Das 16. Prüfmuster zeigt den fehlerfreihen Fall an. Wir sehen also, daß das Prüfmuster einen „Code im Code“ darstellt.

Soll ein Code der Länge L neben den einfachen auch Doppelfehler korrigieren, so sind folgenden Überlegungen anzustellen: Offenbar gibt es nun insgesamt $L + \binom{L}{2}$ Fehlermöglichkeiten, L Einfach– und $\binom{L}{2}$ Zweifachfehler. Damit muß die Anzahl der Prüfstellen k die Bedingung

$$k \geq \operatorname{ld} \left[1 + L + \binom{L}{2} \right]$$

erfüllen. Sollen im allgemeinen Fall bis zu N–fache Fehler korrigiert werden, so muß offenbar folgendes gelten,

$$\boxed{k \geq \operatorname{ld} \sum_{n=0}^{N} \binom{L}{n}} \tag{3.19}$$

Die Summe auf der rechten Seite wird *Hamming–Grenze* genannt.

Will man z. B. einen binären Blockcode der Länge $L = 20$ konstruieren, so daß Einfach– und Doppelfehler korrigierbar sind, also für $N = 2$, dann muß für die Anzahl der Prüfstellen

$$k \geq \operatorname{ld} \left[1 + 20 + \binom{20}{2} \right] = \operatorname{ld} 211$$

gelten. Weil die Anzahl der Paritätsbits ganzzahlig sein muß, wird man $k = 8 = \operatorname{ld} 256$ wählen. Somit verbleiben noch $L - k = 12$ Binärstellen für die Codierung von $2^{12} = 4096$ Buchstaben. In diesem Beispiel werden nicht alle der mit den 8 Prüfstellen codierbaren Fehler genutzt, weil L nicht als $2^k - 1$ darstellbar ist. In dieser Hinsicht wäre es günstiger, wenn die Codewortlänge L gerade so gewählt wird, daß in (3.19) k gleich der Hamming–Grenze ist. Für $N = 2$ müßten wir also

$$k = \operatorname{ld} \left[1 + L + \binom{L}{2} \right] = \operatorname{ld} \left(1 + L + \frac{L!}{2(L-2)!} \right)$$

fordern.

Kapitel 4

Gestörte Kanäle

Im Zusammenhang mit dem Bergerschen Diagramm (Abb. 2.4, S. 33) haben wir bereits davon gesprochen, daß über einen Nachrichtenkanal nur soviel Information übertragen wird, wie die Transinformation angibt. Es besteht nun ein erhebliches betriebswirtschaftliches Interesse, diese Transinformation zu maximieren, würde doch dann ein gegebener Nachrichtenkanal voll ausgenutzt. Dies kann erhebliche Kosten sparen.

Um nun die Transinformation zu maximieren, werden die Sendewahrscheinlichkeiten an den Kanal geeignet angepaßt, man spricht von der *Kanalcodierung*. Der Kanal wird hierbei als gegeben angesehen. Im folgenden diskutieren wir dies genauer, zunächst für Kanäle, die endlich viele Buchstaben übertragen, sogenannte *diskreten Kanäle*. Später werden wir auch kontinuierliche Kanäle betrachten.

4.1 Diskrete gestörte Kanäle

Ein *diskreter Nachrichtenkanal* ist beschrieben durch ein Eingangsalphabet $\{x_m\}_{m=1}^{M}$, ein Ausgangsalphabet $\{y_n\}_{n=1}^{N}$ sowie die bedingten Wahrscheinlichkeiten $\{q_{n|m}\}_{m,n=1}^{M,N}$. Hierbei ist $q_{n|m}$ die Wahrscheinlichkeit, daß y_n empfangen wird, unter der Bedingung, daß x_m gesandt wurde. Ein diskreter Nachrichtenkanal heißt *gedächtnislos*, wenn für beliebige ausgesandte Worte

$$X^{(D)} = x_{m_1} x_{m_2} \dots x_{m_D}$$

und empfangene Worte

$$Y^{(D)} = y_{n_1} y_{n_2} \dots y_{n_D}$$

und alle Wortlängen $D = 1, 2, 3, \ldots$ folgendes gilt,

$$\text{prob}\left\{Y^{(D)}|X^{(D)}\right\} = \prod_{d=1}^{D} q_{n_d|m_d} \quad . \tag{4.1}$$

Hierbei ist $\text{prob}\left\{Y^{(D)}|X^{(D)}\right\}$ die bedingte Wahrscheinlichkeit, $Y^{(D)}$ zu empfangen, unter der Bedingung, daß $X^{(D)}$ gesendet wurde. Die Beziehung (4.1) bedeutet, daß die Wahrscheinlichkeit $q_{n|m}$, bei der Aussendung von x_m den Buchstaben y_n zu empfangen, nicht von den vorher ausgesandten und empfangenen Buchstaben abhängt.

Sind die Wahrscheinlichkeiten $\mathcal{P} \equiv \{p_m\}$, mit denen die Buchstaben $\{x_m\}$ auftreten, gegeben, so folgen aus den Übergangswahrscheinlichkeiten $\{q_{n|m}\}$ die Empfangswahrscheinlichkeiten $\{q_n\}$, denn es gilt

$$q_n = \sum_{m=1}^{M} s_{mn} = \sum_{m=1}^{M} p_m q_{n|m} \quad . \tag{4.2}$$

Hier ist $s_{mn} = p_m q_{n|m}$ die Verbundwahrscheinlichkeit dafür, daß x_m ausgesandt und y_n empfangen wird.

Die Übergangswahrscheinlichkeiten $\{q_{n|m}\}$ werden in der *Übergangsmatrix* (*Rauschmatrix*) zusammengefaßt,

$$\mathcal{Q}_{n|m} = \begin{pmatrix} q_{1|1} & \cdots & q_{n|1} & \cdots & q_{N|1} \\ \vdots & & & & \\ q_{1|m} & \cdots & q_{n|m} & \cdots & q_{N|m} \\ \vdots & & & & \\ q_{1|M} & \cdots & q_{n|M} & \cdots & q_{N|M} \end{pmatrix} \tag{4.3}$$

Für die Summe über alle Wahrscheinlichkeiten einer Zeile der Übergangsmatrix gilt offenbar

$$\sum_{n=1}^{N} q_{n|m} = 1 \quad , m = 1, 2, \ldots, M \quad .$$

Das bedeutet, welcher Buchstabe x_m auch immer gesandt wurde, irgend ein Buchstabe y_n wird mit Sicherheit empfangen werden.

Zu diesen Verbundwahrscheinlichkeiten gehört die Transinformation

$$I_{\mathrm{T}} = \sum_{m,n=1}^{M,N} s_{mn} \,\mathrm{ld}\, \frac{s_{mn}}{p_m q_n} \quad . \tag{4.4}$$

Nach (4.2) hängt die Transinformation zum einen von den Sendewahrscheinlichkeiten $\mathcal{P}$ ab, welche die Quelle charakterisieren, und zum anderen von den bedingten Wahrscheinlichkeiten

$\{q_{n|m}\}$, welche den Kanal beschreiben,

$$I_{\mathrm{T}}\big(\mathcal{P}, \mathcal{Q}_{n|m}\big) = \sum_{m,n=1}^{M,N} p_m q_{n|m} \,\mathrm{ld}\, \frac{p_m q_{n|m}}{p_m \sum_{m^*=1}^{M} p_{m^*} q_{n|m^*}} \quad .$$

Unter der *Kanalkapazität* eines diskreten gedächtnislosen Nachrichtenkanals versteht man das Maximum dieser Transinformation, welches über alle Wahrscheinlichkeitsverteilungen $\mathcal{P}$ des Eingangsalphabetes $\{x_m\}$ bei festen Übergangswahrscheinlichkeiten $\mathcal{Q}_{n|m}$ gebildet wird,

$$\boxed{C_{\mathrm{K}}\big(\mathcal{Q}_{n|m}\big) \equiv \max_{\mathcal{P}} I_{\mathrm{T}}\big(\mathcal{P}, \mathcal{Q}_{n|m}\big)} \tag{4.5}$$

Mit der Buchstaben–Sendefrequenz f_{send} kann die Kanalkapazität auch in Einheiten von bit/s angegeben werden,

$$C_{\mathrm{K}}\,[\mathrm{bit/s}] = C_{\mathrm{K}}\,[\mathrm{bit/Buchstabe}] \times f_{\mathrm{send}}\,[\mathrm{Buchstabe/s}] \quad .$$

Dabei darf aber nicht übersehen werden, daß die den Kanal charakterisierenden Übertragungswahrscheinlichkeiten $q_{n|m}$ im allgemeinen selbst Funktionen von f_{send} sind. In der Praxis gilt immer

$$\lim_{f_{\mathrm{send}} \to \infty} C_{\mathrm{K}}\big(\{q_{n|m}(f_{\mathrm{send}})\}\big) = 0 \quad .$$

Bei endlichen Sendefrequenzen, so daß $C_{\mathrm{K}} > 0$ gilt, steht die Aufgabe, die Sendewahrscheinlichkeiten $\mathcal{P}$ gerade so zu wählen, daß die Transinformation maximiert wird, um die gegebene Kanalkapazität C_{K} voll auszunutzen. Dazu muß eine entsprechende Codierung erfolgen, die sogenannte *Kanalcodierung*. Eine Codierung erfolgt somit im allgemeinen in zwei Schritte (vgl. Abb. 1.1):

1. Quellcodierung, mit dem Ziel der Beseitigung der Redundanz des Signals. Beispielsweise führt der Huffman–Algorithmus zu einer Quellcodierung (s. Kap. 3).

2. Kanalcodierung, mit dem Ziel der vollen Nutzung der Fähigkeit des Kanals, Nachrichten zu übertragen.

Beispiel: Gestörter Binärkanal

Der *gestörte Binärkanal*, überträgt nur zwei Buchstaben, „Low“ (L) und „High“ (H). Er ist durch die Rauschmatrix

$$\mathcal{Q}_{n|m} = \begin{pmatrix} q_{\mathrm{L|L}} & q_{\mathrm{H|L}} \\ q_{\mathrm{L|H}} & q_{\mathrm{H|H}} \end{pmatrix} = \begin{pmatrix} 1-\varepsilon_{\mathrm{L}} & \varepsilon_{\mathrm{L}} \\ \varepsilon_{\mathrm{H}} & 1-\varepsilon_{\mathrm{H}} \end{pmatrix} \tag{4.6}$$

beschrieben. Hierin sind $\varepsilon_{\mathrm{L}} \equiv q_{\mathrm{H|L}}$ und $\varepsilon_{\mathrm{H}} \equiv q_{L|H}$ die Wahrscheinlichkeiten einer fehlerhaften Übertragung von L bzw. von H. In realen Nachrichtenkanälen schwanken diese Fehlerwahrscheinlichkeiten zwischen 10^{-8} und 10^{-4}. Im allgemeinen laufen sie aber von 0 bis 1.

Unter Beachtung von (4.2) können wir nun für die Transinformation (4.4) wie folgt schreiben,

$$\begin{aligned} I_{\mathrm{T}} &= -p_{\mathrm{L}} \operatorname{ld} p_{\mathrm{L}} - p_{\mathrm{H}} \operatorname{ld} p_{\mathrm{H}} \\ &\quad +(1-\varepsilon_{\mathrm{L}}) p_{\mathrm{L}} \operatorname{ld}(1-\varepsilon_{\mathrm{L}}) + \varepsilon_{\mathrm{L}} p_{\mathrm{L}} \operatorname{ld} \varepsilon_{\mathrm{L}} \\ &\quad +\varepsilon_{\mathrm{H}} p_{\mathrm{H}} \operatorname{ld} \varepsilon_{\mathrm{H}} + (1-\varepsilon_{\mathrm{H}}) p_{\mathrm{H}} \operatorname{ld}(1-\varepsilon_{\mathrm{H}}) \ . \end{aligned} \tag{4.7}$$

Um die Kanalkapazität (4.5) zu erhalten, suchen wir das Maximum von I_{T} bei Variation der Sendewahrscheinlichkeiten p_{L} und p_{H}. Dazu setzen wir die entsprechenden partiellen Ableitungen null,

$$\left.\frac{\partial I_{\mathrm{T}}}{\partial p_{\mathrm{L}}}\right|_{p_{\mathrm{L}}=p_{\mathrm{L}}^{\mathrm{opt}},\, p_{\mathrm{H}}=p_{\mathrm{H}}^{\mathrm{opt}}} = 0 \ , \quad \left.\frac{\partial I_{\mathrm{T}}}{\partial p_{\mathrm{H}}}\right|_{p_{\mathrm{L}}=p_{\mathrm{L}}^{\mathrm{opt}},\, p_{\mathrm{H}}=p_{\mathrm{H}}^{\mathrm{opt}}} = 0 \ .$$

Mit den Abkürzungen

$$\begin{aligned} H_{\mathrm{empf|}L} &\equiv -(1-\varepsilon_{\mathrm{L}}) \operatorname{ld}(1-\varepsilon_{\mathrm{L}}) - \varepsilon_{\mathrm{L}} \operatorname{ld} \varepsilon_{\mathrm{L}} \\ H_{\mathrm{empf|}H} &\equiv -\varepsilon_{\mathrm{H}} \operatorname{ld} \varepsilon_{\mathrm{H}} - (1-\varepsilon_{\mathrm{H}}) \operatorname{ld}(1-\varepsilon_{\mathrm{H}}) \\ A_{\mathrm{H}} &\equiv \Delta\,(\varepsilon_{\mathrm{H}} H_{\mathrm{empf|L}} - (1-\varepsilon_{\mathrm{L}}) H_{\mathrm{empf|H}}) \\ A_{\mathrm{L}} &\equiv \Delta\,(\varepsilon_{\mathrm{L}} H_{\mathrm{empf|H}} - (1-\varepsilon_{\mathrm{H}}) H_{\mathrm{empf|L}}) \\ \Delta &\equiv 1/\det \mathcal{Q}_{n|m} = (1-\varepsilon_{\mathrm{H}}-\varepsilon_{\mathrm{L}})^{-1} \end{aligned}$$

erhalten wir (nach einer einfachen aber etwas aufwendigen Rechnung) die Kanalkapazität des binären Nachrichtenkanals,

$$\boxed{C_{\mathrm{K}} = \operatorname{ld}\left[2^{A_{\mathrm{H}}} + 2^{A_{\mathrm{L}}}\right]} \tag{4.8}$$

Die zugehörigen optimalen Sendewahrscheinlichkeiten sind

$$\boxed{\begin{aligned} p_{\mathrm{H}}^{\mathrm{opt}} &= 2^{-C_{\mathrm{K}}\Delta}\ \left[(1-\varepsilon_{\mathrm{L}})2^{A_{\mathrm{H}}} - \varepsilon_{\mathrm{L}}2^{A_{\mathrm{L}}}\right] \\ p_{\mathrm{L}}^{\mathrm{opt}} &= 2^{-C_{\mathrm{K}}\Delta}\ \left[(1-\varepsilon_{\mathrm{H}})2^{A_{\mathrm{L}}} - \varepsilon_{\mathrm{H}}2^{A_{\mathrm{H}}}\right] \end{aligned}} \tag{4.9}$$

Die Kanalkapazität (4.8) sowie die optimalen Sendewahrscheinlichkeiten (4.9) hängen nur von den beiden Fehlerwahrscheinlichkeiten ε_{H} und ε_{L} ab. Die Determinante $\det \mathcal{Q}_{n|m}$ der Rauschmatrix $\mathcal{Q}_{n|m}$ verschwindet, falls $1 = \varepsilon_{\mathrm{H}} + \varepsilon_{\mathrm{L}}$ gilt. In diesem Fall gilt für die Kanalkapazität $C_{\mathrm{K}} = 0$.

Werden beide Symbole H und L gleich häufig gestört, gilt also $\varepsilon_{\mathrm{H}} = \varepsilon_{\mathrm{L}} \equiv \varepsilon$, so nennen wir den Kanal *symmetrisch gestört*. In diesem Fall ist es am günstigsten, wenn $p_{\mathrm{H}} = p_{\mathrm{L}} = 1/2$ gewählt wird. Es gilt dann auch $A_{\mathrm{L}} = A_{\mathrm{H}} = -\varepsilon \operatorname{ld}\varepsilon - (1-\varepsilon)\operatorname{ld}(1-\varepsilon)$, und für die Kanalkapazität erhalten wir somit

$$C_{\mathrm{K}} = 1 - \left[-\varepsilon \operatorname{ld}\varepsilon - (1-\varepsilon)\operatorname{ld}(1-\varepsilon)\right] \ .$$

Die Kanalkapazität nimmt hier ihren Maximalwert an für den ungestörten Kanal, $\varepsilon = 0$. Aber auch im Fall $\varepsilon = 1$, wo regelmäßig beim Senden von L der Buchstabe H empfangen wird und umgekehrt, wird die Kanalkapität maximal. Dies entspricht einer simplen Invertierung im Nachrichtenkanal, wodurch keine Information „verloren geht“.

4.2 Kontinuierliche gestörte Kanäle

In realen Nachrichtenkanälen senden wir in der Regel kontinuierliche (analoge) Signale, selbst dann, wenn nach einer entsprechenden Quellcodierung schon ein digitaler (z. B. binärer) Code vorliegt. Demzufolge ist eine Betrachtung zur Kanalkapazität bei kontinuierlichen Signalen angebracht.

Gegeben sei ein kontinuierliches Signal der Zeitdauer T und der Bandbreite B. Der Kanal selbst habe eben diese Bandbreite. Nach dem Samplingtheorem können wir dann gerade mit $2BT$ Abtastwerten

$$\boldsymbol{x} = (x_1, \ldots, x_t, \ldots, x_{2BT}) \tag{4.10}$$

das Signal beschreiben.[1] Die Abtastwerte fassen wir nun als Koordinaten im $2BT$–dimensionalen

[1] Hierbei setzen wir voraus, daß $2BT$ ganzzahlig ist, was für die folgenden Betrachtungen keine wesentliche Einschränkung ist.

Raum $\mathbb{R}^{2BT}$ auf. Jedes mögliche Signal entspricht dann einem Vektor $\boldsymbol{x}$ in diesem Raum. Dabei hängt die Leistung $P_{\boldsymbol{x}}$ des Signals mit der Länge $\|\boldsymbol{x}\|$ des Signalvektors (4.10) wie folgt zusammen,

$$P_{\boldsymbol{x}} = \frac{1}{2BT}\sum_{t=1}^{2BT} x_t^2 = \frac{\|\boldsymbol{x}\|^2}{2BT} \ . \tag{4.11}$$

Würden diese Abtastwerte fehlerfrei übertragen, so könnte man das Signal beim Empfänger eindeutig rekonstruieren. Dazu müßte aber mit einem jeden Abtastwert unendlich viel Information übertragen werden, stammt dieser Abtastwert doch aus einem Kontinuum möglicher Werte eines Intervalls auf der reellen Zahlenachse. Die Kanalkapazität wäre dann unendlich groß. In realen Kanälen treten jedoch Störungen auf. Dadurch wird die Kanalkapazität endlich, was die folgenden Überlegungen zeigen.

Mögen also im Kanal Störungen auftreten, die wir als additives Rauschen $\boldsymbol{r}$ beschreiben,

$$\boldsymbol{x} \longrightarrow \boldsymbol{x} + \boldsymbol{r} = (x_1 + r_1,\ x_2 + r_2, \ldots,\ x_{2BT} + r_{2BT}) \ .$$

Das Rauschen habe selbst wieder die Bandbreite B, mit der Rauschleistung

$$P_{\boldsymbol{r}} = \frac{\|\boldsymbol{r}\|^2}{2BT} \ . \tag{4.12}$$

Alle bandbegrenzten Signale der Leistung $P_{\boldsymbol{x}}$, die mit einem Rauschen von höchstens der Leistung $P_{\boldsymbol{r}}$ gestört sind, liegen in einer $2BT$–dimensionalen Kugelschale

$$K_{\|\boldsymbol{x}\|\pm\|\boldsymbol{r}\|} = \{\boldsymbol{y} \in \mathbb{R}^{2BT} : \|\boldsymbol{y}\|_{\min} < \|\boldsymbol{y}\| < \|\boldsymbol{x}\| + \|\boldsymbol{r}\|\} \tag{4.13}$$

Hier setzen wir $\|\boldsymbol{y}\|_{\min} = \|\boldsymbol{x}\| - \|\boldsymbol{r}\|$, falls $\|\boldsymbol{x}\| > \|\boldsymbol{r}\|$, und $\|\boldsymbol{y}\|_{\min} = 0$, falls $\|\boldsymbol{x}\| \leq \|\boldsymbol{r}\|$.

Um das Volumen dieser Kugelschale zu bestimmen, erinnern wir zunächst an die allgemeine Formel

$$V_N(R) = C_N R^N \ , \text{ mit } \quad C_N = \frac{\pi^{N/2}}{\Gamma(1 + N/2)} \ , \tag{4.14}$$

für das Volumen einer N–dimensionalen Kugel vom Radius R. [2] Das Volumen einer N–dimensionalen

[2] Für den Vorfaktor in der Gleichung (4.14) gilt

$$C_N = \frac{2\pi}{N} C_{N-2}$$

für $N = 3, 4, 5, \ldots$, wobei $C_1 = 2$ und $C_2 = \pi$. Somit erhalten

$$V_N(R) = \begin{cases} \left(\frac{\pi^{N/2}}{(N/2)!}\right) R^N & : \ N \text{ gerade} \\ \frac{2\cdot(2\pi)^{(N-1)/2}}{3\cdot 5\cdot 7 \ldots N} R^N & : \ N \text{ ungerade} \ . \end{cases}$$

Kugelschale zwischen den Radii R und $R-\delta$, mit $0<\delta<R$, unterscheidet sich für große Dimensionen N kaum von Volumen einer Kugel vom Radius R, denn es gilt

$$\begin{aligned}\frac{V_N(R)-V_N(R-\delta)}{V_N(R)} &= \frac{C_N R^N - C_N(R-\delta)^N}{C_N R^N} \\ &= 1-\left(1-\frac{\delta}{R}\right)^N \quad .\end{aligned}$$

Folglich erhalten wir im Grenzfall

$$\lim_{N\to\infty}\frac{V_N(R)-V_N(R-\delta)}{V_N(R)} = 1 \ ,$$

das Volumen der Kugelschale geht also in das der Kugel über.

Wir fragen nun nach der Anzahl M von Signalen in unserer Kugelschale (4.13), die noch unterschieden werden können, wenn wir eine Rauschleistung bis zu $P_{\boldsymbol{r}}$ zulassen. Offenbar ist M näherungsweise durch das Verhältnis des Volumens der Kugelschale zum Volumen der Rauschkugel $K_{\|\boldsymbol{r}\|}$ gegeben. Wir setzen große Zeitabschnitte, $T\to\infty$, voraus und können deshalb anstelle des Volumens der Kugelschale das der Kugel vom Radius $\|\boldsymbol{x}\|+\|\boldsymbol{r}\|$ setzen,

$$M = \frac{V_{2BT}(\|\boldsymbol{x}\|+\|\boldsymbol{r}\|)}{V_{2BT}(\|\boldsymbol{r}\|)} = \left(\frac{\|\boldsymbol{x}\|+\|\boldsymbol{r}\|}{\|\boldsymbol{r}\|}\right)^{2BT} \quad ,$$

Mit (4.11) und (4.12) finden wir

$$M = \left(\frac{\sqrt{P_{\boldsymbol{x}}}+\sqrt{P_{\boldsymbol{r}}}}{\sqrt{P_{\boldsymbol{r}}}}\right)^{2BT} \quad . \tag{4.15}$$

Setzen wir nun jedes Signal als gleich wahrscheinlich voraus, so kann pro Zeiteinheit die Information $C_{\mathrm{K,kont}} = \frac{1}{T}\operatorname{ld} M$ übertragen werden. Mit (4.15) können wir dafür auch wie folgt schreiben:

$$\boxed{C_{\mathrm{K,kont}} = 2B\operatorname{ld}\left(1+\sqrt{\frac{P_{\boldsymbol{x}}}{P_{\boldsymbol{r}}}}\right)} \tag{4.16}$$

Ist der Radius R fest vorgegeben, so findet man, daß das Volumen einer N–dimensionalen Kugel mit wachsender Dimension N gegen null strebt,

$$\lim_{N\to\infty} V_N(R) = 0 \ .$$

Damit kommt aber auch jeder Punkt der Kugel mit wachsender Dimension N der Kugeloberfläche beliebig dicht, was offenbar unserer Anschauung widerspricht. Geben wir uns also eine beliebig dünne Kugelschale vor, dann liegen mit wachsenden N immer mehr Punkte in der Kugelschale. Für eine unendlich große Dimension sind es fast alle Punkte.

Ist also die Bandbreite $B > 0$ gegeben, so können wir über den Kanal immer dann Information übertragen, wenn $P_{\boldsymbol{x}}/P_{\boldsymbol{r}} > 0$ gilt, selbst dann noch, wenn für $P_{\boldsymbol{x}}/P_{\boldsymbol{r}} \gtrsim 0$ das eigentliche Signal tief im Rauschen liegt und darin fast „untergeht“.

Kapitel 5

Praktische Codierverfahren

5.1 Vektorcodierung

Die *Vektorcodierung* oder auch *Vektorquantisierung* ist ein universelles Verfahren, um Folgen zufälliger Vektoren mit einer vorgegebenen mittleren Genauigkeit zu approximieren. Ausgangspunkt hierfür ist eine Trainingssequenz, die für die zu codierenden Signale statistisch repräsentativ ist. Aus dieser wird ein *Codebuch* erzeugt, dessen Einträge heißen *Codebuchvektoren*. Das Codebuch ist ein Reservoir von Repräsentanten, welche alle möglichen zufälligen Vektoren der Trainingssequenz im Mittel möglichst genau approximieren. In der Praxis wählt man die Anzahl der Codebuchvektoren bedeutend kleiner, als die Anzahl der möglichen zufälligen Vektoren. Daraus resultiert letztlich eine Datenreduktion, dies allerdings im allgemeinen zum Preis einer gewissen Ungenauigkeit. Linde, Buzzo und Gray [10] haben im Jahre 1980 ein Verfahren vorgeschlagen, um aus einer endlichen Anzahl von Trainingsvektoren die Codebuchvektoren zu bestimmen, den sogenannten *LBG–Algorithmus*.

Vektorquantisierer spielen eine Rolle z. B. bei der redundanzarmen Codierung von Sprachsignalen. Mit dem Verfahren können weitreichende statistische Abhängigkeiten berücksichtigt werden, indem hinreichend hoch–dimensionale Vektoren betrachtet werden. Darüber hinaus können Vektorquantisierer bei verschiedenen Aufgaben der Klassifizierung von zufälligen Merkmalsvektoren verwendet werden.

Prinzip

Gegeben sei eine Folge

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t, \ldots, \boldsymbol{x}_T \tag{5.1}$$

von D–dimensionalen Trainingsvektoren,

$$(x_{t,1}, \ldots, x_{t,d}, \ldots, x_{t,D}) \equiv \boldsymbol{x}_t \in \mathbb{R}^D \ .$$

Wir können uns diese zum Beispiel aus einem zeitdiskreten skalaren Signal

$$x_1, \ldots, x_D, \ x_{D+1}, \ldots, x_{2D}, \ \ldots, \ x_{(T-1)D+1} \ldots, x_{TD} \tag{5.2}$$

entstanden denken, wenn wir immer D aufeinander folgende Werte zusammenfassen,

$$\boldsymbol{x}_t = x_{(t-1)D+1}, \ldots, x_{tD}, \quad t = 1, 2, \ldots, T \ .$$

Wir partitionieren nun $\mathbb{R}^D$ mit K paarweise disjunkten Boxen,

$$\begin{aligned} \beta &\equiv \{B_k\}_{k=1}^K \ , \\ B_{k_1} \bigcap B_{k_2} &= \emptyset \quad \text{falls } k_1 \neq k_2 \ , \\ \bigcup_{k=1}^K B_k &= \mathbb{R}^D \ . \end{aligned}$$

Aus jeder Box B_k wählen wir einen Repräsentanten $\boldsymbol{v}_k$, den *Codebuchvektor*. Die Menge aller Codebuchvektoren bildet das *Codebuch*,

$$V \equiv \{\boldsymbol{v}_k\}_{k=1}^K \ . \tag{5.3}$$

Somit kann die Trainingssequenz (5.1) oder auch eine beliebige andere Folge von Vektoren auf eine Folge von Codebuchindizes abgebildet werden:

$$\boldsymbol{x}_t \xrightarrow{\boldsymbol{x}_t \in B_{k_t}} k_t \ .$$

Die Folge

$$k_1, k_2, \ldots, k_t, \ldots, k_T$$

kann nun beispielsweise binär codiert über einen Nachrichtenkanal zum Empfänger übertragen werden. Kennt dieser das Codebuch (5.3), so kann er daraus die Folge

$$\boldsymbol{v}_{k_1}, \boldsymbol{v}_{k_2}, \ldots, \boldsymbol{v}_{k_t}, \ldots, \boldsymbol{v}_{k_T} \tag{5.4}$$

rekonstruieren. Diese Folge ist eine gewisse Approximation der ursprünglichen Folge (5.1), wenn $\boldsymbol{x}_t \approx \boldsymbol{v}_{k_t}$ gilt.

Um den praktischen Gewinn dieses Verfahrens zu verdeutlichen, nehmen wir an, das ursprünliche Signal (5.2) würde mit einer Amplitudenauflösung von N bit gewonnen sein, also bei einer Unterscheidung von 2^N Amplitudenstufen. Dann müßten bei einer direkten Übertragung NTD bit über den Nachrichtenkanal gesandt werden. Bei einer Vektorcodierung mit K Codebuchvektoren sind aber etwa $T \operatorname{ld} K$ bit zu übertragen. Eine Bitersparnis ergibt sich also, falls

$$ND > \operatorname{ld} K \ .$$

Wäre $ND = \operatorname{ld} K$, so könnten wir alle D–dimensionalen Vektoren, die bei einer Amplitudenauflösung der Koordinaten von N bit kombinatorisch möglich sind, in das Codebuch aufnehmen. Damit hätten wir zwar keine Ungenauigkeiten mehr, aber auch keine Bitersparnis bei der Signalübertragung.

Treten im Signal starke statistische Abhängigkeiten auf, dann würde nur ein Teil aller kombinatorisch möglichen Vektoren mit signifikant von null verschiedenen Wahrscheinlichkeiten auftreten. In der Praxis heißt dies oft, daß einige dieser Vektoren gar nicht vorkommen. In diesem Fall bringt die Vektorcodierung eine Bitersparnis. Wird nun die Anzahl K der Codebuchvektoren möglichst klein gewählt, können ihre Indizes mit kurzen (binären) Codeworten verschlüsselt werden. Somit wird letztlich die Bitrate des Vektorcoders klein gehalten. Andererseits darf diese Anzahl nicht zu gering werden, würde doch dann die Auswahl an Codebuchvektoren allzu gering werden, um noch den mittleren Approximationsfehler

$$T^{-1} \sum_t \|\boldsymbol{x}_t - \boldsymbol{v}_{k_t}\| \tag{5.5}$$

in akzeptablen Grenzen halten zu können. Bei der Entwicklung eines Vektorquantisierers steht somit vor allem die Aufgabe, ein Codebuch zu finden, das bei gegebener Anzahl K von Codebuchvektoren den mittleren Quantisierungsfehler minimiert.

Bei der Entwicklung eines Codebuches gehen wir von einem festen gegebenen Fehlermaß $\|\cdot\|$ aus. Dies kann der euklische Abstand

$$\|\boldsymbol{x}_t - \boldsymbol{v}_{k_t}\| = \sum_{d=1}^{D} (x_{t,d} - v_{k_t,d})^2$$

sein. Die Abbildung 5.1 illustriert die sich daraus ergebende Partitionierung für zwei verschiedene Codebücher.

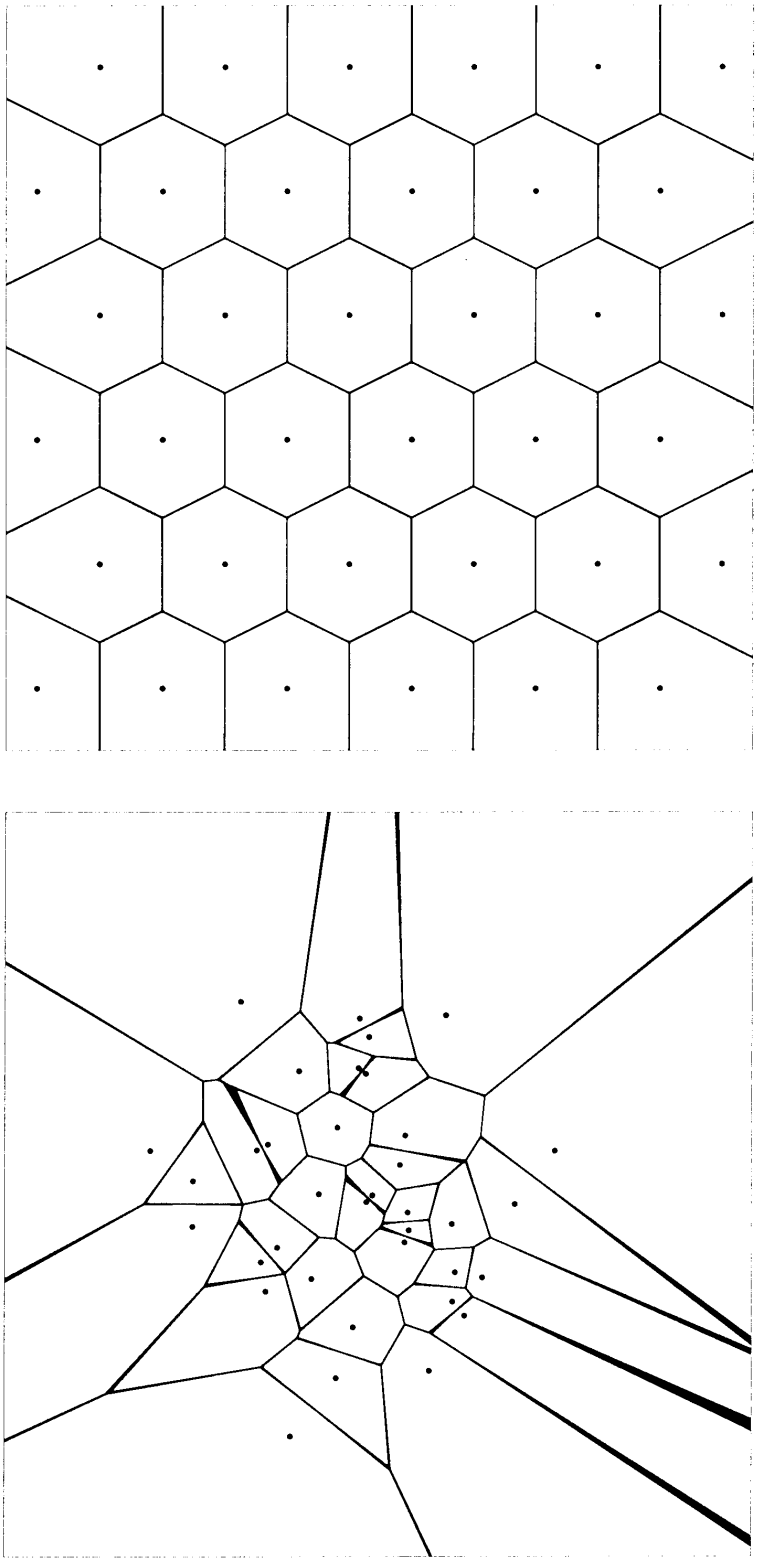

Abb. 5.1: Partitionierung der Ebene R^2 zu zwei unterschiedlich verteilten Codebuchbüchern. Die Codebuchvektoren $v_k \in \mathrm{R}^2$ sind durch Punkte markiert. Die zugehören Boxen (Cluster) B_k fassen alle Vektoren $x \in \mathrm{R}^2$ zusammen, die an v_k dichter als an einem anderen Codebuchvektor v_{k^*}, $k \neq k^*$, liegen. Auf den gezeichneten Boxengrenzen liegen alle Vektoren x, die mindestens zu zwei Codebuchvektoren den gleichen Abstand haben

Allgemein sind aber auch andere Abstandsmaße möglich, die der konkreten Signalklasse und dem Verwendungszweck des Coders angepaßt sind. So werden z. B. bei der Sprachverarbeitung gewisse hörphysiologisch motivierte Metriken verwendet.

Darüber hinaus muß bei der Entwicklung des Codebuches darauf geachtet werden, daß für die Entscheidung

$$„\ \boldsymbol{x}_t \in B_{k_t}\ ?\ “$$

ein möglichst effektiver Algorithmus zur Verfügung steht.

Optimales Codebuch bei gegebener Wahrscheinlichkeitsdichte

Für die Ableitung eines optimalen Codebuches machen wir den Ansatz, daß das Signal–Rausch–Verhältnis

$$\mathrm{SNR} \equiv 10 \lg \frac{\sum_t \|\boldsymbol{x}_t\|^2}{\sum_t \|\boldsymbol{x}_t - \boldsymbol{v}_{k_t}\|^2}$$

maximiert werden soll. Offenbar wird dies erreicht, wenn der mittlere Quantisierungsfehler (5.5) minimal ist. Sei $p(\boldsymbol{x})$ die D–dimensionale Wahrscheinlichkeitsdichte, welche die Verteilung der Vektoren der als stationär und ergodisch vorausgesetzten Folge (5.1) beschreibt. Dann können wir den mittleren Quantisierungsfehler auch als Scharmittel schreiben,

$$\lim_{T\to\infty} T^{-1} \sum_t \|\boldsymbol{x}_t - \boldsymbol{v}_{k_t}\| = \sum_{k=1}^{K} \int_{B_k} \|\boldsymbol{x} - \boldsymbol{v}_k\|^2 p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \equiv E_\beta(V) \tag{5.6}$$

$E_\beta(V)$ ist eine Funktion in den DK Koordinaten

$$\boldsymbol{v}_k = (v_{k,1}, v_{k,2}, \ldots, v_{k,d}, \ldots, v_{k,D})$$

der Codebuchvektoren. Um das Minimum von $E_\beta(V)$ über alle Codebuchvektoren zu finden, setzen wir nun die entsprechenden partiellen Ableitungen von $E_\beta(V)$ gleich null,

$$\begin{aligned} 0 &= \frac{\partial E_\beta(V)}{\partial v_{k,d}} \\ &= \int_{B_k} \frac{\partial \|\boldsymbol{x} - \boldsymbol{v}_k\|^2}{\partial v_{k,d}} p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \\ &= -\int_{B_k} 2(x_d - v_{k,d}) p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\ . \end{aligned}$$

Daraus folgt

$$v_{k,d} = \frac{\int_{B_k} x_d p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}}{\int_{B_k} p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}}\ . \tag{5.7}$$

Aus

$$\frac{\partial^2 E_\beta(V)}{\partial v_{k,d}^2} = 2\int_{B_k} p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} > 0$$

folgt, daß (5.7) tatsächlich das SNR minimiert. Danach sind also die optimalen Codebuchvektoren $\boldsymbol{v}_k$ gerade die „statistischen Schwerpunkte“ der Vektoren der Trainingssequenz, welche in die Boxen B_k, $k = 1, 2, \ldots, K$, fallen.

Die allgemeine Lösung dieses Optimierungsproblems setzt die Kenntnis der D–dimensionalen Dichte $p(\boldsymbol{x})$ voraus. Aber selbst bei dieser Kenntnis ist das Problem im allgemeinen noch recht kompliziert, zumal von vorn herein nicht klar ist, wie die Boxen B_k am günstigsten gewählt werden. Nur sehr einfache Modellsituationen erlauben eine analytische Lösung. In der Praxis werden deshalb andere Wege beschritten.

Optimales Codebuch aus LBG–Algorithmus

Linde, Buzo und Gray [10] haben den nach ihnen benannten *LBG–Algorithmus* zur Entwicklung eines optimalen Codebuches aus einer Trainingssequenz vorgeschlagen. Dabei handelt es sich um eine Methode, welche keine explizite Kenntnis der Wahrscheinlichkeitsdichte $p(\boldsymbol{x})$ voraussetzt. Letztlich wird die entsprechende Information der Trainingssequenz entnommen.

Ausgangspunkt des LBG–Algorithmus ist eine Trainingssequenz (5.1) der Länge T und eine Norm $\|\boldsymbol{x}\|$, welche eine Metrik $\|\boldsymbol{x} - \boldsymbol{y}\|$ impliziert. Der LBG–Algorithmus ist iterativ, mit den folgenden Schritten (Abbildung 5.2 stellt diese Schritte übersichtlich graphisch dar.):

1. Schritt (Initialisierung): Wir wählen zunächst ein Codebuch

$$V^{(0)} = \left\{\boldsymbol{v}_k^{(0)}\right\}_{k=1}^{K} \quad ,$$

das wir Start–Codebuch nennen, zufällig aus (auswürfeln). Die Vektoren in $V^{(0)}$ sollen alle verschieden sein, und es empfielt sich, sie „in die Nähe“ der Punkte der Trainingssequenz (5.1) zu legen. Man kann die Start–Vektoren $\boldsymbol{v}_k^{(0)}$ auch zufällig aus der Trainingssequenz auswählen.

Der Parameter

$$A^{(0)} = T^{-1}\sum_{t=1}^{T} \min_k \left\{\left\|\boldsymbol{x}_t - \boldsymbol{v}_k^{(0)}\right\|\right\}$$

wird zur Konstruktion eines Abbruchkriteriums für den LBG–Algorithmus verwendet, was wir weiter unten noch genauer erläutern werden.

Der Iterationszähler m wird zunächst auf null gesetzt.

2. Schritt (Konstruktion neuer Codebuchvektoren, $m := m + 1$): In der Box

$$B_k^{(m)} = \left\{ \boldsymbol{x} \ : \ \left\| \boldsymbol{x} - \boldsymbol{v}_k^{(m-1)} \right\| < \left\| \boldsymbol{x} - \boldsymbol{v}_{k^*}^{(m-1)} \right\| \quad \forall \quad k \neq k^* \right\}$$

fassen wir alle Punkte $\boldsymbol{x} \in \mathbb{R}^D$ zusammen, die am Codebuchvektor $\boldsymbol{v}_k^{(m-1)}$ dichter, als an einem beliebig anderen Codebuchvektor $\boldsymbol{v}_{k^*}^{(m-1)} \in V^{(m-1)}$ liegen. Den neuen Codebuchvektor zur Box $B_k^{(m)}$ gewinnen wir als Schwerpunkt aller Vektoren der Trainingssequenz, welche in die Box fallen,

$$\boldsymbol{v}_k^{(m)} = \frac{1}{\#\{\boldsymbol{x}_t \in B^{(m)}\}} \sum_{t \,:\, \boldsymbol{x}_t \in B_k^{(m)}} \boldsymbol{x}_t$$

Dabei ist $\#\{\boldsymbol{x}_t \in B^{(m)}\}$ die Anzahl dieser Vektoren.

Die neuen Codebuchvektoren $\boldsymbol{v}_k^{(m)}$, $k = 1, 2, \ldots, K$, werden im neuen Codebuch $V^{(m)}$ zusammengefaßt.

3. Schritt (Beurteilung der Güte des neuen Codebuches): Die Größe

$$\min_k \left\{ \left\| \boldsymbol{x}_t - \boldsymbol{v}_k^{(m)} \right\| \right\}$$

ist der Abstand des Vektors $\boldsymbol{x}_t$ der Trainigssequenz zu jenem neuen Codebuchvektor $\boldsymbol{v}_k^{(m)} \in V^{(m)}$, der $\boldsymbol{x}_t$ im Sinne unserer Metrik am besten approximiert, und folglich ist

$$A^{(m)} = T^{-1} \sum_{t=1}^{T} \min_k \left\{ \left\| \boldsymbol{x}_t - \boldsymbol{v}_k^{(m)} \right\| \right\}$$

der über die Trainingssequenz gemittelte Wert dieser kleinsten Abstände. Je kleiner $A^{(m)}$, desto besser können wir, in einem statistischen Sinne, die Vektoren der Trainingssequenz durch Vektoren aus dem aktuellen Codebuch $V^{(m)}$ approximieren. In diesem Sinne ist $A^{(m)}$ ein „Gütemesser“ für das Codebuch.

Wir erwarten, daß $A^{(m)}$ für fortschreitende Iterationszahl fällt. Dann mißt der Quotient

$$Q \equiv \frac{A^{(m-1)} - A^{(m)}}{A^{(m-1)}} \ ,$$

wie stark sich die Güte des Codebuches mit dem $m$ten Iterationsschritt verbessert hat. Offenbar ist Q ein relatives Qualitätsmaß, und es gilt:

Verbesserung?		Güte
keine	$A^{(m-1)} = A^{(m)}$	$Q = 0$
gering	$A^{(m-1)} \gtrsim A^{(m)}$	$0 < Q \ll 1$
stark	$A^{(m-1)} \gg A^{(m)}$	$Q \lesssim 1$
schlechter	$A^{(m-1)} < A^{(m)}$	$Q < 0$

Der Algorithmus wird abgebrochen, wenn mit einem Iteratiosschritt nur noch eine geringfügige Verbesserung erreicht wird. Damit kann $0 \leq Q < \varepsilon$ als Abbruchbedingung dienen, mit einer (zunächst) willkürlich gewählten Abbruchschwelle $\varepsilon \ll 1$. Für $\varepsilon < Q$ wiederholen wir das Prozedere ab dem 2. Schritt. Für $Q < 0$ sollte der Algorithmus mit einem veränderten Startcodebuch ausgeführt werden.

Mit dem LBG–Algorithmus findet man zu gegebener Anzahl K von Codebuchvektoren im allgemeinen kein globales optimales Codebuch. In der Praxis kann man sich damit behelfen, den Algorithmus mehrere Male mit verschiedenen Startcodebüchern auszuführen, und dann das beste Codebuch zu verwenden, für das $A^{(m)}$ bei Abbbruch des Algorithmus minimal wird.[1]

Schnelle Suche

Rechenzeit–Einschränkungen gibt es für den LBG–Algorithmus in der Regel nicht, da in den meisten Anwendungsfällen die Codebücher Off–line erzeugt werden. Allerdings ist es für die Nutzung eines Codebuches V im On–line–Betrieb wichtig, möglichst schnell den besten Repräsentanten aus V für einen Vektor $\boldsymbol{x}_t$ zu finden. Mit unserer bisherigen Vorgehensweise müßten wir zur Auffindung der besten Approximation $\boldsymbol{v}_{k_t}$ von $\boldsymbol{x}_t$, also von

$$k_t : \|\boldsymbol{x}_t - \boldsymbol{v}_{k_t}\| \leq \|\boldsymbol{x}_t - \boldsymbol{v}_k\|\,, \quad k = 1, 2, \ldots, K \tag{5.8}$$

genau K Abstandsberechnungen und $K - 1$ Vergleiche ausführen. Eine solche Vektorcodierung bzw. Vektorquantisierung (VQ) wird *full search VQ* (FSVQ) genannt. Dabei kann K einige hundert oder bei höheren Approximations–Genauigkeiten bzw. größeren Vektordimensionen D gar

[1] Der Algorithmus stellt eigentlich die Iteration eines nichtlinearen dynamischen Systems dar, nämlich einer DK–dimensionale Differenzengleichung für die Koordinaten der K Codebuchvektoren, mit dem Startcodebuch $V^{(0)} \in \mathbb{R}^{DK}$ als Anfangsbedingung und der Trainingssequenz als Satz von Kontrollparametern. In solchen Systemen kann man im allgemeinen nicht damit rechnen, daß die Codebuchvektoren auf einen stabilen Fixpunkt im Codebuchraum $\mathbb{R}^{DK}$ zulaufen.

einige tausend betragen. Somit stellt sich die Frage nach schnellen Suchalgorithmen. Durch eine geeignete Strukturierung des Codebuches kann der Suchaufwand tatsächlich erheblich reduziert werden, was wir nun genauer ausführen.

Bei der Baumsuche eines „besten“ Codebuchvektors (tree structured VQ = TSVQ) wird eine für die Suche günstige Codebuch–Struktur schon beim Entwurf des Codebuches durch geeignete Bedingungen erzwungen. Zur Entwicklung eines baumstrukturierten Codebuches geht man wie folgt vor:

1. Schritt: Erzeugung eines Anfangscodebuches

$$V \equiv \{\boldsymbol{v}_{k_1}\}_{k_1=1}^{K}$$

mit dem LBG–Algorithmus zur Trainingssequenz (5.1).

2. Schritt: Einteilung der Trainingssequenz (5.1) in K Teilsequenzen

$$\{\boldsymbol{x}_{k_1 t}\}_{t=1}^{T_{k_1}}, \quad \text{so daß } \boldsymbol{x}_{k_1 t} \in B_{k_1},\ k_1 = 1, 2, \ldots, K \ .$$

Dabei ist B_{k_1} die zu $\boldsymbol{v}_{k_1} \in V$ gehörige Box. Eine jede Teilsequenz wird nun als Trainingssequenz des LBG–Algorithmus verwendet, was zu weiteren K Codebüchern

$$V_{k_1} \equiv \{\boldsymbol{v}_{k_1 k_2}\}_{k_2=1}^{K} \quad , \quad k_1 = 1, 2, \ldots, K \ ,$$

führt.

m–ter Schritt: Es werden Teilsequenzen der Trainingssequenz gebildet, die aus allen Vektoren $\boldsymbol{x}_t$ bestehen, welche in der Box $B_{k_1 \ldots k_{m-1}}$ liegen. Zu dieser Trainingssequenz wird mit dem LBG–Algorithmus das Codebuch

$$V_{k_1 \ldots k_{m-1}} \equiv \left\{\boldsymbol{v}_{k_1 k_2 \ldots k_{m-1} k_m}\right\}_{k_m=1}^{K}$$

gebildet.

Mit diesem m–ten Schritt ist der LBG–Algorithmus also insgesamt

$$1 + K + K^2 + \ldots + K^{m-1} = \frac{K^m - 1}{K - 1}$$

mal angewandt. Auf dieser m–ten Stufe erhalten wir somit K^m Codebuchvektoren. Dabei ist zu beachten, daß die Teil–Trainings-

sequenzen in der Regel von Schritt zu Schritt kürzer werden, so daß sie immer schlechter die tatsächliche Wahrscheinlichkeitsverteilung repräsentieren.

Um nun einen Eingangsvektor $\boldsymbol{x}_t$ zu codieren, könnten wir einerseits wie in (5.8) verfahren, auf der Grundlage des Codebuches $V_{k_1 \ldots k_m}$ gemäß dem FSVQ. Dazu wären K^m Abstandsberechnungen und $K^m - 1$ Vergleiche nötig. Wir können nun aber auch iterativ vorgehen, indem wir zunächst

$$k_1 : \|\boldsymbol{x}_t - \boldsymbol{v}_{k_1}\| \leq \|\boldsymbol{x}_t - \boldsymbol{v}_k\|, \quad k = 1, 2, \ldots, K$$

bestimmen. Dann

$$k_2 : \|\boldsymbol{x}_t - \boldsymbol{v}_{k_1 k_2}\| \leq \|\boldsymbol{x}_t - \boldsymbol{v}_{k_1 k}\|, \quad k = 1, 2, \ldots, K$$

und so weiter, bis schließlich

$$k_m : \left\|\boldsymbol{x}_t - \boldsymbol{v}_{k_1 k_2 \ldots k_{m-1} k_m}\right\| \leq \left\|\boldsymbol{x}_t - \boldsymbol{v}_{k_1 k_2 \ldots k_{m-1} k}\right\|,$$

für $k = 1, 2, \ldots, K$. Dazu sind nur noch $K \times m$ Abstandsberechnungen nötig! Schon für $K = 2$ und $m = 10$ ist die relative Aufwandsersparnis gegenüber dem FSVQ erheblich:

$$\frac{K \times m}{K^m} = \frac{20}{1024} \approx 0,02 \ .$$

Bei diesem Verfahren müssen wir aber auch alle anderen Codebücher der vorherigen Stufen speichern. Das sind insgesamt

$$K + K^2 + \ldots + K^m = \frac{K^{m+1} - K}{K - 1}$$

Codebuchvektoren.

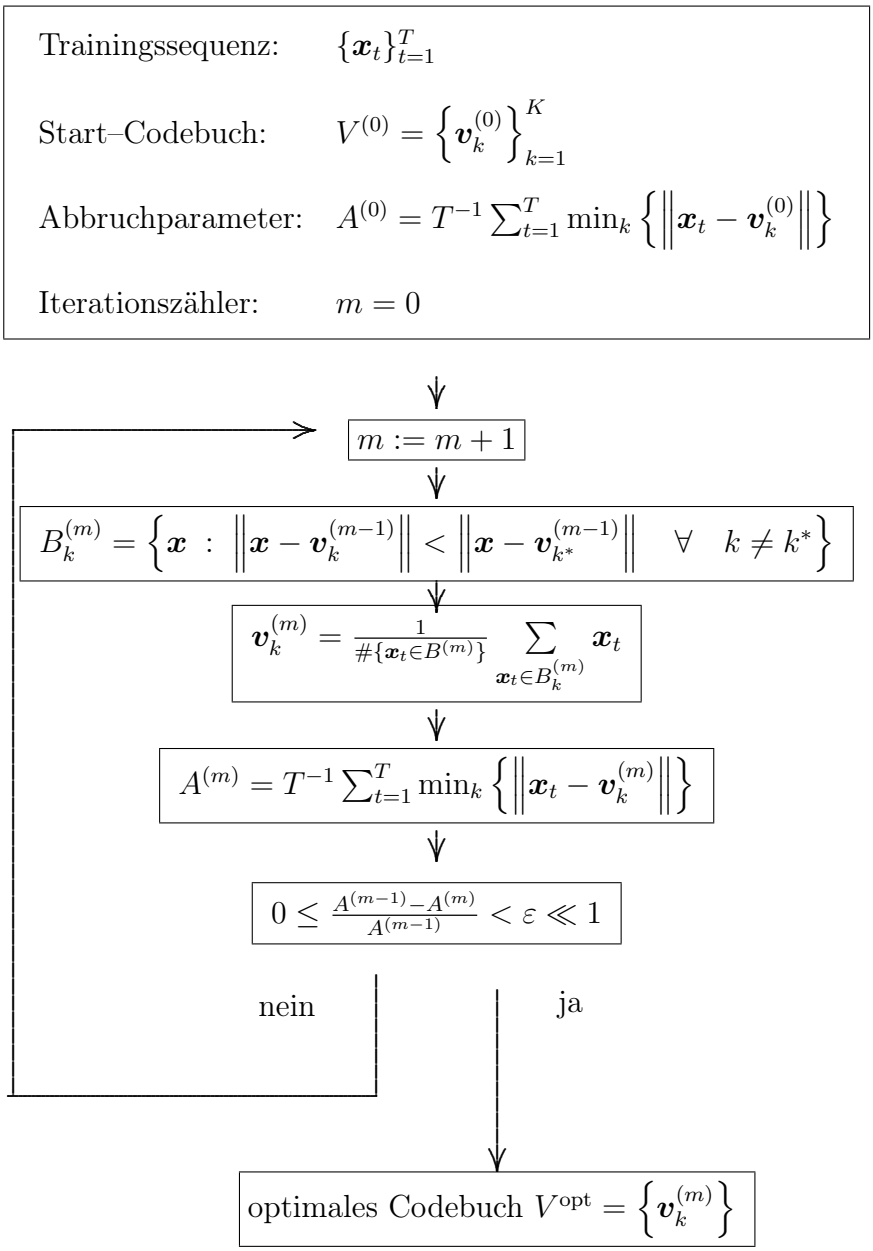

Abb. 5.2: LBG–Algorithmus

Kapitel 6

Verschlüsselung von Information

6.1 Einleitung

Um bei der Informationsübertragung vom Sender zum Empfänger das Abhören durch Dritte zu erschweren oder gar unmöglich zu machen, muss die Nachricht verschlüsselt werden. Dazu gibt es in der Geschichte der Signalübertragung viele verschiedene Vorschläge. Besonders in der englischsprachigen Literatur ist es heute üblich, den Sender, Empfänger und Lauscher mit "Alice", "Bob"und "Eve"zu benennen. Ziel ist es, eine Verschlüsselung zu finden, so dass Eve ohne Knntnis des Schlüssels keine praktikable Möglichkeit hat, aus der verschlüsselten Nachricht den Klartext zu bekommen.

Jede Nachricht kann letztlich durch eine Folge von Binärzeichen dargestellt werden, den *Bitstrom*. Bei einer perfekten Quellcodierung (s. Kapitel 3, S. 55 ff.) ist die Folge von Binärzeichen vollständig statistisch unabhängig, und die Zeichen 0 und 1 treten gleich wahrscheinlich auf. Mit der Einführung fehlererkennender und korrigierbarer Codes (s. Abschnitt 3.3, S. 70 ff.) wird der Bitstrom wieder redundant, das heißt, es treten Abhängigkeiten auf, und mit der Kanalcodierung (s. Kapitel 4, S. 79 ff.) werden die Binärzeichen im Allgemeinen ungleich wahrscheinlich. Abhängigkeiten im Bitstrom treten im Allgemeinen auch deshalb auf, weil die Quellcodierung in der Praxis zumeist nicht perfekt ist. Die Lauscherin Eve könnte solche Abhängigkeiten nutzen, um die Botschaft zu entschlüsseln. Deshalb müssen sich Alice und Bob eine Verschlüsselungssystem vereinbaren.

Es stellt sich somit die Frage, wie aus einem gegebenen Bitstrom durch einfache Algorithmen eine perfekte Verschlüsselung vorgenommen werden kann.

6.2 Perfekte Verschlüsselung

Gegeben seien die zu verschlüsselnde Botschaft

$$\boldsymbol{x} \equiv x_1 x_2 \dots x_t \dots x_T \ , \text{ mit } a_t \in \{0,1\}$$

und eine Folge

$$\boldsymbol{k} \equiv k_1 k_2 \dots k_t \dots k_T \ , \text{ mit } k_t \in \{0,1\} \ ,$$

der sogenannte *Schlüssel* — eine binäre Zufallsfolge, die vollkommen statistisch unabhängig ist (i.i.d.), mit der Wahrscheinlichkeitsverteilung $(p_{\boldsymbol{k},0}, p_{\boldsymbol{k},1})$

Wir verschlüsseln nun $\boldsymbol{x}$ mit Hilfe von $\boldsymbol{k}$, so dass wir eine Folge

$$\boldsymbol{y} \equiv y_1 y_2 \dots y_t \dots y_T \ , \text{ mit } y_t = x_t \oplus k_t \ .$$

erhalten. Darin bezeichnet $\oplus$ die logische Antivalenz,

$$x \oplus k = \begin{cases} 0 & : \quad x = k \\ 1 & : \quad x \neq k \end{cases}$$

Über den Nachrichtenkanal wird nun $\boldsymbol{y}$ anstelle von $\boldsymbol{x}$ übertragen. Bob empfängt diese und nutzt denselben Schlüssel $\boldsymbol{k}$, den Eve zur Verschlüsselung verwandte, um die Klarschrift $\boldsymbol{x}$ zu bekommen, indem er

$$\boldsymbol{x} \equiv x_1 x_2 \dots x_t \dots x_T \ , \text{ mit } x_t = y_t \oplus k_t$$

bildet.

Wir betrachten ein Beispiel:

$$\begin{aligned} \text{Alice:} \quad \boldsymbol{x} &= 000011000011000011000011 \\ \boldsymbol{k} &= 011010100011010100101010 \\ \boldsymbol{y} &= 011001100000010111101001 \\ \text{Bob:} \quad \boldsymbol{x} &= 000011000011000011000011 \end{aligned}$$

Es stellt sich nun die Frage, ob die Lauscherin Eve, die $\boldsymbol{y}$ abhört, ohne den Schlüssel $\boldsymbol{k}$ auf die Klarschrift $\boldsymbol{x}$ schließen könnte. Dies muss jedoch verneint werden, wie man sich wie folgt überlegt:

Seien $(p_{\boldsymbol{x},0}, p_{\boldsymbol{x},1})$ und $(p_{\boldsymbol{y},0}, p_{\boldsymbol{y},1})$ die Wahrscheinlichkeitsverteilungen von 0 und 1 in $\boldsymbol{x}$ bzw. $\boldsymbol{y}$. Dann unterscheiden wir folgende Fälle:

x_t	k_t	$y_t = x_t \oplus k_t$	$s_{\boldsymbol{xy},mn}$
0	0	0	$p_{\boldsymbol{x},0} \cdot p_{\boldsymbol{k},0} = s_{\boldsymbol{xy},00}$
0	1	1	$p_{\boldsymbol{x},0} \cdot p_{\boldsymbol{k},1} = s_{\boldsymbol{xy},01}$
1	0	1	$p_{\boldsymbol{x},1} \cdot p_{\boldsymbol{k},0} = s_{\boldsymbol{xy},11}$
1	1	0	$p_{\boldsymbol{x},1} \cdot p_{\boldsymbol{k},1} = s_{\boldsymbol{xy},10}$

Darin bezeichnet $s_{\boldsymbol{xy},mn}$ die Verbundwahrscheinlichkeit, dass Alice das Symbol $m \in \{0,1\}$ sendet und Eve das Symbol $n \in \{0,1\}$ beobachtet. Die Wahrscheinlichkeiten, mit denen Eve im Bitstrom $\boldsymbol{y}$ das Zeichen $y_t = n \in \{0,1\}$ beobachtet, sind

$$p_{\boldsymbol{y},0} = s_{\boldsymbol{xy},00} + s_{\boldsymbol{xy},10} = p_{\boldsymbol{x},0} \cdot p_{\boldsymbol{k},0} + p_{\boldsymbol{x},1} \cdot p_{\boldsymbol{k},1} \tag{6.1}$$

$$p_{\boldsymbol{y},1} = s_{\boldsymbol{xy},01} + s_{\boldsymbol{xy},11} = p_{\boldsymbol{x},0} \cdot p_{\boldsymbol{k},1} + p_{\boldsymbol{x},1} \cdot p_{\boldsymbol{k},0} \ . \tag{6.2}$$

Die bedingten Wahrscheinlichkeiten, dass Eve n beobachtet, unter der Bedingung, dass Alice $m = n$ gesandt hat, ist dann

$$\begin{aligned} p_{\boldsymbol{y},0|0} &= \frac{s_{\boldsymbol{xy},00}}{p_{\boldsymbol{x},0}} = \frac{p_{\boldsymbol{x},0} \cdot p_{\boldsymbol{k},0}}{p_{\boldsymbol{x},0}} = p_{\boldsymbol{k},0} \\ p_{\boldsymbol{y},1|1} &= \frac{s_{\boldsymbol{xy},11}}{p_{\boldsymbol{x},1}} = \frac{p_{\boldsymbol{x},1} \cdot p_{\boldsymbol{k},0}}{p_{\boldsymbol{x},1}} = p_{\boldsymbol{k},0} \end{aligned}$$

Wegen der Normierungen $p_{\boldsymbol{y},0|0} + p_{\boldsymbol{y},1|0} = 1$, $p_{\boldsymbol{y},0|1} + p_{\boldsymbol{y},1|1} = 1$ sowie $p_{\boldsymbol{k},0} + p_{\boldsymbol{k},1} = 1$ folgen

$$\begin{aligned} p_{\boldsymbol{y},1|0} &= 1 - p_{\boldsymbol{y},0|0} = 1 - p_{\boldsymbol{k},0} = p_{\boldsymbol{k},1} \\ p_{\boldsymbol{y},0|1} &= 1 - p_{\boldsymbol{y},1|1} = 1 - p_{\boldsymbol{k},0} = p_{\boldsymbol{k},1} \end{aligned}$$

Wir betrachten nun den Spezialfall $p_{\boldsymbol{k},0} = p_{\boldsymbol{k},1} = 1/2$. Damit lassen sich (6.1) und (6.2) unter Beachtung der Normierung $p_{\boldsymbol{x},0} + p_{\boldsymbol{x},1} = 1$ wie folgt weiterschreiben,

$$\begin{aligned} p_{\boldsymbol{y},0} &= \frac{1}{2} \cdot p_{\boldsymbol{x},0} + \frac{1}{2} \cdot p_{\boldsymbol{x},1} = \frac{1}{2} \\ p_{\boldsymbol{y},1} &= \frac{1}{2} \cdot p_{\boldsymbol{x},0} + \frac{1}{2} \cdot p_{\boldsymbol{x},1} = \frac{1}{2} \ . \end{aligned}$$

Darüber hinaus erhalten wir nun

$$s_{\boldsymbol{xy},mn} = p_{\boldsymbol{x},m} \cdot p_{\boldsymbol{y},n} \ , \quad \text{für alle } m,n \in \{0,1\} \ .$$

Dies bedeutet, die Symbole x_t und y_t sind *statistisch unabhängig.*[1] Man überlegt sich auch leicht, dass für $p_{\boldsymbol{k},0} = p_{\boldsymbol{k},1} = 1/2$ und wegen der vorausgesetzten vollständigen Unabhängigkeit des

[1] Beachte, dass die statische Unabhängigkeit nur für den Spezialfall $p_{\boldsymbol{k},0} = p_{\boldsymbol{k},1} = 1/2$ gezeigt wurde. Würde zum Beispiel $p_{\boldsymbol{k},0} = 0$, $p_{\boldsymbol{k},1} = 1$ gelten, dann wäre y_t nur das Inverse von x_t, was Eve die Botschaft leicht entschlüsseln ließe.

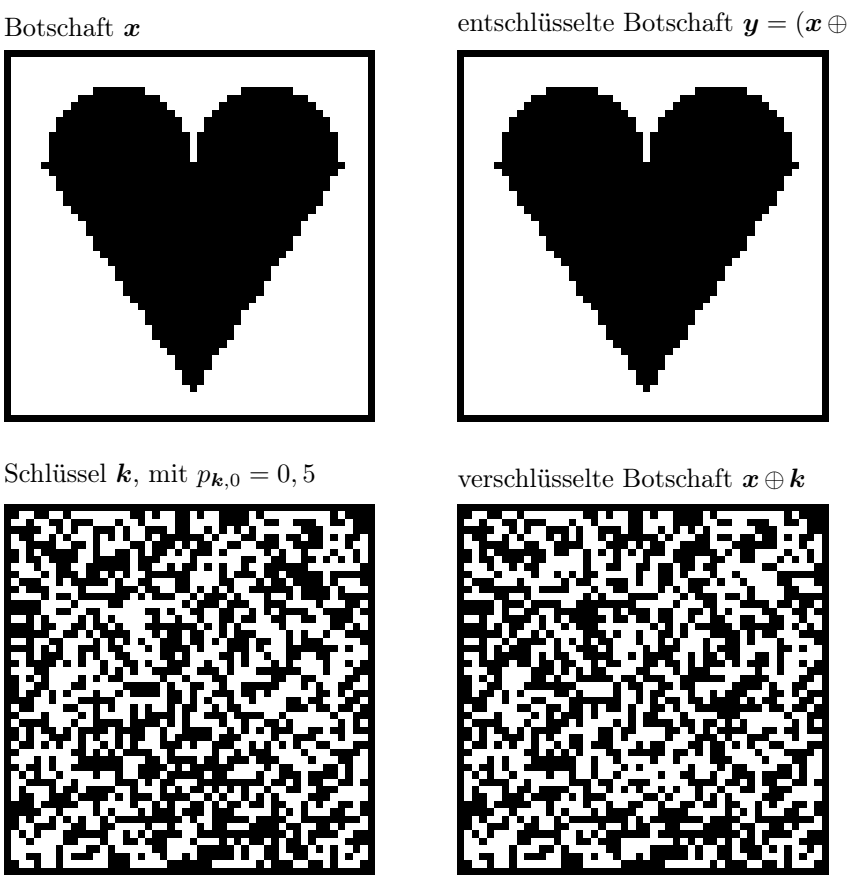

Abb. 6.1: Sichere Verschlüsselung eines Bildes: Die verschlüsselte Botschaft ist gleich dem Schlüssel, an den Stellen, wo das Originalbild schwarz ist, und invers dem Schlüssel, wo das Originalbild weiß ist

Bitstroms $\boldsymbol{k}$ auch $\boldsymbol{y}$ vollständig statistisch unabhängig ist, selbst dann, wenn $\boldsymbol{x}$ irgendwelche Abhängigkeiten aufweist, zum Beispiel konstant ist, $\boldsymbol{x} = 000000\ldots$

Die Abildungen 6.1 und 6.2 zeigen die Verschüsselung eines Schwarz–Weiß–Bildes. Wird als Schlüssel ein zufälliges Pixelmuster mit gleichen Wahrscheinlichkeiten für Weiß und Schwarz ($p_{\boldsymbol{k},0} = p_{\boldsymbol{k},1} = 0,5$) gewählt, dann ist die Verschlüsselung absolut sicher. Weichen die Wahrscheinlichkeiten jedoch von der Gleichverteilung ab, so kann Eve bei der Betrachtung des verschlüsselten Bildes Allice Botschaft erraten, umso eher, je stärker sich $p_{\boldsymbol{k},0}$ und $p_{\boldsymbol{k},1}$ unterscheiden.

6.3 Schlüssel nur einmal verwenden !

Für die oben beschriebene perfekte Verschlüsselung muss Bob zur Entschlüsselung den von Alice verwandten Schlüssel $\boldsymbol{k}$ kennen. Dieser müsste also zuvor über den Nachrichtenkanal gesandt

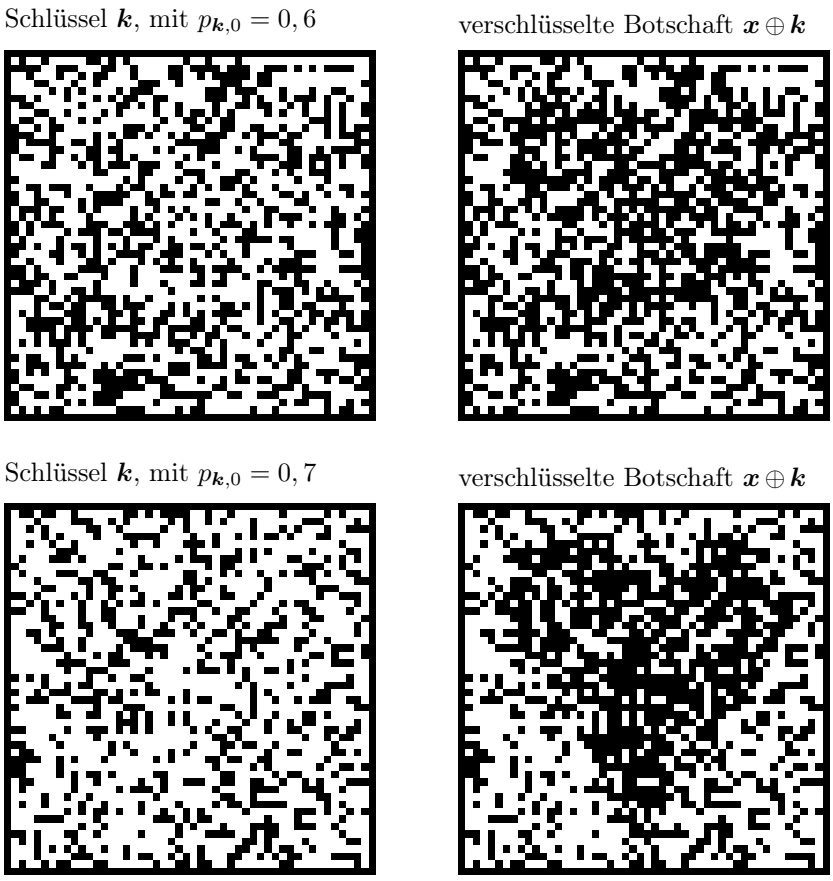

Abb. 6.2: Unsichere Verschlüsselung des Originalbildes aus Abb. 6.1 mit Schlüsseln ungleicher Wahrscheinlichkeiten. Die verschlüsselte Botschaft lässt das Motiv der Botschaft (Herz) erraten

werden, wobei Eve ihn abhören könnte. Alice und Bob könnten einen sicheren Weg zum Schlüsselaustausch wählen, so dass der Schlüssel nicht über einen öffentlichen Kanal gesandt werden müsste. Dann hätte Eve keine Gelegenheit, den Schlüssel abzulauschen. Das Problem hierbei ist jedoch, dass der sichere Schlüsselaustausch etwa durch Botschafter oder eine persönliche Absprache recht aufwändig ist. Hinzu kommt, dass ein Schlüssel nur einmal verwendet werden darf, denn es gilt Folgendes: Sind

$$\boldsymbol{x}_i = x_{i,1}x_{i,2}\ldots x_{i,T}$$

und

$$\boldsymbol{x}_j = x_{j,1}x_{j,2}\ldots x_{j,T}$$

zwei Botschaften, die mit dem selben Schlüssel $\boldsymbol{k}$ verschlüsselt werden, so dass die verschlüsselten Bitströme $\boldsymbol{y}_i$ und $\boldsymbol{y}_j$ über den Kanal gesandt werden, dann kann Eve

$$y_{i,t} \oplus y_{j,t} = (x_{i,t} \oplus k_t) \oplus (x_{j,t} \oplus k_t) = x_{i,t} \oplus x_{j,t}$$

bilden, was aus folgender Tabelle ersichtlich ist:

$x_{i,t}$	$x_{j,t}$	k_t	$y_{i,t}$	$y_{j,t}$	$y_{i,t} \oplus y_{j,t}$	$x_{i,t} \oplus x_{j,t}$
0	0	0	0	0	0	0
0	0	1	1	1	0	0
0	1	0	0	1	1	1
0	1	1	1	0	1	1
1	0	0	1	0	1	1
1	0	1	0	1	1	1
1	1	0	1	1	0	0
1	1	1	0	0	0	0

Aus $\boldsymbol{x}_i \oplus \boldsymbol{x}_j$ kann dann Eve ohne Kenntnis des Schlüssels $\boldsymbol{k}$ etwas über die Klarschriften $\boldsymbol{x}_i$ und $\boldsymbol{x}_j$ erraten.

Würden beispielsweise Alice und Bob mit dem selben Schlüssel dieselbe Nachricht, also $\boldsymbol{x}_i = \boldsymbol{x}_j$ zweimal austauschen und Alice diese abhören, dann würde Eve $x_{i,t} \oplus x_{j,t} = 0$ für alle t erhalten. Damit wüsste Eve allerdings nur, dass die selbe Botschaft zweimal ausgetauscht wurde.

Kapitel 7

Verschiedenes

7.1 Spezielle Codes

Gray–Code

Der *Gray–Code* ist ein spezieller Binärcode. Im Unterschied zum üblichen Dualcode ist bei ihm der Hammingabstand benachbarter Codeworte genau eins. Für die Zahlen 0 bis 7 illustriert dies die folgende Tabelle:

Zahl Z	Dualcode $D_2D_1D_0$	Gray–Code $G_2G_1G_0$
0	000	000
1	001	001
2	010	011
3	011	010
4	100	110
5	101	111
6	110	101
7	111	100

Interpretiert man die Buchstaben 0 und 1 im l–stelligen Dual–Codewort $D_{l-1}, \ldots, D_1, D_0$ als Dezimalzahl, dann gelangt man wie folgt zur Dezimalzahl Z:

$$Z = D_{l-1}2^{l-1} + \ldots + D_1 2^1 + D_0 2^0 \ .$$

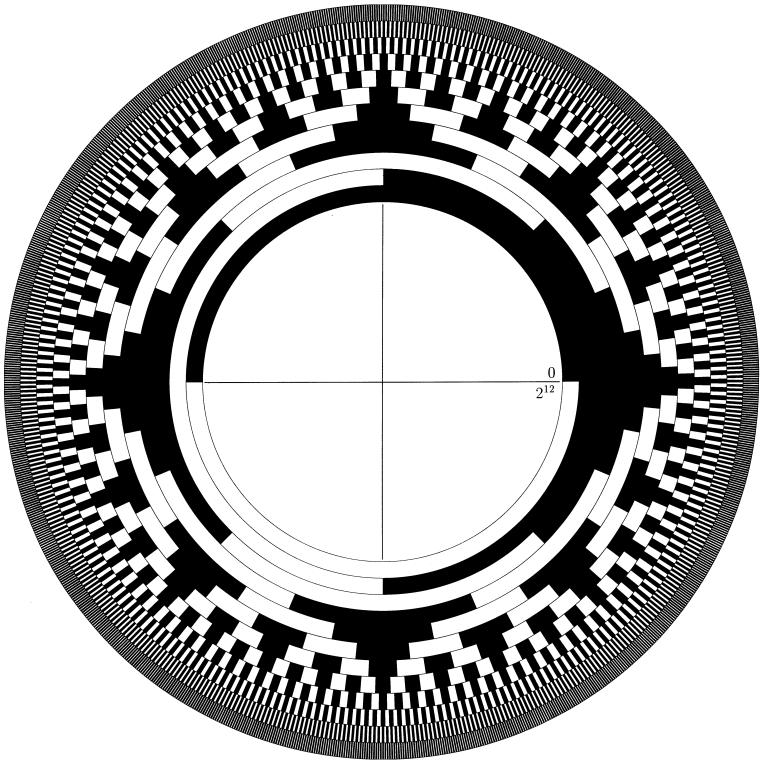

Abb. 0.1: Gray–Winkel–Code

Abb. 7.1: 12–Bit–Winkelcodescheibe im Gray–Code

Vom Gray– zum Dualcode gelangt man rekursiv über die logische Funktion

$$D_i = \begin{cases} D_{i+1} \oplus G_i & : \quad i = l-2, \ldots, 1, 0 \\ G_{l-1} & : \quad i = l-1 \ . \end{cases}$$

Darin bezeichnet $\oplus$ die logische Operation Antivalenz,

$$x \oplus y = \begin{cases} 0 & : \quad x = y \\ 1 & : \quad x \neq y \ . \end{cases}$$

Gray–Codes werden z. B. bei der Codierung von Auslenkwinkeln mittels einer *Winkelcodescheibe* benutzt, wobei der Auslenkwinkel mit einer radial angeordneten Lichtschrankenleiste abgegriffen wird. Durch die Verwendung des Graycodes wird die Justierung der Leiste weniger kritisch als bei der Verwendung eines Dualcodes, denn beim Übergang von einem zum benachbarten Winkelsegment sollten alle Lichtschranken, die hier schalten müssen, exakt gleichzeitig schalten. Beim Gray–Code ist dies immer nur eine Stelle, Dual–Code aber bis zu l viele. Da in der Praxis niemals alle Lichtschranken gleichzeitig schalten, können beim Dualcode unerwünschte Spikes auftreten, nämlich dann, wenn die Folgeschaltungen (Digital–Analog–Wandler,

Verstärker) eine hinreichend große Bandbreite haben, so daß die hochfrequenten Spikes auch übertragen bzw. verstärkt werden.

Literaturverzeichnis

[1] Bongard, M. M. (1963): Über den Begriff der nützlichen Information, in: Probleme der Kybernetik 6 S. 91–130, (Akademie–Verlag, Berlin, 1966) (1963 in russ.)

[2] Campbell, L. L. (1965): A coding theorem and Rényi's entropy, Information and Control **8**, 423–429.

[3] Campbell, L. L. (1966): Definition of entropy by means of a coding problem, Z. Wahrscheinlichkeitstheorie verw. Geb. **6**, 113–118.

[4] Chintschin, A.J. (1953): Der Begriff der Entropie in der Wahrscheinlichkeitsrechnung, [in: Arbeiten zur Informationstheorie I, 2. berichtigte Auflage, (VEB Deutscher Verlag der Wissenschaften, Berlin, 1961) S. 7–29]; [Originalarbeit (russ.) in Uspechi mat. nauk **11** (1953) 3–20]

[5] Faddejew, D. K. (1956): Zum Begriff der Entropie eines endlichen Wahrscheinlichkeitsschemas [in: Arbeiten zur Informationstheorie I, 2. berichtigte Auflage, (VEB Deutscher Verlag der Wissenschaften, Berlin, 1961) S. 86–90] [Originalarbeit (russ.) in Uspechi mat. nauk **11** (1956) 227–231]

[6] Hartley, R. V. (1928): Transmission of Information, Bell Syst. Techn. J. **7**, 535–563.

[7] Jaglom, A. M., und I. M. Jaglom (1984): *Wahrscheinlichkeit und Information*, (VEB Deutscher Verlag der Wissenschaften, Berlin)

[8] Küpfmüller, K. (1954): Die Entropie der deutschen Sprache, Fernmeldetechnische Zeitschrift **7**, 255–272.

[9] Kullback, S. (1959): *Information Theory and Statistics* (John Wiley, New York)

[10] Linde, Y., A. Buzo, and R.M. Gray (1980): An Algorithm for Vector Quantizer Design, IEEE Trans. Com **28**, 84-95.

[11] Mainzer (1997): *Thinking in the Complex*, (Springer, Berlin [u.a.])

[12] Mildenberger, O. (1990): *Informationstheorie und Codierung*, (Vieweg, Braunschweig, Wiesbaden)

[13] Pötschke, D., und F. Sobik (1980): *Mathematische Informationstheorie*, (Akademie–Verlag, Berlin)

[14] Rényi, A. (1961): On measures of entropy and information, Proc. Fourth Berkley Symp. Math. Statist. Probab. **1**, 547–561, (Univ. of Calif. Press, Berkeley)

[15] Rényi, A. (1977): *Wahrscheinlichkeitsrechnung mit einem Anhang über Informationstheorie*, (VEB Deutscher Verlag der Wissenschaften Berlin)

[16] Schützenberger, M. P. (1954): Contribution aux applications statistiques de la théorie de l'information, Publ. Inst. Statist., Univ. Paris **3**, 3.

[17] Shannon, C. E. (1948): *A mathematical theory of communication*, Bell System Technical Journal **27**, 379, 623.

[18] Shannon, C. E., and W. Weaver (1949): The mathematical theory of communication, Univ. Illinois Press, Illinois [dt. Übers.: *Mathematische Grundlagen der Informationstheorie.* (R. Oldenbourg Verlag, München, Wien, 1976)]

[19] Völz, H. (1982): *Information I, Studie zur Vielfalt und Einheit der Information*, (Akademie–Verlag, Berlin)

[20] Wheeler, J. A. (1989): *Information, Physics, Quantum: The Search for Links*, Proc. 3rd International Symposium on the Foundation of Quantum Mechanics, p. 354–368

Index