

Kai Bruchlos

Erläuterungen zum Quellencodierungssatz
von Shannon

THM-Hochschulschriften Band 31

Kai Bruchlos

Erläuterungen zum Quellencodierungssatz
von Shannon

THM-Hochschulschriften Band 31

THM-Hochschulschriften Band 31

© 2024 Kai Bruchlos

Technische Hochschule Mittelhessen

Fachbereich Mathematik, Naturwissenschaften, Datenverarbeitung

Herausgeber der THM-Hochschulschriften:

Der Präsident der Technischen Hochschule Mittelhessen

Alle Rechte vorbehalten, Nachdruck, auch auszugsweise, nur mit schriftlicher Genehmigung und Quellenangabe.

Die Hochschulschriften sind online abrufbar:

www.thm.de/bibliothek/thm-hochschulschriften

ISSN (Print) 2568-0846

ISSN (Online) 2568-3020

Vorwort

Ziel dieser Abhandlung ist es, den Begriff der Entropie der Informationstheorie inhaltlich zu begründen und zu erläutern und den Zusammenhang zwischen verschiedene Aussagen zum Codieraufwand und dem Quellencodierungssatz aufzuzeigen.

Nachdem im ersten Kapitel das Ziel der Informationstheorie dargestellt worden ist, wird sich im zweiten Kapitel „Technik“ der Codierung zugewandt und die Grundlagenbegriffe für die mittlere Codewortlänge — auch als empirische Entropie bezeichnet — und die (theoretische) Entropie definiert. Im dritten Kapitel „Bewertung“ wird zunächst der Bewertungsmaßstab für den Informationsgehalt festgelegt, um dann für den Erwartungswert des Informationsgehaltes, der Entropie, grundlegende Eigenschaften anzuführen. Im abschließenden vierten Kapitel wird dann der Codieraufwand mit Hilfe des Quellencodierungssatzes abgeschätzt.

Für Begriffe der Stochastik verwende ich die Notation von Georgii 2015. – Es werden nur die Aussagen bewiesen, zu denen mir aus der Literatur kein Beweis bekannt ist.

Bedanken möchte ich mich bei meinem Kollegen Prof. Dr. Christian Schulze für das Lesen des Manuskriptes, die Hinweise und Anmerkungen.

Inhaltsverzeichnis

1	Informationstheorie	5
2	Technik	6
2.1	Codierung	6
2.2	Mittlere Codewortlänge	8
3	Bewertung	12
3.1	Die Entropie einer Informationsquelle	12
3.2	Eigenschaften der Entropie	15
4	Codieraufwand	18
	Literaturverzeichnis	20

1 Informationstheorie

Die **Informationstheorie** beschäftigt sich mit der Frage, *wie Nachrichten so codiert werden können, dass möglichst wenige ihrer Informationen verloren gehen. Dabei soll nicht Fehlertoleranz, sondern Effizienz ein Qualitätsziel sein, d.h. die Codewörter sollen möglichst kurz sein.*¹ Im Gegensatz dazu hat die **Codierungstheorie** zum Ziel, Daten so zu codieren, *dass Fehler bei deren Verarbeitung erkannt und möglicherweise sogar korrigiert werden können.*² Die Codierungstheorie verwendet unter anderem Methoden der Algebra und Zahlentheorie. Die Informationstheorie ist ein Teilgebiet der Stochastik.

Neben dem **technischen Aspekt** des Qualitätszieles der Effizienz, möglichst kurze Codewörter zu verwenden, gibt es noch den **bewertenden Aspekt**. Dabei geht es um die Frage, was unter „Information“ verstanden werden soll. Nicht die Bedeutung einer Nachricht (Semantik) soll wichtig sein, sondern die Wahrscheinlichkeitsverteilung der Zeichen.³

*Es ist daher wichtig, darauf hinzuweisen, daß die Theorie nichts über die Bedeutung, den Inhalt oder den Wert einer Mitteilung (einer „Information“) aussagt. Sie [die Informationstheorie] gibt nur Aufschluß über Zusammenhänge, die durch die statistische Struktur bestimmt sind.*⁴

Ziel des bewertenden Aspektes der Informationstheorie ist die Quantifizierung des „Informationsgehaltes“ aus Sicht der Nachrichtentechnik: Welche Zeichen sind wichtig?

Die Informationstheorie ist von Claude Shannon 1948 mit dem Aufsatz *A mathematical theory of communication* begründet worden.

¹Witt 2007, S. 273. Vgl. Shannon 1948, S. 384 f.

²Witt 2007, S. 243.

³Vgl. Shannon 1948, S. 379.

⁴Topsøe 1974, S. 5.

2 Technik

In diesem Kapitel beschäftigen wir uns mit dem technischen Aspekt der Informationstheorie: Die Codeworte sollen möglichst kurz sein.

2.1 Codierung

Zunächst benötigen wir einige Grundbegriffe:

Definition 2.1.1:¹ (i) Ein **Alphabet** oder **Zeichenvorrat** A ist eine nicht leere, endliche Menge. Die Elemente von A heißen **Symbole** oder **Zeichen**.

(ii) Ein **Wort (der Länge $k \in \mathbb{N}$)** über einem Alphabet A ist ein Tupel (a_1, \dots, a_k) von Elementen aus A , also $(a_1, \dots, a_k) \in A^k$.

(iii) Die **Menge aller Wörter über dem Alphabet A** (unterschiedlicher, aber endlicher Länge) ist $A^* := \bigcup_{i=0}^N A^i$, $N \in \mathbb{N}$.² Dabei bedeutet $A^0 := \emptyset$ das „leere Wort“.

Definition 2.1.2:³ Seien A und B zwei Alphabete. Eine **Codierung** c (über den Alphabeten A und B) ist eine injektive Abbildung

$$c : A \rightarrow B^* .$$

Das Bild von A unter c , also die Menge $c(A) \subset B^*$ heißt **Code**, $c(a)$ ist ein **Codewort**, A heißt **Senderalphabet** oder **Quelle(nalphabet)**, B **Codealphabet** oder **Kanalalphabet**. Gilt $|B| = 2$, so liegt ein **binärer Code** vor, bei $|B| = 3$ ein **ternärer Code**.

Bemerkung 2.1.3:⁴ (i) Gilt für alle Zeichen einer Quelle, dass die Sendung eines Zeichen nicht von dem vorher gesendeten Zeichen abhängt, dann

¹Vgl. Schulz 2003, S. 1, 6 f.

²Schulz 2003 definiert auf Seite 7 die Menge aller Wörter als $A^* := \bigcup_{i=0}^{\infty} A^i$, wobei A^{∞} aus Folgen aus A besteht. Da die Wörter eine endliche Länge haben sollen, ist diese Definition nicht zutreffend. Die hier verwendete Definition mit N findet sich auf Seite 32 Schulz 2003. Die Situation beschreibt am besten $A^* := \bigcup_{i=0}^{\infty} A^i \setminus A^{\infty}$.

³Vgl. Schulz 2003, S. 32; Witt 2007, S. 243.

⁴Vgl. Schulz 2003, S. 43; Witt 2007, S. 274

handelt es sich um eine **Quelle ohne Gedächtnis**. Es gibt keinen kausalen Zusammenhang zwischen den gesendeten Zeichen. Formale Darstellung für eine Quelle ohne Gedächtnis mit n Zeichen: Das Senden eines Zeichens a_j , $2 \leq j \leq n$ ist unabhängig von den gesendeten Zeichen a_i , $1 \leq i < j$. – In einer gesprochenen Sprache ist dies meist anders. Beispielsweise folgt in der deutschen Sprache auf zwei Vokale fast immer ein Konsonant. Dementsprechend sprechen wir von einer **Quelle mit Gedächtnis**, wenn es mindestens ein Zeichen gibt, dessen Sendung von dem vorher gesendeten Zeichen abhängt.

(ii) Eine Codierung c heißt auch **Einzelzeichen-Codierung**. Diese ist sinnvoll bei Quellen ohne Gedächtnis. Bei Quelle mit Gedächtnis bietet es sich an, sich bedingende Zeichen zusammenzufassen, also Paare oder Tripel, allgemeiner Wörter über dem Quellenalphabet zu codieren: $c^* : A^* \rightarrow B^*$.⁵ Es liegt dann eine **Wort-Codierung** vor.

Im Folgenden betrachten wir nur spezielle Wort-Codierungen:

Definition 2.1.4:⁶ Sei eine Codierung $c : A \rightarrow B^*$ gegeben. Die Wort-Codierung $c^* : A^* \rightarrow B^*$ heißt **buchstabenweise (Wort-)Codierung**, wenn

$$\begin{aligned} c^*(a_1 a_2 \dots a_k) &:= c^*((a_1, a_2, \dots, a_k)) &:= (c(a_1), c(a_2), \dots, c(a_k)) \\ &=: c(a_1) c(a_2) \dots c(a_k) \end{aligned}$$

ist. Eine **Nachricht** ist ein Wort $a_1 a_2 \dots a_k := (a_1, \dots, a_k)$ der Quelle A , die dazugehörige **kodierte Nachricht** das Wort $c^*(a_1 a_2 \dots a_k)$ des Codealphabets B .

Die Codierung c muss injektiv sein, da ansonsten die **Dekodierung** — also die eindeutige Zuordnung des Codewortes $c(a)$ dem Zeichen a — nicht möglich ist. Mathematisch folgt aus der Injektivität von c , dass $c^{-1} : c(A) \rightarrow A$ eine Funktion ist.⁷

Im Allgemeinen folgt nun aber aus der Injektivität von c nicht die Injektivität von c^* .⁸ Dies bedeutet, dass eine kodierte Nachricht nicht **dekodierbar** sein muss. Dieses Problem kann auf unterschiedliche Weise behoben werden. Beim **Morse-Code** gibt es ein Trennzeichen, die Pause, zwischen den Codeworten.⁹ Haben alle Codeworte $c(a)$ dieselbe Länge — es liegt ein sogenannter **Block-Code** vor —, dann ist c^* injektiv.¹⁰ Ein weiteres Beispiel für eine injektive buchstabenweise Codierung c^* basiert auf dem Präfix-Code:

⁵Vgl. Witt 2007, S. 243.

⁶Vgl. Schulz 2003, S. 29, 32; Witt 2007, S. 243, 267.

⁷Vgl. Heuser Teil 1 2003, S. 106.

⁸Schulz 2003, S. 32; Witt 2007, S. 267

⁹Vgl. Krenzel 2005, S. 115.

¹⁰Schulz 2003, S. 32.

Definition 2.1.5:¹¹ Sei eine Codierung $c : A \rightarrow B^*$ gegeben. c heißt **Präfix-Code**, wenn kein Code-Wort aus $c(A)$ Anfangsstück (Präfix¹²) eines anderen Codewortes aus $c(A)$ ist.

Satz 2.1.6:¹³ Jede buchstabenweise Wort-Codierung c^* eines Präfix-Codes ist injektiv. In diesem Sinne ist jeder Präfix-Code dekodierbar.

Proposition 2.1.7:¹⁴ Jeder Block-Code ist ein Präfix-Code.

Wir betrachten im Folgenden nur Präfix-Codes. Dies ist keine Einschränkung:¹⁵

Satz 2.1.8:¹⁶ Ist für eine Codierung c die Wort-Codierung c^* injektiv, dann gibt es einen Präfix-Code mit genau denselben Wortlängen wie c .

2.2 Mittlere Codewortlänge

Nachdem wir uns mit Codierungen c beschäftigt und begründet haben, warum wir nur Präfix-Codes betrachten, kommen wir jetzt zum Qualitätsziel, dass Codeworte möglichst kurz sein sollen (Datenkompression). Dafür gibt es verschiedene Methoden.¹⁷ Shannon schlägt mit Blick auf die Bewertung der Information — den Informationsgehalt — vor, die statistischen Eigenschaften der Quelle heranzuziehen:

*The main point at issue is the effect of statistical knowledge about the source in reducing the required capacity of the channel, by the use of proper encoding of the information.*¹⁸

Wie sich noch zeigen wird (Satz 2.2.10), liefert die **statistische Codierung** das beste Ergebnis:

*Häufig zu erwartende Symbole werden durch kurze Codewörter, seltenere Symbole durch längere Codewörter beschrieben.*¹⁹

¹¹Vgl. Krengel 2005, S. 114; Schulz 2003, S. 37; Witt 2007, S. 268.

¹²In der Linguistik ist ein Präfix eine Worterweiterung, die dem Wortstamm vorangestellt wird. Beim Wort „ablegen“ ist „ab“ das Präfix und „legen“ der Wortstamm.

¹³Witt 2007, S. 268, Satz 19.1.

¹⁴Witt 2007, S. 268, Folgerung 19.1.

¹⁵Vgl. Krengel 2005, S. 115.

¹⁶Jacobs und Jungnickel 2004, S. 133, Satz 1.7. Ein sofort entzifferbarer A-B-Code ist ein Präfix-Code: S. 130, Definition 1.2. Eindeutig entzifferbarer A-B-Code c bedeutet, dass c^* injektiv ist: S. 130, Definition 1.4.

¹⁷Vgl. Schulz 2003, S. 39 f.

¹⁸Shannon 1948, S. 384 f. In Deutsch: *Ein wesentlicher Ausgangspunkt ist der Effekt, daß durch die Kenntnis der statistischen Eigenschaft der Quelle die benötigte Kanalkapazität reduziert werden kann, indem man die Nachricht auf die günstigste Weise codiert.* Shannon und Weaver 1949, S. 49.

¹⁹Schulz 2003, S. 40.

Für die statistische Codierung führen wir jetzt Zeichen und Wahrscheinlichkeiten zusammen:

Definition 2.2.1:²⁰ Sei $A := \{a_1, \dots, a_n\}$ ein Alphabet und $(A, \mathcal{P}(A), P)$ ein Wahrscheinlichkeitsraum. Die Matrix

$$Q := \begin{pmatrix} a_1 & \dots & a_n \\ p_1 & \dots & p_n \end{pmatrix}$$

heißt **diskrete Informationsquelle**, wobei $p_i := P(a_i)$ für $i = 1, \dots, n$ ist.

Bemerkung 2.2.2: (i) Da das Alphabet A nur endlich viele Elemente umfasst, ist A eine sogenannte diskrete Menge. Q wird dementsprechend eine diskrete Informationsquelle genannt und $(A, \mathcal{P}(A), P)$ ein diskreter Wahrscheinlichkeitsraum.²¹

(ii) Einem Alphabet wird nun nicht irgendein Wahrscheinlichkeitsmaß P zugeordnet. $P(a_i)$ soll die Wahrscheinlichkeit dafür sein, dass das Zeichen a_i gesendet wird. Shannon spricht von „probabilities of occurrence“,²² Schulz 2003 auf Seite 29 von der „Signalwahrscheinlichkeit des Signals a_i “.

Lemma 2.2.3:²³ Sei Q eine diskrete Informationsquelle. Dann ist $0 \leq p_i \leq 1$ für $i = 1, \dots, n$ und $\sum_{i=1}^n p_i = 1$.

Bemerkung 2.2.4: $p_i = 0$ bedeutet, dass das Zeichen a_i nie gesendet wird. Ist hingegen $p_i = 1$, dann wird nur das Zeichen a_i gesendet und sonst kein anderes, also $p_j = 0$ für $j = 1, \dots, i-1, i+1, \dots, n$.

Ist nun A eine Quelle ohne Gedächtnis, dann wird dies wie folgt modelliert:

Definition 2.2.5:²⁴ Sei Q eine diskrete Informationsquelle und $a_1 \dots a_k$ eine Nachricht der Quelle A der Länge $k \in \mathbb{N}$. Sei $(A^k, \mathcal{P}(A^k), P^k)$ der Produkt-Wahrscheinlichkeitsraum des Wahrscheinlichkeitsraumes $(A, \mathcal{P}(A), P)$. Gilt

$$P^k(a_1 \dots a_k) = P(a_1) \cdot \dots \cdot P(a_k),$$

dann heißt Q **diskrete Informationsquelle ohne Gedächtnis**. Gilt die Gleichung nicht, dann heißt Q **diskrete Informationsquelle mit Gedächtnis**.

Bemerkung 2.2.6: Der Produkt-Wahrscheinlichkeitsraum $(A^k, \mathcal{P}(A^k), P^k)$ ist das Modell für die wiederholte Ausführung eines Zufallsexperimentes — hier k mal —, dessen Durchführungsbedingungen durch den Wahrscheinlichkeitsraum $(A, \mathcal{P}(A), P)$ festgelegt sind. Siehe hierzu Georgii 2015, S. 7, Beispiel 1.2.

²⁰Vgl. Schulz 2003, S. 29; Witt 2007, S. 273 f. Siehe auch Shannon 1948, S. 385.

²¹Vgl. Georgii 2015, S. 13, S. 21.

²²Shannon 1948, S. 392)

²³Vgl. Georgii 2015, S. 15, Satz 1.11.

²⁴Vgl. Schulz 2003, S. 29; Witt 2007, S. 273 f.

Satz 2.2.7:²⁵ Sei Q eine diskrete Informationsquelle ohne Gedächtnis und $X_i : A^k \rightarrow A$, $(a_1, \dots, a_k) \mapsto a_i$ die i -te Projektion für $i = 1, \dots, k$. Dann sind die Zufallsvariablen X_1, \dots, X_k stochastisch unabhängig.

Bemerkung 2.2.8: (i) In der Wahrscheinlichkeitstheorie heißen zwei Ereignisse C und D **stochastisch unabhängig**, wenn $P(C \cap D) = P(C) \cdot P(D)$ gilt.²⁶ Die stochastische Unabhängigkeit hier für Ereignisse, sprich Nachrichten zu formulieren, ist technisch aufwendig. Sehr viel eleganter geht es über die Zufallsvariablen X_i , die jeder Nachricht das i -te Zeichen zuordnen. Damit wird praktisch die „stochastische Unabhängigkeit der Zeichen“ dargestellt.

(ii) Die nicht vorhandene Kausalität zwischen dem Senden der Zeichen einer Quelle ohne Gedächtnis wird hier als stochastische Unabhängigkeit interpretiert. Das kann so gesehen werden. An sich ist aber stochastische Unabhängigkeit etwas anderes als kausale Unabhängigkeit. Siehe hierzu die Beispiele 3.15 und 3.16 sowie die anschließenden Ausführungen in Georgii 2015, S. 71 f.

(iii) Die Formel $P^k(a_1 \dots a_k) = P(a_1) \cdot \dots \cdot P(a_k)$ bedeutet, dass bei einer diskreten Informationsquelle ohne Gedächtnis sich die Wahrscheinlichkeit einer Nachricht als Produkt der Wahrscheinlichkeiten der vorkommenden Zeichen berechnet. Der Grund dafür ist die stochastische Unabhängigkeit der i -ten Projektionen (Zeichen) X_i .

(iv) Das hier verwendete Modell für eine Quelle ohne Gedächtnis ist das stochastische Standardmodell „(aus einer Urne) Ziehen mit Zurücklegen“: Die Zeichen der Quelle, das Alphabet befinden sich in einem Beutel (Urne). Dabei gibt es von einem Zeichen a_i meist mehrere Exemplare entsprechend der Wahrscheinlichkeit $P(a_i)$. Beispiel:

Das Alphabet besteht aus den Zeichen $\clubsuit, \heartsuit, \spadesuit, \bullet$ mit den Wahrscheinlichkeiten $P(\clubsuit) = 0,5$, $P(\heartsuit) = 0,2$, $P(\spadesuit) = 0,2$ und $P(\bullet) = 0,1$. Dann befinden sich in dem Beutel 5 \clubsuit , 2 \heartsuit , 2 \spadesuit und 1 \bullet .

Es wird aus dem Beutel ein Zeichen gezogen und notiert, welches Zeichen es ist. Dies entspricht dem Senden des Zeichens a_i . Danach wird das Zeichen wieder in den Beutel gelegt: Ziehen mit Zurücklegen. Das Ziehen eines Zeichens hängt somit nicht davon ab, welche Zeichen vorher gezogen worden sind. Es handelt sich also um eine Quelle ohne Gedächtnis, vergleiche Bemerkung 2.1.3, (i).

Besteht nun eine Nachricht aus k Zeichen, dann bedeutet dies in dem „Urnenmodell Ziehen mit Zurücklegen“, dass k mal ein Zeichen aus dem Beutel gezogen wird. Das zugehörige Wahrscheinlichkeitsmaß ist das sogenannte Produktmaß P^k mit der stochastischen Unabhängigkeit der Projektionen X_i . Siehe hierzu Georgii 2015, S. 75 f., Beispiel 3.22.

²⁵Georgii 2015, S. 75, Beispiel 3.22 und Korollar 3.21, (a). Beachte Georgii 2015, S. 64 f., Satz 3.8.

²⁶Vgl. Georgii 2015, S. 71.

Zur Berechnung der durchschnittlichen Codewortlänge verwenden wir die Wahrscheinlichkeit des Auftretens eines Zeichens:

Definition 2.2.9:²⁷ Sei Q eine diskrete Informationsquelle, $c : A \rightarrow B^*$ eine Codierung und $l : A \rightarrow \mathbb{N}_0$ die Abbildung, die jedem Zeichen a_i die Länge seines Codewortes $c(a_i)$ zuordnet, also $l_i := |c(a_i)|$. Dann heißt

$$\bar{l}_c := \bar{l}_c(Q) := \sum_{i=1}^n p_i \cdot l_i$$

mittlere Codewortlänge (empirische Entropie, mittlerer Codieraufwand) (der Codierung c).

Es stellt sich nun die Frage, ob es eine Codierung mit kleinster mittlerer Codewortlänge gibt. Bei injektiver buchstabenweiser Codierung — jede Nachricht ist dekodierbar — gibt es solch eine Codierung, den **Huffman-Code**, der eine statistische Codierung ist:²⁸

Satz 2.2.10:²⁹ Sei Q eine diskrete Informationsquelle, $c : A \rightarrow B^*$ eine Codierung mit injektiver buchstabenweiser Wort-Codierung c^* . Dann gilt $\bar{l}_{\text{Huffman}} \leq \bar{l}_c$. Insbesondere gilt die Abschätzung für einen Präfix-Code c .

²⁷Vgl. Schulz 2003, S. 42; Witt 2007, S. 274.

²⁸Zum Huffman-Code siehe Schulz 2003, S. 41 f.; Witt 2007, S. 275 ff.

²⁹Witt 2007, S. 278, Satz 20.1 in Verbindung mit Satz 2.1.8. Vgl. Schulz 2003, S. 43, Satz 6.5.

3 Bewertung

Wir kommen jetzt zum bewertenden Aspekt der Informationstheorie, der Quantifizierung des „Informationsgehaltes“.

3.1 Die Entropie einer Informationsquelle

Gemäß der Idee von Shannon¹ erfolgt die Quantifizierung des „Informationsgehaltes“ mit der Wahrscheinlichkeitsverteilung der Zeichen:

Der Zahlenwert der Information soll nur von der Wahrscheinlichkeit der Quellensignale abhängen.²

Und zwar in dem Sinne, dass ein Zeichen des Alphabets um so wichtiger ist je kleiner seine (Sende-)Wahrscheinlichkeit. Dazu ein

Beispiel 3.1.1: Es soll ein Wort erraten werden. Zunächst wird die Wortlänge angegeben und ein Buchstabe vorgegeben. Wird das Wort nicht erraten, dann wird ein weiterer Buchstabe angezeigt und so weiter. – Das Erraten des Wortes

– Y _ _ _ _ _

ist nun wesentlich leichter, als das Erraten des Wortes

N _ _ _ _ _ .

Dies liegt an der sehr unterschiedlichen Wahrscheinlichkeit des Auftretens von Y und N . Das „Senden“ von Y hat einen deutlich höheren Informationsgewinn als das „Senden“ von N .

Übertragen auf den hiesigen Kontext betrachten wir die zwei diskreten Informationsquellen

$$Q_1 := \begin{pmatrix} a_1 & a_2 & a_3 & a_4 \\ 0,25 & 0,25 & 0,25 & 0,25 \end{pmatrix}$$

und

$$Q_2 := \begin{pmatrix} b_1 & b_2 & b_3 & b_4 \\ 0,91 & 0,03 & 0,03 & 0,03 \end{pmatrix} .$$

¹Shannon 1948, S. 392.

²Schulz 2003, S. 47.

Aus jeder Quelle wird nun ein Zeichen gesendet. Bei Quelle Q_1 besteht sehr große Ungewissheit über das gesendete Zeichen, bei Quelle Q_2 wird mit sehr hoher Wahrscheinlichkeit das gesendete Zeichen b_1 sein. Bei Quelle Q_1 gibt es maximale Wahlfreiheit zwischen den Zeichen und damit einen sehr hohen Informationsgewinn nach dem Senden des Zeichens. Für den Empfänger ist das Zeichen besonders wertvoll. Bei Quelle Q_2 verhält es sich genau anders herum: Praktisch keine Wahlfreiheit und keinen Informationsgewinn. Der Empfänger rechnet mit keinem anderen Zeichen als mit b_1 .³

Zur mathematischen Umsetzung des reziproken Verhaltens von Wahrscheinlichkeit und Wichtigkeit bietet sich die Logarithmusfunktion an:

Definition 3.1.2:⁴ Sei die diskrete Informationsquelle

$$Q := \begin{pmatrix} a_1 & \dots & a_n \\ p_1 & \dots & p_n \end{pmatrix}$$

gegeben. Die Funktion

$$I : A \rightarrow \mathbb{R}_0^+, \quad a_i \mapsto -\log_2(p_i),$$

wobei $\log_2(0) := 0$ sein soll, heißt **Informationsgehalt**(, **Informationsmaß**). $I(a_i)$ ist der **Informationsgehalt des Zeichens** a_i mit der Maßeinheit **bit**.

Bemerkung 3.1.3: (i) Es kann eine andere Basis gewählt werden. Wegen $\log_b x = \log_a x / \log_a b$ ändert sich der Informationsgehalt nur um eine Konstante.

(ii) Es ist zwischen Bit und bit zu unterscheiden. Bit ist eine Binärziffer, bit die von Shannon definiert Maßeinheit pro Zeichen.⁵

(iii) Die Verwendung der Wahrscheinlichkeitsverteilung bei der Definition des Informationsgehaltes deutet bereits den Zusammenhang zur statistischen Codierung bzw. dem Huffman-Code an.

Die folgende Aussage zeigt die Umsetzung von „ein Zeichen des Alphabets ist um so wichtiger, je kleiner dessen Wahrscheinlichkeit ist“:

Lemma 3.1.4:⁶ Gilt $0 < p_i < p_j < 1$, so ist $-\log_2(p_i) > -\log_2(p_j)$.

Nachdem wir den Informationsgehalt eines Zeichens festgelegt haben, stellt sich die Frage nach dem Informationsgehalt einer Quelle. Dieser sollte das gewogene Mittel der Informationsgehalte der Zeichen sein, was in der Wahrscheinlichkeitstheorie dem Erwartungswert⁷ entspricht. Dafür muss I aber eine Zufallsvariable sein:

³Vgl. Shannon und Weaver 1949, S. 25.

⁴Vgl. Schulz 2003, S. 48; Witt 2007, S. 282. Auf Seite 392 spricht Shannon 1948 von *measure*.

⁵Shannon 1948, S. 396.

⁶Heuser Teil 1 2003, S. 166.

⁷Siehe Georgii 2015, S. 101.

Proposition 3.1.5:⁸ I ist eine Zufallsvariable vom Messraum $(A, \mathcal{P}(A))$ in den Messraum $(\mathbb{R}_0^+, \mathcal{B}_{\mathbb{R}_0^+})$.

Nun zum Erwartungswert des Informationsgehaltes, dem Maßstab für den mittleren Informationsgehalt einer Quelle:

Definition 3.1.6:⁹ Sei $A := \{a_1, \dots, a_n\}$ ein Alphabet, $(A, \mathcal{P}(A), P)$ ein Wahrscheinlichkeitsraum und Q die zugehörige diskrete Informationsquelle. Dann heißt der Erwartungswert des Informationsgehaltes, also

$$H(Q) := H(p_1, \dots, p_n) := \mathbb{E}(I) = - \sum_{i=1}^n p_i \cdot \log_2(p_i)$$

die **Entropie von Q** . Ihre Maßeinheit ist bit pro Zeichen.

Bemerkung 3.1.7: (i) Shannon nennt H „the entropy of the set of probabilities p_1, \dots, p_n “¹⁰.

(ii) Da der Erwartungswert des Informationsgehaltes formal der Entropie der Thermodynamik entspricht, nennt Shannon H Entropie.¹¹ Inhaltlich beschreiben beide Größen aber sehr Unterschiedliches.

(iii) In der Definition ist bewusst nicht gefordert, dass es sich um eine Quelle ohne Gedächtnis handelt. Shannon ganz entsprechend gilt die Entropie für Quellen mit und ohne Gedächtnis.

(iv) Ein große Entropie bedeutet, dass vor dem Senden eine große Wahlfreiheit zwischen den Zeichen besteht und damit eine große Unsicherheit darüber, welches Zeichen gesendet wird.¹² Nach dem Senden des Zeichens ist der Informationsgewinn beim Empfänger groß.

(v) Shannon hat die Entropie anders eingeführt: Es werden Bedingungen an die Funktion H gestellt und dann gezeigt, dass H die in der Definition 3.1.6 angegebene Funktion ist.¹³ So wird auch heute oft noch vorgegangen.¹⁴ Aus mathematischer Sicht ist dies reizvoll, da aus den Bedingungen die Existenz und — bis auf eine Konstante — Eindeutigkeit der Entropie H folgt. Der Nachteil dieses Ansatzes ist, dass die Bedingungen nicht so leicht zu begründen sind. Dagegen ist die Definition der Entropie als Erwartungswert des Informationsgehaltes wesentlich einsichtiger.

Insbesondere für die bedingte Entropie¹⁵, die wir hier nicht betrachten, gibt es noch die folgende

⁸Georgii 2015, S. 22, Beispiel 1.24.

⁹Vgl. Mathar 1996, S. 26. Siehe auch Schulz 2003, S. 49; Witt 2007, S. 283; Shannon 1948, S. 393.

¹⁰Shannon 1948, S. 393 f.

¹¹Shannon 1948, S. 393.

¹²Topsøe 1974, S. 5.

¹³Shannon 1948, S. 392 f.

¹⁴Schulz 2003, S. 47 ff; Witt 2007, S. 281 ff.

¹⁵Vgl. Mathar 1996, S. 27.

Notation 3.1.8:¹⁶ Sei $A := \{a_1, \dots, a_n\}$ ein Alphabet, $(A, \mathcal{P}(A), P)$ ein Wahrscheinlichkeitsraum und Q die zugehörige diskrete Informationsquelle. Mit der Zufallsvariable $X : A \rightarrow A, a_i \mapsto a_i$ wird $H(X) := H(Q)$ erklärt.

Bemerkung 3.1.9: (i) Die Festlegung der Zufallsvariablen als Identität ist ein Standardvorgehen. Siehe hierzu Behnen und Neuhaus 2003, Seite 65, Bemerkung 5.5.

(ii) Allgemeiner legt Topsøe 1974 auf Seite 74 die Zufallsvariable X fest: Sei $(\Omega, \mathcal{F}, \mathcal{P})$ ein Wahrscheinlichkeitsraum. Dann ist $X : \Omega \rightarrow A$.

(iii) Zum Einsatz der Notation $H(X)$ siehe Krengel 2005, S. 117 f., Mathar 1996, S. 27, Topsøe 1974, S. 50 f. und Shannon 1948, S. 394 ff.

3.2 Eigenschaften der Entropie

Der „Informationsgehalt“ soll so bestimmt werden, dass ein Zeichen des Alphabets um so wichtiger ist je kleiner dessen (Sende-)Wahrscheinlichkeit. Lemma 3.1.4 zeigt, dass dies für Zeichen gilt. Analog soll dies natürlich nun auch für Alphabete gelten, d.h. die Entropie soll um so größer sein, je geringer die Abweichungen der Wahrscheinlichkeiten der einzelnen Zeichen ist. Dem ist so:

Satz 3.2.1:¹⁷ Die Entropie $H(p_1, \dots, p_n)$ ist am größten, wenn alle Wahrscheinlichkeiten gleich sind:

$$0 \leq H(p_1, \dots, p_n) \leq H\left(\frac{1}{n}, \dots, \frac{1}{n}\right)$$

Korollar 3.2.2:¹⁸ Es gilt $H\left(\frac{1}{n}, \dots, \frac{1}{n}\right) = \log_2 n$, speziell $H\left(\frac{1}{2}, \frac{1}{2}\right) = 1$.

Satz 3.2.3:¹⁹ Es ist $H(Q) = 0$ genau dann, wenn $p_i = 1$ gilt für genau ein $i \in \{1, \dots, n\}$.

Beweis: „ \Leftarrow “: Wegen $p_i = 1$ gilt $p_j = 0$ für $j = 1, \dots, i-1, i+1, \dots, n$ (Bemerkung 2.2.4). Da $\log_2(0) := 0$ gesetzt ist und $\log_2(p_i) = 0$, folgt die Behauptung $H(Q) = 0$.

„ \Rightarrow “: Sei $\sum_{i=1}^n -p_i \cdot \log_2(p_i) = 0$. Es gilt $0 < -p_i \cdot \log_2(p_i)$ für $0 < p_i < 1$, $i \in \{1, \dots, n\}$. Dies bedeutet, dass entweder ein p_i gleich 1 ist oder gleich 0 für alle $i \in \{1, \dots, n\}$. Der Fall $p_1 = \dots = p_n = 0$ kann nicht auftreten, da ansonsten $P(A) = 0$ wäre anstatt $P(A) = 1$.²⁰ — $p_1 = \dots = p_n = 0$ bedeutet, dass kein Zeichen gesendet wird. — Damit muss es ein $i \in$

¹⁶Vgl. Mathar 1996, S. 25.

¹⁷Topsøe 1974, S. 28, Satz 2. Vgl. Mathar 1996, S. 32, Satz 3.1. a); Shannon 1948, S. 394. Siehe auch Witt 2007, S. 284, Satz 20.2.

¹⁸Topsøe 1974, S. 26, Satz 2. Vgl. Shannon 1948, S. 394.

¹⁹Vgl. Shannon 1948, S. 394.

²⁰Additivität und Normierung des Wahrscheinlichkeitsmaßes P .

$\{1, \dots, n\}$ geben mit $p_i = 1$. – Ergänzung: Es kann nur genau ein p_i gleich 1 sein. Alle anderen Wahrscheinlichkeiten sind dann gleich 0. Ansonsten wäre $P(A) > 1$.

□

Diese letzte Aussage bedeutet, dass die Entropie nur dann gleich 0 ist, wenn das gesendete Zeichen von vornherein fest steht.

Die mathematische Rechtfertigung für die Festlegung $\log_2(0) := 0$ ist die folgende Aussage:

Proposition 3.2.4: *Es gilt $\lim_{p_i \rightarrow 0} p_i \cdot \log_2(p_i) = 0$ für $i \in \{1, \dots, n\}$.*

Beweis: Mit der Regel von de L'Hospital²¹ (3. Gleichheitszeichen) gilt mit $x \in \mathbb{R}^+$ und $b \in \mathbb{R}_{>1}$ ²²

$$0 = \lim_{x \rightarrow 0} -x = \lim_{x \rightarrow 0} \frac{\frac{1}{x}}{-\frac{1}{x^2}} = \lim_{x \rightarrow 0} \frac{\log_b(x)}{\frac{1}{x}} = \lim_{x \rightarrow 0} x \cdot \log_b(x),$$

womit wir die allgemeinere Aussage $\lim_{x \rightarrow 0} x \cdot \log_b(x) = 0$ gezeigt haben.

□

Abschließend benötigen wir noch Aussagen zur Entropie von Produkträumen. Dafür führen wir den folgenden Begriff ein:

Definition 3.2.5: Seien k Wahrscheinlichkeitsräume $(A_1, \mathcal{P}(A_1), P_1), \dots, (A_k, \mathcal{P}(A_k), P_k)$ mit zugehörigen diskreten Informationsquellen

$$Q_i := \begin{pmatrix} a_{i1} & \cdots & a_{in_i} \\ p_{i1} & \cdots & p_{in_i} \end{pmatrix}, \quad i = 1, \dots, k$$

gegeben. Sei ferner der Produkt-Wahrscheinlichkeitsraum $(A_1 \times \cdots \times A_k, \mathcal{P}(A_1 \times \cdots \times A_k), P)$ mit

$$P((a_{1i_1}, \dots, a_{ki_k})) := p_{1i_1} \cdots p_{ki_k} = P_1(a_{1i_1}) \cdots P_k(a_{ki_k})$$

für $i_1 = 1, \dots, n_1, \dots, i_k = 1, \dots, n_k$ gegeben. Die Matrix

$$Q_1 \times \cdots \times Q_k := \begin{pmatrix} (a_{11}, \dots, a_{k1}) & \cdots & (a_{1n_1}, \dots, a_{kn_k}) \\ P((a_{11}, \dots, a_{k1})) & \cdots & P((a_{1n_1}, \dots, a_{kn_k})) \end{pmatrix}$$

heißt **diskrete Produktinformationsquelle** (ohne Gedächtnis). Gilt $(A_1, \mathcal{P}(A_1), P_1) = \cdots = (A_k, \mathcal{P}(A_k), P_k)$, dann wird abkürzend

$$Q^k := \underbrace{Q \times \cdots \times Q}_{k\text{-mal}}$$

²¹Heuser Teil 1 2003, S. 287, Satz 50.1.

²²Zur Definition der Logarithmusfunktion siehe Heuser Teil 1 2003, S. 165, Satz und Definition 25.4.

geschrieben, wobei $Q := Q_1 = \dots = Q_K$ ist.

Bemerkung 3.2.6: (i) Erläuterung zur Notation: Es ist $a_{ji} \in A_j$.

(ii) Bei $A_1 \times \dots \times A_k$ handelt es sich um ein Alphabet, womit $Q_1 \times \dots \times Q_k$ eine diskrete Informationsquelle ist.

(iii) Festzuhalten ist, dass P ein Wahrscheinlichkeitsmaß ist (Georgii 2015, S. 75, Korollar 3.21). Die Funktionsvorschrift von P ist gleichbedeutend mit stochastischer Unabhängigkeit. Vergleiche hierzu Definition 2.2.5 und Bemerkung 2.2.8. Siehe hierzu auch Krengel 2005, S. 117, Satz 8.5.

(iv) Liegen nur zwei Wahrscheinlichkeitsräume $(A_1, \mathcal{P}(A_1), P_1)$ und $(A_2, \mathcal{P}(A_2), P_2)$ vor, dann wird abkürzend $p(a_1, a_2) := P((a_1, a_2))$ für $a_1 \in A_1$ und $a_2 \in A_2$ sowie

$$Q_1 \times Q_2 = \left(\begin{array}{ccc} (a_{11}, a_{21}) & \dots & (a_{1n_1}, a_{2n_2}) \\ p(a_{11}, a_{21}) & \dots & p(a_{1n_1}, a_{2n_2}) \end{array} \right)$$

geschrieben.

Satz 3.2.7:²³ Für die Entropie der diskreten Produktinformationsquelle $Q_1 \times Q_2$ gilt

$$H(Q_1 \times Q_2) = H(Q_1) + H(Q_2) .$$

Korollar 3.2.8:²⁴ Sei Q eine diskrete Informationsquelle ohne Gedächtnis. Dann gilt $H(Q^k) = k \cdot H(Q)$.

Bemerkung 3.2.9: (i) $H(Q^k)$ kann als Entropie der Wörter der Länge k (des Alphabetes A) verstanden werden .

(ii) Wesentliche Voraussetzung ist hier, dass es sich um eine Quelle ohne Gedächtnis handelt. Siehe Bemerkungen 2.1.3, (i) und 2.2.8, (iv).

(iii) Die Aussage des Satzes zeigt die Additivität der Entropie in dem Sinne, „daß unabhängige Wiederholungen die Summe der Funktionswerte der einzelnen Versuche entspricht“²⁵.

²³Witt 2007, S. 284, Satz 20.3, a). ; Schulz 2003, S. 50, Hilfssatz 7.7, (a). Vgl. Shannon 1948, S. 394 f., 3.

²⁴Witt 2007, S. 284, Satz 20.3, b); Schulz 2003, S. 50, Hilfssatz 7.7, (b). Vgl. Topsøe 1974, S. 25, Lemma 3.

²⁵Topsøe 1974, S. 25.

4 Codieraufwand

Wir kommen jetzt zu zentralen Ergebnissen der Informationstheorie. Mit Hilfe der Entropie — der Theorie — werden Aussagen zur Qualität von Codierungen — der Praxis, der Anwendung — getroffen:

Satz 4.0.1:¹ **Quellencodierungssatz** Sei Q eine diskrete Informationsquelle und c ein dekodierbarer Code, d.h. c^* ist injektiv.

(i) Es gilt $H(Q) \leq \bar{l}_c(Q)$.

(ii) Es gibt einen dekodierbaren Code \tilde{c} mit $\bar{l}_{\tilde{c}}(Q) < H(Q) + 1$.

Bemerkung 4.0.2: (i) Voraussetzung für Satz 4.0.1 ist, dass c^* eine buchstabenweise Wort-Codierung ist.

(ii) Wesentliche Grundlage für den Beweis von Satz 4.0.1 ist die Ungleichung von Kraft, siehe Krengel 2005, S. 115, Satz 8.1.

Mit Satz 2.2.10 gilt das

Korollar 4.0.3: Schranken für die mittlere Codewortlänge Es ist

$$H(Q) \leq \bar{l}_{\text{Huffman}}(Q) < H(Q) + 1 .$$

Hieraus folgt mit Korollar 3.2.8 das

Korollar 4.0.4:² **Fundamentalsatz** Ist Q jetzt eine diskrete Informationsquelle ohne Gedächtnis, dann gilt

$$H(Q) \leq \frac{1}{k} \cdot \bar{l}_{\text{Huffman}}(Q^k) < H(Q) + \frac{1}{k} .$$

Bemerkung 4.0.5: (i) Bauer und Goos 1991 fassen den Fundamentalsatz verbal wie folgt, wobei L die mittlere Codewortlänge ist: „Jede Nachrichtenquelle kann so codiert werden, daß die Differenz $L - H$ beliebig klein wird.“³ Oder anders ausgedrückt: Wenn ich den besten Code aller decodierbaren Codes nehme, nämlich den Huffman-Code, dann kann ich mich der Entropie

¹Krengel 2005, S. 117, Satz 8.4 in Verbindung mit Satz 2.1.8. Vgl. Witt 2007, S. 285, Satz 20.4; Schulz 2003, S. 51, Satz 7.8.

²Witt 2007, S. 287, Satz 20.5. Vgl. Schulz 2003, S. 52, Fundamentalsatz 7.9; Topsøe 1974, S. 25, Satz 1. Beachte Krengel 2005, S. 118; Shannon 1948, S. 397, Theorem 3, S. 401, Theorem 9.

³Bauer und Goos 1991, S. 333.

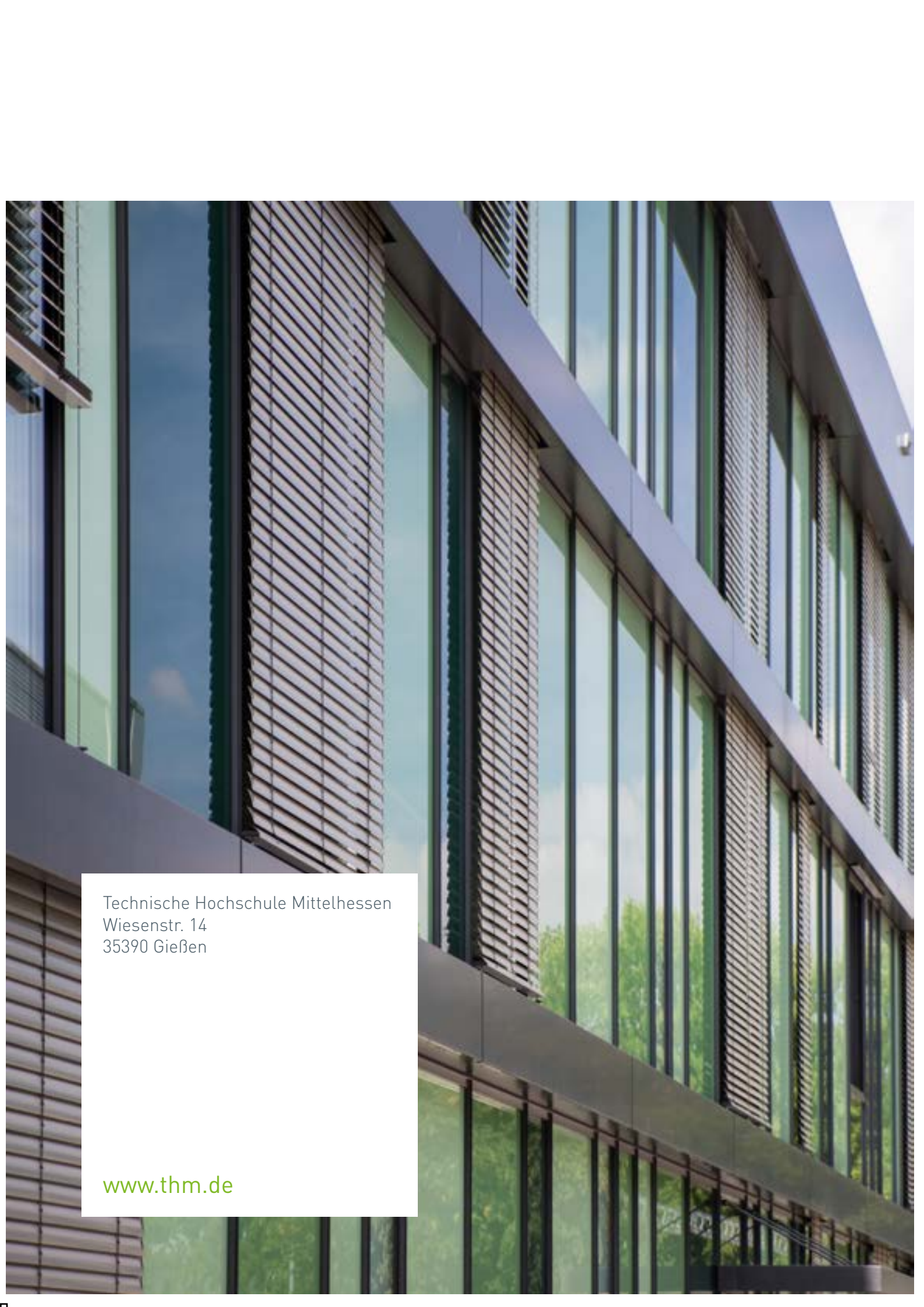
durch Wortverlängerung der Quellworte A^k über einem Alphabet A beliebig annähern.

(ii) Das Korollar 4.0.4 wird auch als Hauptsatz bezeichnet. Dies entspricht nicht der üblichen mathematischen Konvention: Ein Hauptsatz verbindet Theorien miteinander, etwa der Hauptsatz der Differential- und Integralrechnung. Hingegen ist die Aussage eines Fundamentalsatzes Grundlage für eine Theorie, etwa der Fundamentalsatz der Arithmetik.

(iii) Die Grundlage für Korollar 4.0.4 bilden die beiden Theoreme 3 und 4 von Shannon 1948, Seite 397.

Literaturverzeichnis

- Bauer, Friedrich L. und Gerhard Goos. *Informatik 1. Eine einführende Übersicht*. 4. verbes. Aufl. Berlin, Heidelberg: Springer, 1991.
- Behnen, Konrad und Georg Neuhaus. *Grundkurs Stochastik*. 4. neu bearb. u. erw. Aufl. Heidenau: PD-Verlag, 2003.
- Georgii, Hans-Otto. *Stochastik*. 5. Aufl. Berlin, Boston: Walter de Gruyter, 2015.
- Heuser, Harro. *Lehrbuch der Analysis. Teil 1*. 15. durchges. Aufl. Wiesbaden: Teubner, 2003.
- Jacobs, Konrad und Dieter Jungnickel. *Einführung in die Kombinatorik*. 2. völlig neu bearb. und erweiter. Aufl. Berlin, New York: Walter de Gruyter, 2004.
- Krengel, Ulrich. *Einführung in die Wahrscheinlichkeitstheorie und Statistik*. 8. erweiter. Aufl. Wiesbaden: Springer, 2005.
- Mathar, Rudolf. *Informationstheorie: Diskrete Modelle und Verfahren*. 1. Aufl. Stuttgart: Teubner, 1996.
- Schulz, Ralph-Hardo. *Codierungstheorie*. 2. aktual. u. erw. Aufl. Wiesbaden: Vieweg, 2003.
- Shannon, Claude E. „A mathematical theory of communication.“ *The Bell System Technical Journal*. XXVII (1948): 379 - 423.
- Shannon, , Claude E. und Warren Weaver. *Mathematische Grundlagen der Informationstheorie*. Übers. Helmut Dreßler (Münche, Wien: Oldenbourg, 1976). Illinois: University of Illinois Press, 1949.
- Topsøe, Flemming. *Informationstheorie: Eine Einführung*. 1. Aufl. Stuttgart: Teubner, 1974.
- Witt, Kurt-Ulrich. *Algebraische Grundlagen der Informatik*. 3. überarb. u. erw. Aufl. Wiesbaden: Vieweg, 2007.

A photograph of a modern building facade featuring large glass windows and metal panels. The glass reflects the sky and surrounding greenery. The metal panels have a textured, grid-like pattern. The building is viewed from a low angle, looking up.

Technische Hochschule Mittelhessen
Wiesenstr. 14
35390 Gießen

www.thm.de