SKRIPTUM
zur Lehrveranstaltung
INFORMATIONSTHEORIE

von
Ferdinand Österreicher

Institut für Mathematik
der
Universität Salzburg

Salzburg Juni 2008

Contents

EINLEITUNG

Die Informationstheorie beschäftigt sich mit mathematischen Problemen, die bei der Speicherung, Umformung und der Übermittlung von Information auftreten.

Ein Übermittlungs- oder Kommunikationssystem ist schematisch folgendermaßen aufgebaut.

Abbildung 1: Schematische Darstellung eines Kanals (bitte einfügen)

Sowohl die Erzeugung der Information durch die Quelle, als auch die Störung des Übertragungsvorgangs durch Geräusche im Kanal sind stochastische Vorgänge, weswegen die von *Claude E. Shannon* und *Norbert Wiener* begründete Informationstheorie naturgemäß stochastisch ist. Die Art der Information kann dabei Wellenform oder digitalen Charakter besitzen. Im vorliegenden Skriptum beschäftigen wir uns, *Shannon* folgend, ausschließlich mit Information von digitalem Charakter.

Das nachstehende detailliertere Schema des Kommunikationssystems entspricht der technisch vielfach realisierten Aufspaltung des Kodierers in jeweils zwei Komponenten. Dabei hat der Quellkodierer die Aufgabe, die gegebene Information möglichst ökonomisch in eine binäre Form zu pressen. Der Kanalkodierer fügt hingegen wieder gezielt binäre Information dazu, damit die bei der Übertragung der Information unvermeidlichen Fehler weitgehend korrigiert oder zumindest erkannt werden können. Auf die gruppentheoretischen Algorithmen, die bei der Kanalkodierung verwendet werden, wird im vorliegenden Skriptum nicht eingegangen.

Abbildung 2: Detaillierte schematische Darstellung eines Kanals (bitte einfügen)

In Kapitel 1 wird der Themenbereich der Quellkodierung abgehandelt. Darin wird der zentrale Begriff der *Entropie* zusammen mit der von *Kullback* und *Leibler* stammenden *I-Divergenz* eingeführt. In Abschnitt 1.1 wird die *Hartley'sche Formel* motiviert und axiomatisch untersucht. In Abschnitt 1.2.1 werden die grundlegenden Eigenschaften optimaler Fragestrategien motiviert und die mittlere Codewortlänge durch die *Entropie* nach unten abgeschätzt. Dabei wird erstmals von der *I-Divergenz* Gebrauch gemacht. Abschnitt 1.2.2 dient der axiomatischen Untersuchung der Entropie. In Abschnitt 1.3.1 wird der *Huffman Code* betrachtet. Die Abschnitte 1.3.2 und 1.3.3 dienen der Charakterisierung der *eindeutigen Dekodierbarkeit* mit Hilfe der *Fano-Kraft'schen Ungleichung.*

Abschnitt 1.4 gewährt einen Ausblick zur Wärmelehre. *Ludwig Boltzmann* folgend, wird der von *Rudolf Clausius* im Zusammenhang mit dem 2. Hauptsatz der Wärmelehre eingeführte Begriff der Entropie stochastisch gefasst und das *Maximum Entropie Prinzip* vorgestellt. Als Anwendungen der einschlägigen Lösungen - der sogenannten *Gibbsverteilungen* - werden die *barometrische Höhenformel* und *Maxwell'sche Geschwindigkeitsverteilung* behandelt.

Kapitel 2 ist dem Themenbereich der Kanalkodierung gewidmet. In Abschnitt 2.1 werden nach einer knappen Motivation der *Kanalkapazität* Hilfsmittel, wie gemeinsame und bedingte Entropie und wechselseitige Information, erarbeitet und grundlegende Definitionen und Aussagen präsentiert. In Abschnitt 2.2 wird der Kanalkodierungssatz für den binären symmetrischen Kanal behandelt. Als Vorbereitung wird eine geeignete Form der Tail Inequality bewiesen. Ein relativ breiter Raum wird der Motivation des Kanalkodierungssatzes eingeräumt, ehe dessen Beweis durchgeführt wird. In Abschnitt 2.3 wird die Umkehrung des Kanalkodierungssatzes behandelt. Die *Ungleichung von Fano* und die *Data Processing Inequality* schaffen die Voraussetzungen für den kurzen Beweis.

1 QUELLKODIERUNG

1.1 HARTLEYS FORMEL

1.1.1 Motivation

Beispiel 1: (a) Durch "Ja-Nein"-Fragen herausfinden, an welchem Tag eines bestimmten Monats jemand Geburtstag hat. Fragestrategie: Sukzessives Halbieren von Mengen liefert $32/2 = 16, 16/2 = 8, 8/2 = 4,\ 4/2 = 2,\ 2/2 = 1$ und demnach $32 = 2^5$. Die Anzahl der nötigen Fragen ist also $\lceil \log_2(31) \rceil = \log_2(32) = 5$.

(b) Mit den Fingern einer Hand bis 31 zählen oder die Binärdarstellung einer Zahl $x \in \{1, ..., 31\}$. Die Binärdarstellung von x ist ein Vektor $\varkappa = (\varkappa_1(x), ..., \varkappa_5(x)) \in \{0, 1\}^5$ mit der Eigenschaft

$$x = \sum_{i=1}^{5} \varkappa_i(x)\, 2^{5-i}\,.$$

Sei $\Omega_m = \{x_1, ..., x_m\}$ eine endliche Menge mit m Elementen. Dann ist die Anzahl der Fragen - mit den beiden möglichen Antworten "Ja" und "Nein" -, die nötig sind, um ein Element zu bestimmen, gleich

$$\lceil \log_2(m) \rceil\,.$$

Um ein Element $(y_1, ..., y_n) \in \Omega_m^n$ der n-ten Potenz von Ω_m zu bestimmen, ist die Anzahl der nötigen Fragen demnach $\lceil \log_2(m^n) \rceil$. Für die Anzahl der pro Koordinate des Vektors $(y_1, ..., y_n)$ nötigen Fragen gilt somit wegen $x \leq \lceil x \rceil < x + 1$ und $\log(m^n) = n \log(m)$

$$\log_2(m) \leq \frac{1}{n} \lceil \log_2(m)^n \rceil < \log_2(m) + \frac{1}{n}\,.$$

Der Grenzwert für $n \to \infty$ liefert die *Hartley'sche Formel*

$$I(\Omega_m) = \log_2(m)\,,$$

welche den Informationsgehalt angibt, der jedem Element eine Menge mit m Elementen innewohnt.

Anmerkung 1: Gäbe es anstelle von 2 möglichen Antworten $b \in \mathbb{N}\backslash\{1, 2\}$ mögliche Antworten, so würde man anstelle der Basis 2 die Basis b verwenden.

1.1.2 Axiomatik

Proposition 1: Eine reellwertige Funktion I, welche die Eigenschaften

(I1) $I(\Omega_{m \times n}) = I(\Omega_m) + I(\Omega_n) \quad \forall\, m, n \in \mathbb{N}$
(I2) $I(\Omega_m) < I(\Omega_{m+1}) \quad \forall\, m \in \mathbb{N}$

(I3) $I(\Omega_b) = 1$ für ein $b \in \mathbb{N}\backslash\{1\}$

erfüllt, besitzt die Form

$$I(\Omega_m) = \log_b(m) \; .$$

Anmerkung 1: Setzt man in (I1) $n = 1$ ein, so erhält man

$$I(\Omega_m) = I(\Omega_m) + I(\Omega_1)$$

und somit - wegen $I(\Omega_m) \in \mathbb{R}$ - $I(\Omega_1) = 0$. (I2) impliziert $I(\Omega_m) > 0 \;\; \forall\, m \in \mathbb{N}\backslash\{1\}$.

Beweis: Seien $n, r \in \mathbb{N}$ gegeben. Dann gibt es einen Exponenten $s(r) \in \mathbb{N}$ derart, dass gilt

$$b^{s(r)} \leq n^r < b^{s(r)+1} \; ,$$

oder, gleichbedeutend,

$$s(r) \leq r \log_b(n) < s(r) + 1 \; ,$$

und somit nach Division durch r ,

$$\left| \log_b(n) - \frac{s(r)}{r} \right| \leq \frac{1}{r} \; . \tag{1}$$

Setzt man nun $f(m) = I(\Omega_m)$, dann gilt wegen (I1)

$$f(n^r) = r f(n)$$

und somit wegen (I2) und (I3)

$$s(r) \leq r f(n) < s(r) + 1$$

und daher

$$\left| f(n) - \frac{s(r)}{r} \right| \leq \frac{1}{r} \; . \tag{2}$$

Die Anwendung der Dreiecksungleichung und die Berücksichtigung von (2) und (1) ergibt

$$\begin{aligned} |f(n) - \log_b(n)| &= \left| (f(n) - \frac{s(r)}{r}) + (\frac{s(r)}{r} - \log_b(n)) \right| \\ &\leq \left| f(n) - \frac{s(r)}{r} \right| + \left| \frac{s(r)}{r} - \log_b(n) \right| \leq \frac{2}{r} \; . \end{aligned}$$

Indem man r gegen ∞ gehen lässt, erhält man schließlich $f(n) = \log_b(n)$. $\square$

Anmerkung 2 [1]: Die obige Proposition gilt auch dann, wenn (I2) durch die Eigenschaft (I2'):

$\cdot \; \lim_{m \to \infty} (I(\Omega_m) - I(\Omega_{m-1})) = 0$

ersetzt wird.

[1] Hinsichtlich des relativ aufwändigen Beweises sei auf [25], S 438 verwiesen.

1.2 SHANNONS ENTROPIE

1.2.1 Motivation von Entropie und I-Divergenz

Gegeben sei

eine Menge	$\Omega = \{x_1, ..., x_m\}$	mit zugehöriger
Wahrscheinlichkeitsverteilung	$P = (p_1, ..., p_m)$.	

Zunächst werde ein Element $X \in \Omega$ gemäß der Wahrscheinlichkeitsverteilung P gewählt. Dann gehe es darum, X durch Fragen der Form "ist X Element der Teilmenge $E\,(\subset \Omega)$ ", welche wahrheitsgemäß mit "Ja" oder "Nein"zu beantworten sind, herauszufinden - und zwar so, dass die durchschnittliche Anzahl der Fragen minimal ist.

Sei also $A : \Omega \mapsto \mathbb{N}$ die (zufällige) Anzahl der Frage, dann sind die Fragen so zu stellen, dass $E_P\,(A)$ minimal ist.

Beispiel 1: Werfen eines regelmäßigen Oktaeders mit den Augenzahlen $1, 2, 3, 4, 5, 6, 7, 8$. Wir bilden im Folgenden auf unterschiedliche Weise stets $m = 4$ Ereignisse:

(a)

		$\{1,2\}$	$\{3,4\}$	$\{5,6\}$	$\{7,8\}$
P	$=$	$(\frac{1}{4},$	$\frac{1}{4},$	$\frac{1}{4},$	$\frac{1}{4})$
A	$=$	$(2$	2	2	$2)$

Die Anzahl der Fragen ist $A = 2$. Dies ist im Einklang mit der Hartley'schen Formel zur Basis $b = 2$, nämlich $2 = \log_2(4)$.

(b)

		$\{1,2,3,4\}$	$\{5,6\}$	$\{7\}$	$\{8\}$
P	$=$	$(\frac{1}{2},$	$\frac{1}{4},$	$\frac{1}{8},$	$\frac{1}{8})$
A	$=$	$(1$	2	3	$3)$

Die durchschnittliche Anzahl der Fragen ist

$$E_P(A) = \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{8} \times 3 = \frac{7}{4} < 2\,.$$

(c)

		$\{1,2,3\}$	$\{4,5,6\}$	$\{7\}$	$\{8\}$
P	$=$	$(\frac{3}{8},$	$\frac{3}{8},$	$\frac{1}{8},$	$\frac{1}{8})$
A	$=$	$(1$	2	3	$3)$

Die durchschnittliche Anzahl der Fragen ist

$$E_P(A) = \frac{3}{8} \times 1 + \frac{3}{8} \times 2 + \frac{1}{8} \times 3 + \frac{1}{8} \times 3 = \frac{15}{8} \in (\frac{7}{4}, 2)\,.$$

Anmerkung 1: In allen Fällen (a), (b) und (c) sind die Ereignisse stets nach nichtsteigenden Wahrscheinlichkeiten geordnet, d.h. es gilt $p_1 \geq p_2 \geq p_3 \geq p_4$. Für die Anzahlen der nötigen Fragen gilt offensichtlich stets

$$a_1 \leq a_2 \leq a_3 = a_4\,.$$

Darüber hinaus lässt sich $E_P(A)$ folgendermaßen mit Hilfe der a_i, $i \in \{1,2,3,4\}$ darstellen

$$\begin{aligned} E_P(A) &= p_1 \log_2 (2^{a_1}) + p_2 \log_2 (2^{a_2}) + p_3 \log_2 (2^{a_3}) + p_4 \log_2 (2^{a_4}) \\ &= p_1 \log_2 \left(\frac{1}{(\frac{1}{2})^{a_1}} \right) + p_2 \log_2 \left(\frac{1}{(\frac{1}{2})^{a_2}} \right) + p_3 \log_2 \left(\frac{1}{(\frac{1}{2})^{a_3}} \right) + p_4 \log_2 \left(\frac{1}{(\frac{1}{2})^{a_4}} \right) \end{aligned}$$

und es gilt

$$(\frac{1}{2})^{a_1} + (\frac{1}{2})^{a_2} + (\frac{1}{2})^{a_3} + (\frac{1}{2})^{a_4} = 1\,.$$

Diese Beobachtung entspricht dem folgenden allgemeingültigen Sachverhalt für die

Anzahlen a_i der Fragen für eine optimale Fragestragtegie

Proposition 1: Es seien die Elemente von Ω so geordnet, dass gilt

$$p_1 \geq p_2 \geq ... \geq p_{m-1} \geq p_m\,.$$

Bei optimaler Fragestrategie gelten für die Anzahlen $a_i = A(x_i)$, $i \in \{1, ..., m\}$ folgende Eigenschaften

(i)

$$a_1 \leq a_2 \leq ... \leq a_{m-1} \leq a_m\,,$$

(ii)

$$a_{m-1} = a_m\,,$$

wobei ohne Einschränkung der Allgemeinheit angenommen werden kann, dass die Ermittlung von x_{m-1} und x_m durch ein- und dieselbe Frage erfolgt.

(iii)

$$\sum_{i=1}^{m} (\frac{1}{2})^{a_i} = 1\,.$$

Beweis der Aussage (i): Wir gehen von der optimalen Fragestrategie S^* - mit den Anzahlen $a_1, ..., a_k, ..., a_l, ..., a_m$ der Fragen - über zu einer Strategie S' mit $a_1', ..., a_k', ..., a_l', ..., a_m'$, sodass $a_i' = a_i$ für $i \in \{1, ..., m\} \backslash \{k, l\}$ und $a_k' = a_l$, $a_l' = a_k$. Dann gilt für die Differenz der mittleren Anzahl der Fragen

$$\begin{aligned} E_P[A(S') - A(S^*)] &= \sum_{i=1}^{m} (a_i' - a_i)p_i = (a_k' - a_k)p_k + (a_l' - a_l)p_l \\ &= (a_l - a_k)p_k + (a_k - a_l)p_l = (a_l - a_k)(p_k - p_l)\,, \end{aligned}$$

wobei wegen der Optimalität von S^*, $E[A(S')] - E[A(S^*)] \geq 0$ ist. Also gilt $(a_l - a_k)(p_k - p_l) \geq 0$, woraus aufgrund von $p_k > p_l$ $a_l \geq a_k$ folgt. $\square$

Anmerkung 1 (Fortsetzung): Die Eigenschaft (i) ist unmittelbar einleuchtend: Häufige Elemente werden durch wenige Fragen, seltene durch relativ viele Fragen bestimmt. Man denke in diesem Zusammenhang auch an den Morsecode. Dabei wird etwa der in der englischen Sprache häufige Buchstabe "E" durch einen Punkt " $\cdot$ " kodiert, während der seltene Buchstabe "Z" durch die Zeichenfolge " $- - \cdot - \cdot$ " kodiert wird.

Die Eigenschaft (ii) wird bei der Konstruktion des *Huffman Codes* die entscheidende Rolle spielen. (Vgl. dazu Abschnitt 1.3.1.)

Die Eigenschaft (iii) ist eine Verschärfung der *Fano-Kraft'schen Ungleichung,* bei welcher in (iii) anstelle des Gleichheitszeichens das "$\leq$"-Zeichen steht, und die die entscheidende Bedingung für die *Präfix-Freiheit* eines Codes ist. (Vgl. dazu Korollar 1 und die Abschnitte 1.3.2 und 1.3.3.)

Im Folgenden beschäftigen wir uns mit

Schranken für den Erwartungswert der Anzahl der Fragen

Anmerkung 2: Durch die optimale Fragestrategie wird der Wahrscheinlichkeitsverteilung $P = (p_1, ..., p_m)$ gemäß $q_i^* = (\frac{1}{2})^{a_i}$, $i \in \{1, ..., m\}$ eine weitere Wahrscheinlichkeitsverteilung

$$Q^* = (q_1^*, ..., q_m^*)$$

zugeordnet und es gilt

$$E_P\,(A) = \sum_{i=1}^{m} p_i \log_2(\frac{1}{q_i^*})\,.$$

Beobachtungen anhand des obigen Beispiels legen überdies folgende Vermutung nahe

$$(\,\log_2(m) \geq\,)\ \sum_{i=1}^{m} p_i \log_2(\frac{1}{q_i^*}) \geq \sum_{i=1}^{m} p_i \log_2(\frac{1}{p_i})\,,$$

oder, gleichbdeutend,

$$\sum_{i=1}^{m} p_i \log_2(\frac{p_i}{q_i^*}) \geq 0\,.$$

Definition 1: Sei $P = (p_1, ..., p_m)$ eine Wahrscheinlichkeitsverteilung. Dann heißt die Größe

$$H(P) = \sum_{i=1}^{m} p_i \log_2(\frac{1}{p_i}) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

die *Entropie* von P . Dabei wird nachstehende Vereinbarung berücksichtigt

$$0 \times \log_2(0) = \lim_{u \downarrow 0} u \log_2(u) = 0\,.$$

Definition 2: Seien $P = (p_1, ..., p_m)$ und $Q = (q_1, ..., q_m)$ zwei Wahrscheinlichkeitsverteilungen, wobei gelte $q_i > 0 \;\; \forall \; i \in \{1, ..., m\}$. Dann heißt die durch

$$I(P\|Q) = \sum_{i=1}^{m} p_i \log_2(\frac{p_i}{q_i})$$

definierte Größe die *I-Divergenz* von P und Q .

Untersuchung des Spezialfalles $m = 2$ [2]

Die Funktion

$$q \mapsto g(p,q) = -\left(p \ln q + (1-p)\ln(1-q)\right)\;,\; q \in (0,1)$$

nimmt wegen

$$\frac{\partial\, g(p,q)}{\partial\, q} = -\frac{p}{q} + \frac{1-p}{1-q} = \frac{q-p}{q\,(1-q)} \left\{ \begin{array}{lll} < 0 & \text{für} & q < p \\ = 0 & \text{für} & q = p \\ > 0 & \text{für} & q > p \end{array} \right.$$

ihr Minimum für $q = p$ an. Die Funktion

$$p \mapsto h(p) = g(p,p) = -\left(p \ln p + (1-p)\ln(1-p)\right)\;,\; p \in [0,1]$$

des Minimums ist wegen

$$h'(p) = -\ln p + \ln(1-p) = \ln(\frac{1}{p} - 1) \left\{ \begin{array}{lll} > 0 & \text{für} & p < \frac{1}{2} \\ = 0 & \text{für} & p = \frac{1}{2} \\ < 0 & \text{für} & p > \frac{1}{2} \end{array} \right.$$

[2] Im Hinblick auf die Beziehung $\log_2(x) = \ln(x)/\ln(2)$ werden wird dabei und beim Beweis des folgenden Lemmas anstelle des Logarithmus zur Basis 2 den nartürlichen Logarithmus verwenden.

und

$$h''(p) = -\frac{1}{p\,(1-p)} < 0$$

konkav und nimmt ihr Maximum für $p = \frac{1}{2}$ an.

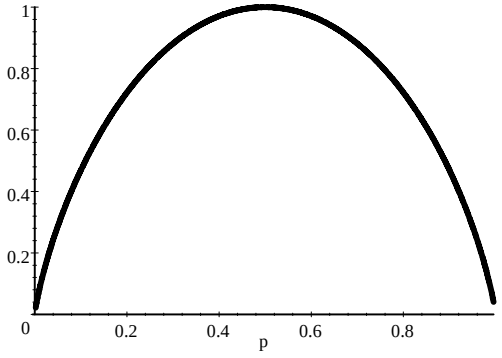


Abbildung 1: Graph der Funktion $h(p)$

Nachdem wir uns von der Gültigkeit unserer Vermutungen für den Spezialfall $m = 2$ überzeugt haben, wenden wir uns nun dem allgemeinen Fall zu. Dessen Behandlung beruht auf folgendem Lemma.

Lemma 1: Seien P und Q zwei Wahrscheinlichkeitsverteilungen auf $\{0, ..., m\}$, wobei dies der Träger von Q ist. Dann gilt

$$I(P \parallel Q) \geq 0 \quad \text{mit Gleichheit genau dann, wenn } \; Q = P \; \text{ist.}$$

Beweis: Bezeichne $T(P)$ den Träger von P. Dann gilt wegen $0 \ln 0 = 0$ und $\ln u \geq 1 - \frac{1}{u}$ - mit Gleichheit genau dann, wenn $u = 1$ ist -

$$\begin{aligned}
\sum_{i=0}^{m} p_i \ln(\frac{p_i}{q_i}) &= \sum_{i \in T(P)} p_i \ln(\frac{p_i}{q_i}) \geq \\
&\geq \sum_{i \in T(P)} p_i (1 - \frac{q_i}{p_i}) = \\
&= \sum_{i \in T(P)} (p_i - q_i) = 1 - \sum_{i \in T(P)} q_i \geq 0 \,.
\end{aligned}$$

Gleichheit gilt also genau dann, wenn $q_i = p_i \quad \forall\, i \in T(Q)$ und $T(Q) = \{0, ..., m\}$ ist. □

Proposition 2: Sei $P = (p_1, ..., p_m)$ eine Wahrscheinlichkeitsverteilung auf $\Omega = \{1, ..., m\}$ und erfüllen die Anzahlen $a_i \in \mathbb{N}$, $i \in \{1, ..., m\}$ der Fragen einer bestimmten Fragestrategie die sogenannte *Fano-Kraft'sche Ungleichung*

$$\sum_{i=1}^{m} (\frac{1}{2})^{a_i} \leq 1 \,. \tag{3}$$

Dann gilt für den Erwartungswert $E_P(A) = \sum_{i=1}^{m} p_i \cdot a_i$ der Anzahl der Fragen

$$E_P(A) \geq H(P) \,,$$

wobei Gleichheit genau dann zutrifft, wenn in (3) Gleichheit gilt und für alle $i \in \{1, ..., m\}$ $p_i = (\frac{1}{2})^{a_i}$ ist.

Beweis: Es seien $Q = (q_0, q_1, ..., q_m)$ die durch

$$q_i = \begin{cases} 1 - \sum_{j=1}^{m} (\frac{1}{2})^{a_j} & \text{für} \quad i = 0 \\ \\ (\frac{1}{2})^{a_i} & \text{für} \quad i \in \{1, ..., m\} \end{cases}$$

definierte Wahrscheinlichkeitsverteilung auf $\Omega_0 = \{0, 1, ..., m\}$ und $P_0 = (0, p_1, ..., p_m)$ die Fortsetzung von P auf Ω_0. Die Behauptung ergibt sich daraus durch Anwendung des obigen Lemmas.

Proposition 3 (Extremaleigenschaften der Entropie): Es gilt

$$0 \leq H_2(P) \leq \log_2(m)$$

mit Gleichheit

(i) im ersten Fall genau dann, wenn P eine Punktverteilung ist, d.h. wenn es ein Element $x_0 \in \Omega$ gibt, sodass $P = (1_{\{x_0\}}(x) : x \in \Omega)$ ist, und
(ii) im zweiten Fall genau dann,wenn $P = P_m$ die Gleichverteilung ist.

Beweis: (i) Wegen $p_i \leq 1$ ist jeder Summand $p_i \log_2(p_i)$ von $-H(P)$ kleiner oder gleich 0 und somit $H(P) \geq 0$, wobei Gleichheit nur dann gilt, wenn ein $p_i = 1$ und alle anderen gleich 0 sind.

(ii) Wendet man das Lemma auf den Spezialfall $Q = P_m = (\frac{1}{m}, ..., \frac{1}{m})$ an, so ergibt sich

$$I(P \parallel P_m) = \sum_{i=0}^{m} p_i \ln(\frac{p_i}{\frac{1}{m}}) = \log_2(m) - H(P) \geq 0$$

mit Gleichheit genau dann, wenn $P = P_m$ ist.

1.2.2 Axiomatik

Seien $m \in \mathbb{N}$ beliebig und $\mathcal{V}_m$ die Menge aller Wahrscheinlichkeitsverteilungen $P = (p_1, ..., p_m)$ auf $\Omega_m = \{x_1, ..., x_m\}$.

Proposition 1: Eine Funktion $H : \mathcal{V}_m \mapsto \mathbb{R}$, welche die Eigenschaften (H0)-(H3) besitzt, hat die Form

$$H(P) = -\sum_{i=1}^{m} p_i \log_2(p_i) \ . \tag{4}$$

(H0) H hängt nur von den Elementarwahrscheinlichkeiten p_i , $i \in \{1, ..., m\}$ von P, nicht jedoch von deren Anordnung ab. D.h. es gilt

$$H((p_{\pi_1}, ..., p_{\pi_m})) = H((p_1, ..., p_m))$$

für alle Permutationen $(\pi_1, ..., \pi_m)$ von $\{1, ..., m\}$.

(H1) Für $p = p_{m-1} + p_m > 0$ gilt

$$H((p_1, ..., p_{m-2}, p_{m-1}, p_m)) = H((p_1, ..., p_{m-2}, p)) + pH((\frac{p_{m-1}}{p}, \frac{p_m}{p})) .$$

(H2) Die Funktion $H((p, 1-p))$, $p \in [0, 1]$ ist stetig.

(H3) $H((\frac{1}{2}, \frac{1}{2})) = 1$.

Anmerkung 1 [3]**:** Es gelten

$$H(1) = 0 \quad \text{und} \quad H(p_1, ..., p_n, 0) = H(p_1, ..., p_n) \ ,$$

was im Folgenden gezeigt und in der Regel stillschweigend berücksichtigt wird.

(H1) angewandt auf den Spezialfall $m = 2$, $p_1 = 0$ und $p_2 = 1$ ergibt

$$H(0, 1) = H(1) + H(0, 1)$$

und somit $H(1) = 0$. (H1) angewandt auf den Spezialfall $m = 3$, $p_1 = 0$, $p_2 = p_3 = \frac{1}{2}$ ergibt unter Berücksichtigung von (H0)

$$\begin{aligned} 0 &= H(0, \frac{1}{2}, \frac{1}{2}) - H(\frac{1}{2}, \frac{1}{2}, 0) = \left(H(0, 1) + H(\frac{1}{2}, \frac{1}{2})\right) - \left(H(\frac{1}{2}, \frac{1}{2}) + \frac{1}{2}H(1, 0)\right) \\ &= \frac{1}{2}H(0, 1) \end{aligned}$$

und somit auch $H(1, 0) = 0$. Berücksichtigt man dies und wendet nochmals (H1) an, so erhält man schließlich

$$\begin{aligned} H(p_1, ...p_{n-1}, p_n, 0) &= H(p_1, ..., p_{n-1}, p_n) + p_n H(1, 0) \\ &= H(p_1, ..., p_{n-1}, p_n) \ . \end{aligned}$$

[3] Im Weiteren wird bei $H((p_1, ..., p_n))$ auf eine Klammer verzichtet.

Lemma 1: Es gilt folgende Verallgemeinerung von (H1): Seien $Q = (q_1, ..., q_m)$ eine Wahrscheinlichkeitsverteilung auf $\Omega_m = \{1, ..., m\}$ und $P^{(i)}$ Wahrscheinlichkeitsverteilungen auf $\{(i-1) \cdot n + 1, ..., i \cdot n\}$[4], $i \in \Omega_m$. Dann ist die Entropie der Mischverteilung $\sum_{i=1}^{m} q_i P^{(i)}$ gleich

$$H(\sum_{i=1}^{m} q_i P^{(i)}) = H(Q) + \sum_{i=1}^{m} q_i H(P^{(i)}) . \tag{5}$$

Anmerkung 2: **(a)** Im Fall, dass Q und die Einschränkungen von $P^{(i)}$ auf die Trägermengen $\{(i-1) \cdot n + 1, ..., i \cdot n\}$ die Gleichverteilungen $Q_m = (\frac{1}{m}, ..., \frac{1}{m})$ bzw. $P_n = (\frac{1}{n}, ..., \frac{1}{n})$ sind, ist $\sum_{i=1}^{m} q_i P^{(i)} = P_{m \cdot n}$ und (5) erhält die Form

$$H(P_{m \cdot n}) = H(P_m) + H(P_n) \ .$$

Daher lässt sich die Eigenschaft (5) in der Notation mittels $g(n) = H(P_n)$ als Verallgemeinerung der Eigenschaft (I1) betrachten.

(b): Für den etwas allgemeineren Fall, dass alle $P^{(i)} = P = (p_1, ..., p_n)$ sind, nimmt (5) die Form

$$H(Q \times P) = H(Q) + H(P)$$

an, wobei $Q \times P = (q_i \cdot p_j : (i,j) \in \{1, ..., m\} \times \{1, ..., n\})$ das Produkt der Verteilungen Q und P ist.

Anmerkung 3: Im Beweis von Proposition 1 (und auch später) wird der Spezialfall für $m = 2$ angewandt, nämlich: Sind $n, k \in \mathbb{N}$, $Q = (q, 1-q)$ eine Wahrscheinlichkeitsverteilung,

$$P^{(1)} = (p_1^{(1)}, ..., p_n^{(1)}, 0, ..., 0) \quad \text{und} \quad P^{(2)} = (0, ..., 0, p_{n+1}^{(2)}, ..., p_{n+k}^{(2)})$$

zwei Wahrscheinlichkeitsverteilungen auf $\{1, ..., n+k\}$ und

$$qP^{(1)} + (1-q)P^{(2)} = (q \cdot p_1^{(1)}, ..., q \cdot p_n^{(1)}, (1-q) \cdot p_{n+1}^{(2)}, ..., (1-q) \cdot p_{n+k}^{(2)})$$

die zugehörige Mischverteilung. Dann gilt

$$H(qP^{(1)} + (1-q)P^{(2)}) = H(Q) + qH(P^{(1)}) + (1-q)H(P^{(2)}) . \tag{6}$$

Unter Zuhilfenahme von (6) ergibt sich mittels vollständiger Induktion

$$H(P_N) = \sum_{i=0}^{N-2} (1 - \frac{i}{N}) H(\frac{1}{N-i}, 1 - \frac{1}{N-i}) = \sum_{n=2}^{N} \frac{n}{N} H(\frac{1}{n}, 1 - \frac{1}{n}) \ \forall N \geq 2 . \tag{7}$$

[4] Anstelle "Wahrscheinlichkeitsverteilungen auf $\{(i-1) \cdot n + 1, ..., i \cdot n\}$" kann es wegen Anmerkung 1 gleichermaßen heißen: "Wahrscheinlichkeitsverteilungen auf $\{1, ..., m \cdot n\}$ mit dem Träger $\{(i-1) \cdot n + 1, ..., i \cdot n\}$".

Beweis von Proposition 1: Wir gehen von folgendem Spezialfall von (6) aus

$$(\frac{1}{n},...,\frac{1}{n}) = \frac{t}{n}(\frac{1}{t},...,\frac{1}{t},0,...,0) + \frac{n-t}{n}(0,...,0,\frac{1}{n-t},...,\frac{1}{n-t}) \; .$$

Dann gilt mit $g(n) = H(P_n)\,,\; n \in \mathbb{N}\,,$

$$g(n) = H(\frac{t}{n}, 1-\frac{t}{n}) + \frac{t}{n}g(t) + (1-\frac{t}{n})g(n-t) \; . \tag{8}$$

Aus dem Spezialfall von (8) für $t = 1$ lässt sich, wie im Beweis von Lemma 2 ausgeführt wird, die Eigenschaft (I2') folgern. Damit sind für $g(n) = H(P_n)$ die Eigenschaften (I1), (I2') und (H3)=(I3) für $b = 2$ erfüllt, weswegen nach Anmerkung 2 in Abschnitt 1.1.2 gilt $g(n) = \log_2(n)$.

Daraus ergibt sich für $H\left(\frac{t}{n}, 1-\frac{t}{n}\right)$

$$\begin{aligned} H(\frac{t}{n}, 1-\frac{t}{n}) &= \log_2(n) - \left(\frac{t}{n}\log_2(t) + (1-\frac{t}{n})\log_2(n-t)\right) \\ &= -\left(\frac{t}{n}\log_2(\frac{t}{n}) + (1-\frac{t}{n})\log_2(1-\frac{t}{n})\right) \end{aligned}$$

und somit

$$H(p, 1-p) = -\left(p\log_2(p) + (1-p)\log_2(1-p)\right) \tag{9}$$

für rationale $p \in (0,1)$. Die Stetigkeitsvoraussetzung (H2) zieht die Gültigkeit von (9) für alle $p \in (0,1)$ nach sich.

Der Beweis von (4) erfolgt mittels vollständiger Induktion nach m : Anwendung von (6) auf die Wahrscheinlichkeitsverteilung $P = (p_1,...,p_m,p_{m+1})$ ergibt nämlich mit $q = p_{m+1} < 1$ und $p = 1-q$

$$\begin{aligned} H(p_1,...,p_m,p_{m+1}) &= H(p,q) + pH(\frac{p_1}{p},...,\frac{p_m}{p},0) + qH(0,...,0,1) \\ &= H(p,q) + pH(\frac{p_1}{p},...,\frac{p_m}{p}) + qH(1) \\ &= pH(\frac{p_1}{p},...,\frac{p_m}{p}) + H(p,q) \\ &= -[p(\sum_{i=1}^{m}\frac{p_i}{p}\log_2(\frac{p_i}{p})) + p\log_2(p) + q\log_2(q)] \\ &= -\sum_{i=1}^{m+1} p_i\log_2(p_i) \;\; . \;\; \square \end{aligned}$$

Lemma 2: Sei $g(n) = H(P_n)\,,\; n \in \mathbb{N}$. Dann gilt

$$\lim_{n\to\infty} g(n) - g(n-1) = 0 \; .$$

Beweis: Der Spezialfall von (8) für $t = 1$ ist wegen $g(1) = 0$

$$g(n) = H(\frac{1}{n}, 1 - \frac{1}{n}) + (1 - \frac{1}{n})g(n-1)$$

oder, gleichbedeutend,

$$g(n) - g(n-1) + \frac{1}{n}g(n-1) = H(\frac{1}{n}, 1 - \frac{1}{n}) \,. \tag{10}$$

Indem man für $n \geq 2$

$$d_n = g(n) - g(n-1) \quad \text{und} \quad \delta_n = H(\frac{1}{n}, 1 - \frac{1}{n}) \quad (\text{wobei } d_2 = \delta_2 = 1 \text{ ist})$$

setzt, lässt sich $g(n-1)$ wegen $g(1) = 0$ als teleskopierende Summe $g(n-1) = \sum_{i=2}^{n-1} d_i$ darstellen, sodass (10) die folgende Form erhält

$$d_n + \frac{1}{n}\sum_{i=2}^{n-1} d_i = \delta_n \,. \tag{11}$$

Da wegen (H2) $\lim_{n\to\infty} \delta_n = H(0,1) = 0$ gilt, garantiert diese Beziehung die Konvergenz $\lim_{n\to\infty} d_n = 0$ - und somit die Behauptung - wenn man die Konvergenz

$$\lim_{N\to\infty} \frac{1}{N+1}\sum_{n=2}^{N} d_n = 0$$

des zugehörigen Cesáro-Mittels zeigt. Dies geschieht wie folgt: Multipliziert man (11) mit n und summiert über alle $n \in \{2, ..., N\}$, so erhält man wegen (7)

$$\sum_{n=2}^{N}(nd_n + \sum_{i=2}^{n-1} d_i) = \sum_{n=2}^{N}(nd_n + d_n(N-n)) = N\sum_{n=2}^{N} d_n = N\,H(P_N) = \sum_{n=2}^{N} n\delta_n \,.$$

Division durch $N(N+1)$ und Berücksichtigung von $\frac{n}{N+1} < 1 \; \forall\, n \in \{2, ..., N\}$ ergibt die Abschätzung

$$\left|\frac{1}{N+1}\sum_{n=2}^{N} d_n\right| = \left|\frac{1}{N}\sum_{n=2}^{N}\frac{n}{N+1}\delta_n\right| \leq \frac{1}{N}\sum_{n=2}^{N}|\delta_n| \;.$$

Da wegen (H2) $\lim_{n\to\infty}|\delta_n| = 0$ gilt, konvergiert die obere Schranke (als Cesáro-Mittel "umsomehr") gegen 0, womit alles gezeigt ist. □

Beweis von Lemma 1: Wir beweisen zunächst folgende Tatsache.

H(1') Sei $P = (p_1, ..., p_n, p_{n+1}, ..., p_{n+k})$ eine Wahrscheinlichkeitsverteilung derart, dass $p_{n+i} > 0 \; \forall\, i \in \{1, ..., k\}$. Dann gilt mit $q = \sum_{i=1}^{k} p_{n+i}$

$$H(p_1, ..., p_n, p_{n+1}, ..., p_{n+k}) = H(p_1, ..., p_n, q) + qH(\frac{p_{n+1}}{q}, ..., \frac{p_{n+k}}{q}) \,.$$

Der Beweis erfolgt durch vollständige Induktion nach k, wobei der Induktionsanfang für $k = 2$ durch (H1) gegeben ist. Der Induktionsschritt von $k \mapsto k+1$ verläuft wie folgt: Anwendung von (H1) mit $p = p_{n+k} + p_{n+k+1} > 0$, Berücksichtigung der Induktionsvoraussetzung und abermalig Anwendung von (H1) liefert mit $r = \sum_{i=1}^{k-1} p_{n+i} + p$

$$\begin{aligned} H(p_1, ..., p_n, p_{n+1}, ..., p_{n+k}, p_{n+k+1}) &= H(p_1, ..., p_n, p_{n+1}, ..., p_{n+k-1}, p) + pH(\frac{p_{n+k}}{p}, \frac{p_{n+k+1}}{p}) \\ &= H(p_1, ..., p_n, r) + r[H(\frac{p_{n+1}}{r}, ..., \frac{p_{n+k-1}}{r}, \frac{p}{r}) + \\ &\quad \frac{p}{r} H(\frac{p_{n+k}}{p}, \frac{p_{n+k+1}}{p})] \\ &= H(p_1, ..., p_n, r) + rH(\frac{p_{n+1}}{r}, ..., \frac{p_{n+k}}{r}, \frac{p_{n+k+1}}{r}) \; . \end{aligned}$$

Der Beweis des Lemmas erfolgt schließlich durch vollständige Induktion nach m. Die Induktion erfolgt mittels (H1') für $k = n$ nachdem (H0) berücksichtigt wird.
□

1.3 AUSSAGEN ZUR QUELLKODIERUNG

1.3.1 Der Huffman-Code

Im Folgenden sei X eine Zufallsvariable mit Verteilung P und zugehörigen Trägermenge Ω. $a(x,S)$ sei die Anzahl der nötigen Fragen, um ein Element $x \in \Omega$ mit Hilfe einer Fragestragtegie S zu bestimmen. Weiters bezeichne $\mathcal{S}_\Omega$ die Menge aller möglichen Fragestrategien.

Definition 1: Die Größe

$$H^*(P) = \min_{S \in \mathcal{S}_\Omega} E_P\left[a\left(X,S\right)\right]$$

heißt die *wirkliche Entropie* von P. Eine Fragestrategie $S^* \in \mathcal{S}_\Omega$ heißt *optimal*, wenn gilt

$$H^*(P) = E_P\left[a\left(X,S^*\right)\right] .$$

Der folgende Satz nützt die Eigenschaft (ii) einer optimalen Fragestrategie (vgl. Proposition 1 in 1.2.1), um aus einer optimalen Fragestrategie für eine geeignete Verteilung auf einer Menge Ω_{m-1} mit $m-1$ Elementen eine solche für eine gegebenen Verteilung auf einer Menge Ω_m mit m zu konstruieren. Aufgrund derselben Eigenschaft benötigt übrigens eine optimale Fragestrategie für jede (nichtentartete) Verteilung auf einer Menge Ω_2 mit 2 Elementen genau eine Frage. Aufgrund dessen lässt sich mit Hilfe vollständiger Induktion nach m
die Existenz einer optimalen Strategie S^* für jede mögliche Verteilung P zeigen.

Darüber hinaus garantiert der Satz auch die rekursive Berechnung der wirklichen Entropie $H^*(P)$.

Satz 1 (Huffman) : Sei $m \geq 3$ und sei X eine Zufallsvariable mit Wertebereich $\Omega_m = \{x_1, ..., x_{m-2}, x_{m-1}, x_m\}$ und Verteilung $P_m = (p_1, ..., p_{m-2}, p_{m-1}, p_m)$, für welche $p_1 \geq ... \geq p_{m-2} \geq p_{m-1} \geq p_m > 0$ gilt.
Sei ferner $x = \{x_{m-1}, x_m\}$, $\Omega_{m-1} = \{x_1, ..., x_{m-2}, x\}$, $p = p_{m-1} + p_m$ und $P_{m-1} = (p_1, ..., p_{m-2}, p)$ die Einschränkung $P_X|\Omega_{m-1}$ von P_m auf Ω_{m-1}.
Dann gilt

(a) $H^*(P_m) = H^*(P_{m-1}) + p$

(b) Aus einer optimalen Strategie $S^*(P_{m-1})$ für P_{m-1} erhält man dadurch eine optimale Strategie $S^*(P_m)$ für P_m, dass man ausgehend von $x = \{x_{m-1}, x_m\}$ durch eine weitere Frage bestimmt, welches der beiden Elemente x_{m-1} oder x_m gewählt wurde.

Beweis:
Seien $a_1, ..., a_{m-2}, a_{m-1}$ die Codewortlängen der optimalen Strategie $S^*(P_{m-1})$ für P_{m-1}, und sei S_m die Strategie auf Ω_m, die aus $S^*(P_{m-1})$ dadurch hervorgeht, dass man - ausgehend von x - zur Ermittlung des tatsächlich gewählten Elements aus $x = \{x_{m-1}, x_m\}$ eine weitere Frage stellt. Dann besitzt S_m die Codewortlängen $a_1, ..., a_{m-2}, a_{m-1} + 1, a_{m-1} + 1$, und es gilt aufgrund der Definition von $H^*(P_m)$

$$\begin{aligned} H^*(P_m) &\leq E[a(X, S_m)] \\ &= \sum_{i=1}^{m-2} p_i \cdot a_i + p_{m-1} \cdot (a_{m-1} + 1) + p_m \cdot (a_{m-1} + 1) \\ &= \sum_{i=1}^{m-2} p_i \cdot a_i + p \cdot a_{m-1} + p = H^*(P_{m-1}) + p\,. \end{aligned}$$

Seien umgekehrt $b_1, ..., b_{m-2}, b_{m-1}, b_m$ die Codewortlängen der optimalen Strategie $S^*(P_m)$. Dann gilt wegen Proposition 1(ii) in 1.2.1 $b_{m-1} = b_m$ und dass die Ermittlung des gewählten Elements aus $\{x_{m-1}, x_m\}$ durch ein- und dieselbe Frage erfolgt.

Sei ferner S_{m-1} die Strategie auf Ω_{m-1}, die aus $S^*(P_m)$ dadurch hervorgeht, dass lediglich die Frage zur Ermittlung des gewählten Elements aus $x = \{x_{m-1}, x_m\}$ unterbleibt.

Dann besitzt S_{m-1} die Codewortlängen $b_1, ..., b_{m-2}, b_{m-1} - 1$ und es gilt aufgrund der Definition von $H^*(P_{m-1})$

$$\begin{aligned} H^*(P_{m-1}) &\leq E[b(X \mid \Omega_{m-1}, S_{m-1})] \\ &= \sum_{i=1}^{m-2} p_i \cdot b_i + (p_{m-1} + p_m) \cdot (b_{m-1} - 1) = \sum_{i=1}^{m} p_i \cdot b_i - p \\ &= H^*(P_m) - p\,. \end{aligned}$$

Zusammenfassend gilt

$$H^*(P_m) \leq E[a(X, S_m)] = H^*(P_{m-1}) + p \leq H^*(P_m)\,.$$

Somit ist die aus $S^*(P_{m-1})$ konstruierte Strategie S_m optimal für P_m. □

Beispiel 1:

$$\begin{aligned} \Omega_7 &= \{\ x_1,\ x_2,\ x_3,\ x_4,\ x_5,\ x_6,\ x_7\ \} \\ P_7 &= (\ 20,\ 20,\ 18,\ 17,\ 15,\ 6,\ 4\ )\ /\ 100 \end{aligned}$$

Abbildung 1: Rekursiver Aufbau des Huffman Codes (bitte einfügen)

Mittlere Codewortlänge des Hufman Codes

$$H^*(P) = (10 + 25 + 35 + 40 + 60 + 100) \,/\, 100 = 2.7$$

Abbildung 2: Codebaum der Huffman Codes (bitte einfügen)

Codewörter des Huffman Codes

$$\begin{array}{rcl} \varkappa(x_1) & = & (0,0) \\ \varkappa(x_2) & = & (0,1) \\ \varkappa(x_3) & = & (1,0,0) \\ \varkappa(x_3) & = & (1,0,1) \\ \varkappa(x_3) & = & (1,1,0) \\ \varkappa(x_6) & = & (1,1,1,0) \\ \varkappa(x_7) & = & (1,1,1,1) \end{array}$$

Abschließend nehmen wir noch einige Begriffsklärungen hinsichtlich Codebäumen vor.

Definition 2: Ein orientierter, zusammenhängender Graph heißt *(binärer) Codebaum*, wenn gelten

(a) Es gibt genau eine Ecke (die *Wurzel*), zu der keine Kante führt

(b) Von jeder Ecke gehen höchstens zwei Kanten aus.

Ecken ohne ausgehende Kanten heißen *Blätter*, alle anderen Ecken nennt man *innere Punkte*. Die *Tiefe* $a(x)$ eine Blattes x ist die Anzahl der inneren Punkte auf dem Pfad, der von der Wurzel zu x führt. Man nennt einen Codebaum *vollständig*, wenn von jedem inneren Punkt zwei Kanten ausgehen. Innere Punkte, von denen genau zwei Kanten ausgehen, nennt man *Gabeln*.

Anmerkung 1: Es gelten folgende Entsprechungen

Innerer Punkt ... Frage

Ausgehende Kante ... Antwort

Blatt ... Element der Grundmenge (möglicher Ausgang des entsprechenden Versuches)

Tiefe eines Blattes ... Anzahl der Fragen, die gestellt werden, um das zugehörige Element zu bestimmen

Ein Codebaum, der eine optimale Fragestrategie beschreibt (ein sogenannter *optimaler Codebaum)* ist vollständig.

1.3.2 Die Fano-Kraft'sche Ungleichung

Satz 1: Genügen die Zahlen $a_1, a_2, ..., a_m \in \mathbb{N}$ der *Fano-Kraft*'schen Ungleichung

$$\sum_{i=1}^{m} (\frac{1}{2})^{a_i} \leq 1 \,,$$

so gibt es einen zugehörigen binären Codebaum.

Beweis: Seien

$j_0 = \min\{a_i : i \in \{1, ..., m\}\}$ und $j_1 = \max\{a_i : i \in \{1, ..., m\}\}$

die minimale bzw. maximale Codewortelänge und

$$w_j = |\{i \in \{1, ..., m\} : a_i = j\}|$$

die Anzahl der Codewörter der Länge $j \in \{j_0, ..., j_1\}$. Dann gilt

$$\sum_{i=1}^{m} (\frac{1}{2})^{a_i} = \sum_{j=j_0}^{j_1} w_j (\frac{1}{2})^j \,,$$

somit wegen (*) und $w_j \geq 0$

$$\sum_{j=j_0}^{k} w_j (\frac{1}{2})^j \leq 1 \quad \forall\, k \in \{j_0, ..., j_1\}$$

und durch Multiplikation mit 2^k und Umformung schließlich

$$w_k \leq 2^k - 2^{k-j_0} \cdot w_{j_0} - 2^{k-(j_0+1)} \cdot w_{j_0+1} - \ldots - 2^1 \cdot w_{k-1} \quad \forall\, k \in \{j_0, ..., j_1\} \,.$$

Nun zur Konstruktion des binären Codebaumes: Zunächst baut man einen binären Codebaum bis zur Tiefe j_0 auf. Von den 2^{j_0} Ecken besetzt man

$$w_{j_0} \; (\leq 2^{j_0})$$

mit Codewörten. Sofern $j_1 > j_0$ ist, setzt man auf den restlichen $2^{j_0} - w_{j_0}$ Ecken den Aufbau des Codebaumes durch Errichtung von Gabeln fort. Auf diese Weise erhält man

$$2 \cdot (2^{j_0} - w_{j_0}) = 2^{j_0+1} - 2 \cdot w_{j_0}$$

Ecken der Tiefe $j_0 + 1$. Von diesen besetzt man

$$w_{j_0+1} \; (\leq 2^{j_0+1} - 2 \cdot w_{j_0})$$

mit Codewörtern. Für die restlichen $2^{j_0+1} - 2 \cdot w_{j_0} - w_{j_0+1}$ Ecken setzt man den Aufbau des Codebaumes fort, bis schließlich die Tiefe j_1 des Baumes erreicht und alle

$$w_{j_1} \; (\leq 2^{j_1} - 2^{j_1-j_0} \cdot w_{j_0} - 2^{j_1-(j_0+1)} \cdot w_{j_0+1} - \ldots - 2^1 \cdot w_{j_1-1})$$

verbleibenden Ecken der Tiefe j_1 durch Codewörter besetzt sind. $\square$

Die in der folgenden Proposition angegebene obere Schrank für die wirkliche Entropie $H^*(P)$ ist eine unmittelbare Folgerung des obigen Satzes und der elementaren Beziehung $x \leq \lceil x \rceil < x + 1$.

Proposition 1: Es gilt

$$H^*(P) < H(P) + 1 .$$

Beweis: Wegen

$$1 = \sum_{i=1}^{m} p_i = \sum_{i=1}^{m} (\frac{1}{2})^{\log_2(\frac{1}{p_i})} \geq \sum_{i=1}^{m} (\frac{1}{2})^{\lceil \log_2(\frac{1}{p_i}) \rceil}$$

gilt die *Fano-Kraft'*sche Ungleichung für $a_i' = \lceil \log_2 \left(\frac{1}{p_i} \right) \rceil$, $i \in \{1, ..., m\}$. Somit gibt es einen Code $S'(P)$ für die Verteilung P mit den Codewortlängen $a_i' = a(x_i, S'(P))$. Für diesen gilt aufgrund der Definition von $H^*(P)$

$$\begin{aligned} H^*(P) &= \inf_{S \in \mathcal{S}} \sum_{i=1}^{m} p_i \cdot a(x_i, S(P)) \leq \sum_{i=1}^{m} p_i \cdot a(x_i, S'(P)) \\ &= \sum_{i=1}^{m} p_i \cdot \lceil \log_2(\frac{1}{p_i}) \rceil < \sum_{i=1}^{m} p_i \cdot (\log_2(\frac{1}{p_i}) + 1) \\ &= H(P) + 1 . \ \square \end{aligned}$$

Indem wir Proposition 2 in 1.2.1 und die obige Proposition zusammenfassen, erhalten wir folgenden Satz.

Satz 2: Es gilt

$$H(P) \leq H^*(P) < H(P) + 1 .$$

Zusammen mit Anmerkung 2(b) in 1.2.2 ergibt sich daraus die nachstehende Folgerung.

Korollar 1: Für die Folge $(P^n)_{n \in \mathbb{N}}$ der n-ten Potenzen der Verteilung P gilt

$$\lim_{n \to \infty} \frac{1}{n} H^*(P^n) = H(P) .$$

Beweis: Anmerkung 2(b) in 1.2.2 nimmt für $Q = P^{n-1}$ die Form

$$H(P^n) = H(P^{n-1} \times P) = H(P^{n-1}) + H(P)$$

an. Mit Hilfe dieses Rekursionsschritts ergibt sich durch vollständige Induktion $H(P^n) = nH(P)$. Anwendung von Satz 2 auf P^n und Division durch n ergibt somit

$$H(P) \leq \frac{1}{n} H^*(P^n) < H(P) + \frac{1}{n},$$

woraus sich die Behauptung durch Grenzübergang ergibt.

Interpretation: Dieser Sachverhalt lässt folgende Interpretation der Entropie einer Verteilung P zu: $H(P)$ ist - für große n - ungefähr gleich dem n-ten Teil des Erwartungswerts der Anzahl der Fragen, die bei optimaler Fragestrategie nötig sind, um ein gemäß der Verteilung P^n erzeugtes Element $(y_1, ..., y_n) \in \Omega^n$ zu bestimmen.

1.3.3 Arithmetische Kodierung

Die folgende konstruktive Alternative zu Formulierung und Beweis von Satz 1 in 1.3.2 geht auf eine eigene Skizze aus dem Jahr 1992 zurück. Nach der Ausarbeitung stellte sich jedoch heraus, dass *J. Karush* bereits im Jahr 1961 eine ähnliche Vorgangsweise gewählt hatte.

Formulierung und Beweis des nachstehenden Satzes beruhen auf der Tatsache, dass jeder Folge $\alpha = (\alpha_1, \ldots, \alpha_m) \in \{0, \ldots, D-1\}^m$ gemäß

$$a = \sum_{i=1}^{m} \alpha_i \cdot D^{-i}$$

genau ein Element aus $\{0, \ldots, D^m - 1\} \cdot D^{-m} \subseteq [0, 1)$ entspricht, womit auch das Intervall $[a, a + D^{-m})$ eine Teilmenge von $[0, 1)$ ist.

Definition 1: Seien $\alpha = (\alpha_1, \ldots, \alpha_m) \in \{0, \ldots, D-1\}^m$; $\beta = (\beta_1, \ldots, \beta_n) \in \{0, \ldots, D-1\}^n$. Dann heißt α *Präfix* von β, wenn gilt

$$m \leq n \quad \text{und} \quad (\alpha_1, \ldots, \alpha_m) = (\beta_1, \ldots, \beta_m).$$

Das folgende Lemma stellt die Argumente für den Satz bereit.

Lemma 1 (Charakterisierung der Präfixeigenschaft): Es gelten

$$\begin{array}{lllll} (i) & \alpha \text{ ist Präfix von } & \beta & \Longrightarrow & [b, b + D^{-n}) \subseteq [a, a + D^{-m}), \\ (ii) & \alpha \text{ ist nicht Präfix von } & \beta & \text{und} & \\ & \beta \text{ ist nicht Präfix von } & \alpha & \Longrightarrow & [b, b + D^{-n}) \cap [a, a + D^{-m}) = \emptyset. \end{array}$$

Beweis: (i) : Sei $b = \sum_{i=1}^{n} \beta_i \cdot D^{-i}$. Dann gilt voraussetzungsgemäß und wegen $\beta_i \geq 0$

$$b - a = \sum_{i=m+1}^{n} \beta_i \cdot D^{-i} \geq 0$$

und mithin $b \geq a$. Zudem gilt wegen $\beta_i \leq D-1$ und $\sum\limits_{i=m+1}^{\infty} (D-1) \cdot D^{-i} = D^{-m}$

$$b - a = \sum_{i=m+1}^{n} \beta_i \cdot D^{-i} \leq \sum_{i=m+1}^{n} (D-1) \cdot D^{-i} = D^{-m} - D^{-n}$$

und mithin $b + D^{-n} \leq a + D^{-m}$.
(ii): Sei $i_0 = \min\{i \geq 1 : \alpha_i \neq \beta_i\}$. Dann gilt für den Fall $\beta_{i_0} > \alpha_{i_0}$ wegen $\beta_i \geq 0$, $\alpha_i \leq D-1$ und wiederum $\sum\limits_{i=m+1}^{n} (D-1) \cdot D^{-i} = D^{-m} - -D^{-n}$

$$b - a \geq (\beta_{i_0} - \alpha_{i_0}) \cdot D^{-i_0} - \sum_{i=i_0+1}^{m} (D-1) \cdot D^{-i} = (\beta_{i_0} - \alpha_{i_0} - 1) \cdot D^{-i_0} + D^{-m} \geq D^{-m}$$

und daher $b \geq a + D^{-m}$. Für den Fall $\alpha_{i_0} > \beta_{i_0}$ erhalt man ganz analog $a \geq b + D^{-n}$. $\square$

Satz 1 (Alternative zu Satz 1 in 1.3.2): Seien $D \in \mathbb{N} \setminus \{1\}$ und $\Omega = \{x_1, \ldots, x_m\}$. Das zugehörige Tupel $(n_1, \ldots, n_m) \in \mathbb{N}^m$ sei monoton nicht fallend und erfülle die *Fano-Kraft'sche Ungleichung*

$$(*) \qquad \sum_{k=1}^{m} D^{-n_k} \leq 1 \, .$$

Sei ferner $k \in \{1, \ldots, m\}$. Dann besitzt

$$(**) \qquad \sum_{i=1}^{n_k} \varepsilon_{ki} \cdot D^{-i} = \sum_{j=1}^{k-1} D^{-n_j}$$

genau eine Lösung $C(x_k) = (\varepsilon_{k1}, \ldots, \varepsilon_{kn_k}) \in \{0, \ldots, D-1\}^{n_k}$ und $(C(x_1), \ldots, C(x_m))$ ist ein präfixfreier Code mit den Codebuchstaben $0, \ldots, D-1$ und den Codewortlängen $n_1, \ldots, n_m$.

Umgekehrt erfüllt jeder solche Code die Fano-Kraft'sche Ungleichung.

Beweis: Sei $S_k = \sum\limits_{j=1}^{k-1} D^{-n_j}$, $k \in \{1, \ldots, m+1\}$. Wegen $n_1 \leq \ldots \leq n_{k-1} \leq n_k$ ist jeder Summand der Summe S_k - und damit die Summe selbst - ein Vielfaches von D^{-n_k} . Für $k \in \{1, ..., m\}$ gilt wegen $(*)$ $S_k < S_{m+1} \leq 1$. Aufgrund dessen besitzt $(**)$ für jedes $k \in \{1, ..., m\}$ genau eine Lösung

$$C(x_k) = (\varepsilon_{k1}, \ldots, \varepsilon_{kn_k}) \in \{0, \ldots, D-1\}^{n_k} \, .$$

Wegen $0 = S_1 < S_2 < \cdots < S_m < S_{m+1} \leq 1$ definieren die Intervalle $[S_k, S_{k+1})$, $k \in \{1, \ldots, m\}$ eine Partition von $[0, S_{m+1}) \subseteq [0, 1)$. Somit gilt

$$[S_j, S_{j+1}) \cap [S_k, S_{k+1}) = \emptyset \quad \text{für} \quad j \neq k \, .$$

Aufgrund von (i) des Lemmas kann demnach weder $C(x_j)$ Präfix von $C(x_k)$ noch $C(x_k)$ Präfix von $C(x_j)$ sein.

Nun zur Umkehrung: Da gemäß (ii) des Lemmas alle den Codewörtern zugeordneten Intervalle $[a_j, a_j + D^{-n_j}) \subseteq [0,1)$ disjunkt sind und somit

$$\bigcup_{j=1}^{m} [a_j, a_j + D^{-n_j}) \subseteq [0,1)$$

ist, gilt für die Summe der Längen dieser Intervalle $(*)$. $\square$

Beispiel für die Anwendung von Satz 1 ($D = 2$ und $m = 8$):

Angabe:

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	
n_k	2	2	3	3	4	4	4	4	
	2^{-2}	$+2^{-2}$	$+2^{-3}$	$+2^{-3}$	$+2^{-4}$	$+2^{-4}$	$+2^{-4}$	$+2^{-4}$	$= \quad 1$

Auswertung der Summen S_k :

k	1	2	3	4	5	6	7	8
S_k	0	2^{-2}	2^{-1}	2^{-1}	2^{-1}	2^{-1}	2^{-1}	2^{-1}
				$+2^{-3}$	$+2^{-2}$	$+2^{-2}$	$+2^{-2}$	$+2^{-2}$
						$+2^{-4}$	$+2^{-3}$	$+2^{-3}$
								$+2^{-4}$
n_k	2	2	3	4	4	4	4	4

Code: Die Codewörter sind daher

$C(x_1)$	$C(x_2)$	$C(x_3)$	$C(x_4)$	$C(x_5)$	$C(x_6)$	$C(x_7)$	$C(x_8)$
0	0	1	1	1	1	1	1
0	1	0	0	1	1	1	1
		0	1	0	0	1	1
				0	1	0	1

Schließlich wird ein Mathematica-Programm zur Konstruktion eines *Präfix-freien Codes* mit vorgegebenen Codewortlängen angegeben.

1.4 AUSBLICK: ENTROPIE UND WÄRMELEHRE

1.4.1 Boltzmanns Entropiebegriff

nach *Boltzmann*s Arbeit [1]

"Über die Beziehung zwischen dem zweiten Hauptsatz der Wärmetheorie und der Wahrscheinlichkeitsrechnung respektive den Sätzen über das Wärmegleichgewicht"

In der genannten Arbeit geht es zunächst darum, die Größe

$$\frac{n!}{n_1! \cdot \ldots \cdot n_m!} \quad \text{für } n \in \mathbb{N} \text{ und } n_1, ..., n_m \in \mathbb{N}_0 \, : \, \sum_{i=1}^{m} n_i = n$$

mit Hilfe der *Stirling*'schen Formel

$$n! \sim n^n \cdot e^{-n} \cdot \sqrt{2\pi n}$$

für große n und n_i zu approximieren. Anwendung derselben ergibt

$$\frac{n!}{n_1! \cdot \ldots \cdot n_m!} \cong \left[\prod_{i=1}^{m} (\frac{n_i}{n})^{n_i} \cdot (2\pi)^{(m-1)/2} \cdot (\frac{n_1 \cdot \ldots \cdot n_m}{n})^{1/2} \right]^{-1}$$

und somit für den Logarithmus

$$\ln(\frac{n!}{n_1! \cdot \ldots \cdot n_m!}) \cong -n \left[\sum_{i=1}^{m} \frac{n_i}{n} \cdot \ln(\frac{n_i}{n}) + \frac{1}{2n} \cdot \Big((m-1)\ln(2\pi) + \ln(\frac{n_1 \cdot \ldots \cdot n_m}{n}) \Big) \right] .$$

Daraus folgt schließlich in erster Näherung

$$\ln(\frac{n!}{n_1! \cdot \ldots \cdot n_m!}) \cong -n \cdot \sum_{i=1}^{m} \frac{n_i}{n} \cdot \ln(\frac{n_i}{n}) = n \cdot H(Q) \, ,$$

wobei $H(Q)$ die (jetzt mittels des natürlichen Logarithmus definierte) Entropie der Verteilung $Q = (q_i = \frac{n_i}{n}, \, i \in \{1, ..., m\})$ ist.

Anmerkung 1: Sei $P_m = \left(\frac{1}{m}, ..., \frac{1}{m}\right)$ die Gleichverteilung auf $\{1, ..., m\}$, dann erhält man den Logarithmus der Wahrscheinlichkeit

$$W = W(X_1 = n_1, \cdots, X_m = n_m) = \frac{n!}{n_1! \cdot \ldots \cdot n_m!} \cdot (\frac{1}{m})^n$$

eines gemäß einer Multinomialverteilung M_{n,P_m} mit den Parametern n und P_m verteilten Zufallsvektors $(X_1, ..., X_m)$ aus der obigen Näherung

$$\ln W \cong n \cdot (H(Q) - \ln(m)) \; (= -n \cdot I\,(Q \| P_m)\,) \, .$$

Daraus lässt sich unschwer die von *Max Planck* in [2] formulierte Beziehung

$$\boxed{S = k \cdot \ln W + const.}$$

für den Zusammenhang der Entropie S und der Wahrscheinlichkeit W erkennen, wobei k die *Boltzmannkonstante* ist.

1.4.2 Das Maximum-Entropie-Prinzip

In *Boltzmann*s Arbeit geht es vorwiegend darum, die Wahrscheinlichkeit W und somit in erster Näherung die Entropie $H(Q)$ einer Verteilung $Q = (q_i\ ,\ i \geq 1)$ bei vorgegebenem Erwartungswert

$$\sum_{i\geq 1} q_i \cdot i = \mu \tag{12}$$

mit $\mu > 1$ größtmöglich zu machen. Der Bequemlichkeit halber wählen wir anstelle des endlichen Wertebereiches $\{1, ..., m\}$ den Wertebereich $\mathbb{N}$ der natürlichen Zahlen.

Boltzmann bedient sich der Methode der *Langrange*'schen Multiplikatoren, bei welcher neben (12) auch die Nebenbedingung

$$\sum_{i\geq 1} q_i = 1 \tag{13}$$

zu berücksichtigen ist. Daher ist die Zielfunktion

$$H(Q, \lambda, s) = -\sum_{i\geq 1} q_i \cdot \ln(q_i) + \lambda \cdot (\sum_{i\geq 1} q_i - 1) - s \cdot (\sum_{i\geq 1} q_i \cdot i - \mu)$$

mit $\lambda, s \in \mathbb{R}$, deren partiellen Ableitung der nach q_i

$$\frac{\partial H(Q, \lambda, s)}{\partial q_i} = -\ln(q_i) - 1 + \lambda - s \cdot i$$

ist. Nullsetzen und Auflösen nach q_i ergibt

$$q_i = e^{\lambda - 1} \left(e^{-s}\right)^i = d \cdot q^i\ ,$$

wobei $q = e^{-s}$ und $d = e^{\lambda - 1}$ gesetzt wird. Aufgrund von $\lambda, s \in \mathbb{R}$ und der Positivität der Exponentialfunktion gilt zunächst $q_i > 0\ \forall\, i \in \mathbb{N}$. Wegen (13) muss $s > 0$ und $d = p \cdot q^{-1}$ mit $p = 1 - q$ gelten. Daher ist die gesuchte Lösung die Geometrischen Verteilung

$$Q^* = \left(pq^{i-1} :\ i \in \mathbb{N}\right)$$

mit dem Parameter $p = 1/\mu$. Letzteres weil der Erwartungswert von Q^* bekanntlich $1/p$ ist und die für die vorliegende Extremalaufgabe charakteristische Nebenbedingung (12) gilt.

Tatsächlich gilt die folgende, in gewissem Sinn zu Proposition 2 in 1.2.1 duale Aussage, welche wir mit Hilfe der I-Divergenz nachprüfen.

Proposition 1: Für die geometrische Verteilung $Q^* = \left(pq^{i-1}\ ,\ i \in \mathbb{N}\right)$ mit dem Parameter $p = 1/\mu < 1$ gilt

$$\begin{aligned} H(Q^*) &= \max\{H(Q) : Q \in \mathcal{P}(\mathbb{N})\ ,\ \sum_{i=1}^{\infty} q_i \cdot i = \mu\} \\ &= \mu \ln \mu - (\mu - 1) \ln (\mu - 1)\ . \end{aligned}$$

Beweis: Wegen $\sum_{i\geq 1} q_i^* = \sum_{i\geq 1} q_i = 1$ und $\sum_{i\geq 1} q_i^* \cdot i = \sum_{i\geq 1} q_i \cdot i = \mu$ gilt

$$\begin{aligned}
\sum_{i\geq 1} q_i^* \cdot \ln(q_i^*) &= \sum_{i\geq 1} q_i^* \cdot (\ln(pq^{-1}) + i \cdot \ln(q)\ ) \\
&= (\sum_{i\geq 1} q_i^*) \cdot \ln(pq^{-1}) + (\sum_{i\geq 1} i \cdot q_i^*) \cdot \ln(q) \\
&= (\sum_{i\geq 1} q_i) \cdot \ln(pq^{-1}) + (\sum_{i\geq 1} i \cdot q_i) \cdot \ln(q) \\
&= \sum_{i\geq 1} q_i \cdot \ln(q_i^*)\,.
\end{aligned}$$

Somit ist

$$\begin{aligned}
H(Q^*) - H(Q) &= \sum_{i\geq 1} q_i \ln(q_i) - \sum_{i\geq 1} q_i^* \ln(q_i^*) \\
&= \sum_{i\geq 1} q_i \ln(q_i) - \sum_{i\geq 1} q_i \ln(q_i^*) = \sum_{i\geq 1} q_i \ln(\frac{q_i}{q_i^*}) \\
&= I(Q||Q^*)
\end{aligned}$$

und wegen $I(Q||Q^*) \geq 0$ daher

$$H(Q) \leq H(Q^*)$$

mit Gleichheit genau dann, wenn $Q = Q^*$ ist. Der maximale Wert $H(Q^*)$ der Entropie ist aufgrund des Obigen und $p = 1/\mu = 1 - q$

$$\begin{aligned}
H(Q^*) &= \sum_{i\geq 1} q_i^* \cdot \ln(\frac{q}{p}) + \sum_{i\geq 1} i\, q_i^* \cdot \ln(\frac{1}{q}) \\
&= \ln(\frac{q}{p}) + \mu \ln(\frac{1}{q}) \\
&= \mu \ln \mu - (\mu - 1) \ln(\mu - 1)\ . \qquad \square
\end{aligned}$$

1.4.3 Gibbsverteilungen

Satz 1 (Satz über die Maximum-Entropie-Verteilung): Seien

$\Omega = \{1, ..., m\}$ eine endliche Grundmenge

$\mathcal{V} = \{\, Q = (q_1, ..., q_m) \in [0,1]^m : \sum_{i=1}^m q_i = 1 \,\}$ die Menge der Wahrscheinlichkeitsverteilungen über Ω

$u : \Omega \to [0, \infty)$ eine streng monoton wachsende Energiefunktion

$e \in (u_1, u_m)$

$\mathcal{V}_e = \{\, Q \in \mathcal{V} : E_Q(U) = e \,\}$

Weiters sei

$$\mathcal{G} = \{\, Q^*(s) = (q_1^*(s), ..., q_m^*(s)),\, s \in \mathbb{R} \;\text{ mit }\; q_i^*(s) = \frac{e^{-su_i}}{Z(s)},\, i \in \{1, ..., m\}\,\}\,,$$

die Menge der *Boltzmann-Gibbs-Verteilungen.* Dabei ist $Z(s) = \sum_{i=1}^m e^{-su_i}$ die sogenannte *Zustandssumme.*

Dann gelten

(a) $\exists \mid s_e \in \mathbb{R} : Q^*(s_e) \in \mathcal{V}_e$

(b) $H(Q) \leq H(Q^*(s_e))\ \forall Q \in \mathcal{V}_e$ mit Gleichheit genau dann, wenn $Q = Q^*(s_e)$ ist

(c) $H(Q^*(s_e)) = \ln(Z(s_e)) + s_e \cdot e$

Beweis: (a) Es sei weiters $\bar{u} = \frac{1}{m}\sum_{i=1}^m u_i$. Dann gelten

$\lim_{s\nearrow\infty} Q^*(s) = \big(1_{\{1\}}(i),\, i \in \{1, ..., m\}\big)$	und	$\lim_{s\nearrow\infty} E_{Q^*(s)}(U) = u_1$
$Q^*(0) = \big(\frac{1}{m}, ..., \frac{1}{m}\big)$	und	$E_{Q^*(0)}(U) = \bar{u}$
$\lim_{s\searrow-\infty} Q^*(s) = \big(1_{\{m\}}(i),\, i \in \{1, ..., m\}\big)$	und	$\lim_{s\searrow-\infty} E_{Q^*(s)}(U) = u_m$

Wir gehen im Weiteren von der Zielfunktion

$$s \mapsto \ln(Z(s)) + s \cdot e$$

aus. Deren erste und zweite Ableitungen sind

$$\frac{d}{ds}\left[\ln\left(Z\left(s\right)\right) + s \cdot e\right] = e - \frac{\sum_{i=1}^m u_i \cdot e^{-su_i}}{Z(s)} = e - E_{Q^*(s)}(U)$$

und

$$\begin{aligned}\frac{d^2}{ds^2}\left[\ln\left(Z\left(s\right)\right) + s \cdot e\right] &= \frac{\sum_{i=1}^m u_i^2 \cdot e^{-su_i}}{Z(s)} - \frac{\left(\sum_{i=1}^m u_i \cdot e^{-su_i}\right)^2}{Z^2(s)} \\ &= E_{Q^*(s)}(U^2) - \big(E_{Q^*(s)}(U)\big)^2 \\ &= V_{Q^*(s)}(U) > 0\,.\end{aligned}$$

Da aufgrund des letzteren $E_{Q^*(s)}(U)$ streng monoton fallend ist, gibt es genau ein $s = s_e \in \mathbb{R}$, sodass gilt

$$E_{Q^*(s_e)}(U) = e\,.$$

Damit ist (a) bewiesen.

(b) Seien $Q \in \mathcal{V}_e$ und $s_e \in \mathbb{R}$ derart, dass $Q^*(s_e) \in \mathcal{V}_e$.

Dann gilt wegen $I\left(Q \parallel Q^*\left(s_e\right)\right) \geq 0\,,\ Q^*\left(s_e\right) \in \mathcal{V}_e$ und $Q, Q^*\left(s_e\right) \in \mathcal{V}_e$

$$\begin{aligned}
H(Q) &= \sum_{i\geq 1} q_i \ln(\frac{1}{q_i}) \\
&= \sum_{i\geq 1} q_i \ln(\frac{q_i^*(s_e)}{q_i} \cdot \frac{1}{q_i^*(s_e)}) \\
&= -I\left(Q \parallel Q^*\left(s_e\right)\right) + \sum_{i\geq 1} q_i \ln(\frac{1}{q_i^*(s_e)}) \\
&\leq \sum_{i\geq 1} q_i \ln(\frac{1}{q_i^*(s_e)}) \\
&= (\sum_{i\geq 1} q_i) \ln\left(Z(s_e)\right) + s_e \cdot (\sum_{i\geq 1} u_i \cdot q_i) \\
&= (\sum_{i\geq 1} q_i^*(s_e)) \ln\left(Z(s_e)\right) + s_e \cdot (\sum_{i\geq 1} u_i \cdot q_i^*(s_e)) \\
&= \sum_{i\geq 1} q_i^*(s_e) \ln(\frac{1}{q_i^*(s_e)}) \\
&= H(Q^*(s_e))\,,
\end{aligned}$$

wobei Gleichheit genau dann gilt, wenn $Q = Q^*(s_e)$ ist.

(c) Aus der dritten Zeile von unten ergibt sich zudem

$$H(Q^*(s_e)) = \ln\left(Z(s_e)\right) + s_e \cdot e\,. \ \square$$

Aus dem Beweis von Teil (a) ergibt sich folgende Tatsache.

Korollar 1: Die Zielfunktion $s \mapsto \ln(Z(s)) + s \cdot e$ ist strikt konvex und nimmt für $s = s_e$ ihr Minimum $H_{\max}(e)$ an. Es gilt somit $\forall\, s \in \mathbb{R}$

$$\begin{aligned}
\ln(Z(s)) + s \cdot e &\geq \ln(Z(s_e)) + s_e \cdot e \\
&= H_{\max}(e)
\end{aligned}$$

mit Gleichheit genau dann, wenn $s = s_e$ ist.

Bekanntlich ist eine konvexe Funktion Hüllkurve ihrer Tangenten und - gegebenenfalls - ihrer Asymptoten. Für die vorliegenden Zielfunktion gilt folgende Behauptung.

Proposition 1: Seien $0 \leq u_1 < ... < u_m$ und $e \in (u_1, u_m)$. Dann ist die Zielfunktion

$$\ln\left(Z\left(s\right)\right) + s \cdot e \geq \left\{ \begin{array}{l} s \cdot (e - u_1) \\ \ln(m) + s \cdot (e - \bar{u}) \\ s \cdot (e - u_m) \end{array} \right. \quad \forall\, s \in \mathbb{R}\,,$$

wobei $(e-u_1)\cdot s$ für $s>0$ und $(e-u_m)\cdot s$ für $s<0$ deren Asymptoten sind.

Beweis: Wegen $Z(s) = e^{-s\cdot u_1}\cdot[1+\sum_{i=2}^{m} e^{-s\cdot(u_i-u_1)}]$, $u_i-u_1>0 \quad \forall\, i\in\{2,...,m\}$ und $\ln(1+x)\le x$ gilt für $s>0$

$$\begin{aligned}\ln(Z(s))+s\cdot u_1 &= \ln(1+\sum_{i=2}^{m} e^{-s\cdot(u_i-u_1)})\\ &\le \ln(1+(m-1)\cdot e^{-s\cdot(u_2-u_1)})\\ &\le (m-1)\cdot e^{-s\cdot(u_2-u_1)}\,.\end{aligned}$$

Zudem gilt offensichtlich $\ln(Z(s))+s\cdot u_1>0$, weswegen zusammenfassend gilt

$$0<\frac{\ln(Z(s))}{s}+u_1\le(m-1)\cdot\frac{e^{-s\cdot(u_2-u_1)}}{s}\,.$$

Da $\lim_{s\nearrow\infty}\frac{e^{-s\cdot(u_2-u_1)}}{s}=0$ gilt, ist der erste Teil der Aussage gezeigt. Der dritte Teil der Aussage ergibt sich durch Herausheben des Faktors $e^{-s\cdot u_m}$ und Grenzübergang des Paramters $s\searrow-\infty$. $\square$

Beispiel 1: $\Omega=\{0,1\},\ P(X=1)=\frac{e^{-s}}{Z(s)}=e$ mit $Z(s)=1+e^{-s}$

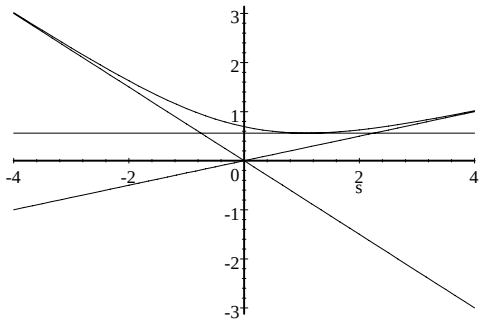


Abbildung 1: Abbildung der Funktion $s\mapsto\ln(Z(s))+e\cdot s$ für $e=1/4$

1.4.4 Thermodynamische Parametrisierung

Hinsichtlich der näheren Untersuchung der Zielfunktion $s\mapsto\ln(Z(s))+s\cdot e$ und der Funktionen $e\mapsto s_e=s(e)$ und $e\mapsto H_{\max}(e)$ ist es zweckmäßig, die Fälle $e=\bar{u},\ e<\bar{u}$ und $e>\bar{u}$ zu unterscheiden.

Fall $e=\bar{u}$: Für diesen Fall gilt $s(\bar{u}_m)=0,\ Q^*(0)=\left(\frac{1}{m},...,\frac{1}{m}\right)$,
$\frac{d}{ds}\left[\ln(Z(s))+s\cdot\bar{u}\right]|_{s=0}=0$ und somit

$$\ln(Z(s))+s\cdot\bar{u}\ge\ln(Z(0))=\ln(m)=H_{\max}(\bar{u})\,.$$

Da $\ln(m)$ der maximale Wert der Entropie einer Verteilung mit einem Träger mit m Elementen ist, wirkt sich die auferlegte Bedingung $e = \bar{u}$ nicht einschränkend aus.

Fall $e \in (u_1, \bar{u})$: Die Funktion $\ln(Z(s)) + s \cdot e$ nimmt wegen

$$\frac{d}{ds}\left[\ln\left(Z\left(s\right)\right) + s \cdot e\right] |_{s=0} = e - \bar{u} < 0 \quad \text{und} \quad \ln\left(Z\left(s\right)\right) + s \cdot e > (e - u_1) \cdot s \quad \text{für} \quad s > 0$$

ihr Minimum $H_{\max}(e) \in (0, \ln(m))$ für $s(e) \in (0, \infty)$ an.

Sei nun $u_1 < e_1 < e_2 < \bar{u}$. Wegen $s(e_1), s(e_2) \in (0, \infty)$ und $\ln(Z(s)) + s \cdot e_2 > \ln(Z(s)) + s \cdot e_1$ für $s > 0$ gilt

$$\begin{aligned} \ln(Z(s(e_2))) + s\left(e_2\right) \cdot e_2 &> \ln(Z(s(e_2))) + s\left(e_2\right) \cdot e_1 > \\ &> \ln(Z(s(e_1))) + s\left(e_1\right) \cdot e_1 \,. \end{aligned}$$

Das bedeutet, dass mit der Energie e die Entropie $H_{\max}(e) = \ln(Z(s(e))) + s(e) \cdot e$ wächst. Dies steht in Einklang mit der Thermodynamik [5]. Weil $E_{Q^*(s)}(U)$ streng monoton fällt, gilt zudem $s(e_1) > s(e_2)$ (> 0). Wegen der ersten beiden Aussagen von $(*)$ gilt weiters

$$\begin{array}{lcll} e \searrow u_1 & \Leftrightarrow & s(e) \nearrow \infty & \text{mit} \quad H_{\max}(e) \searrow 0 \\ e \nearrow \bar{u} & \Leftrightarrow & s(e) \searrow 0 & \text{mit} \quad H_{\max}(e) \nearrow \ln(m) \,. \end{array}$$

Bereits dies lässt vermuten, dass der Parameter $s > 0$ umgekehrt proportional zur absoluten Temperatur $T > 0$ ist [6]. Schließlich gilt jedoch für die Ableitung von $H_{\max}(e)$ gemäß

$$\frac{d}{de}\left[\ln(Z(s(e))) + s(e) \cdot e\right] = \left(-E_{Q^*(s(e))}(U) + e\right) \cdot s'(e) + s(e) = s(e)$$

die Beziehung

$$\frac{d\, H_{\max}(e)}{d\, e} = s(e) \,.$$

Zusammen mit dem aus der Thermodynamik bekannten Sachverhalt [7]

$$\frac{d\, E}{d\, S} = k \cdot T \,,$$

[5] Vgl. [21], S 167

[6] Der *Nernst*'sche Satz - vielfach auch 3. Haupsatz der Wärmelehre genannt - besagt übrigens:

Am absoluten Nullpunkt besitzt die Entropie für alle im thermodynamischen Gleichgewicht befindlichen Systeme unabhängig von den anderen Zustandsgrößen und unabhängig vom Phasenzustand der Systeme den Wert Null.

[7] Dieser Sachverhalt wurde vom ungarischen Mathematiker *John von Neumann* (1903 – 1957) im Zusammenhang mit der Themodynamik quantenmechanischer Gesamtheiten entdeckt.

wobei E die Energie, S die Entropie und T die absolute Temperatur eines thermodynamischen Systems sowie $k > 0$ die *Boltzmann*-Konstante sind, zeigt diese Beziehung, dass

$$s = \frac{1}{k \cdot T}$$

tatsächlich die der physikalischen Wirklichkeit angemessene Interpretation des Parameter $s > 0$ ist.

Fall $e \in (\bar{u}, u_m)$: Die Funktion $\ln(Z(s)) + s \cdot e$ nimmt wegen

$$\frac{d}{ds}\left[\ln\left(Z\left(s\right)\right) + s \cdot e\right]|_{s=0} = e - \bar{u} > 0 \quad \text{und} \quad \ln(Z(s)) + s \cdot e > -s \cdot (u_m - e) \quad \text{für} \quad s < 0$$

tatsächlich ihr globales Minimum $H_{\max}(e) \in (0, \ln(m))$ für $s(e) \in (-\infty, 0)$ an.

Sei nun $\bar{u} < e_1 < e_2 < u_m$. Wegen $s(e_1)$, $s(e_2) \in (-\infty, 0)$ und $\ln(Z(s)) + s \cdot e_2 < \ln(Z(s)) + s \cdot e_1$ für $s < 0$ gilt

$$\begin{aligned} \ln(Z(s(e_2))) + s\left(e_2\right) \cdot e_2 &< \ln(Z(s(e_1))) + s\left(e_1\right) \cdot e_2 \\ &< \ln(Z(s(e_1))) + s\left(e_1\right) \cdot e_1 \,. \end{aligned}$$

Das würde jedoch bedeuten, dass die Entropie $H_{\max}(e) = \ln(Z(s(e))) + s(e) \cdot e$ mit wachsender Energie e fällt. Da dies im Widerspruch zur Thermodynamik steht, kommt diesem Fall keine physikalische Bedeutung zu.

1.4.5 Anwendungsbeispiele

Da wir dabei Verteilungen mit Dichten verwenden werden - also einen "Quantensprung" machen - sind einige Vorbemerkungen zur sogenannten *differentiellen Entropie* (der Entropie für absolutstetige Verteilungen) angebracht.

Stetige Gleichverteilung auf $[0, a]$

$X \sim U_{[0,a]} \;:=\; q(x) = \frac{1}{a} 1_{[0,a]}(x)\,,\; x \in \mathbb{R}$

$$h(Q) = -\int_0^\infty q(x)\ln(q(x))dx = -\int_0^a \frac{1}{a}\ln(a)dx = \ln(a) = \begin{cases} > 0 & \text{für} \quad a > 1 \\ = 0 & \text{für} \quad a = 1 \\ < 0 & \text{für} \quad a < 1 \end{cases} .$$

Anmerkung 1: Die differentielle Entropie $h(Q)$ kann somit auch negative Werte annehmen. Um sie von der Entropie für diskrete Verteilungen zu unterscheiden, bezeichnen wir sie mit "klein h"

Exponentialverteilung

$X \sim Ex_\mu \;:=\; q(x) = \frac{e^{-x/\mu}}{\mu} 1_{[0,\infty)}(x)\,,\; x \in \mathbb{R}$

$E_Q(X) = \int_0^\infty xq(x)dx = \mu \int_0^\infty \frac{x}{\mu} e^{-x/\mu} d\left(\frac{x}{\mu}\right) = \mu \cdot 1 = \mu\,.$

$$\begin{aligned} h(Q) &= -\int_0^\infty q(x)\ln(q(x))dx \\ &= \int_0^\infty \frac{e^{-x/\mu}}{\mu} \ln\left(\mu \cdot e^{x/\mu}\right) dx = \int_0^\infty \frac{e^{-x/\mu}}{\mu}(\ln(\mu) + \frac{x}{\mu})dx \\ &= \ln(\mu) + \frac{1}{\mu}E_Q(X) = \ln(\mu) + 1\,. \end{aligned}$$

Normalverteilung

$X \sim N(\mu, \sigma^2) \;:=\; q(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2}\,,\; x \in \mathbb{R}$

$V_Q(X) = \int_{-\infty}^\infty x^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} dx = \sigma^2$

$$\begin{aligned} h(Q) &= -\int_0^\infty q(x)\ln(q(x))dx = \\ &= \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} \cdot \ln(\sqrt{2\pi}\sigma \cdot e^{x^2/2\sigma^2})dx \\ &= \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} \cdot \frac{1}{2}(\ln\left(2\pi\sigma^2\right) + \frac{x^2}{\sigma^2})dx \\ &= \frac{1}{2}\left(\ln\left(2\pi\sigma^2\right) + 1\right)\,. \end{aligned}$$

Anmerkung 2: (a) Als Maßstab für die differentielle Entropie kann der Parameter (die Intervalllänge)

$$a = e^{h(Q)}$$

der Gleichverteilung $U_{[0,a]}$ mit derselben Entropie $h(Q)$ dienen.

(b) Im eigentlichen Kerngebiet der Informationstheorie würde man naturgemäß als Basis des Logarithmus $b = 2$ wählen.

(c) Die Eigenschaften der differentiellen Entropie sind denen der Entropie für diskrete Verteilungen verschieden: Sie kann - wie gesagt - auch negative Werte annehmen. Die Eigenschaften der I-Divergenz sind jedoch für beide Verteilungstypen gleich. Dieser Umstand erlaubt die Übertragung des Teils (b) des Beweises unseres Satzes auf Verteilungen mit Dichten.

Das stetige Analogon der Aussage, dass die Geometrische Verteilung unter allen Verteilungen auf $\mathbb{N}$ mit gegebenem Erwartungswert die größte Entropie besitzt, ist folgende Behauptung.

Behauptung 1: Sei X eine Zufallsvariable mit absolutstetiger Verteilung Q

auf $[0,\infty)$ mit $E_Q(X) = \mu \in (0,\infty)$ und sei $Q^*(\mu) = Ex_\mu$. Dann gilt

$$h(Q) \leq h(Q^*(\mu)) = \ln(\mu) + 1\,.$$

Das folgende Anwendungsbeispiel ist eine unmittelbare Folgerung.

Beispiel 1: Die barometrische Höhenformel

Seien

$\Omega = [0,\infty)$

$u(x) = mgx$ ... die potentielle Energie des Teilchens eines idealen Gases mit Masse m und Höhe x

$\mathcal{V}_e = \left\{ Q = (q(x),\, x \geq 0) \text{ mit } \int_0^\infty mgx\, q(x)dx = e \right\}$, $e > 0$

$\mathcal{G} = \left\{ Q_T^* = (q_T^*(x) = \frac{mg}{kT} e^{-\frac{mg}{kT}x},\, x \geq 0) \right\}$, $T > 0$

Weiters sei $T = T(e)$ derart, dass $E_{Q_T^*}(U) = e$ mit

$$E_{Q_T^*}(U) = \int_0^\infty mgx\, q_T^*(x)dx = mg\, E_{Q^*(T)}(X) = mg \cdot \frac{kT}{mg} = kT\,.$$

Dann gilt

$$h(Q) \leq h(Q_T^*) = \ln(\frac{kT}{mg}) + 1 \;\; \forall\, Q \in \mathcal{V}_{kT}\,.$$

Bevor wir Beispiel 2 besprechen, welches im $\mathbb{R}^3$ "spielt", formulieren wir den wesentlichen Teil der Aussage über $\mathbb{R}$.

Behauptung 2: Sei X eine Zufallsvariable mit absolutstetiger Verteilung Q auf $\mathbb{R}$ mit $V_Q(X) = \sigma^2 \in (0,\infty)$ und sei $Q^*(\sigma^2) = N(\mu,\sigma^2)$. Dann gilt

$$h(Q) \leq h(Q^*(\sigma^2)) = \frac{1}{2}\left(\ln(2\pi\sigma^2) + 1\right)\,.$$

Beispiel 2: Die Maxwellsche Geschwindigkeitsverteilung

Seien

$\Omega = \mathbb{R}^3$

$u(v_x, v_y, v_z) = \frac{m}{2}\left(v_x^2 + v_y^2 + v_z^2\right)$ ... kinetische Energie des Teilchens eines idealen Gases mit Masse m und Geschwindigkeitsvektor (v_x, v_y, v_z)

$\mathcal{V}_e = \left\{ Q = \left(q\,(v_x, v_y, v_z) : (v_x, v_y, v_z) \in \mathbb{R}^3\right) \text{ mit } E_Q\,(U) = e \right\}$

$\mathcal{G} = \left\{ Q_T^* = \left(q_T^*\,(v_x, v_y, v_z) = (\frac{m}{2\pi kT})^{3/2} e^{-\frac{m}{2kT}\left(v_x^2+v_y^2+v_z^2\right)} : (v_x, v_y, v_z) \in \mathbb{R}^3 \right) \right\}$,

$T > 0$

Weiters sei $T = T(e)$ derart, dass $E_{Q_T^*}(U) = e$ mit

$$\begin{aligned} E_{Q_T^*}(U) &= \int\int\int \frac{m}{2}\left(v_x^2 + v_y^2 + v_z^2\right) \cdot q_T^*\left(v_x, v_y, v_z\right) dv_x\, dv_y\, dv_z \\ &= 3\frac{m}{2}\int_{-\infty}^{\infty} v_x^2 \cdot (\frac{m}{2\pi kT})^{1/2} e^{-\frac{m}{2kT}v_x^2} \\ &= 3\frac{m}{2}\cdot\frac{kT}{m} = \frac{3}{2}kT\,. \end{aligned}$$

Dann gilt

$$h(Q) \leq h(Q_T^*) = \frac{3}{2}(\ln(2\pi\frac{kT}{m}) + 1) \quad \forall\, Q \in \mathcal{V}_{\frac{3}{2}kT}\,.$$

2 KANALKODIERUNG

2.1 KANAL UND KANALKAPAZITÄT

2.1.1 Motivation

Da wir uns im Folgenden vorwiegend mit dem binären symmetrischen Kanal beschäftigen, ziehen wir diesen auch zur Motivation der Begriffe *Kanal* und *Kanalkapazität* heran.

Beispiel 1 (Binärer symmetrischer Kanal): $A_X = A_Y = \{0, 1\}$,

$$\mathbb{P} = \begin{pmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{pmatrix},$$

wobei wir in der Regel davon ausgehen, dass über den Kanal nicht bloß ein Symbol, sondern eine Folge von Symbolen übertragen wird und Fehler unabhängig voneinander auftreten.

Angenommen, wir schicken die $0-1$-Folge

1	0	1	1	0	1	0	0	0	1	1	0	1	1	1	0	0	1	0	1	1	1	1
					F													F				

über den Kanal, wobei Fehler an jenen Stellen auftreten mögen, welche durch "F" markiert sind. Der Empfänger erhält daher die Folge

1	0	1	1	0	0	0	0	0	1	1	0	1	1	1	0	0	1	1	1	1	1	1

aus der dieser keineswegs feststellen kann, ob bzw. wenn ja, an welchen Stellen Übertragungsfehler aufgetreten sind. Um die gesandte $0-1$-Folge aus der empfangenen wiederherstellen zu können, müsste der Empfänger über die $0-1$-Folge der Fehlerstellen verfügen. Weil Fehler mit der Wahscheinlichkeit ε auftreten, ist zur Wiederherstellung der Folge pro Symbol zusätzlich der Informationsgehalt

$$H(\varepsilon, 1-\varepsilon) = -\left(\varepsilon \log_2 \varepsilon + (1-\varepsilon)\log_2(1-\varepsilon)\right)$$

vonnöten. Da der Informationsgehalt pro Symbol der wiederhergestellten Folge - wie der der gesandten - ein Bit ist, kann über den Kanal maximal der Informationsgehalt

$$C = 1 - H(\varepsilon, 1-\varepsilon) = I\left(P_\varepsilon \| P_{1/2}\right)$$

übertragen werden. Dabei ist $I\left(P_\varepsilon \| P_{1/2}\right)$ die I-Divergenz der Verteilungen $P_\varepsilon = (\varepsilon, 1-\varepsilon)$ und $P_{1/2} = \left(\frac{1}{2}, \frac{1}{2}\right)$. Die Größe C ist die Kapazität des binären symmetrischen Kanals.

Allgemein versteht man unter der Kapazität eines Kanals den maximalen Informationsgehalt, der sich pro Symbol über diesen fehlerlos übertragen lässt.

2.1.2 Gemeinsame und bedingte Entropie, wechselseitige Information

Definition 1: Sei (X,Y) ein Paar von diskreten Zufallsvariablen mit gemeinsamer Verteilung $P_{(X,Y)} = (p(x,y) : x \in W_X,\, y \in W_Y)$, wobei $W_X = \{x : p(x) = \sum_y p(x,y) > 0\}$ bzw. W_Y die Trägermengen der Randverteilungen P_X bzw. P_Y von X und Y sind. Dann heißt

$$\begin{aligned} H\left((X,Y)\right) &= -E_{P_{(X,Y)}}\left(\log_2 p(X,Y)\right) \\ &= -\sum_{(x,y)\in W_X \times W_Y} p(x,y)\log_2\left(p(x,y)\right) \end{aligned}$$

die *gemeinsame Entropie* von X und Y und

$$\begin{aligned} H(Y\,/\,X) &= \sum_{x\in W_X} p(x)\cdot H(Y\,/\,X=x) \\ &= -\sum_{x\in W_X} p(x)\cdot \sum_{y\in W_Y} p(y\,/\,x)\log_2 p(y\,/\,x) = \\ &= -E_{P_{(X,Y)}}\left(\log_2 p(Y\,/\,X)\right) \end{aligned}$$

die durch X *bedingte Entropie* von Y.

Proposition 1 (Kettenregel): Es gilt

$$\begin{aligned} H\left((X,Y)\right) &= H(Y) + H(X\,/\,Y) \\ &= H(X) + H(Y\,/\,X) = H\left((Y,X)\right)\,. \end{aligned}$$

Beweis: Aufgrund der Definition der bedingten Wahrscheinlichkeit, der Additivität des Logarithmus $\log(u\cdot v) = \log(u) + \log(v)$ und der Additivität des Erwartungswerts gilt

$$\begin{aligned} H\left((X,Y)\right) &= -E_{P_{(X,Y)}}\left(\log_2 p(X,Y)\right) \\ &= -E_{P_{(X,Y)}}\left(\log_2\left(p(X)\cdot p(Y\,/\,X)\right)\right) \\ &= -E_{P_{(X,Y)}}\left(\log_2 p(X) + \log_2 p(Y\,/\,X)\right) \\ &= -E_{P_X}\left(\log_2 p(X)\right) - E_{P_{(X,Y)}}\left(\log_2 p(Y\,/\,X)\right) \\ &= H(X) + H(Y\,/\,X)\,. \;\square \end{aligned}$$

Proposition 2: Es gilt $H(Y\,/\,X) \geq 0$ mit Gleichheit genau dann, wenn $Y = f(X)$ mit $f : W_X \mapsto W_Y$ ist.

Beweis: Dies gilt wegen

$$H(Y\,/\,X) = \sum_{x\in W_X} p(x)\cdot H(Y\,/\,X=x)\,,$$

$p(x) > 0 \;\; \forall\; x \in W_X$ und $H(Y\,/\,X=x) \geq 0$ mit Gleichheit genau dann, wenn $Y = f(x)$.

Definition 2: Seien (X,Y) , $P_{(X,Y)}$, P_X und P_Y wie oben. Dann heißt

$$\begin{aligned} I(X;Y) &= I(P_{(X,Y)} \| P_X \times P_Y) \\ &= E_{P_{(X,Y)}} \left(\log_2 \left(\frac{p(X,Y)}{p(X) \cdot p(Y)} \right) \right) \\ &= \sum_{(x,y) \in W_X \times W_Y} p(x,y) \log_2 \left(\frac{p(x,y)}{p(x) \cdot p(y)} \right) \end{aligned}$$

die *wechselseitige Information* von X und Y .

Proposition 3 (Darstellung mittels Entropie): Es gilt

$$\begin{aligned} I(X;Y) &= H(X) - H(X \,/\, Y) \\ &= H(X) + H(Y) - H\left((X,Y)\right) \\ &= H(Y) - H(Y \,/\, X) = I(Y;X) \end{aligned}$$

Beweis: Wie im obigen Beweis und vermittels der Kettenregel gilt

$$\begin{aligned} I(X;Y) &= E_{P_{(X,Y)}} \left(\log_2 \left(\frac{p(X,Y)}{p(X) \cdot p(Y)} \right) \right) \\ &= E_{P_{(X,Y)}} \left(\log_2 \left(\frac{p(X \,/\, Y)}{p(X)} \right) \right) \\ &= -E_{P_{(X,Y)}} \left(\log_2 p(X) - \log_2 p(X \,/\, Y) \right) \\ &= -\left(E_{P_X} \left(\log_2 p(X) \right) - E_{P_{(X,Y)}} \left(\log_2 p(X \,/\, Y) \right) \right) \\ &= H(X) - H(X \,/\, Y) \\ &= H(X) + H(Y) - H(X,Y) \,, \end{aligned}$$

wobei zuletzt die Kettenregel $H\left((X,Y)\right) = H(Y) + H(X \,/\, Y)$ berücksichtigt wird. □

Proposition 4 (Wertebereich der wechselseitigen Information): Es gilt

$$0 \leq I(X;Y) \leq \min\left\{H(X), H(Y)\right\} \left(\leq \frac{1}{2} \left(H(X) + H(Y) \right) \right) ,$$

mit Gleichheit in der (i) ersten und (ii) zweiten Ungleichung genau dann, wenn

X und Y unabhängig sind

bzw. wenn gilt

$X = g(Y)$ mit $g : W_Y \mapsto W_X$ falls $\min = H(X)$
$Y = f(X)$ mit $f : W_X \mapsto W_Y$ falls $\min = H(Y)$

Beweis: (i) Die erste Aussage ist eine unmittelbare Folgerung der Eigenschaft der I-Divergenz

$$I(X;Y) = I(P_{(X,Y)} \| P_X \times P_Y) \geq 0$$

mit Gleichheit genau dann, wenn $P_{(X,Y)} = P_X \times P_Y$.

(ii) Die zweite Aussage ist eine unmittelbare Folgerung aus der Darstellung von $I(X;Y)$ mittels der bedingten Entropie, deren Nichtnegativität und der Charakterisierung derselben. $\square$

Anmerkung 1: Für die Größe

$$\varrho(X;Y) = \frac{1}{2}\,(H(X) + H(Y)) - I(X;Y)$$

gilt

$$\begin{aligned} \varrho(X;Y) &= \frac{1}{2}\,(H(X,Y) - I(X;Y)) \\ &= H(X,Y) - \frac{1}{2}\,(H(X) + H(Y)) \\ &= \frac{1}{2}\,(H(X\,/\,Y) + H(Y\,/\,X)) \\ &\geq 0 \end{aligned}$$

mit

$$\text{Gleichheit} \;\Leftrightarrow\; \exists\,\text{Bijektion } f : W_X \mapsto W_Y \;\text{ mit }\; Y = f(X)\,.$$

2.1.3 Definitionen und Aussagen

Definition 1: Ein *(homogener diskreter) gedächtnisloser Kanal* ist ein Tripel $(A_X, A_Y, \mathbb{P})$ bestehend aus einem *Eingangsalphabet* A_X , einem *Ausgangsalphabet* A_Y und einer *Übergangsmatrix*

$$\mathbb{P} = (P(x,y))_{(x,y)\in A_X \times A_Y}\ .$$

Dies ist eine Matrix mit den Eigenschaften $P(x,y) \geq 0 \quad \forall\,(x,y) \in A_X \times A_Y$ und $\sum_{y\in A_Y} P(x,y) = 1 \quad \forall\, x \in A_X$. Die Bedeutung eines Elements von $\mathbb{P}$ ist somit die der bedingten Wahrscheinlichkeit

$$P(x,y) = P(Y = y\,/\,X = x)\,.$$

Die Zeilenvektoren sind somit Wahrscheinlichkeitsverteilungen

$$P(x) = (\,P(x,y) : y \in A_Y\,)\ ,\ x \in A_X\,.$$

Eine *(homogene) gedächtnislose Quelle* wird - wie üblich - durch eine Wahrscheinlichkeitsverteilung

$$P_X = (\,p(x) : x \in A_X\,)$$

beschrieben. Die gemeinsame Verteilung des Zufallspaares (X,Y) wird im vorliegenden Fall mit Hilfe von P_X und $\mathbb{P}$ durch

$$P_{(X,Y)} = (p(x,y) = p(x) \cdot P(x,y) : (x,y) \in A_X \times A_Y)$$

definiert. Die (Rand-)Verteilung von Y ist somit durch

$$P_Y = P_X \cdot \mathbb{P} = (q(y) = \sum_{x \in A_X} p(x) \cdot P(x,y) : y \in A_Y)$$

gegeben.

Definition 2: Seien $(A_X, A_Y, \mathbb{P})$ ein (homogener) gedächtnisloser Kanal und $\mathcal{V}_X$ die Menge aller Eingangsverteilungen (d.i. die Menge aller Wahrscheinlichkeitsverteilungen P_X auf A_X), dann heißt die Größe

$$\begin{aligned} C &= \sup_{P_X \in \mathcal{V}_X} I(X;Y) \\ &= \sup_{P_X \in \mathcal{V}_X} I(P_{(X,Y)} \parallel P_X \times P_Y) = \\ &= \sup_{P_X \in \mathcal{V}_X} \left(H\left(\sum_{x \in A_X} p(x) \cdot P(x) \right) - \sum_{x \in A_X} p(x) H\left(P(x)\right) \right) \end{aligned}$$

die *Kapazität des Kanals*.

Anmerkung 1: (i) Wegen $I(X;Y) \leq \min\{H(X), H(Y)\}$ und $H(X) \leq \log_2(|A_X|)$ gilt

$$C \leq \log_2 \left(\min \{ |A_X| , |A_Y| \} \right) .$$

(ii) Da bei festem $\mathbb{P}$ $P_Y = \sum_{x \in A_X} p(x) \cdot P(x)$ und $\sum_{x \in A_X} p(x) H\left(P(x)\right)$ lineare Funktionen von P_X sind und $P_Y \mapsto H(P_Y)$ konkav ist, ist

$$I(X;Y) = H\left(P_Y\right) - \sum_{x \in A_X} p(x) H\left(P(x)\right)$$

eine konkave und somit stetige Funktion von P_X . Weil zudem das Simplex $\mathcal{V}_X$ beschränkt und abgeschlossen ist, wird das Supremum daher angenommen. Somit können wir anstelle des Supremums das Maximum verwenden.

Interpretation: Die Kanalkapazität ist ein Maß für die Fähigkeit eines Kanals, Informationen zuverlässig zu übermitteln. Sie ist etwa vergleichbar mit der Leitfähigkeit eines Widerstandes in einem elektrischen Netz.

Beispiele von Kanälen

Beispiel 0 (Kanal mit nichtüberlappenden Ausgängen): $A_X = \{1,2,3,4\}$, $A_Y = \{1,2,3,4,5,6,7\}$,

$$\mathbb{P} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{3} & \frac{2}{3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{3}{4} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

$$C = \log_2(|A_X|) = 2\,.$$

Beispiel 1 (Binärer symmetrischer Kanal): $A_X = A_Y = \{0,1\}$,

$$\mathbb{P} = \begin{pmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{pmatrix},$$

$$C = 1 - h(\varepsilon)\,.$$

Beispiel 2 (Binärer symmetrischer Löschkanal): $A_X = \{0,1\}$, $A_X = \{0,*,1\}$ und

$$\mathbb{P} = \begin{pmatrix} 1-\varepsilon & \varepsilon & 0 \\ 0 & \varepsilon & 1-\varepsilon \end{pmatrix},$$

$$C = 1 - \varepsilon\,.$$

Zur Berechnung der Kanalkapazität

Die Berechnung der Kanalkapazität ist im Allgemeinen schwierig bzw. kompliziert. Zu den Kanälen, für welche die Berechnung der Kanalkapazität einfach ist, zählen die oben angegebenen Beispiele. Der folgende Sachverhalt vereinfacht gelegentlich die Berechnung.

Sachverhalt: Seien $P_1, ..., P_k$ Wahrscheinlichkeitsverteilungen auf einer Menge $\Omega = \{1, ..., m\}$, welche paarweise orthogonal sind (d.h. voneinander verschiedene Trägermengen besitzen) und sei $\alpha = (\alpha_1, ..., \alpha_k)$ eine Wahrscheinlichkeitsverteilung auf $\{1, ..., k\}$. Dann ist die Entropie der Mischverteilung $\sum_{i=1}^{k} \alpha_i \cdot P_i$ gleich

$$H\left(\sum_{i=1}^{k} \alpha_i \cdot P_i\right) = H(\alpha) + \sum_{i=1}^{k} \alpha_i H(P_i)\,.$$

Definition 3: *Ein Kanal mit nichtüberlappenden Ausgängen* ist ein Kanal, für welchen die durch $\mathbb{P}$ gegebenen Wahrscheinlichkeitsverteilungen

$$P(x) = (P(x,y) : y \in A_Y) , \;\; x \in A_X$$

paarweise orthogonal sind. (In diesem Fall kann von den erhaltenen Symbolen fehlerfrei auf die gesandten zurückgeschlossen werden.)

Proposition 1: Die Kanalkapazität für einen Kanal mit nichtüberlappenden Ausgängen ist

$$C = \log_2 (|A_X|)$$

und es gibt eine Funktion $g : W_Y \mapsto W_X$ derart, dass gilt $X = g(Y)$, wobei der Rückschluss fehlerfrei erfolgt.

Beweis: Da die Wahrscheinlichkeitsverteilungen $P(x)$, $x \in A_X$ paarweise orthogonal sind, können wir vom obigen Sachverhalt Gebrauch machen. Die Entropie von P_Y ist daher

$$H \left(\sum_{x \in P_X} p(x) P(x) \right) = H (P_X) + \sum_{x \in P_X} p(x) H (P(x)) \; .$$

Somit ist

$$I(X;Y) = H \left(\sum_{x \in P_X} p(x) P(x) \right) - \sum_{x \in P_X} p(x) H (P(x)) = H (P_X)$$

und demzufolge

$$C = \max_{P_X \in \mathcal{V}_X} H (P_X) = \log_2 (|A_X|) \; . \; \square$$

Definition 4: Die Matrix $\mathbb{P}$ eines (homogenen) gedächtnislosen Kanals $(A_X, A_Y, \mathbb{P})$ heißt bis auf Permutationen zeilengleich oder kurz π-*zeilengleich*, wenn die Wahrscheinlicheitsverteilungen $P(x) = (P(x,y) : y \in A_Y) , \; x \in A_X$ Permutationen voneinander sind.

$\mathbb{P}$ heißt π-spaltengleich, wenn die Spalten $(P(x,y) : x \in A_X)$ von $\mathbb{P}$ Permutationen voneinander sind.

Ein Kanal $(A_X, A_Y, \mathbb{P})$ heiße *symmetrisch*, wenn $\mathbb{P}$ π-zeilen- und -spaltengleich ist.

Die Übergangsmatrix heißt *schwach symmetrisch*, wenn $\mathbb{P}$ π-zeilengleich und wenn alle Spaltensummen $\sum_{x \in A_X} P(x,y) = c$ sind, wobei c eine positive Konstante ist.

Die Übergangsmatrix $\mathbb{P}$ heißt doppeltstochastisch, wenn - neben den Zeilensummen - auch alle Spaltensummen $\sum_{x \in A_X} P(x,y) = 1$ sind.

Proposition 2: (i) Die Kapazität eines symmetrischen Kanals oder eines Kanals mit schwach symmetrischer Übergangsmatrix $\mathbb{P}$ ist

$$C = \log_2(|A_Y|) - H(P(x_1)) = I\left(P(x_1) \| P_{|A_Y|}\right) ,$$

wobei $H(P(x_1))$ die Entropie der Wahrscheinlichkeitsverteilung $P(x_1) = (P(x_1, y) : y \in A_Y)$ und $I\left(P(x_1) \| P_{|A_Y|}\right)$ die I-Divergenz von $P(x_1)$ und der Gleichverteilung $P_{|A_Y|} = \left(\frac{1}{|A_Y|}, ..., \frac{1}{|A_Y|}\right)$ sind.

(ii) Wenn $\mathbb{P}$ π-zeilengleich und doppeltstochastisch ist, gilt überdies $|A_Y| = |A_X|$.

Beweis: (i) Da $\mathbb{P}$ in beiden Fällen π-zeilengleich ist und die Entropie einer Verteilung gegenüber Permutationen der Wahrscheinlichkeiten invariant ist, gilt nämlich für ein beliebiges $x_1 \in A_X$

$$H(P(x)) = H(P(x_1)) \ \forall\, x \in A_X$$

und somit

$$I(X;Y) = H(Y) - \sum_{x \in A_X} p(x) \cdot H(P(x)) = H(Y) - H(P(x_1)) .$$

Wegen $H(Y) \leq \log_2(|A_Y|)$ ist daher

$$C = \max_{P_X \in \mathcal{V}_X} H(Y) - H(P(x_1)) \leq \log_2(|A_Y|) - H(P(x_1)) = I\left(P(x_1) \| P_{|A_Y|}\right) . \tag{1}$$

Da die Ausgangsverteilung $P_Y = P_X \cdot \mathbb{P}$ der Gleichverteilung $P_X = (\frac{1}{|A_X|}, ..., \frac{1}{|A_X|})$ auf A_X wegen der folgenden Überlegung die Gleichverteilung $P_Y = (\frac{1}{|A_Y|}, ..., \frac{1}{|A_Y|})$ auf A_Y ist, gilt $\max_{P_X \in \mathcal{V}_X} H(Y) = \log_2(|A_Y|)$ und somit Gleichheit in (2): Tatsächlich ist wegen $\sum_{x \in A_X} P(x, y) = c \ \forall\, y \in A_Y$

$$q(y) = \sum_{x \in A_X} \frac{1}{|A_X|} P(x, y) = \frac{1}{|A_X|} \sum_{x \in A_X} P(x, y) = \frac{c}{|A_X|} = \frac{1}{|A_Y|} ,$$

wobei die letzte Gleichheit wegen $\sum_{y \in A_Y} q(y) = 1$ gilt.

(ii) Eine doppeltstochastische Matrix erfüllt $\sum_{x \in A_X} P(x, y) = c = 1 \ \forall\ y \in A_Y$. Daher gilt in diesem Fall $|A_Y| = |A_X|$. □

Proposition 3: Die Kapazität des binären Löschkanals ist $c = 1 - \varepsilon$.

Beweis: Da $\mathbb{P}$ π-zeilengleich ist, gilt

$$C = \max_{P_X \in \mathcal{V}_X} H(Y) - H(P(x_1))$$

mit $H(P(x_1)) = h(\varepsilon) = -\left(\varepsilon \log_2 \varepsilon + (1-\varepsilon)\log_2(1-\varepsilon)\right)$. Da für die Eingangsverteilung $P_X = (p, 1-p)$ die Ausgangsverteilung

$$\begin{aligned} P_Y &= P_X \cdot \mathbb{P} = ((1-\varepsilon)\,p, \varepsilon, (1-\varepsilon)\,(1-p) \\ &= (1-\varepsilon)\,(p, 0, 1-p) + \varepsilon\,(0,1,0) \end{aligned}$$

ist, ist deren Entropie wegen der Orthogonalität von $(p, 0, 1-p)$ und $(0,1,0)$

$$\begin{aligned} H(Y) &= H(P_Y) \\ &= H(\varepsilon, 1-\varepsilon) + (1-\varepsilon) H\,(p, 0, 1-p) + \varepsilon H\,(0,1,0) \\ &= h(\varepsilon) + (1-\varepsilon) h(p)\,. \end{aligned}$$

Wegen $h(p) \leq h(\frac{1}{2}) = 1$ ist die Kanalkapazität somit gleich

$$C = h(\varepsilon) + 1 - \varepsilon - h(\varepsilon) = 1 - \varepsilon\,.$$

Definition 5: Gegeben sei ein gedächtnisloser Kanal $(A_X, A_Y, \mathbb{P})$ mit Eingangs-alphabet A_X, Ausgangsalphabet A_Y und Übergangsmatrix $\mathbb{P}$. Dann nennen wir das Tripel $(A_X^n, A_Y^n, \mathbb{P}^{(n)})$ mit

$$\mathbb{P}^{(n)} = \left(\prod_{i=1}^{n} P\,(x_i, y_i) \right)_{((x_1,\ldots,x_n),(y_1,\ldots,y_n)) \in A_X^n \times A_Y^n}$$

die *n-te Fortsetzung* von $(A_X, A_Y, \mathbb{P})$.

Proposition 4: Es gilt

$$I(\mathbf{X}^n; \mathbf{Y}^n) \leq nC\,.$$

Beweis: Aus der Darstellung von $I(X;Y)$ ergibt sich zunächst

$$H\,(Y_j \,/\, Y_1, \ldots, Y_{j-1}) \leq H\,(Y_j)\ ,\ j \in \{1, \ldots, n\}\,.$$

Gedächtnislosigkeit und Homogenität des Kanals ergeben zudem

$$H\,(Y_j \,/\, Y_1, \ldots, Y_{j-1}, \mathbf{X}^n) = H\,(Y_j \,/\, X_j) \quad \text{und} \quad I(X_j; Y_j) = I(X_1; Y_1)\,.$$

Daher ist

$$\begin{aligned} I(\mathbf{X}^n; \mathbf{Y}^n) &= H(\mathbf{Y}^n) - H\,(\mathbf{Y}^n \,/\, \mathbf{X}^n) \\ &= \sum_{j=1}^{n} \left(H\,(Y_j \,/\, Y_1, \ldots, Y_{j-1}) - H\,(Y_j \,/\, Y_1, \ldots, Y_{j-1}, \mathbf{X}^n) \right) \\ &\leq \sum_{j=1}^{n} \left(H\,(Y_j) - H\,(Y_j \,/\, X_j) \right) \\ &= \sum_{j=1}^{n} I(X_j; Y_j) \\ &= nI(X_1; Y_1) \\ &\leq nC\,. \ \square \end{aligned}$$

2.2 DER KANALKODIERUNGSSATZ

2.2.1 Die Tail Inequality

Lemma 1 (Tail Inequality): Seien $P_p = (p, 1-p)$ und $P_\lambda = (\lambda, 1-\lambda)$ zwei Alternativverteilungen, für deren Parameter $0 \le \lambda \le p \le 1$ gelte und sei

$$I\left(P_\lambda \| P_p\right) = \lambda \log_2\left(\frac{\lambda}{p}\right) + (1-\lambda)\log_2\left(\frac{1-\lambda}{1-p}\right)$$

die I-Divergenz von P_λ und P_p. Schließlich sei $S_n \sim B_{n,\lambda}$- bzw. $B_{n,p}$-verteilt, je nachdem welche der Alternativverteilungen P_λ oder P_p zugrunde liegt. Dann gilt

$$P_p\left(S_n \le n\lambda\right) \le 2^{-nI(P_\lambda \| P_p)} \cdot P_\lambda\left(S_n \le n\lambda\right).$$

Beweis: Für den Fall $0 = \lambda \le p < 1$ ist die Ungleichung scharf. Dann gilt nämlich $P_p\left(S_n \le 0\right) = (1-p)^n$, $P_0\left(S_n \le 0\right) = 1$ und $I\left(P_0 \| P_p\right) = \log_2\left(1/(1-p)\right)$ und somit

$$P_p\left(S_n \le 0\right) = 2^{-nI(P_0 \| P_p)} \cdot P_0\left(S_n \le 0\right) = (1-p)^n.$$

Für den Fall $0 \le \lambda \le p = 1$ ist die Ungleichung ebenfalls scharf. Dann gilt nämlich $P_1\left(S_n \le 0\right) = 0$ und $I\left(P_\lambda \| P_1\right) = \begin{cases} +\infty & \text{für} \quad \lambda < 1 \\ 0 & \text{für} \quad \lambda = 1 \end{cases}$ und somit

$$P_1\left(S_n \le 0\right) = 2^{-nI(P_\lambda \| P_1)} \cdot P_\lambda\left(S_n \le 0\right) = 0.$$

Sei im Weiteren also $0 < \lambda \le p < 1$ oder, gleichbedeutend, $\frac{\lambda}{p}\frac{1-p}{1-\lambda} \le 1$. Dann gilt

$$\begin{aligned}
P_\lambda\left(S_n \le n\lambda\right) &= \sum_{k=0}^{\lfloor n\cdot\lambda \rfloor} \binom{n}{k} \lambda^k (1-\lambda)^{n-k} \\
&= \sum_{k=0}^{\lfloor n\cdot\lambda \rfloor} \binom{n}{k} p^k (1-p)^{n-k} \cdot \left(\frac{\lambda}{p}\right)^k \left(\frac{1-\lambda}{1-p}\right)^{n-k} \\
&= \left(\frac{1-\lambda}{1-p}\right)^n \sum_{k=0}^{\lfloor n\cdot\lambda \rfloor} \binom{n}{k} p^k (1-p)^{n-k} \cdot \left(\frac{\lambda}{p}\frac{1-p}{1-\lambda}\right)^k \ge \\
&\ge \left(\frac{1-\lambda}{1-p}\right)^n \sum_{k=0}^{\lfloor n\cdot\lambda \rfloor} \binom{n}{k} p^k (1-p)^{n-k} \cdot \left(\frac{\lambda}{p}\frac{1-p}{1-\lambda}\right)^{n\cdot\lambda} \\
&= \left(\left(\frac{\lambda}{p}\right)^\lambda \left(\frac{1-\lambda}{1-p}\right)^{1-\lambda}\right)^n \cdot \sum_{k=0}^{\lfloor n\cdot\lambda \rfloor} \binom{n}{k} p^k (1-p)^{n-k} \\
&= 2^{nI(P_\lambda \| P_p)} \cdot P_p\left(S_n \le n\lambda\right)
\end{aligned}$$

und somit die behauptete Ungleichung. $\square$

Folgerung: Sei $H(P_\lambda) = -(\lambda \log_2 \lambda + (1-\lambda)\log_2(1-\lambda))$ die Entropie der Verteilung P_λ. Dann gilt

$$\sum_{k=0}^{\lfloor n\lambda \rfloor} \binom{n}{k} \leq 2^{nH(P_\lambda)} .$$

Beweis: Für den Spezialfall $p = \frac{1}{2}$ hat $I\left(P_\lambda \| P_{1/2}\right)$ bekanntlich die Form

$$I\left(P_\lambda \| P_{1/2}\right) = 1 - H(P_\lambda) .$$

Berücksichtigt man zudem die Abschätzung $P_\lambda(S_n \geq n\lambda) \leq 1$, so erhält man die Ungleichung

$$P_{1/2}(S_n \leq n\lambda) \leq 2^{-n(1-H(P_\lambda))} ,$$

welche zur behaupteten Ungleichung gleichwertig ist. $\square$

Anmerkung 1: Für $0 \leq p \leq \lambda \leq 1$ gilt wegen $S'_n = n - S_n \sim B_{n,1-p}$ bzw. $B_{n,1-\lambda}$ und $I(P_{1-\lambda} \| P_{1-p}) = I(P_\lambda \| P_p)$ unter Anwendung der *Tail Inequality* für $0 \leq 1-\lambda \leq 1-p \leq 1$

$$\begin{aligned} P_p(S_n \geq n\lambda) &= P_p(S'_n \leq n(1-\lambda)) \leq \\ &\leq 2^{-nI(P_{1-\lambda}\|P_{1-p})} \cdot P_\lambda(S'_n \leq n(1-\lambda)) \\ &= 2^{-nI(P_\lambda \| P_p)} \cdot P_\lambda(S_n \geq n\lambda) . \end{aligned}$$

2.2.2 Motivation der Kanalkodierung

Definition 1: Gegeben sei ein gedächtnisloser Kanal mit n-ter Fortsetzung $(A_X^n, A_Y^n, \mathbb{P}^{(n)})$ und seien

- $\{1, ..., M\}$ eine Indexmenge mit $M < |A_X|^n$
- eine (eineindeutige) Kodierungsfunktion $\mathbf{X}^n : \{1, ..., M\} \mapsto A_X^n$,
- eine Dekodierungsfunktion $g : A_Y^n \mapsto \{1, ..., M\}$

Dieses Tripel heißt ein (M, n)-*Code* mit dem *Codebuch*

$$\mathcal{C} = (C_1, ..., C_M) = (\mathbf{X}^n(1), ..., \mathbf{X}^n(M)) \subseteq (A_X^n)^M .$$

Bezeichnungen: Sei $i \in \{1, ..., M\}$ das gesandte Symbol und $\mathbf{c}_i = \left(c_1^{(i)}, ..., c_n^{(i)}\right)$ dessen Kodierung. Dann ist

$$\begin{aligned} \lambda_i(\mathcal{C}) &= P_{\mathcal{C}}(g(\mathbf{Y}^n) \neq i \,/\, \mathbf{X}^n = \mathbf{c}_i) \\ &= \sum_{y^n \in A_Y^n} \prod_{j=1}^{n} P\left(c_j^{(i)}, y_j\right) 1_{\{1,...,M\}\setminus\{i\}}(g(\mathbf{y}^n)) \end{aligned}$$

die *(bedingte) Fehlerwahrscheinlichkeit*,

$$\lambda^{(n)}(\mathcal{C}) = \max\{\lambda_i(\mathcal{C}),\, i \in \{1, ..., M\}\}$$

die *maximale (bedingte) Fehlerwahrscheinlichkeit* und

$$P_e^{(n)}(\mathcal{C}) = \frac{1}{M}\sum_{i=1}^{M} \lambda_i(\mathcal{C}) = P_{\mathcal{C}}\left(g\left(\mathbf{Y}^n\right) \neq I\right)$$

die *mittlere (bedingte) Fehlerwahrscheinlichkeit*. Dabei wird I als auf $\{1, ..., M\}$ gleichverteilt angenommen.

Motivation des Kanalkodierungssatzes

Wir gehen aus von einem binären symmetrischen Kanal $A_X = A_Y = \{0, 1\}$,

$$\mathbb{P} = \begin{pmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{pmatrix} \quad \text{mit} \quad \varepsilon < \frac{1}{2}\,.$$

Dekodieren: Wenn $1\,(0)$ empfangen wird, schließen wir darauf, dass auch $1\,(0)$ gesandt wurde.

Fehler: $1\,(0)$ wird empfangen, obwohl $0\,(1)$ gesandt wurde.

Die Wahrscheinlichkeit einer fehlerhaften Dekodierung - kurz Fehlerwahrscheinlichkeit - ist ε.

Die Verminderung der Fehlerwahrscheinlichkeit kann durch Hinzufügenvon Redundanz erreicht werden. In der primitivsten Form geschieht dies dadurch, dass anstelle eines Symbols $n = 3$ gleiche Symbole gesandt werden, also anstelle von 0 das Tripel $(0, 0, 0)$ und anstelle von 1 das Tripel $(1, 1, 1)$.

Beispiel 1 $n = 3$, $M = 2$: Beispiel eines $(2, 3)$-Codes mit Codebuch

$$\mathcal{C} = \{(0, 0, 0), (1, 1, 1)\}$$

und Dekodierungsfunktion $g : \{0, 1\}^3 \mapsto \{(0, 0, 0), (1, 1, 1)\}$ gegeben durch

$$g((y_1, y_2, y_3)) = \begin{cases} (0, 0, 0) \triangleq 0 & \text{falls} \quad S_3 < \frac{3}{2} \\ (1, 1, 1) \triangleq 1 & \text{falls} \quad S_3 > \frac{3}{2} \end{cases},$$

wobei $S_3 = \sum_{i=1}^{3} y_i$ die Anzahl der 1-er im empfangenen Tripel $\mathbf{y}^3$ ist. Somit sind

$$\left\{S_3 < \frac{3}{2}\right\} = \{(0, 0, 0), (1, 0, 0), (0, 1, 0), (0, 0, 1)\}$$

und

$$\left\{S_3 > \frac{3}{2}\right\} = \{(1, 1, 0), (1, 0, 1), (0, 1, 1), (1, 1, 1)\}\,.$$

Da sich die beiden Codewörter $(0,0,0)$ und $(1,1,1)$ in allen $r=3$ Koordinaten unterscheiden, korrigiert dieser Code genau einen Fehler. Die Wahrscheinlichkeit einer fehlerhaften Dekodierung ist wegen der Symmetrie des Kanals für beide mögliche Fehler (dekodieren zu $(1,1,1)$, obwohl $(0,0,0)$ übertragen wurde, d.h. obwohl 0 das Ausgangssymbol ist bzw. dekodieren zu $(0,0,0)$, obwohl 1 das Ausgangssymbol ist) gleich

$$\begin{aligned} P_{B_{n,\varepsilon}}\left(S_3 > \frac{3}{2}\right) &= 3\varepsilon^2(1-\varepsilon) + \varepsilon^3 = \varepsilon\left(3\varepsilon - 2\varepsilon^2\right) \\ &= \varepsilon\left(1 - 2\,(1-\varepsilon)\left(\frac{1}{2} - \varepsilon\right)\right) < \varepsilon\,. \end{aligned}$$

Beispiel 2 $n=7$, $M=16$: Für $n=7$ und somit $\left|\{0,1\}^7\right| = 128$ ist es möglich, einen $(M,7)$-Code mit $M=16$ zu konstruieren, der es ebenfalls erlaubt, genau einen Fehler zu korrigieren. Wegen $128 = 16 \cdot 2^3$ ist es nämlich tatsächlich möglich, je zwei der 16 Codewörter so anzuordnen, dass sie sich in mindestens $r=3$ Koordinaten unterscheiden. Ein mögliches derartiges Codebuch ist

$$\begin{aligned} \mathcal{C} = \{\,&(0,0,0,0,0,0,0), && (1,1,0,1,0,0,1),\\ &(1,1,1,0,0,0,0), && (1,1,0,0,1,1,0),\\ &(1,0,0,1,1,0,0), && (1,0,1,1,0,1,0),\\ &(1,0,0,0,0,1,1), && (1,0,1,0,1,0,1),\\ &(0,1,0,1,0,1,0), && (0,1,1,1,1,0,0),\\ &(0,1,0,0,1,0,1), && (0,1,1,0,0,1,1),\\ &(0,0,1,1,0,0,1), && (0,0,0,1,1,1,1),\\ &(0,0,1,0,1,1,0), && (1,1,1,1,1,1,1)\,\}\,. \end{aligned}$$

Die zugehörige Fehlerwahrscheinlichkeit ist demgemäß wie oben.

Sendet man anstelle eines Tripels ein n-Tupel gleicher Symbole und dekodiert analog zu 1, falls die Mehrzahl der erhaltenden Symbole 1 ist, so lässt sich aufgrund von

$$\left\{S_n > \frac{n}{2}\right\} = \left\{S_n - n\varepsilon > n\left(\frac{1}{2} - \varepsilon\right)\right\} \subseteq \left\{|S_n - n\varepsilon| > n\left(\frac{1}{2} - \varepsilon\right)\right\}$$

die Fehlerwahrscheinlichkeit $P_{B_{n,\varepsilon}}\left(S_n > \frac{n}{2}\right)$ mit Hilfe der Tschebyscheffschen Ungleichung folgendermaßen abschätzen

$$P_{B_{n,\varepsilon}}\left(S_n > \frac{n}{2}\right) \leq P_{B_{n,\varepsilon}}\left(|S_n - n\varepsilon| > n\left(\frac{1}{2} - \varepsilon\right)\right) \leq \frac{\varepsilon(1-\varepsilon)}{n\left(\frac{1}{2} - \varepsilon\right)^2}\,.$$

Die obere Schranke konvergiert zwar gegen 0, jedoch sehr langsam. Die sogenannte *Tail-Inequality* liefert eine drastische Verschärfung dieser Abschätzung und lässt folgende exponentielle Konvergenz zu

$$P_{B_{n,\varepsilon}}\left(S_n > \frac{n}{2}\right) \leq 2^{-nI\left(P_{1/2} \| P_\varepsilon\right)}\,.$$

Im Weiteren werden wir anstelle von $\left\{S_n > \frac{n}{2}\right\}$ die Untermenge $\{S_n > n \cdot (\varepsilon + \delta)\}$ mit $\varepsilon < \varepsilon + \delta < \frac{1}{2}$ und demzufolge die nachstehende Abschätzung verwenden

$$P_{B_{n,\varepsilon}}\left(S_n > n \cdot (\varepsilon + \delta)\right) \leq 2^{-nI(P_{\varepsilon+\delta} \| P_\varepsilon)} \,. \tag{2}$$

Die betrachteten Beispiele zeigen, dass es für $n = 3$ und $n = 7$ möglich ist, einen $(2, 3)$-Code bzw. einen $(16, 7)$-Code mit der Fehlerwahrscheinlichkeit

$$3(1 - \varepsilon)\varepsilon^2 + \varepsilon^3 < \varepsilon$$

zu finden. Dies, die Tatsache, dass die Anzahl 2^n der $0 - 1$-Folgen der Länge n exponentiell wächst und die Fehlerwahrscheinlichkeiten gemäß der Beziehung (2) exponentiell gegen 0 fallen, lassen hoffen, dass man für große n die Codes konstruieren kann, die gleichzeitig die folgenden beiden Eigenschaften erfüllen:

(i) die Anzahl M_n der Elemente des Codebuches $\mathcal{C}_n$ ist sehr groß

(ii) die Fehlerwahrscheinlichkeit ist nahezu gleich 0.

Die Frage ist nur: Wie rasch wächst die Anzahl M_n der Elemente des Codebuches $\mathcal{C}_n$ mit n, während die Fehlerwahrscheinlichkeit gegen 0 fällt ?

Die grobe Aussage des Kanalkodierungssatzes ist, dass das Wachstum exponentiell ist

$$M_n \cong 2^{nR} \,,$$

wobei die Wachstumsrate R durch die Kapazität C des Kanals nach oben beschränkt ist.

Dieser Sachverhalt zeigt, dass Kapazität C eine ähnlich fundamentale Größe für einen Kanal ist, wie es die Entropie H für eine Quelle ist.

Die Art, wie *Shannon* den Code konstruiert, ist besonders raffiniert. Er konstruiert ihn nämlich per Zufall: Die Elemente des Codebuches $\mathcal{C}_n$ werden unabhängig voneinander und zufällig gemäß der Gleichverteilung auf A^n erzeugt. Man spricht in diesem Zusammenhang von *random coding*.

2.2.3 Zum Beweis

Definition 1: Die Rate R eines (M, n)-Codes ist

$$R = \frac{\log_2(M)}{n}$$

Definition 2: Die Rate R heißt erreichbar (achievable), wenn es eine Folge von $(\lceil 2^{nR} \rceil, n)$-Codes mit Codebuch $\mathcal{C}_n$ gibt, deren maximale Fehlerwahrscheinlichkeit $\lambda^{(n)}(\mathcal{C}_n)$ mit wachsendem n gegen 0 geht, d.h. wenn gilt

$$\lim_{n \to \infty} \lambda^{(n)}(\mathcal{C}_n) = 0 \,.$$

Der Kanalkodierungssatz und dessen Umkehrung, welche das Thema des nächsten Abschnitts ist, lauten wie folgt.

Satz 1 (Kanalkodierungssatz): Sei $(A_X, A_Y, \mathbb{P})$ ein gedächtnisloser Kanal mit Kapazität C . Dann ist jede Rate $R \in (0, C)$ erreichbar.

Umkehrung des Kanalkodierungssatzes: Für jede Folge von $(\lceil 2^{nR} \rceil, n)$-Codes mit Codebuch $\mathcal{C}_n$ und einer Rate $R > C$ gilt

$$\lim_{n \to \infty} \inf \lambda^{(n)}(\mathcal{C}_n) > 0 \,.$$

Wir werden uns beim Beweis des Kanalkodierungssatzes auf den Fall des binären symmetrischen Kanals beschränken. Seien also $n \in \mathbb{N}$, $A_X = A_Y = A = \{0, 1\}$ und

$$\mathbb{P} = \begin{pmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{pmatrix} \quad \text{mit } \varepsilon < \frac{1}{2} \,.$$

Zur Formulierung der Dekodierungsregel treffen wir folgende Vorbereitungen.

Definition 3 (*Hamming* 1950): Seien $\mathbf{x}^n, \mathbf{y}^n \in A^n$. Dann heißt die durch

$$d(\mathbf{x}^n, \mathbf{y}^n) = \sum_{i=1}^{n} |x_i - y_i|$$

definierte Funktion $d : (A^n)^2 \mapsto [0, \infty)$ die *Hamming-Distanz* von $\mathbf{x}^n$ und $\mathbf{y}^n$.

Bezeichnung: Sei $r \in \{0, ..., n\}$. Dann ist

$$S_r(\mathbf{y}^n) = \{\, \mathbf{x}^n \in A^n : d(\mathbf{x}^n, \mathbf{y}^n) \leq r \,\}$$

die *Hamming-Kugel* vom Radius r um den Punkt $\mathbf{y}^n \in A^n$.

Beweis des Kanalkodierungssatzes: Seien $\mathcal{C}_n = (\mathbf{c}_1, ..., \mathbf{c}_M) \in (A^n)^M$ ein Codebuch und die Dekodierungsfunktion $g : A^n \mapsto \{1, ..., M\}$ wie folgt definiert

$$g(\mathbf{y}^n) = \begin{cases} \arg\min\{\, d(\mathbf{y}^n, \mathbf{c}_i) : i \in \{1, ..., M\} \,\} & \text{falls dieses eindeutig ist} \\ 1 \text{ (etwa)} & \text{andernfalls} \,. \end{cases}$$

Bekanntlich ist die Kapazität des binären symmetrischen Kanals

$$C = I((\varepsilon, 1-\varepsilon) \,\|\, P_X^*)$$

mit $P_X^* = \left(\frac{1}{2}, \frac{1}{2}\right)$.

Das Codebuch $\mathcal{C}_n$ wird nun gemäß der Gleichverteilung auf $(A^n)^M$ zufällig gewählt. Es sei somit

$$P_n^*(\mathcal{C}_n = (\mathbf{c}_1, ..., \mathbf{c}_M)) = \prod_{i=1}^{M} P^*(\mathbf{C}_i = \mathbf{c}_i) = \left(\frac{1}{2}\right)^{n \cdot M} \quad \forall\, (\mathbf{c}_1, ..., \mathbf{c}_M) \in (A^n)^M \,.$$

Demgemäß ist der zugehörigen Erwartungswert der mittleren Fehlerwahrscheinlichkeit

$$\begin{aligned} E_{P_n^*}\left(P_e^{(n)}(\mathcal{C}_n)\right) &= E_{P_n^*}\left(\frac{1}{M}\sum_{i=1}^{M}\lambda_i(\mathcal{C}_n)\right) \\ &= \frac{1}{M}\sum_{i=1}^{M} E_{P_n^*}\left(P_{\mathcal{C}_n}\left(g\left(\mathbf{Y}^n\right)\neq i \,/\, \mathbf{X}^n=\mathbf{C}_i\right)\right) \end{aligned}$$

mit

$$\begin{aligned} E_{P_n^*}\left(P_{\mathcal{C}_n}\left(g\left(\mathbf{Y}^n\right)\neq i \,/\, \mathbf{X}^n=\mathbf{C}_i\right)\right) &= \sum_{(c_1,\ldots,c_M)\in(A^n)^M} P_n^*\left(\mathcal{C}_n=(\mathbf{c}_1,\ldots,\mathbf{c}_M)\right)\cdot \\ &\quad \cdot P_{(\mathbf{c}_1,\ldots,\mathbf{c}_M)}\left(g\left(\mathbf{Y}^n\right)\neq i \,/\, \mathbf{X}^n=\mathbf{c}_i\right) . \end{aligned}$$

Im Weiteren werden wir die (bedingte) Fehlerwahrscheinlichkeit

$$P\left(g\left(\mathbf{Y}^n\right)\neq i\right) = P_{(\mathbf{c}_1,\ldots,\mathbf{c}_M)}\left(g\left(\mathbf{Y}^n\right)\neq i \,/\, \mathbf{X}^n=\mathbf{c}_i\right)$$

abschätzen. Mit $r_n \in (0,n)$ kann eine fehlerhafte Dekodierung in folgenden beiden Fällen eintreten:

· entweder es ist $d(\mathbf{c}_i,\mathbf{Y}^n) > r_n$

· oder es ist $d(\mathbf{c}_i,\mathbf{Y}^n) \leq r_n$ und es gibt mindestens ein weiteres $j \in \{1,\ldots,M_n\}\backslash\{i\}$ mit $d(\mathbf{c}_j,\mathbf{Y}^n) \leq r_n$.

Somit gilt

$$\begin{aligned} \{g\left(\mathbf{Y}^n\right)\neq i\} &\subseteq \{d(\mathbf{c}_i,\mathbf{Y}^n) > r_n\} \cup \left(\{d(\mathbf{c}_i,\mathbf{Y}^n)\leq r_n\} \cap \cup_{j\neq i}\{d(\mathbf{c}_j,\mathbf{Y}^n)\leq r_n\}\right) \\ &\subseteq \{d(\mathbf{c}_i,\mathbf{Y}^n) > r_n\} \cup \left(\cup_{j\neq i}\{d(\mathbf{c}_j,\mathbf{Y}^n)\leq r_n\}\right) \end{aligned}$$

und daher

$$\begin{aligned} P\left(g\left(\mathbf{Y}^n\right)\neq i\right) &\leq P\left(d(\mathbf{c}_i,\mathbf{Y}^n) > r_n\right) + P\left(\cup_{j\neq i}\{d(\mathbf{c}_j,\mathbf{Y}^n)\leq r_n\}\right) \\ &\leq P\left(d(\mathbf{c}_i,\mathbf{Y}^n) > r_n\right) + \sum_{j\neq i} P\left(d(\mathbf{c}_j,\mathbf{Y}^n)\leq r_n\right) . \end{aligned}$$

Für die folgende Abschätzung stellt sich folgende Wahl von r_n als zweckmäßig heraus

$$r_n = n\left(\varepsilon+\delta\right) \text{ mit } \varepsilon < \varepsilon+\delta < \frac{1}{2} \text{ und } R < I\left(P_{\varepsilon+\delta}\|P_{1/2}\right) .$$

Da $S_n = d(\mathbf{c}_i,\mathbf{Y}^n)$ $B_{n,\varepsilon}$-verteilt ist, lässt sich der erste Term durch

$$P_{(\mathbf{c}_1,\ldots,\mathbf{c}_M)}\left(d(\mathbf{c}_i,\mathbf{Y}^n) > r_n \,/\, \mathbf{X}^n=\mathbf{c}_i\right) = P_{B_{n,\varepsilon}}\left(S_n > n(\varepsilon+\delta)\right) \leq 2^{-nI(P_{\varepsilon+\delta}\|P_\varepsilon)}$$

abschätzen.

Der Erwartungswert (bzgl. P_n^*) jedes der restlichen $M-1$ Terme lässt sich für gegebene $\mathbf{c}_i$ und $\mathbf{c}_j$ wegen

$$\sum_{(\mathbf{c}_1,\ldots,\mathbf{c}_i,\ldots,\mathbf{c}_j,\ldots,\mathbf{c}_M)\in(A^n)^{M-2}} \prod_{k\in\{1,\ldots,M\}\setminus\{i,j\}}^{M} P^*\left(\mathbf{C}_k=\mathbf{c}_k\right) = P^*\left(\mathbf{C}_i=\mathbf{c}_i\right) P^*\left(\mathbf{C}_j=\mathbf{c}_j\right)$$

und $1_{\{S_{r_n}(\mathbf{c}_j)\}}(\mathbf{y}^n) = 1_{\{S_{r_n}(\mathbf{y}^n)\}}(\mathbf{c}_j)$ zunächst so umformen

$$\begin{aligned}
E &= \sum_{(\mathbf{c}_1,\ldots,\mathbf{c}_M)\in(A^n)^M} \prod_{k=1}^{M} P^*\left(\mathbf{C}_k=\mathbf{c}_k\right) P\left(d(\mathbf{c}_j,\mathbf{Y}^n)\le r_n \,/\, \mathbf{X}^n=\mathbf{c}_i\right) \\
&= \sum_{(\mathbf{c}_i,\mathbf{c}_j)\in(A^n)^2} P^*\left(\mathbf{C}_i=\mathbf{c}_i\right) P^*\left(\mathbf{C}_j=\mathbf{c}_j\right) \sum_{y^n\in A^n} 1_{\{S_{r_n}(\mathbf{c}_j)\}}(\mathbf{y}^n) P(\mathbf{Y}^n=\mathbf{y}^n \,/\, \mathbf{X}^n=\mathbf{c}_i) \\
&= \sum_{(\mathbf{c}_i,\mathbf{y}^n)\in(A^n)^2} P^*\left(\mathbf{C}_i=\mathbf{c}_i\right) P(\mathbf{Y}^n=\mathbf{y}^n \,/\, \mathbf{X}^n=\mathbf{c}_i) \sum_{\mathbf{c}_j\in A^n} 1_{\{S_{r_n}(\mathbf{y}^n)\}}(\mathbf{c}_j) P^*\left(\mathbf{C}_j=\mathbf{c}_j\right) \\
&= \sum_{(\mathbf{c}_i,\mathbf{y}^n)\in(A^n)^2} P^*\left(\mathbf{C}_i=\mathbf{c}_i\right) P(\mathbf{Y}^n=\mathbf{y}^n \,/\, \mathbf{X}^n=\mathbf{c}_i) P^*\left(\mathbf{C}_j\in S_{r_n}\left(\mathbf{y}^n\right)\right) \\
&= \sum_{(\mathbf{c}_i,\mathbf{y}^n)\in(A^n)^2} P^*\left(\mathbf{C}_i=\mathbf{c}_i\right) P(\mathbf{Y}^n=\mathbf{y}^n \,/\, \mathbf{X}^n=\mathbf{c}_i) P^*\left(d(\mathbf{C}_j,\mathbf{y}^n)\le r_n\right) .
\end{aligned}$$

Da $\mathbf{C}_j$ auf A^n gleichverteilt ist, ist $S_n^* = d\left(\mathbf{C}_j,\mathbf{y}^n\right) \sim B_{n,1/2}$. Somit ist

$$\begin{aligned}
P^*\left(d(\mathbf{C}_j,\mathbf{y}^n)\le r_n\right) &= P_{B_{n,1/2}}(S_n^*\le n(\varepsilon+\delta)) \\
&\le 2^{-nI\left(P_{\varepsilon+\delta}\|P_{1/2}\right)} .
\end{aligned}$$

Für $M = M_n = \lceil 2^{nR}\rceil$ lässt sich also der Erwartungswert der $M-1$ letzten Summanden wegen

$$M_n - 1 = \lceil 2^{nR}\rceil - 1 < 2^{nR} \tag{3}$$

durch

$$(M_n-1)\,2^{-nI\left(P_{\varepsilon+\delta}\|P_{1/2}\right)} \le 2^{-n\left(I\left(P_{\varepsilon+\delta}\|P_{1/2}\right)-R\right)}$$

abschätzen. Zusammenfassend gilt also

$$E_{P_n^*}\left(P_E^{(n)}(\mathcal{C}_n)\right) \le 2^{-nI(P_{\varepsilon+\delta}\|P_\varepsilon)} + 2^{-n\left(I\left(P_{\varepsilon+\delta}\|P_{1/2}\right)-R\right)} ,$$

wobei bekanntlich $I\left(P_{\varepsilon+\delta}\|P_\varepsilon\right) > 0$ ist und voraussetzungsgemäß $R < I\left(P_{\varepsilon+\delta}\|P_{1/2}\right) < I\left(P_\varepsilon\|P_{1/2}\right) = C$ gilt.

Daher gibt es für jedes vorgegebene $\eta > 0$ ein $N_0(\eta) \in \mathbb{N}$, sodass der Erwartungswert der mittleren Fehlerwahrscheinlichkeit (hinsichtlich P_n^* auf der Menge aller Codes mit Indexmenge $\{1,\ldots,\lceil 2^{nR}\rceil\}$) kleiner oder gleich η ist, soferne $n \ge N_0(\eta)$ ist.

Somit gibt es für alle $n \geq N_0(\eta)$ einen $\left(\lceil 2^{nR} \rceil, n\right)$-Code mit mittlerer Fehlerwahrscheinlichkeit $\leq \eta$.

Um den Beweis zu beenden, brauchen wir nur noch "mittlere Fehlerwahrscheinlichkeit" durch "maximale Fehlerwahrscheinlichkeit" zu ersetzen. Dazu benötigen wir folgendes einfach zu beweisende Lemma.

Lemma 1: Seien $x_1, ..., x_n \geq 0$ mit Mittelwert

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$$

und $A = \{i \in \{1, ..., n\} : x_i \leq 2\bar{x}_n\}$. Dann gilt

$$|A| > \frac{n}{2} .$$

Beweis: Wegen

$$\begin{aligned} n\,\bar{x}_n &= \sum_{i=1}^{n} x_i \geq \sum_{i \in A^c} x_i \\ &> \sum_{i \in A^c} 2\bar{x}_n = 2\,|A^c|\,\bar{x}_n \end{aligned}$$

gilt $|A^c| < \frac{n}{2}$ und daher wegen $n = |A| + |A^c|$ die Behauptung. $\square$

Sei nun $R' \in (R, C)$, dann gilt wegen (3) für hinreichend große $n \in \mathbb{N}$

$$2M_n < 2\left(2^{nR} + 1\right) \leq 2^{nR'} + 1 .$$

Sei weiters $\eta > 0$ und $N_0'(\eta/2) \in \mathbb{N}$ derart, dass obige Beziehung gilt und für $n \geq N_0'(\eta/2)$ $\mathcal{C}_n' = \{\mathbf{c}_1, ..., \mathbf{c}_{2M_n}\}$ das Codebuch eines $(2M_n, n)$-Codes mit mittlerer Fehlerwahrscheinlichkeit

$$\frac{1}{2M_n} \sum_{i=1}^{2M_n} \lambda_i(\mathcal{C}_n') \leq \frac{\eta}{2}$$

ist. Dabei sind selbstverständlich alle $\lambda_i(\mathcal{C}_n') = P_{\mathcal{C}_n'}(g(\mathbf{Y}^n) \neq i \,/\, \mathbf{X}^n = \mathbf{C}_i) \geq 0$. Aufgrund unseres Lemmas sind $\tilde{M}_n \geq M_n$ dieser Größen

$$\lambda_i(\mathcal{C}_n') \leq \eta .$$

Sei $\{i_1, ..., i_{\tilde{M}_n}\}$ die Menge der zugehörigen Indizes. Dann gilt für den Code mit dem Codebuch $\tilde{\mathcal{C}}_n = \{c_{i_1}, ..., c_{i_{\tilde{M}_n}}\}$ - wegen $\lambda_{i_j}\left(\tilde{\mathcal{C}}_n\right) \leq \lambda_{i_j}(\mathcal{C}_n')$ -

$$\lambda^{(n)}\left(\tilde{\mathcal{C}}_n\right) \leq \max\left\{\lambda_{i_j}(\mathcal{C}_n'),\ j \in \left\{1, ..., \tilde{M}_n\right\}\right\} \leq \eta . \ \square$$

2.3 DIE UMKEHRUNG DES KANALKODIERUNGSSATZES

2.3.1 Die Ungleichung von Fano

In diesem Abschnitt sei (X,Y) ein Zufallsvektor mit gemeinsamer Verteilung $P = (p(x,y) : (x,y) \in W_X \times W_Y)$, wobei die Trägermengen W_X und W_Y der Randverteilungen der Zufallsvariablen X und Y gleich sind, d.h. es gilt $W_X = W_Y$.

Lemma 1 (Ungleichung von Fano): Seien (X,Y) ein Zufallsvektor mit gemeinsamer Verteilung P auf W_X^2 und $p_E = P(X \neq Y)$. Dann gilt mit $M = |W_X|$

$$H(X/Y) \leq H(p_E, 1-p_E) + p_E \log_2(M-1) \ .$$

Beweis: Es sei

$$E = \begin{cases} 1 & \text{für} \quad X \neq Y \\ 0 & \text{für} \quad X = Y \end{cases} = 1_{\{1,\ldots,M\}\setminus\{X\}}(Y) \ .$$

Bekanntlich gilt

$$\begin{aligned} H((X,E)/Y) &= H(X/Y) + H(E/X,Y) \\ &= H(E/Y) + H(X/E,Y) \ . \end{aligned}$$

Für die letzte Größe der ersten Zeile gilt

$\cdot\ H(E/X,Y) = 0$ da E eine Funktion von X und Y ist.

Die beiden Größen der zweiten Zeile lassen sich folgendermaßen nach oben abschätzen:

$\cdot\ H(E/Y) \leq H(E) = H(p_E, 1-p_E)$ (wegen $I(E;Y) = H(E) - H(E/Y) \geq 0$)

$\cdot\ H(X/E,Y) \leq p_E \log_2(M-1)$ denn

$$\begin{aligned} H(X/E,Y) &= P(E=0) \cdot H(X/Y, E=0) + P(E=1) \cdot H(X/Y, E=1) \\ &= P(E=1) \cdot H(X/Y, E=1) \\ &\leq p_E \cdot \log_2(M-1) \ . \end{aligned}$$

Ersteres, weil im Fall $E=0$ $X=Y$ und somit $H(X/Y, E=0) = 0$ ist und letzteres, weil im Fall $E=1$ der Träger der bedingten Verteilung von X, gegeben $Y=y$, die $M-1$-elementige Menge $\{1,\ldots,M\}\setminus\{y\}$ und somit wegen $H(X/Y=y, E=1) \leq \log_2(M-1)$

$$\begin{aligned} H(X/Y, E=1) &= \sum_{y=1}^{M} P(Y=y) \cdot H(X/Y=y, E=1) \leq \\ &\leq \sum_{y=1}^{M} P(Y=y) \cdot \log_2(M-1) = \log_2(M-1) \end{aligned}$$

ist. Zusammenfassend gilt also

$$\begin{aligned} H\left(X\,/\,Y\right)+0 &= H\left(E\,/\,Y\right)+H\left(X\,/\,E,Y\right)\leq \\ &\leq H\left(p_E,1-p_E\right)+p_E\cdot\log_2\left(M-1\right)\ .\ \square \end{aligned}$$

2.3.2 Die Data Processing Inequality

Im Folgenden seien (X,Y) ein Zufallsvektor mit gemeinsamer Verteilung $P=(p(x,y):(x,y)\in W_X\times W_Y)$, $P_X=(p(x)=\sum_y p(x,y):x\in W_X)$ und $P_Y=(q(y)=\sum_x p(x,y):y\in W_Y)$ die zugehörigen Randverteilungen und W_X und W_Y deren Trägermengen.

Weiters seien

$$Z=f(Y)\ \text{ mit }\ f:W_Y\mapsto\mathbb{R}\,,$$

$R=(\,r(x,z)=\sum_{y\in f^{-1}(z)}p(x,y):(x,z)\in W_X\times f\left(W_Y\right)\,)$ die gemeinsame Verteilung von (X,Z) und $R_Z=(\,r(z)=\sum_{y\in f^{-1}(z)}q(y):z\in f\left(W_Y\right)\,)$ die Randverteilung von Z .

Für die Formulierung des nachstehenden Satzes benötigen wir schließlich noch die bedingten Verteilungen von X bei gegebenem Y bzw. Z , nämlich $P(X\,/Y=y)=\left(p(x/y)=\frac{p(x,y)}{q(y)}:x\in W_X\right)$ und $P(X\,/\,Z=z)=\left(r(x/z)=\frac{r(x,z)}{r(z)}:x\in W_X\right)$, $y\in W_Y$, $z\in W_Z$.

Lemma 1 (Data Processing Inequality): Seien (X,Y) ein Zufallsvektor mit gemeinsamer Verteilung P , $f:W_Y\mapsto\mathbb{R}$ und $Z=f(Y)$. Dann ist

$$I\left(X;Y\right)\geq I\left(X;f\left(Y\right)\right)\,,$$

wobei Gleichheit genau dann zutrifft, wenn gilt

$$P(X\,/\,Y=y)=P(X\,/\,Z=z)\ \ \forall\,y\in f^{-1}\left(z\right),\ z\in f(W_Y)\,.$$

Anmerkung 1: Die Bedingung für die Gleichheit ist jedenfalls dann erfüllt, wenn die Funktion f umkehrbar eindeutig ist.

Beweis: Wegen $p(x,y)=q(y)p(x/y)$, $r(x,z)=r(z)r(x/z)$ und der

Additivität des Logarithmus gilt

$$\begin{aligned} I(X;Y) - I(X;f(Y)) &= \sum_{(x,y)\in W_X\times W_Y} p(x,y)\log_2\left(\frac{p(x,y)}{p(x)\cdot q(y)}\right) \\ &\quad - \sum_{(x,z)\in W_X\times f(W_Y)} r(x,z)\log_2\left(\frac{r(x,z)}{p(x)\cdot r(z)}\right) \\ &= \sum_{(x,y)\in W_X\times W_Y} p(x,y)\log_2\left(\frac{p(x/y)}{p(x)}\right) \\ &\quad - \sum_{(x,z)\in W_X\times f(W_Y)} \sum_{y\in f^{-1}(z)} p(x,y)\log_2\left(\frac{r(x/z)}{p(x)}\right) \end{aligned}$$

$$\begin{aligned} &= \sum_{x\in W_X} \sum_{z\in f(W_Y)} \sum_{y\in f^{-1}(z)} q(y)p(x/y)\log_2\left(\frac{p(x/y)}{r(x/z)}\right) \\ &= \sum_{z\in f(W_Y)} \sum_{y\in f^{-1}(z)} q(y) \sum_{x\in W_X} p(x/y)\log_2\left(\frac{p(x/y)}{r(x/z)}\right) \\ &= \sum_{z\in f(W_Y)} \sum_{y\in f^{-1}(z)} q(y)\cdot I\left(P(X\,/\,Y=y)\|P(X\,/\,Z=z)\right). \end{aligned}$$

Dies impliziert zusammen mit $q(y) > 0 \;\forall\, y \in W_Y$ und der Eigenschaft der I-Divergenz die Behauptung. $\square$

2.3.3 Beweis der Umkehrung

Im Folgenden seien - wie gehabt - $(A_X, A_Y, \mathbb{P})$ ein gedächtnisloser Kanal mit Kapazität C, $(A_X^n, A_Y^n, \mathbb{P}^{(n)})$ dessen n-te Fortsetzung und $\mathcal{C}_n = (\mathbf{c}_1, ..., \mathbf{c}_{M_n})$ das Codebuch eines (M_n, n)-Codes mit $M_n = \lceil 2^{nR} \rceil$, $n \in \mathbb{N}$ und $R > C$. Schließlich seien I eine Zufallsvariable, welche auf $\{1, ..., M_n\}$ gleichverteilt ist, $\mathbf{X}^n$ eine eineindeutige Kodierungsfunktion, g eine Dekodierungsfunktion und $P_e^{(n)}(\mathcal{C}_n) = \frac{1}{M}\sum_{i=1}^{M} \lambda_i(\mathcal{C}_n) = P_{\mathcal{C}}\left(g\left(\mathbf{Y}^n\right) \neq I\right)$ die zugehörige mittlere Fehlerwahrscheinlichkeit.

Satz 1 (Umkehrung des Kanalkodierungssatzes): Sei $(A_X, A_Y, \mathbb{P})$ ein gedächtnisloser Kanal mit Kapazität C. Dann gilt für jede Folge von $(\lceil 2^{nR} \rceil, n)$-Codes mit einer Rate $R > C$

$$\lim_{n\to\infty} \inf\, P_e^{(n)}(\mathcal{C}_n) > 0\,.$$

Beweis: Aufgrund der Data Processing Inequality, der Eineindeutigkeit

von $\mathbf{X}^n$ und der bereits festgehaltenen Tatsache $I\left(\mathbf{X}^n;\mathbf{Y}^n\right) \leq nC$ gilt

$$\begin{aligned} H(I) - H(I \,/\, g(\mathbf{Y}^n)) &= I\left(I; g(\mathbf{Y}^n)\right) \\ &= I\left(\mathbf{X}^n\left(I\right); g(\mathbf{Y}^n)\right) \\ &\leq I\left(\mathbf{X}^n\left(I\right); \mathbf{Y}^n\right) \\ &= \frac{1}{M_n}\sum_{i=1}^{M_n} I\left(\mathbf{X}^n\left(i\right); \mathbf{Y}^n\right) \\ &\leq nC\,. \end{aligned}$$

Wegen $H\left(I\right) = \log_2\left(M_n\right) \geq nR$ erhält man demzufolge

$$nR \leq H(I) \leq nC + H(I \,/\, g(\mathbf{Y}^n))\,.$$

Wendet man die Ungleichung von *Fano* auf $H(I \,/\, g(\mathbf{Y}^n))$ an, so ergibt sich folgende Aussage über die mittlere Fehlerwahrscheinlichkeit $p_n = P_e^{(n)}\left(\mathcal{C}_n\right) = P_{\mathcal{C}_n}\left(g\left(\mathbf{Y}^n\right) \neq I\right)$

$$\begin{aligned} H(I \,/\, g(\mathbf{Y}^n)) &\leq H\left(p_n, 1-p_n\right) + p_n \cdot \log_2\left(M_n - 1\right) \\ &< 1 + p_n \cdot nR\,, \end{aligned}$$

wobei $H\left(p_n, 1-p_n\right) \leq 1$ und $\log_2\left(M_n - 1\right) < nR$ berücksichtigt wird. (Letzteres folgt aus $M_n = \lceil 2^{nR} \rceil \in [2^{nR}, 2^{nR} + 1)$.) Zusammenfassend erhält man - nach Division durch n -

$$R \leq C + \frac{1}{n} + p_n R$$

und daher

$$p_n \geq 1 - \frac{C}{R} - \frac{1}{nR}\,.$$

Demzufolge ist wegen $R > C$

$$\lim_{n\to\infty} \inf\, p_n \geq 1 - \frac{C}{R} > 0\,. \;\square$$

2.4 AUSBLICK: RÉNYIS ENTROPIEN ORDNUNG α

2.4.1 Der Index der Übereinstimmung

Definition 1: Seien X, Y unabhängige Zufallsvariable mit Werten aus $\{0, ..., m-1\}$ und der Verteilung $P = (p_0, ..., p_{m-1})$. Dann ist die Wahrscheinlichkeit des Ereignisses $\{X = Y\}$, dass die beiden Zufallsvariablen übereinstimmen,

$$\kappa(P) = P(X = Y) = P^2 = \sum_{i=0}^{m-1} p_i^2 .$$

Dieser Wert heißt der *Index der Übereinstimmung* (der *Kappa-Wert* einer Sprache).[8]

Für den Wertebereich des Index der Übereinstimmung gilt folgende

Proposition 1: Sei $P = (p_0, ..., p_{m-1})$ eine Wahrscheinlichkeitsverteilung. Dann gilt

$$\frac{1}{m} \leq \sum_{i=0}^{m-1} p_i^2 \leq \max(p_i : i \in \{0, ..., m-1\}) \leq 1$$

mit Gleichheit für die erste und letzte Ungleichung jeweils genau dann, wenn gilt

$$\begin{aligned} p_i &= \frac{1}{m}\,,\ i \in \{0, ..., m-1\} \quad \text{bzw.} \\ p_i &= 1_{i_0}(i)\,,\ i \in \{0, ..., m-1\} \ \text{ für ein } \ i_0 \in \{0, ..., m-1\}\ . \end{aligned}$$

Beweis: Die erste Ungleichung mit dem zugehörigen Kriterium für die Gleichheit folgt aus

$$\sum_{i=0}^{m-1} p_i^2 = \frac{1}{m} + \sum_{i=0}^{m-1} \left(p_i - \frac{1}{m}\right)^2 .$$

Die zweite folgt wegen $\sum_{i=0}^{m-1} p_i = 1$ aus

$$\sum_{i=0}^{m-1} p_i^2 = \sum_{i=0}^{m-1} p_i - \sum_{i=0}^{m-1} p_i(1-p_i) = 1 - \sum_{i=0}^{m-1} p_i(1-p_i)\,. \ \square$$

Sprache	κ−Wert		Sprache	κ−Wert
Englisch	0.066		Portugiesisch	0.079
Französisch	0.076		Spanisch	0.078
Deutsch	0.076		Russisch	0.053

[8] Der Index der Übereinstimmung wurde in der Arbeit des US-amerikanischen Kryptoanalytikers *William F. Friedman* (1891 – 1969) in dessen Arbeit *The Index of Coincidence and its Applications in Cryptography* im Jahre 1920 eingeführt (siehe [3]).

Tabelle 1: Die $\boldsymbol{\kappa}-$Werte verschiedener Sprachen (Untere Schranke: $\frac{1}{26} = 0.038$).

Durch Anwendung der streng monoton fallenden Funktion $u \mapsto -\log_2 u\,,\ u \in (0,\infty)$ auf die Aussage von Proposition 1 erhält man

$$\log_2 m \geq -\log_2 \left(\sum_{i=0}^{m-1} p_i^2 \right) \geq -\log_2 \left(\max(p_i : i \in \{0,...,m-1\}) \right) \geq 0\,,$$

was analog zum entsprechenden Sachverhalt

$$\log_2 m \geq -\sum_{i=0}^{m-1} p_i \log_2 (p_i) \geq -\log_2 \left(\max(p_i : i \in \{0,...,m-1\}) \right) \geq 0$$

für die Shannonsche Entropie ist, einschließlich der zugehörigen Aussagen hinsichtlich der Gleichheit in der ersten und letzten Ungleichung.

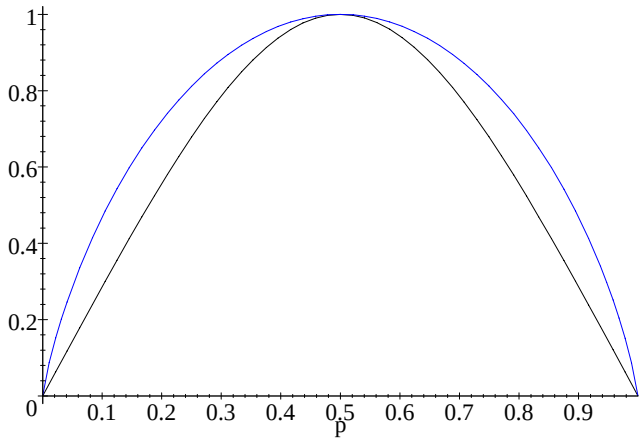


Abbildung 1: Graph der Funktionen $h_1(p)$ und $h_2(p) = -\log_2(p^2 + (1-p)^2)$

Rényis[9] Familie der Entropien der Ordnung α setzen die für $\alpha = 1$ und 2 - und darüber hinaus die für 0 und ∞ - gegebenen Informationsmaße auf $\alpha \in [0,\infty]$ fort.

Definition 2: Sei $P = (p_0,...,p_{m-1})$ eine Wahrscheinlichkeitsverteilung mit Träger $\{0,...,m-1\}$ und sei $\alpha \in [0,\infty]$. Dann heißt die durch

$$H_\alpha(P) = \begin{cases} \frac{1}{1-\alpha} \log_2(\sum_{i=0}^{m-1} p_i^\alpha) & \text{für} \quad \alpha \in [0,\infty)\backslash\{1\} \\ -\sum_{i=0}^{m-1} p_i \log_2 (p_i) & \text{für} \quad \alpha = 1 \\ -\log_2 \left(\max(p_i : i \in \{0,...,m-1\}) \right) & \text{für} \quad \alpha = \infty \end{cases}$$

[9] *Alfréd Rényi* (1921 – 1970), ungarischer Mathematiker

gegebene Größe *Entropie von P der Ordnung α*.

Anmerkung 1: Es gelten

$$\lim_{\alpha\to 1} H_\alpha(P) = H_1(P) \quad \text{und} \quad \lim_{\alpha\to\infty} H_\alpha(P) = H_\infty(P)\,.$$

und

$$H_0(P) \geq H_\alpha(P) \geq H_\infty(P) \geq 0\,,$$

wobei in der ersten Ungleichung Gleichheit genau dann gilt, wenn P die Gleichverteilung ist und in der letzten Ungleichung genau dann, wenn P eine Punktverteilung ist.

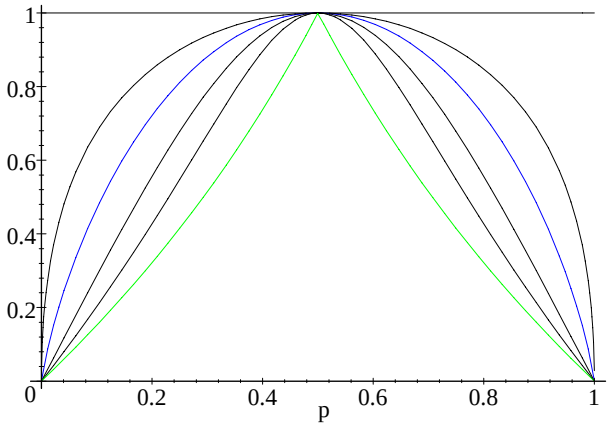


Abbildung 2: Graph der Funktionen $h_\alpha(p)$ für $\alpha \in \{0, 1/2, 1, 2, 4, \infty\}$

Rényi stand in der Tradition der russischen wahrscheinlichkeitstheoretischen Schulen um *Yu. V. Linnik* $(1915-1972)$ und *A.N. Kolmogoroff* $(1903-1987)$.

2.4.2 Die Klasse der homogenen Kolmogoroff-Nagumo-Mittel

Eine allgemeine Klasse von Mitteln, die die drei klassischen Mittel, nämlich das arithmetische Mittel (für $\beta = 1$), das geometrische Mittel (für $\beta = 0$) und das harmonisches Mittel (für $\beta = -1$) als Spezialfälle enthält, wurde von *Kolmogoroff* und *Nagumo*[10] vorgeschlagen. Sie ist - in ihrer positiv homogenen

[10] *Andrej Nikolajewitsch Kolmogoroff*, russischer Mathematiker, Begründer der modernen Wahrscheinlichkeitstheorie
Mitio Nagumo $(1905-....)$, japanischer Mathematiker

Form - gegeben durch

$$M_\beta(x_1,...,x_n) = \begin{cases} \min(x_1,...,x_n) & \text{für} \quad \beta = -\infty \\ (\frac{1}{n}\sum\limits_{j=1}^{n} x_j^\beta)^{1/\beta} & \text{für} \quad \beta \in \mathbb{R}\setminus\{0\} \\ \sqrt[n]{\prod\limits_{j=1}^{n} x_j} & \text{für} \quad \beta = 0\,, \\ \max(x_1,...,x_n) & \text{für} \quad \beta = \infty \end{cases}$$

wobei folgende einschränkenden Voraussetzungen hinsichtlich der Beobachtungswerte $x_1,...,x_n \in \mathbb{R}$ zu treffen sind

$$\begin{array}{ll} x_1,...,x_n \geq 0 & \text{für} \quad \beta \in [0,\infty]\backslash\mathbb{N} \\ x_1,...,x_n > 0 & \text{für} \quad \beta \in [-\infty,0)\,. \end{array}$$

Anmerkung 2: Analog zum arithmetischen Mittel

$$\bar{x}_n = \frac{1}{n}\sum_{j=1}^{n} x_j$$

lässt sich auch die vorliegende Verallgemeinerung mit Hilfe der relativen Häufigkeiten h_i der einzelnen Ausfälle $\omega_i\,,\ i \geq 0$, der Variablen X darstellen:

$$M_\beta(x_1,...,x_n) = \begin{cases} \inf(\omega_i : i \geq 0) & \text{für} \quad \beta = -\infty \\ (\sum\limits_{i\geq 0} \omega_i^\beta \cdot h_i)^{1/\beta} & \text{für} \quad \beta \in \mathbb{R}\setminus\{0\} \\ \prod\limits_{i\geq 0} \omega_i^{h_i} & \text{für} \quad \beta = 0 \\ \sup(\omega_i : i \geq 0) & \text{für} \quad \beta = \infty\,. \end{cases}$$

Die so definierte Klasse von Mitteln hat folgende Eigenschaften

- $M_\beta(cx_1,...,cx_n) = cM_\beta(x_1,...,x_n) \quad \forall c > 0$
  Diese Eigenschaft der positiven Homogenität ist charakteristisch für diese Klasse.
- $M_0(x_1,...,x_n) = \lim\limits_{\beta\to 0} M_\beta(x_1,...,x_n)$
- $M_{-\infty}(x_1,...,x_n) = \lim\limits_{\beta\downarrow -\infty} M_\beta(x_1,...,x_n) \quad \text{und} \quad M_\infty(x_1,...,x_n) = \lim\limits_{\beta\uparrow\infty} M_\beta(x_1,...,x_n)$
- $M_{\beta_1}(x_1,...,x_n) \leq M_{\beta_2}(x_1,...,x_n) \quad \forall \beta_1 < \beta_2\,,$
  wobei Gleichheit jeweils genau dann gilt, wenn $x_1 = ... = x_n$ ist.

Beschränkt man sich nun β auf $[-1,\infty)$ und die Menge der möglichen Ausfälle auf $\{0,...,m-1\}$, wählt für $\beta = \alpha - 1$, ersetzt die relativen Häufigkeiten h_i durch

zugehörige Wahrscheinlichkeiten p_i und legt schließlich die Werte durch $\omega_i = p_i$ fest, so ist im Fall $\alpha \in [0,\infty)\backslash\{1\}$

$$H_\alpha(P) = -\log_2 M_{\alpha-1}(P) = -\log_2(\sum_{i=0}^{m-1} p_i^{\alpha-1} \cdot p_i)^{1/(\alpha-1)} = \frac{1}{1-\alpha}\log_2\left(\sum_{i=0}^{m-1} p_i^\alpha\right)$$

und im Fall $\alpha = 1$

$$H_1(P) = -\log_2 M_0(P) = -\log_2\left(\prod_{i=0}^{m-1} p_i^{p_i}\right) = -\sum_{i=0}^{m-1} p_i \log_2(p_i)\,.$$

Der Fall $\alpha = \infty$ wird analog behandelt.

Anmerkung 3: Es gelten

$$H_\alpha(P) - H_1(P) = \frac{1}{1-\alpha}\cdot I\,(P\|Q_\alpha)\ ,$$

wobei

$$Q_\alpha = (p_i^\alpha / \sum_{j=0}^{m-1} p_j^\alpha : i \in \{0,...,m-1\})\,.$$

Überdies gilt

$$|H_\alpha(P) - H_1(P)| \le -\sum_{i=0}^{m-1} \left|p_i^{\alpha-1} - 1\right| p_i \log_2(p_i)\,.$$

Wir beschränken uns hier darauf, die erste Aussage nachzuprüfen:

$$\begin{aligned} H_\alpha(P) - H_1(P) &= \frac{1}{1-\alpha}\log_2(\sum_{i=0}^{m-1} p_i^\alpha) + \sum_{i=0}^{m-1} p_i \log_2(p_i) \\ &= \frac{1}{1-\alpha}\left(\log_2(\sum_{i=0}^{m-1} p_i^\alpha) + \sum_{i=0}^{m-1} p_i \log_2(\frac{p_i}{p_i^\alpha})\right) \\ &= \frac{1}{1-\alpha}\sum_{i=0}^{m-1} p_i \log_2(\frac{p_i}{p_i^\alpha / \sum_{j=0}^{m-1} p_j^\alpha}) = \frac{1}{1-\alpha} I\,(P\|Q_\alpha)\ . \end{aligned}$$

Proposition 2 (Eigenschaften der Funktion $\alpha \mapsto H_\alpha(P)$): Sei $m \ge 2$, P eine von der Gleichverteilung verschiedene Wahrscheinlichkeitsverteilung mit Träger $\{0,...,m-1\}$. Dann ist die Funktion $\alpha \mapsto H_\alpha(P)$ (i) stetig, (ii) streng monoton fallend und (iii) streng konvex.

Anmerkung 4 (zu (ii)): Es gilt

$$\frac{\partial\, H_\alpha(P)}{\partial\,\alpha} = -\frac{1}{(1-\alpha)^2}\cdot I\,(Q_\alpha\|P)\ ,\quad \alpha \in (0,\infty)\backslash\{1\}$$

und

$$\lim_{\alpha \to 1} \frac{\partial H_\alpha(P)}{\partial \alpha} = -\frac{1}{2} V_P\left(\log_2(P)\right) ,$$

wobei $V_P\left(\log_2(P)\right)$ die Varianzen der Zufallsgrößen $i \mapsto \log_2(p_i)$ bezeichnet.

Proposition 3 (Additivität der Entropie der Ordung α für unabhängige Quellen): Sind $P = (p_0, ..., p_{m-1})$ und $Q = (q_0, ..., q_{n-1})$ zwei Wahrscheinlichkeitsverteilungen und ist

$$P \times Q = (p_i \cdot q_j : (i,j) \in \{0, ..., m-1\} \times \{0, ..., n-1\})$$

deren Produktverteilung, dann gilt für alle $\alpha \in [0, \infty]$

$$H_\alpha\left(P \times Q\right) = H_\alpha\left(P\right) + H_\alpha\left(Q\right) .$$

Beweis: Für $\alpha \in [0, \infty) \backslash \{1\}$ ist dies eine unmittelbare Folge von

$$\sum_{i=0}^{m-1} \sum_{j=0}^{m-1} \left(p_i \cdot q_j\right)^\alpha = \sum_{i=0}^{m-1} p_i^\alpha \cdot \sum_{j=0}^{m-1} q_j^\alpha$$

und der Additivität des Logarithmus. Die Gültigkeit für $\alpha = 1$ und $\alpha = \infty$ ergibt sich daraus durch Grenzübergang für $\alpha \to 1$ bzw. $\alpha \to \infty$. $\square$

2.4.3 Caesar- und Vigenère Code

Julius Caesar wird zugeschrieben, dass er dadurch Nachrichten vor dem Zugriff Unbefugter zu schützten versuchte, indem er jeden Buchstaben der Nachricht durch denjenigen Buchstaben ersetzte, der im Alphabet drei Stellen hinter dem betreffenden Buchstaben steht:

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
D E F G H I J K L M N O P Q R S T U V W X Y Z A B C

So wird zum Beispiel aus dem nachstehenden Text das folgende Kryptogramm:

o m n i a	g a l l i a	e s t	d i v i s a	i n	p a r t e s	t r e s
RPQLD	JDOOLD	HVW	GLYLVD	LQ	SDUWHV	WUHV

Wir verwenden hier, wie im Folgenden, die Kleinbuchstaben für das Alphabet der Nachricht und die Großbuchstaben für das Alphabet des Kryptogramms. Unter einem Caesar-Code versteht man heute die beschriebene Art der Chiffrierung, wobei die (jeweils feste) Anzahl s der Positionen, um die jeder Buchstabe verschoben wird, beliebig sein kann. Diese Anzahl wird als "Schlüssel " des Caesar-Codes bezeichnet.

Wir werden der Bequemlichkeit halber folgende Identifikationen treffen $\{a, ..., z\} \doteq \{0, ..., 25\}$ und $\{A, ..., Z\} \doteq \{0, ..., 25\}$, sodass dem Chiffrieren (hier mit dem Schlüssel $s = 3$):

a b c	...	x y z
DEF	...	ABC

die Verschiebung $k \rightarrow i = (k+s) \mod 26$ entspricht, und dem Dechiffrieren:

ABC	XYZ
x y z	u v w

die Verschiebung $i \rightarrow k = (i + 26 - s) \mod 26$.

Seien

$P = (p_i,\ i \in \{0, ..., 25\})$

die "Wahrscheinlichkeitsverteilung" der Buchstaben des Alphabets der jeweiligen Sprache, in der die Nachricht abgefasst ist, und

$Q = (q_i,\ i \in \{0, ..., 25\})$

die Verteilung der relativen Häufigkeiten der Buchstaben des Kryptogramms.

Dann ist es unsere Absicht, den Schlüssel s^* zu finden, mit dem das Kryptogramm abgefasst wurde und Hilfe dessen wir das Kryptogramm auch dechiffrieren können. Zu diesem Zweck gehen wir von der Verteilung P zur Verteilung

$$P_{-s} = (p_{i-s} := p_{(i+26-s) \bmod 26},\ i \in \{0, ..., 25\})$$

mit dem Parameter $s \in \{0, ..., 25\}$ über. Sei weiters

$$x_1, ..., x_n \in \{0, ..., 25\} \doteq \{A, ..., Z\}$$

die Buchstabenfolge des Kryptogramms. Wir betrachten die n-te Wurzel der "Wahrscheinlichkeit" für das Auftreten der beobachteten Folge bei Verwendung des möglichen Schlüssels s. Das ist das geometrische Mittel der Wahrscheinlichkeiten $p_{x_1}, ..., p_{x_n}$, nämlich

$$\begin{aligned}
\left(\prod_{j=1}^{n} p_{x_j - s}\right)^{\frac{1}{n}} &= \left(\prod_{j=1}^{n}\prod_{i=0}^{25} p_{i-s}^{1_{\{i\}}(x_j)}\right)^{\frac{1}{n}} \\
&= \left(\prod_{i=0}^{25} p_{i-s}^{\sum_{j=1}^{n} 1_{\{i\}}(x_j)}\right)^{\frac{1}{n}} \\
&= \prod_{i=0}^{25} p_{i-s}^{\frac{1}{n}\sum_{d=1}^{n} 1_{\{i\}}(x_j)} \\
&= \prod_{i=0}^{25} p_{i-s}^{q_i}
\end{aligned}$$

Es ist nun naheliegend, zu vermuten, dass wir den Schlüssel dadurch finden können, daß wir diese "mittlere Wahrscheinlichkeit" durch geeignete Wahl des Parameters s maximal machen. Es solches Verfahren nennt man "Maximum-Likelihood-Verfahren".

Gleichbedeutend dazu ist es, den Logarithmus dieser mittleren Wahrscheinlichkeit zu maximieren. Nun ist aber

$$\begin{aligned}
\ln\left(\prod_{i=0}^{25} p_{i-s}^{q_i}\right) &= \sum_{i=o}^{25} q_i \ln p_{i-s} \\
&= \sum_{i=0}^{25} q_i \ln \frac{p_{i-s}}{q_i} \cdot q_i \\
&= -\sum_{i=0}^{25} q_i \ln \frac{q_i}{p_{i-s}} - \sum_{i=o}^{25} q_i \ln \frac{1}{q_i} \\
&= -I(Q \parallel P_{-s}) - H(Q)
\end{aligned}$$

Somit läuft hier das Maximum-Likelihood-Prinzip darauf hinaus, die I-Divergenz

$$I(Q \parallel P_{-s})$$

der Wahrscheinlichkeitsverteilungen Q und P_{-s} durch geeignete Wahl des Parameters s zu minimieren.

Anmerkung 5: Sei $Q_s = (q_{(i+s) \bmod 26},\ i \in \{0, ..., 25\})\,,\ \ s \in \{0, ..., 25\}$. Dann ist

$$I(Q \parallel P_{-s}) = I(Q_s \parallel P)\,.$$

Bei unserer Anwendung des Maximum-Likelihood-Prinzips muss man also Q so lange verschieben, bis Q_s und P (hinsichtlich ihrer I-Divergenz) möglichst gut übereinstimmen. Die Hoffnung, durch dieses Verfahren den richtigen Schlüssel s^* zu finden, ist freilich nur dann gerechtfertigt, wenn die Länge des Kryptogramms (der Stichprobenumfang) hinreichend groß ist. Aussagen in diesem Zusammenhang macht *Shannon* in seiner bahnbrechenden Arbeit [6].

Hätte sich Caesar entschlossen, nur jene Buchstaben in seiner Manier zu ersetzen, welche an einer ungeraden Position des Klartexts stehen, also den 1., 3., 5., ... Buchstaben und die Buchstaben an den geraden Positionen unverändert zu lassen, so wäre aus dem obigen Klartext folgendes Kryptogramm entstanden:

o m n i a	g a l l i a	e s t	d i v i s a	i n	p a r t e s	t r e s
RMQID	GDLOID	EVT	GIYIVA	LN	SAUTHS	WRHS

Diese Chiffre hat die Eigenschaft, dass jeder Buchstabe des Kryptogramms von zwei verschiedene Buchstaben des Klartexts herrühren kann. So kann beispielsweise der Buchstabe "E" des Kryptogamms von einem "b" oder einem "e" herrühren, je nachdem, ob der Buchstabe an einer ungeraden oder einer geraden Position des Klartexts gestanden hat. Daher kann erwartet werden, dass die relative Häufigkeit eines "E" im Kryptogramm das arithmetische Mittel

$$h_E = \frac{h_b + h_e}{2}$$

der relativen Häufigkeiten eines "b" und eines "e" der Sprache ist, in welcher der Klartext abgefasst wurde. Damit werden die Häufigkeiten der einzelnen Buchstaben der verwendeten Sprache "eingeebnet". Für die deutsche Sprache ergibt sich

$$h_E = \frac{h_b + h_e}{2} = \frac{0.017 + 0.173}{2} = 0.095\,.$$

Da ein Schlüssel nun zudem nicht mehr bloß eine Zahl $s \in \{0, ..., 25\}$ ist, sondern ein Paar $(s_1, s_2) \in \{0, ..., 25\}^2$ - wobei s_1 und s_2 die Anzahl der Positionen angibt, um die jeder Buchstabe an einer ungeraden bzw. geraden Positionen verschoben wird - wird eine Häufigkeitsanlalyse zur Auffindung des Schlüssels beträchtlich erschwert. Unsere fiktive Modifikation des Caesar-Codes ist ein sehr einfaches Beispiel eines *Vigenère-Codes* [11]. Bei einem solchen ist der Schlüssel ein "Schlüsselwort", d.i. ein Tupel $(s_1, ..., s_l) \in \{0, ..., 25\}^l$, wobei nicht nur die einzelnen Komponenten des l-Tupels, sondern auch die Codewortlänge l unbekannt ist. Dies führte dazu, dass sich der *Vigenère-Code* mehr als zweihundert Jahre jeglichem Angriff erfolgreich widersetzte und daher als der unbrechbare Code schlechthin galt (siehe [22], Kapitel 2: Le Chiffre indéchiffrable).

Das Obige motiviert, die folgende Größe zu studieren, welche den "Ausgleich" zwischen einer Wahrscheinlichkeitsverteilung P_0 und deren Shift P_c misst.

Definition 3: Seien $c \in \{0, ..., m-1\}$ und X und Y unabhängige Zufallsvariable mit Werten aus $\{0, ..., m-1\}$. X habe die Verteilung $P_0 = (p_0, ..., p_{m-1})$ und Y die Verteilung $P_c = (p_{i+c} = p_{(i+c) \bmod m}\,,\ i \in \{0, ..., m-1\})$. Dann ist die Wahrscheinlichkeit des Ereignisses $\{X = Y\}$, dass die beiden Zufallsvariablen übereinstimmen,

$$\kappa_c(P_0) = P(Y = X) = P_0 \cdot P_c = \sum_{i=0}^{m-1} p_i\, p_{i+c}\,.$$

Wir nennen $\kappa_c(P_0)$ den *κ_c-Wert der Wahrscheinlichkeitsverteilung* P_0.

Anmerkung 6: Es gilt $\kappa_0(P) = \kappa(P)$.

Für den Wertebereich des κ_c-Werts gilt

Proposition 4: Es gilt

$$0 \le \min(p_i : i \in \{0, ..., m-1\}) \le \sum_{i=0}^{m-1} p_i\, p_{i+c} \le \sum_{i=0}^{m-1} p_i^2\,.$$

Dabei gilt in der zweiten Ungleichung Gleichheit genau dann, wenn $p_{i+c} = p_i\,,\ i \in \{0, ..., m-1\}$ ist.

Beweis: Die untere Schranke ist offenkundig. Nun zur oberen: Anwendung der Cauchy-Schwarz'schen Ungleichung auf den Ausdruck $\sum_{i=0}^{m-1}(p_i - \frac{1}{m})(p_{i+c} -$

[11] *Blaise de Vigenère* $(1523 - 1596)$, u.a. französischer Botschafter in Rom

$\frac{1}{m}$) ergibt

$$\begin{aligned}\left|\sum_{i=0}^{m-1} p_i\, p_{i+c} - \frac{1}{m}\right| &= \left|\sum_{i=0}^{m-1}(p_i - \frac{1}{m})(p_{i+c} - \frac{1}{m})\right| \\ &\leq \sqrt{\sum_{i=0}^{m-1}(p_i - \frac{1}{m})^2 \times \sum_{i=0}^{m-1}(p_{i+c} - \frac{1}{m})^2} = \sum_{i=0}^{m-1}(p_i - \frac{1}{m})^2 = \sum_{i=0}^{m-1} p_i^2 - \frac{1}{m}\,,\end{aligned}$$

oder, anders ausgedrückt,

$$\frac{1}{m} - \sum_{i=0}^{m-1}\left(p_i - \frac{1}{m}\right)^2 \leq \sum_{i=0}^{m-1} p_i\, p_{i+c} \leq \sum_{i=0}^{m-1} p_i^2 = \frac{1}{m} + \sum_{i=0}^{m-1}\left(p_i - \frac{1}{m}\right)^2 .$$

Dabei gilt Gleichheit genau dann, wenn

$$p_{i+c} - \frac{1}{m} = a \cdot (p_i - \frac{1}{m}) + b\,, \quad i \in \{0, ..., m-1\}$$

mit $a, b \in \mathbb{R}$ ist. Aufgrund von $\sum_{i=0}^{m-1}(p_{i+c} - \frac{1}{m}) = \sum_{i=0}^{m-1}(p_i - \frac{1}{m}) = 0$ ergeben sich $b = 0$, $a \in \{-1, +1\}$. Gleichheit wird in der zweiten Ungleichung für $a = +1$ und in der ersten für $a = -1$ angenommen. □

Anmerkung 7: (a) Die untere Schanke in der obigen Ungleichungskette ist selbstverständlich nur dann nicht trivial, wenn gilt

$$\sum_{i=0}^{m-1}\left(p_i - \frac{1}{m}\right)^2 < \frac{1}{m}\,.$$

(b) Proposition 4 zeigt, dass der Effekt der polyalphabetischen Substitution, zu der auch der Vigenère Code zählt, darin besteht, dass die Häufigkeitsverteilung "ausgeglichen" wird, denn - wie wir bereits wissen - werden je zwei oder mehr Buchstaben des Klartexts auf ein- und denselben Buchstaben des Kryptogramms abgebildet.

In Proposition 1 wurde eine Aussage über den Wertebereich von $\kappa(P_0^{(m)})$ getroffen. Um eine Vermutung über den Wertebereich der Größe $\kappa_1(P_0^{(m)})$ zu erhalten, ist es zweckmäßig, geeignete Spezialfälle zu betrachten:

- Ist $P_0^{(m)}$ eine Punktverteilung, so ist $P_0^{(m)} \cdot P_1^{(m)} = 0$,
- ist $P_0^{(m)} = \left(\frac{1}{m}, ..., \frac{1}{m}\right)$ die Gleichverteilung, so auch $P_1^{(m)} = \left(\frac{1}{m}, ..., \frac{1}{m}\right)$ und $\kappa_1(P_0^{(m)}) = \kappa(P_0^{(m)}) = \frac{1}{m}$,
- ist $P_0^{(m)} = \left(\frac{1}{2}, \frac{1}{2}, 0, ..., 0\right)$, so ist $P_1^{(m)} = \left(0, \frac{1}{2}, \frac{1}{2}, 0, ..., 0\right)$ und es gilt

$$\kappa_1(P_0^{(m)}) = P_0^{(m)} \cdot P_1^{(m)} = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}\,.$$

Seien nun $m \geq 2$ und $\mathcal{V}_m$ die Menge aller Wahrscheinlichkeitsverteilungen auf $\{0, ..., m-1\}$ und bezeichnen

$$\kappa_1(m) = \min(\kappa_1(P^{(m)}) : P^{(m)} \in \mathcal{V}_m)$$

und

$$\bar{\kappa}_1(m) = \max(\kappa_1(P^{(m)}) : P^{(m)} \in \mathcal{V}_m)\,.$$

Aufgrund der obigen Überlegungen gelten $\kappa_1(m) = 0$ und $\bar{\kappa}_1(m) \geq \max(\frac{1}{m}, \frac{1}{4})$. In der Tat gilt folgende

Proposition 5: Sei $m \in \mathbb{N}\backslash\{1\}$. Dann ist die untere Schranke

$$\kappa_1(m) = 0\,,$$

wobei $\kappa_1(m) = 0$ genau dann gilt, wenn $P_0^{(m)} \perp P_1^{(m)}$, d.h. wenn die Träger von $P_0^{(m)}$ und $P_1^{(m)}$ disjunkt sind. Die obere Schranke ist

$$\bar{\kappa}_1(m) = \max(\frac{1}{m}, \frac{1}{4})\,,$$

wobei Folgendes zutrifft:

$\circ$ Für $m \in \{2, 3\}$ gilt $\bar{\kappa}_1(m) = \frac{1}{m}$ genau dann, wenn $P^{(m)}$ die Gleichverteilung ist.

$\circ$ Für $m = 4$ gilt $\bar{\kappa}_1(4) = \frac{1}{4}$ genau dann, wenn gilt

$$P^{(4)} = (p_0, p_1, \frac{1}{2} - p_0, \frac{1}{2} - p_1),\ p_0, p_1 \in [0, \frac{1}{2}]\,.$$

$\circ$ Für $m \geq 5$ gilt $\bar{\kappa}_0(m) = \frac{1}{4}$ genau dann, wenn $P^{(m)} = (p_0, p_1, \frac{1}{2} - p_0, 0, ..., 0),\ p_0 \in [0, \frac{1}{2}]$ bzw. ein Shift davon modulo m ist.

Beweis: Die Aussage hinsichtlich der unteren Schranke $\kappa_1(m)$ ist offenkundig. Nun zum Nachweis der oberen Schranke:
$m = 2$: Wegen $P^{(2)} = (p_0, p_1)$ ist

$$\kappa_1(P^{(2)}) = p_0p_1 + p_1p_0 = 2p_0p_1\,.$$

Anwendung der geometrisch-arithmetischen Ungleichung und Berücksichtigung von $p_0 + p_1 = 1$ ergibt

$$p_0p_1 \leq \left(\frac{p_0 + p_1}{2}\right)^2 = \frac{1}{4}$$

und somit $2p_0p_1 \leq \frac{1}{2}$, wobei Gleichheit genau dann gilt, wenn $P^{(2)} = (\frac{1}{2}, \frac{1}{2})$ ist.
$m = 3$: Für $P^{(3)} = (p_0, p_1, p_2)$ ist wegen

$$(a + b + c)^2 = a^2 + b^2 + c^2 + 2\,(ab + ac + bc)$$

und $p_0 + p_1 + p_2 = 1$ ist

$$\kappa_1(P^{(3)}) = p_0p_1 + p_1p_2 + p_2p_0 = \frac{1 - \left(p_0^2 + p_1^2 + p_2^2\right)}{2}$$

und mithin wegen $p_0^2 + p_1^2 + p_2^2 \geq \frac{1}{3}$

$$p_0p_1 + p_1p_2 + p_2p_0 = \frac{1 - \left(p_0^2 + p_1^2 + p_2^2\right)}{2} \leq \frac{1 - \frac{1}{3}}{2} = \frac{1}{3}\,.$$

Dabei gilt Gleichheit genau dann, wenn $P^{(3)} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ ist.
$m = 4$: Wegen $P^{(4)} = (p_0, p_1, p_2, p_3)$ ist

$$\begin{aligned} \kappa_1(P^{(4)}) &= p_0p_1 + p_1p_2 + p_2p_3 + p_3p_0 = p_0\,(p_1 + p_3) + p_2\,(p_1 + p_3) \\ &= (p_0 + p_2)\,(p_1 + p_3)\ . \end{aligned}$$

Anwendung der geometrisch-arithmetischen Ungleichung und Berücksichtigung von $p_0 + p_1 + p_2 + p_3 = 1$ ergibt in diesem Fall

$$(p_0 + p_2)\,(p_1 + p_3) \leq \left(\frac{p_0 + p_1 + p_2 + p_3}{2}\right)^2 = \frac{1}{4}\,.$$

Dabei gilt Gleichheit genau dann, wenn $p_0 + p_2 = p_1 + p_3 = \frac{1}{2}$, also wenn

$$P^{(4)} = (p_0, p_1, \frac{1}{2} - p_0, \frac{1}{2} - p_1)\,,\ \ p_0, p_1 \in [0, \frac{1}{2}] \text{ ist.}$$

Sei nun $m \geq 4$ und $P^{(m)} = (p_0, p_1, p_2, ..., p_{m-2}, p_{m-1})$. Dann ist

$$\kappa_1(P^{(m)}) = (p_0 + p_2)p_1 + \sum_{i=2}^{m-3} p_ip_{i+1} + (p_{m-2} + p_0)\,p_{m-1}$$

Schritt 1: Ist $\tilde{P}^{(m)} = (p_0, p_1 + p_{m-1}, p_2, ..., p_{m-2}, 0)$. Dann ist

$$\begin{aligned} \kappa_1(\tilde{P}^{(m)}) &= (\tilde{p}_0 + \tilde{p}_2)\tilde{p}_1 + \sum_{i=2}^{m-3} \tilde{p}_i\tilde{p}_{i+1} + (\tilde{p}_{m-2} + \tilde{p}_0)\,\tilde{p}_{m-1} \\ &= (p_0 + p_2)\,(p_1 + p_{m-1}) + \sum_{i=2}^{m-3} p_ip_{i+1} \end{aligned}$$

und somit

$$\begin{aligned} \tilde{\Delta}_m &= \kappa_1(\tilde{P}^{(m)}) - \kappa_1(P^{(m)}) \\ &= (p_0 + p_2)\,(p_1 + p_{m-1}) - ((p_0 + p_2)p_1 + (p_{m-2} + p_0)\,p_{m-1}) \\ &= p_{m-1}(p_2 - p_{m-2})\,. \end{aligned}$$

Schritt 2: Ist andererseits $\hat{P}^{(m)} = (p_0, 0, p_2, ..., p_{m-2}, p_1 + p_{m-1})$. Dann ist

$$\kappa_1(\hat{P}^{(m)}) = \sum_{i=2}^{m-3} p_i p_{i+1} + (p_{m-2} + p_0)(p_1 + p_{m-1})$$

und somit

$$\begin{aligned} \hat{\Delta}_m &= \kappa_1(\hat{P}^{(m)}) - \kappa_1(P^{(m)}) \\ &= (p_{m-2} + p_0)(p_1 + p_{m-1}) - ((p_0 + p_2)p_1 + (p_{m-2} + p_0)p_{m-1}) \\ &= p_1(p_{m-2} - p_2). \end{aligned}$$

Somit lässt sich durch "Reduzieren einer Wahrscheinlichkeit auf 0 (und gleichzeitiges Zusammenlegen zweier Wahrscheinlichkeiten)" der κ_1-Wert vergrößern (genauer: nicht verkleinern).

Für $m = 4$ ist $p_2 = p_{m-2}$ und daher $\tilde{\Delta}_m = \hat{\Delta}_m = 0$. Also bleibt in diesem Fall der κ_1-Wert unverändert, wenn man eine Wahrscheinlichkeit auf 0 reduziert.

Sei nun $m \geq 5$ und sei bereits eine Wahrscheinlichkeit (etwa durch Anwendung von Schritt 1 oder 2) gleich 0. Dann können wir durch zyklisches Verschieben der Wahrscheinlichkeiten erreichen, dass $p_2 = 0$ ist. Indem wir nun Schritt 2 anwenden, können wir auch $\hat{p}_1 = 0$ setzen und zudem $\kappa_1(\hat{P}^{(m)}) \geq \kappa_1(P^{(m)})$ erreichen. Nun gilt $\hat{p}_1 = \hat{p}_2 = 0$. Daher können wir $\hat{p}_1 = 0$ streichen und somit $\hat{P}^{(m)} = (\hat{p}_0, \hat{p}_1, \hat{p}_2, ..., \hat{p}_{m-1})$ durch $\hat{P}^{(m-1)} = (\hat{p}_0, \hat{p}_2, ..., \hat{p}_{m-1})$ ersetzen, ohne dass der κ_1-Wert geändert würde. Auf die hier beschriebene Art kann erreicht werden, dass m durch $m-1$ ersetzt und zugleich der κ_1-Wert erhöht (nicht verringert) wird. So kommen wir ausgehend von jedem $m \geq 5$ schrittweise zu $m = 4$ und $P^{(4)} = (p_0, \frac{1}{2}, \frac{1}{2} - p_0, 0)$ mit $p_0 \in [0, \frac{1}{2}]$, wobei bei jedem Schritt der κ_1-Wert nicht verringert wird. $\square$

Literatur

Originalartikel

References

[1] *Boltzmann, L.* (1877): Über die Beziehung zwischen dem zweiten Hauptsatz der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung respektive den Sätzen über das Wärmegleichgewicht. Sitzungsberichte der mathematisch-naturwissenschaftlichen Classe der Kaiserlichen Akademie der Wissenschaften Wien, II. Abtg., 373-435.

[2] *Planck, M.* (1901): Ueber das Gesetz der Energieverteilung im Normalspectrum. Annalen der Physik **4**[4], 553-563.

[3] *Friedman, W.F.* (1920): The Index of Coincidence and its Applications in Cryptography. Riverbank Publication No. 22

[4] *Hartley, R.V.L.* (1928): Transmission of information. Bell. Syst. Tech. J. **7**, 537-563.

[5] *Shannon, C.E.* (1948)*:* A mathematical theory of communication. Bell. Syst. Tech. J. **27**, 379-423 and 623-656.

[6] *Shannon, C.E.* (1949)*:* Communication theory of secrecy systems. Bell. Syst. Tech. J. **28**, 656-715.

[7] *Kraft, R.G.* (1949): A device for quantizing, grouping and coding amplitude modulated pulses. Master's Thesis. Department of Electrical Engineering, MIT, Cambridge, MA

[8] *Hamming, R.V.* (1950): Error detecting and error correcting codes. Bell Sys. Tech. Journal **29**, 147-160.

[9] *Shannon, C.E.* (1951): Prediction and Entropy of printed English. Bell Sys. Tech. Journal **30**, 50-64.

[10] *Kullback, S.* and *R.A. Leibler* (1951): On information and sufficiency. Ann. Math. Stat. **22**, 79-86

[11] *Huffman, D.* (1952): A method for the construction of minimum redundancy codes. Proc. IRE **40**, 1098-1101.

[12] *Fano, R.M.* (1952): Class notes for transmission of information. Course 6.574, MIT; Cambridge, MA.

[13] *McMillan, B.* (1953): The basic theorems of information theory. Ann. Math. Stat. **24**, 196-219.

[14] *McMillan, B.* (1956): Two inequalities implied by unique decipherability. IEEE Trans. Inform. Theory **IT-2**, 115-116.

[15] *Breiman, L.* (1957): The individual ergodic theorems of information theory. Ann. Math. Stat. **28**, 809-811.

[16] *Karush, J.* (1961): A simple proof of an inequality of McMillan. IRE Trans. Inform. Theory **IT-7**, 118.

Populärwissenschaftliche Literatur

[17] *Rényi, A.* (1982): Tagebuch über die Informationstheorie. Birkhäuser Verlag, Basel

[18] *Tarassow, L.V.* (1984): Wie der Zufall will ? Vom Wesen der Wahrscheinlichkeit. Spektrum Akademischer Verlag, Heidelberg-Berlin-Oxford

[19] *Kreuzer, F.* (1985): Neue Welt aus Null und Eins: Hirn und Computer - natürliche und künstliche Intelligenz. ORF und Franz Deuticke Verlag, Wien

[20] *Lucky, R.W.* (1989): Silicon dreams: information, man and machine. St. Martin's Press, New York

[21] *Ruelle, D.* (1994): Zufall und Chaos. Springer Verlag, Berlin-Heidelberg-New York

[22] *Singh, S.* (1999): Geheime Botschaften - Die Kunst der Verschlüsselung von der Antike bis in die Zeit des Internet. Carl Hanser Verlag, München - Wien

Lehrbücher, Skripten und Vortragsunterlagen

[23] *Zemanek, H.* (1959): Elementare Informationstheorie. Verlag R. Oldenburg, Wien-München

[24] *Kullback, S.* (1967): Information Theory and Statistics. Dover Publications, New York

[25] *Rényi, A.* (1977): Wahrscheinlichkeitsrechnung - Mit einem Anhang über Informationstheorie. (5. Auflage), VEB Deutscher Verlag der Wissenschaften, Berlin

[26] *Topsøe, F.*: (1974): Informationstheorie. Teuber Studienbücher, Stuttgart

[27] *Nemetz, T.* (1980): Informationstheorie. Skriptum nach der gleichnamigen Vorlesung, Frankfurt am Main

[28] *Welch, D.* (1988): Codes and cryptography. Clarendon Press, Oxford

[29] *Cover, Th.M.* and *J.A. Thomas* (1991): Elements of information theory. Wiley & Sons, New York

[30] *Österreicher, F.* (1991): Informationstheorie. Skriptum, Salzburg

[31] *Österreicher, F.* (1992): Über eine Klasse von Informationsmaßen und deren Anwendungen, Vortragsunterlagen, Marburg/Lahn

Einschlägige Diplomarbeiten aus Salzburg

[32] *Ch. Haas*: Anwendungsaspekte der Information und des Informationsgewinns α-ter Ordnung (März 1992)

[33] *G. Fischer*: Caesar Code, Vigenère Code und Shannons Zugang zur Kryptographie (Dezember 1996)

[34] *Ch. Radauer*: Läufe in binären Zufallsfolgen - Beurteilung und Kompression von Zufallsdaten (Februar 1998)

[35] *D. Steinkogler*: Information und natürliche Sprache - Schätzung der Entropie einer geschriebenen Sprache (Mai 1998)

[36] *K. Brandstätter*: Ein Satz von Chernoff und die Normalapproximation der Binomialverteilung (Informationstheoretische Aspekte der Normalapproximation der Binomialverteilung) (August 2002)

[37] *A. Meschtscherjakov*: Verlustfreie Kompression - Lempel-Ziv-Algorithmen (August 2003)

[38] *M. Hemetsberger:* Nutzung der Kryptologie für den Unterricht - Cäsar- und Vigenère-Code (Juni 2008)

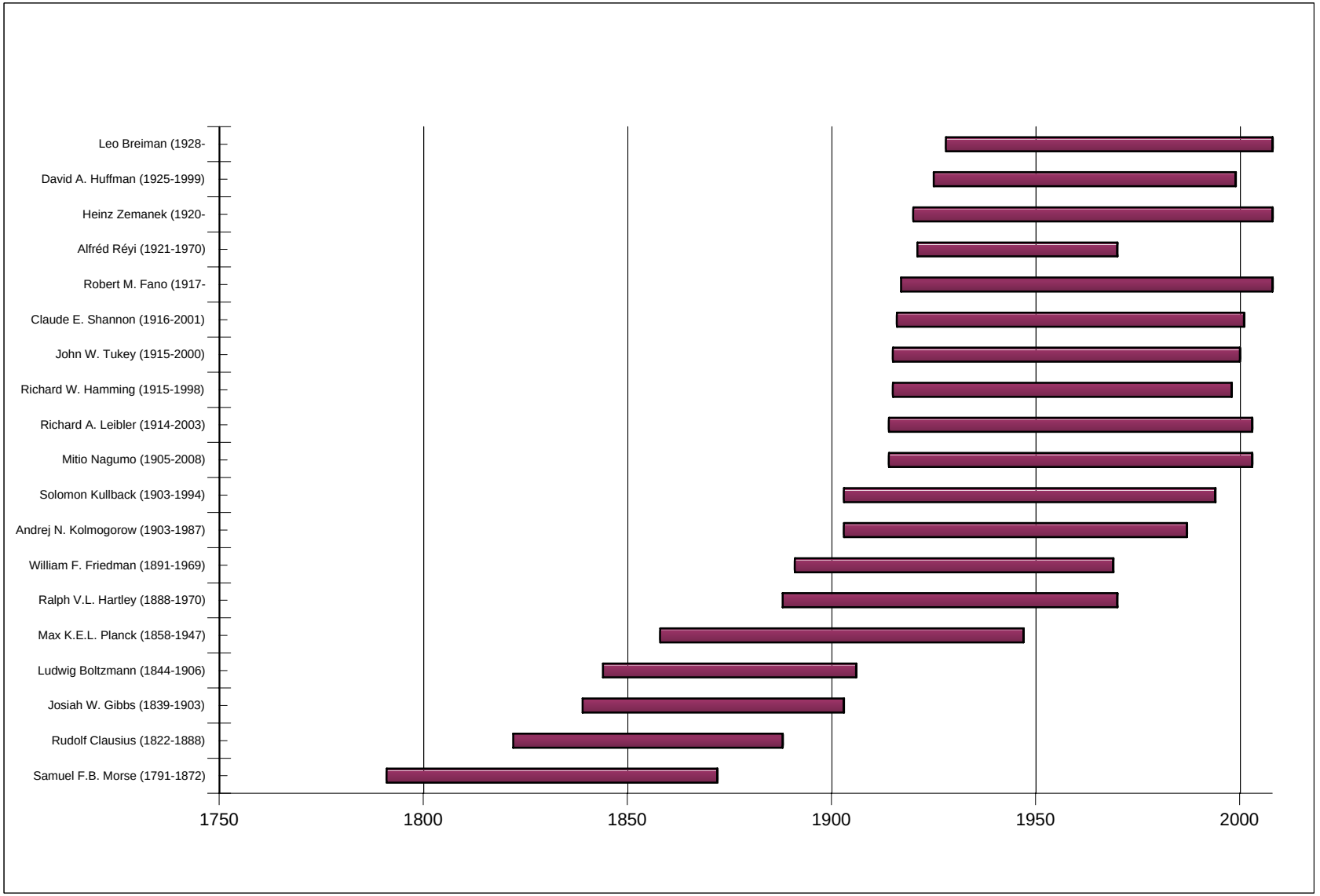
Literatur aus dem Umfeld der Informationstheorie

[39] *Miller, G.A.* (1956): The magical number seven, plus or minus two: some limits on our capacity for processing information. The Psychological Review **63**, 81-97.

[40] *Oberliesen, R.* (1982): Information, Daten und Signale - Geschichte technischer Informationsverarbeitung. Deutsches Museum und Rowohlt Taschenbuch Verlag, Reinbek bei Hamburg

[41] *Pierce, J.R.* (1989): Klang - Mit den Ohren der Physik. Spektrum Akademischer Verlag, Heidelberg-Berlin-Oxford

[42] *Pierce, J.R.* and *A.M. Noll* (1992): Signale - Die Geheimnisse der Telekommunikation. Spektrum Akademischer Verlag, Heidelberg-Berlin-Oxford


Leo Breiman (1928-
David A. Huffman (1925-1999)
Heinz Zemanek (1920-
Alfréd Réyi (1921-1970)
Robert M. Fano (1917-
Claude E. Shannon (1916-2001)
John W. Tukey (1915-2000)
Richard W. Hamming (1915-1998)
Richard A. Leibler (1914-2003)
Mitio Nagumo (1905-2008)
Solomon Kullback (1903-1994)
Andrej N. Kolmogorow (1903-1987)
William F. Friedman (1891-1969)
Ralph V.L. Hartley (1888-1970)
Max K.E.L. Planck (1858-1947)
Ludwig Boltzmann (1844-1906)
Josiah W. Gibbs (1839-1903)
Rudolf Clausius (1822-1888)
Samuel F.B. Morse (1791-1872)
1750
1800
1850
1900
1950
2000