15. Entropie und Information

Das Verständnis der Entropie als Maß für den mittleren Informationsgehalt geht auf Claude E. Shannon (1916-2001) zurück. Mit seiner einflussreichen Arbeit *A Mathematical Theory of Communication* (1948) prägte er die moderne Informationstheorie.

Für die statistische Deutung der Entropie betrachteten wir Z verschiedene Zustände, über die insgesamt N Teilchen verteilt sind. Der Makrozustand ist dann durch die Z Besetzungszahlen n_i vollständig definiert (Vgl. Kapitel Boltzmann-Verteilung). Es gilt

$$\sum_{i=1}^{Z} n_i = N \qquad \text{Anzahl der Mikrozustände: } W = \frac{N!}{n_1!n_2!...n_Z!}$$

Wir wollen nun erraten, in welchem der W möglichen Mikrozustände sich das System (d.h. in welchem Zustand jedes einzelne Teilchen) befindet bei gegebenem Makrozustand.

Frage: Wie oft müssen wir bei W *gleich wahrscheinlichen* Antworten mindestens raten, bis wir die richtige getroffen haben?

Antwort: Wenn nur JA oder NEIN als Antwort erlaubt sind, mindestens $\log_2 W$ -mal!

Begründung: Die Anzahl der erforderlichen Fragen hängt davon ab, wie geschickt die Fragen gestellt werden. Es stellt sich heraus, dass es optimal ist, wenn mit jeder Frage genau die Hälfte der verbleibenden Möglichkeiten ausgeschlossen wird.

Vergleiche: Erraten eines der Felder eines Schachbrettes, indem man a) jedes einzelne Feld abfragt (im Schnitt 64/2 = 32 Fragen) b) abfragt, ob das Feld in einer der Hälften des verbliebenen Teils enthalten ist (immer $\log_2 64 = 6$ Fragen)

Allgemeinere Deutung dieses Formalismus:

N-malige Wiederholung eines Experiments, das Z mögliche Ereignisse E_i als Ausgang hat.

Dabei sollen n_i Experimente den Ausgang E_i haben (Makrozustand).

Beispiele für Experimente:

- Messen des Zustandes eines Teilchens (Z mögliche zustände, s. oben)
- Würfeln mit einem Z-seitigem Würfel (bei symmetrischem Würfel sind alle p_i gleich)
- Lesen des nächsten Buchstaben eines gedruckten Textes mit einem Alphabet von Z Buchstaben.

Wir wollen nun wieder den Mikrozustand erraten – bei dieser Deutung entspricht das dem Erraten des Ausgangs für jedes der N Experimente. Gemäß obiger Überlegung brauchen wir mindestens $\log_2 W$ ja-nein-Fragen.

In diesem Kontext ist es natürlicher den einzelnen Ausgängen Wahrscheinlichkeiten statt Häufigkeiten n_i zuzuordnen:

$$p_i \equiv \frac{n_i}{N} \qquad \text{mit} \qquad \sum_{i=1}^{Z} p_i = 1 \qquad \text{und damit } n_i = N p_i$$

Den „Makrozustand" können wir dann durch die zahlen $\{p_1, p_2, \ldots p_Z, N\}$ definieren.

Wir drücken nun $\log_2 W$ in Termen der Wahrscheinlichkeiten aus (Wechseln der Basis des Logarithmus kommt der Multiplikation mit einer Konstanten gleich):

$$\log_2 W = \frac{1}{\ln 2} \cdot \ln W = \frac{1}{\ln 2}\left(\ln \frac{N!}{n_1! n_2! \ldots n_Z!}\right) = \frac{1}{\ln 2}\left(\ln N! - \sum_{i=1}^{Z} \ln n_i!\right) \quad \text{(Stirling-Formel)}$$

$$\approx \frac{1}{\ln 2}\left(N \ln N - N - \sum_{i=1}^{Z} n_i \ln n_i + \overbrace{\sum_{i=1}^{Z} n_i}^{=N}\right) = \frac{1}{\ln 2}\left(N \ln N - \sum_{i=1}^{Z} n_i \ln n_i\right) \quad \text{(einsetzen der } p_i \text{)}$$

$$= \frac{1}{\ln 2}\left(N \ln N - \sum_{i=1}^{Z} N p_i \ln N p_i\right) = \frac{1}{\ln 2}\left(N \ln N - N \ln N \underbrace{\sum_{i=1}^{Z} p_i}_{=1} - N \sum_{i=1}^{Z} p_i \ln p_i\right)$$

$$= -\frac{N}{\ln 2} \sum_{i=1}^{Z} p_i \ln p_i$$

Die Anzahl der mindestens nötigen ja-nein-Fragen *im Schnitt pro Experiment* (bzw. pro Teilchen) ist also gegeben durch: $\frac{S}{N} = -\lambda \sum_i p_i \ln p_i$ mit $\lambda = \frac{1}{\ln 2} \approx 1{,}44$ anstatt k_B

Wir führen nun die **Shannon-Entropie** als Informationsmaß für eine Wahrscheinlichkeitsverteilung p_i ein:

$$\boxed{I = -\lambda \sum_i p_i \ln p_i}$$

Beispiel: Wir werfen eine Münze: Bevor wir hinschauen haben wir einen Mangel an Information, danach einen Informationsgewinn („Maß der beseitigten Unsicherheit“).

$$p_{\mathrm{Zahl}} = \frac{1}{2}\ ,\quad p_{\mathrm{Kopf}} = \frac{1}{2} \qquad \Rightarrow \qquad I = -\lambda\left(\frac{1}{2}ln\frac{1}{2} + \frac{1}{2}ln\frac{1}{2}\right) = -\lambda \ln\frac{1}{2} = \lambda \cdot \ln 2$$

Wählt man $\lambda = 1/\ln 2$ ergibt sich $I = 1$ „Informationsgewinn von $1\,\mathrm{bit}$“.
Diese Wahl bedeutet, I ist die minimale durchschnittliche Anzahl der nur mit Ja oder Nein zu beantwortenden Fragen, die man pro Experiment stellen muß um die gewünschte Information zu erhalten.

$I = 1\,\mathrm{bit}$ heißt, nur eine Frage erforderlich: F: Kopf? A: Ja $\Rightarrow$ Kopf , A: Nein $\Rightarrow$ Zahl

Beispiel: Würfeln mit 8-seitigem symmetrischen Würfel: $I = -\dfrac{1}{ln 2} 8 \cdot \dfrac{1}{8} ln\dfrac{1}{8} = -\dfrac{1}{ln 2} ln\dfrac{1}{2^3} = 3$

Es sind 3 Fragen erforderlich:

F1: X > 4 ? A1: Nein
F2: X > 2 ? A2: Ja
F3: X = 4 ? A3: Nein $\Rightarrow$ X=3

Beispiel: Welche Augenzahl man mit zwei Würfeln gleichzeitig würfelt:

Summe =2 , 1 Möglichkeit: 1.Würfel =1, 2.Würfel =1 $\Rightarrow$ $n_2 = 1$

Summe =3 , 2 Möglichkeiten: 1.Würfel =1, 2.Würfel =2 oder
1.Würfel=2, 2.Würfel=1 $\Rightarrow$ $n_3 = 2$

Summe = 4: 3 Möglichkeiten $\Rightarrow$ $n_4 = 3$

Summe = 5: 4 Möglichkeiten .

Summe = 6: 5 Möglichkeiten .

Summe = 7: 6 Möglichkeiten .

Summe = 8: 5 Möglichkeiten

Summe = 9: 4 Möglichkeiten

Summe = 10: 3 Möglichkeiten

Summe = 11: 2 Möglichkeiten

Summe = 12: 1 Möglichkeit

insgesamt 36 Möglichkeiten für "Mikrozustände", aber nur 12 "Makrozustände"

N=36 $\Rightarrow$ $p_{\mathrm{i}} = n_{\mathrm{i}}/36$

$$I=-\frac{1}{\ln 2}\sum_{i=1}^{12} p_i \ln p_i = -\frac{1}{\ln 2}\left(\frac{1}{36}\ln\frac{1}{36}+\frac{2}{36}\ln\frac{2}{36}+...\right)=3{,}27 \text{ bit}$$

D.h. mindestens 3,27… Fragen im schnitt pro Experiment notwendig.

Besitzt die Shannon-Entropie alle Eigenschaften, die wir von einem Informationsmaß erwarten?

***1. Positivität*:** Bevor ein Ereignis stattfindet, haben wir stets einen Mangel an Information, nach dem Eintreten des Ereignisses einen Gewinn.

$$I \overset{!}{\geq} 0: \quad I=-\frac{1}{\ln 2}\sum_i p_i \ln p_i \geq 0 \ , \text{ wegen } \ p_i \leq 1$$

2. Information hat ein Minimum: $I=0$

wenn $p_m=1$, und alle $p_{i\neq m}=0$, d.h. ein Ereignis E_m tritt mit Sicherheit ein:
für $p_i=0$ ist zwar $ln\, p_i$ nicht definiert, aber $p_i \ln p_i\Big|_{p_i\to 0}\to 0$, daher:

$$\lim_{\substack{p_m \to 1\\ p_{i\neq m}\to 0}} I(p_i)=\lim_{\substack{p_m \to 1\\ p_{i\neq m}\to 0}} -\lambda(1\cdot\ln 1)-\lambda\sum_{i\neq m}\underbrace{p_i \ln p_i}_{\to 0}=0$$

3. I hat ein Maximum

Am wenigsten können wir ein Ereignis voraussagen (der Informationsgewinn ist am größten), wenn alle Ereignisse gleichwahrscheinlich sind. (Analog zum Entropiemaximierung ohne Energieerhaltung im Kapitel Boltzmann-Verteilung)

$$\text{Alle } p_i=\frac{1}{N}: \qquad I=-\frac{1}{ln 2}N\cdot\frac{1}{N}ln\frac{1}{N}=\frac{ln N}{ln 2} \tag{$*$}$$

Mit Lagrange'schen Multiplikator μ : $\qquad I'=-\sum_i p_i \ln p_i+\mu\sum p_i$

$$\frac{\partial I'}{\partial p_j}=-\ln p_j-1+\mu=0 \qquad \Rightarrow \quad p_j=e^{\mu-1}, \qquad \text{d.h. alle } p_j=\frac{1}{N}$$

4. I steigt mit der Zahl der möglichen gleich wahrscheinlichen Ereignisse (folgt aus (∗))

5. Additivität:

Es werden zwei voneinander unabhängige Ereignisse aus zwei Ereignismengen $x_i = \{1, 2\ldots, m\}$ und $y_j = \{1, 2\ldots, n\}$ betrachtet (z. B. gleichzeitiges Werfen von einer Münze und einem Würfel.)

Wir erwarten, daß sich die Informationen addieren (Zahl der Fragen addiert sich).

$p(i, j) = p_i \cdot p_j$: Wahrscheinlichkeit für das gleichzeitige Auftreten für das Ereignis *i* (aus *x*) und Ereignis *j* (aus *y*), wenn *x* und *y* unabhängig (Definition).

$$I = -\frac{1}{\ln 2}\sum_i \sum_j \underbrace{p_i p_j}_{p(i,j)} \ln \underbrace{p_i p_j}_{p(i,j)} = -\frac{1}{\ln 2}\sum_i \sum_j p_i p_j (\ln p_i + \ln p_j)$$

$$= -\frac{1}{ln\, 2}\left[\sum_i \sum_j p_i p_j \, ln\, p_i + \sum_i \sum_j p_i p_j \, ln\, p_j\right]$$

$$= -\frac{1}{\ln 2}\left[\sum_j p_j \sum_i p_i \ln p_i + \sum_i p_i \sum_j p_j \ln p_j\right]$$

$$= -\frac{1}{\ln 2}\left[\sum p_i \ln p_i + \sum p_j \ln p_j\right]$$

$$= I(x) + I(y)$$

Gesprochene Sprache

Ereignisse: Auftreten von Buchstaben in einer Folge, Symbole, Zahl der Symbole: 27, 26 Buchstaben + Leerzeichen (ohne Umlaute)

Falls gleichwahrscheinlich

$$I = -\frac{1}{\ln 2}\sum_i p_i \ln p_i = \frac{\ln 27}{\ln 2} = 4{,}76 \,\text{bit/Zeichen}$$

Die Buchstaben treten allerdings nicht gleichwahrscheinlich auf:

im Englischen

$$\left.\begin{array}{ll} \text{Leerzeichen} & 0{,}2 \\ \text{E} & 0{,}105 \\ \text{T} & 0{,}072 \\ \text{O} & 0{,}0654 \\ \text{A} & 0{,}063 \\ \text{N} & 0{,}059 \\ \text{I} & 0{,}055 \\ \vdots & \\ \text{Q} & 0{,}001 \end{array}\right\} \qquad I = 4.04 \text{ bit/Zeichen}$$

Ein Satz, der N Buchstaben enthält, enthält damit im Mittel die Information

$$I = -\frac{N}{\ln 2}\sum_i p_i \ln p_i = N \cdot 4{,}04\,\text{bit}$$

Der aktuelle Wert für I ist allerdings deutlich kleiner als 4,04 bit/Zeichen, weil es Korrelationen zwischen den Buchstaben gibt: Z.B. folgt im Deutschen nach "q" niemals "m", aber sehr oft "u".

Das Informationsmaß I ist "inhaltlich neutral", d.h. es hat nichts mit der Bedeutung der Information zu tun - es bezieht sich nur auf die „Vorhersagbarkeit“ eines Ereignisses. Je seltener ein Ereignis, desto größer dessen Informationsgehalt.
Vergleiche: Ein Postangestellter, der dasselbe Telegramm an mehrere Personen verschickt. Er wird sich nur darum kümmern, ob die Zeichen korrekt angegeben sind, die wirkliche Information, d.h. die Bedeutung hängt vom Empfänger ab. Text: *Tante gestorben* jemand freut sich, weil er vielleicht ein großes Erbe antreten kann; jemand anderes unglücklich usw.
Oder: Karten ziehen von einem Stapel von 32 Karten I=5 bit, aber Bedeutung für das Spiel ist völlig unterschiedlich, je nach Karte und Spiel.

Wir hatten bereits den Informationsgehalt von gleichzeitigen Ereignissen betrachtet, die voneinander unabhängig sind ($\rightarrow$ Additivität). Welche Regeln gelten, wenn sie nicht unabhängig sind? Für diese Überlegungen benötigen wir einen einfachen

Hilfssatz:

p_i sei die Wahrscheinlichkeitsverteilung, die uns interessiert, q_i sei eine andere Wahrscheinlichkeitsverteilung. Es gilt: $\sum_i p_i = 1$, $\sum_i q_i = 1$ $(*)$

Wir betrachten die Funktion: $\varphi = \sum_i p_i \ln q_i$

Dann gilt: $\sum p_i \ln q_i \leq \sum_i p_i \ln p_i$ (Gleichheit wenn alle $q_i = p_i$)

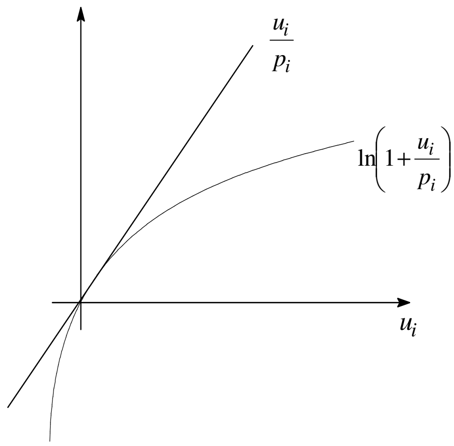
Beweis:

Sei $q_i = p_i + u_i$ aus $(*)$ folgt $\sum_i u_i = 0$

$$\varphi = \sum_i p_i \ln(p_i + u_i) = \sum_i p_i \ln\left[p_i\left(1 + \frac{u_i}{p_i}\right)\right]$$

$$\varphi = \sum_i p_i \ln p_i + \sum_i p_i \ln\left(1 + \frac{u_i}{p_i}\right)$$

$\ln\left(1 + \frac{u_i}{p_i}\right) = 0$ für $u_i = 0$ ansonsten gilt: $\ln\left(1 + \frac{u_i}{p_i}\right) \leq \frac{u_i}{p_i}$



$$\sum_i p_i \ln q_i = \sum_i p_i \ln p_i + \sum_i p_i \ln\left(1+\frac{u_i}{p_i}\right) \le \sum_i p_i \ln p_i + \underbrace{\sum_i p_i \cdot \frac{u_i}{p_i}}_{=0}$$

$$\sum p_i \ln q_i \le \sum_i p_i \ln p_i$$ **q.e.d.**

Aus diesem Satz folgt sofort:

$$-\lambda \sum_i p_i \ln q_i \ge -\lambda \sum_i p_i \ln p_i$$ (Gleichheit wenn alle $q_i = p_i$)

Betrachten wir jetzt wieder zwei Ereignismengen (Korreliert oder Unkorreliert):

$$x = \{1, 2 \ldots, m\} \quad , \quad y = \{1, 2 \ldots, n\}$$

$p(i, j)$ sei die Wahrscheinlichkeit für das simultane Auftreten von $x = i$, $y = j$

Es gilt:

1. $\sum_{i,j} p(i, j) = 1$
2. $p_i = \sum_{j=1}^{n} p(i, j)$: Wahrscheinlichkeit für das Auftreten von *i* (in *x*) unabhängig von *j* (in *y*)

   $p_j = \sum_{i=1}^{m} p(i, j)$: Wahrscheinlichkeit für das Auftreten von j (in y) unabhängig von i (in x)
3. $\sum_i p_i = 1$; $\sum_j p_j = 1$ und demzufolge auch: $\sum_i p_i \sum_j p_j = \sum_{i,j} p_i \cdot p_j = 1$

Angenommen Ereignisse aus x und y korrelieren. Was passiert, wenn wir allein Ereignisse *x* und allein Ereignisse *y* registrieren?

Erwartung: wir verzichten wir auf Kenntnisse (Korrelationen bleiben unberücksichtigt), daher Informationsgewinn größer (weil Kenntnisse geringer). Zur Quantifizierung dieses Sachverhaltes betrachten wir die

Verbundinformation

$$I(x, y) = -\lambda \sum_{i,j} p(i, j) \ln p(i, j)$$

Information für ein Ereignis allein aus *x*:

$$I(x) = -\lambda \sum_{i=1}^{m} p_i \ln p_i = -\lambda \sum_i \sum_j p(i, j) \ln p_i$$

Information für ein Ereignis allein aus *y*:

$$I(y) = -\lambda \sum_{j=1}^{n} p_j \ln p_j = -\lambda \sum_i \sum_j p(i, j) \ln p_j$$

Wir vergleichen die Summe von $I(x)$ und $I(y)$ mit $I(x, y)$:

$$I(x) + I(y) = -\lambda \sum_i \sum_j p(i, j)[\ln p_i + \ln p_j] = -\lambda \sum_i \sum_j \overbrace{p(i, j)}^{p} \ln \overbrace{(p_i \cdot p_j)}^{q}$$

$p(i, j)$ und $p_i \cdot p_j$ sind voneinander verschieden, wenn es sich um Ereignisse handelt, die nicht voneinander unabhängig sind: Setzt man $p_i \cdot p_j = q$ folgt aus dem obigen Hilfssatz:

$$-\lambda \sum_i \sum_j p(i, j) \ln(p_i p_j) \geq -\lambda \sum_i \sum_j p(i, j) \ln p(i, j)$$

$$\boxed{I_x + I_y \geq I(x, y)}$$

(Das Gleichheitszeichen gilt für unabhängige Ereignisse.) Die Berücksichtigung von Korrelationen führt also dazu, dass der erzielte Informationsgewinn geringer wird.

Bedingte Information

für zwei gleichzeitige Ereignisse *i* und *j* aus den Mengen *x* und *y*

Die bedingten Wahrscheinlichkeiten sind folgendermaßen definiert:

$$\underbrace{p(i,j)}_{\substack{\textit{Wahrscheinlichkeit,}\\ \textit{daß sowohl i als}\\ \textit{auch j auftreten}}} = \underbrace{p_i}_{\substack{\textit{Wahrscheinlichkeit,}\\ \textit{daß i auftritt}}} \cdot \underbrace{p_i(j)}_{\substack{\textit{Wahrscheinlichkeit,}\\ \textit{daß j auftritt unter der}\\ \textit{Bedingung, daß i}\\ \textit{aufgetreten ist}}}$$

$p_i(j)$: bedingte Wahrscheinlichkeit

(falls $p(i,j) = p_i p_j$ gilt $p_i p_j = p_i p_i(j)$ und damit $p_i(j) = p_j$)

$$p_i(j) = \frac{p(i,j)}{p_i}$$

Wir betrachten wie bisher die Verbundwahrscheinlichkeit

$$I(x,y) = -\lambda \sum_{i,j} p(i,j) \ln p(i,j)$$

und ersetzen im Logarithmus: $p(i,j)$ durch $p_i p_i(j)$

$$I(x,y) = -\lambda \sum_i \sum_j p(i,j) \ln \left[\underbrace{p_i p_i(j)}_{p(i,j)} \right] = -\lambda \sum_i \underbrace{\sum_j p(i,j)}_{p_i} \ln p_i - \lambda \sum_i \sum_j p(i,j) \ln p_i(j)$$

$$= \underbrace{-\lambda \sum_i p_i \ln p_i}_{I_x} \quad \underbrace{-\lambda \sum_i \sum_j p(i,j) \ln p_i(j)}_{\equiv I_x(y)}$$

mit der Einführung der **bedingten Information** $I_x(y) \equiv -\lambda \sum_i \sum_j p(i,j) \ln p_i(j)$

können wir also schreiben $\boxed{I(x,y) = I_x + I_x(y)}$

Informationsgewinn beim Messen von x und y = Informationsg. beim Messen von x alleine

\+ Informationsg. beim Messen von y bei

bereits bekanntem x.

Denn x bedingt ja y in einem gewissen Maße, wenn die Ereignisse nicht völlig unabhängig sind! Aus der vorher abgeleiteten Ungleichung folgert man:

$I_x + I_y \geq I(x,y) = I_x + I_x(y)$ oder: $\boxed{I_y \geq I_x(y)}$

Der Informationsgewinn, wenn wir nur y messen ist größer, als wenn wir y bei bereits bekanntem x messen, wenn x und y nicht unabhängig sind.

Korrelationen und Redundanz in der Sprache

Das Auftreten "bedingter Wahrscheinlichkeiten" ist typisch für Wahrnehmung der Sprache (sowohl gesprochen als auch geschrieben).
Zeitliche Reihenfolge: 1. Zeichen, dann 2. Zeichen usw.

Beim Hören (Lesen) des m.ten Zeichens ist das $(m-1)$ te bereits bekannt. Frage: enthält es das m.te Zeichen überhaupt noch Informationen (im Extremfall nicht).

In jedem Fall verringert sich der Informationsgehalt durch Betrachtung bedingter Wahrscheinlichkeiten: auf „Q“ folgt in der Regel „U“, auf „C“ folgt oft „H“ usw.

Das sind **Paarkorrelationen**. Berücksichtigen wir weitere, **langreichweitige Korrelationen**, sinkt der Informationsgewinn weiter. Wissen wir z. B. bereits „WOHNUN“ können wir leicht ergänzen „WOHNUNG“.

Bei Betrachtung kürzerer Zeichenketten hätten wir Z. B. „...UN“ ergänzen können zu „UND“, „UNO“ usw.

Wir betachten zunächst "Paarkorrelationen": Lesen eines Buchstabens: Ereignis x, Lesen des darauffolgenden Buchstabens: Ereignis y

$$I_x(y) = -\lambda \sum_{i,j} p(i,j) \ln p_i(j) \qquad (*)$$

$$I_x(y) = -\lambda \sum_i \sum_j p_i p_i(j) \ln p_i(j)$$

Wenn keine Korrelationen existieren, dann $p_i(j) = p_j$

$$I_x(y) = -\lambda \sum_{i,j} p_i \cdot p_j \ln p_i(j) = -\lambda \sum p_j \ln p_j = I(y)$$

Englische Sprache:

a) nur 27 Zeichen (ohne Berücksichtigung der Wahrscheinlichkeitsverteilung)
$I_0 = 4{,}76$ bit/Zeichen

b) Unter Berücksichtigung der Wahrscheinlichkeitsverteilung: $I_1 = 4{,}04$ bit/Zeichen

c) Unter Berücksichtigung von Paarkorrelationen: $I_2 = 3{,}32$ bit/Zeichen (erfordert Berechnung der bedingten Wahrscheinlichkeiten $p_i(j)$).

Dieses Konzept kann verallgemeinert werden, indem man Zeichenketten betrachtet („Blöcke" der Länge *N*):

$b_i(N-1)$ sei ein Block von Zeichen der Länge *(N*-1*)*, *i* dessen Nummer

N=4: „SCH",....."URS", „WOI"
i=1 *i*=2 *i*=3 u.s.w.

$p\{b_i(N-1)\}$: Wahrscheinlichkeit, dass dieser Block in einem ("langen") Text auftritt.

Wir ergänzen beim Lesen einen Block aus $N-1$ Zeichen durch ein *N*-tes Zeichen.

$b_{ij}(N) = (b_i(N-1), j)$: Block aus *N* Zeichen, charakterisiert durch Nummer *i*, ergänzt durch ein weiteres Zeichen *j*.

alle Zeichenketten durchnumeriert, *j:* Nummer des nächsten Zeichens

Anwendung des Konzeptes der bedingten Wahrscheinlichkeiten.

$$p\{b_{ij}(N)\} = p\{b_i(N-1), j\} = p\{b_i(N-1)\} p_{b_{i(N-1)}}(j)$$

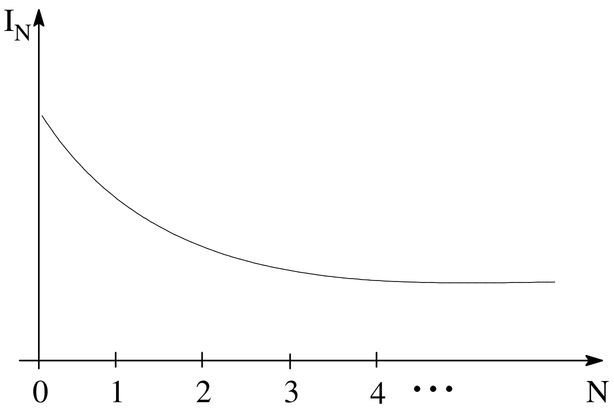
Daraus ergibt sich der durchschnittliche Informationsgehalt eines Zeichens (bei Kenntnis der vorangehenden $N-1$ Zeichen wie folgt (vgl. Gleichung (*)).

$$I_N = -\lambda \sum_{i,j} p\{b_i(N-1)j\} \ln p_{b_{i(N-1)}}(j)$$

bedingte Information

I_N sinkt monoton mit *N*. In obigem Beispiel:

$I_2 = 3{,}32$ bit/Zeichen
$I_3 = 3{,}1$ bit/Zeichen
$\vdots$



$I = \underbrace{\lim}_{N\to\infty} I_N$, dieser Limes ist schwer zu berechnen, da man etwas über die Häufigkeiten langer Zeichenketten in noch längeren Texten wissen muß.

Abschätzungen ergeben: $I = 2{,}14$ bit/Zeichen

Definition der Redundanz $R = \dfrac{\Delta I}{I_0}$ mit $\Delta I = I_0 - I$

Abfall der Information bei Berücksichtigung von Korrelationen (+Normierung)

$$R = \frac{I_0 - I}{I_0} = 1 - \frac{I}{I_0}$$

im vorliegenden Beispiel: $R = \frac{4{,}76 - 2{,}14}{4{,}76} = 0{,}55$

Signalübertragung

$$I = -\lambda \sum_{i=1}^{N} p_i \ln p_i$$

N: Zahl der Ereignisse; p_i: Wahrscheinlichkeit, daß Ereignis i eintritt.

p_i Wahrscheinlichkeit für das Auftreten einzelner Buchstaben oder ganzer Texte

Texte haben eine bestimmte zeitliche Länge.

Zeitdauer t Länge und die zugehörige Information I, die übertragen wird.

Rate der Informationsübertragung

$C = \frac{\Delta I}{\Delta t}$; hier: $\Delta t = t$ $\Delta I = I$

Anzahl $N(t)$ der Texte, die im Zweitraum$(0,t)$ übertragen werden können. Falls Texte gleichwahrscheinlich, gilt: $p = \frac{1}{N(t)}$

$$I = -\lambda \sum_{i=1}^{N} \frac{1}{N(t)} \ln \frac{1}{N(t)}$$

$$I = \lambda \ln N(t)$$