II Entropie und Information

§1 *Die Shannon-Information*

Zwei Aussagen A und B sollen erfragt werden, z.B. geht es nach rechts oder nach links zum Bahnhof. Als Quizfrage formuliert, darf die Antwort nur ja oder nein sein, und es genügt offentsichtlich eine Frage, um die Information herauszubekommen. Hat man zwischen vier Möglichkeiten A, B, C, D zu entscheiden, so muß man genau zwei Fragen stellen, um die richtige Information zu erfragen: Trifft A oder B zu, wenn ja fragt man weiter, trifft A zu, wenn ja so ist es A, wenn nein so ist es B. In beiden Fällen ist man am Ziel. Wenn die erste Frage negativ entschieden wird, dann muß man die Entscheidung zwischen C und D suchen, was auch eine weitere Frage kostet. In jedem Fall muß man also zwei Fragen stellen.

Die Information kann man also quantifizieren, indem man die Anzahl der Alternativfragen zählt, die man braucht, um die "Wahrheit" zu erfragen. Nach den beiden Beispielen wird man vermuten, daß die Information I sich als Logarithmus zur Basis 2 der Anzahl der Möglichkeiten N definieren lassen sollte

$$\boxed{I = \log_2 N} \quad , \qquad (*)$$

wobei $\log_2 x = \ln x / \ln 2$ ist. Es ist zunächst unklar, ob dies eine sinnvolle Definition für die Zahl der Möglichkeiten N ist, wenn N nicht eine Potenz von zwei ist.

Zur Illustration dieses Problems kann der Würfel dienen. Um zu erfragen, wieviel Augen geworfen worden sind, fragt man zuerst, ob es zwischen 1-3 oder zwischen 4-6 war. Die zweite Frage entscheidet z.B zwischen 1 und 2&3. Ist es die 1 gewesen, so ist man mit zwei Fragen am Ziel. Im anderen Fall muß man mit einer dritten Frage zwischen 2 und 3 entscheiden lassen. Da dieser Fall zwei mal häufiger im Durchschnitt auftreten wird als der günstigere mit zwei Fragen, kommt man auf durchschnittlich $2\,\frac{2}{3}$ Fragen. Dieses "rationales Ergebnis" ist mit dem "irrationalen" $\ln 6 / \ln 2 = 2.585$ zu vergleichen! Der Fehler ist also nicht so groß. Der Punkt ist einfach, und dieses Argument werden wir in Zukunft bei allen statistischen Betrachtungen in der Thermodynamik verwenden, daß diese Informationformel $(*)$ eigentlich nur für große N gültig ist. Wenn man also das Ergebnis vom Wurf zweier Würfel, einem schwarzen und einem weißen erfragt, dann gibt es 36 Möglichkeiten, die mit gleicher Wahrscheinlichkeit vorkommen. Beschränkt man sich auf die ersten $32 = 2^5$ Möglichkeiten, so genügen 5 Fragen. Mit der Wahrscheinlichkeit $4/36 = 1/9$ steckt die "Wahrheit" aber in den 4 übriggebliebenen Fällen, so daß man noch weitere zwei Fragen stellen muß. Die Information ist also $5 + 2/9 = 5,222$, was also schon ziemlich nahe bei $\log_2 36 = 5,170$ liegt. Die "Frage", ob dies die optimale Strategie zum Erkunden der Information war, bleibt natürlich offen ...

Ma sieht jedoch sofort, daß die Formel $(*)$ für große N gültig sein muß, denn sie interpoliert zwischen $N = 2^I$ für ganze I. Ist I groß, so muß die Interpolation einfach richtig werden. Wie wir gerade bemerkt haben, ist sie für N=6 so schlecht nicht.

Wirklich brauchbar ist die Formel $(*)$ für die Information noch nicht, denn sie bezieht sich nur auf den Spezialfall, daß alle Möglichkeiten gleich wahrscheinlich sind. Die Information des Textes, den Sie gerade lesen, bestände damit aus allen Buchstaben, Leerzeichen, Kommas, Wagenrückläufen etc., die alle mit gleicher Wahrscheinlichkeit benutzt würden und obendrein statistisch auftauchten. Diesen Eindruck müssen Computer oder Fernschreibern haben, sofern sie keine Wörterbücher, keine Syntax etc. kennen, um die Information auf Fehler zu überprüfen. Kompilieren Sie jedoch ein Programm, dann sieht die Sache anders aus, dann wird wirklich Information verarbeitet, die Texte sind dann für die Rechenmaschine alles andere als zufällig ...

Shannon hat die obige Formel entscheident verbessert, in dem er die Wahrscheinlichkeiten p_n für das Auftreten z.B. der Buchstaben mit der Nummer n berücksichtigt. Sein Vorschlag ist

$$\boxed{I = -\textstyle\sum_{n=1}^{N} p_n \log_2 p_n} \quad , \qquad (**)$$

der sich auf die alte Formel $(*)$ reduziert, falls alle Wahrscheinlichkeiten gleich sind, dh. $p_n = 1/N$. Folgen wir dem obigen Überlegungen, so läßt sich die Shannonformel leicht herleiten.

Wie wir vermuten können, wird diese Gleichung auch nur für große N wirklich gültig sein. Der Einfachheit betrachten wir zunächst einen Text, der nur mit zwei Zeichen geschrieben ist. Das eine Zeichen kommt mit der Wahrscheinlichkeit p_1 vor und das andere mit der Wahrscheinlichkeit p_2. Die Summe der Wahrscheinlichkeiten ist 1. Ist der Text K Zeichen lang, dann besteht er aus $K_1 = p_1 K$ und $K_2 = p_1 K$ Zeichen. K muß also groß sein, und für K_1 und K_2 müßte man eigentlich ganze Zahlen wählen. Die Anzahl verschiedener Texte N ist durch den Binomialkoeffizienten gegeben

$$N_{Anzahl} \;=\; \binom{K}{K_1} \;=\; \frac{K!}{K_1!\,K_2!} \;\approx\; \frac{K^K}{K_1^{K_1}\,K_2^{K_2}} \quad ,$$

den man mit Hilfe der Stirlingformel $K! \approx (K/e)^K$ approximieren kann. Mit dem genäherten Ausdruck erhält man für I oder I/K

$$\frac{I}{K} \;=\; -\frac{K_1}{K}\log_2\frac{K_1}{K} - \frac{K_2}{K}\log_2\frac{K_2}{K} \;=\; -p_1\log_2 p_1 - p_2\log_2 p_2 \quad .$$

Damit hat man einen Ausdruck gewonnen, den wir schon von der Mischungsentropie kennen.

Hat man N verschiedene Buchstaben mit Wahrscheinlichkeit p_n, dann ist die Anzahl der Texte der Länge K in dem $K_i = p_i\,K$ Buchstaben vom Typ i vorkommen

$$N_{Anzahl} \;=\; \frac{K!}{K_1!\,K_2!\cdots K_N!} \quad ,$$

woraus sich analog wie vorher mit Hilfe der Stirlingformel die vollständige Shannonformel ergibt.

Literatur: A.M. Jaglom & I.M. Jaglom, Wahrscheinlichkeit und Information, VEB Deutscher Verlag der Wissenschaften, Berlin 1960, übersetzt aus dem Russischen.

§2 *Eigenschaften der Shannonentropie*

Einige formelle Eigenschaften, die später wichtig sind sollen hier zusammengestellt werden. Die Ausgangsformel

$$I = -\sum_{n=1}^{N} p_n \log_2 p_n$$

soll für zwei verschiedene verschiedene Wahrscheinlichkeiten p_i und q_k berechnet werden, die sich auf unabhängige Möglichkeiten $1 \le i \le M$ und $1 \le k \le N$ beziehen. In diesem Falle ist die Wahrscheinlichkeit $W_{i,\,k} = p_i\,q_k$ und man erhält für die Information die Summe

$$I_{ges} \;=\; -\sum_{i=1}^{M}\sum_{k=1}^{N} p_i\,p_k \log_2(p_i\,p_k) \;=\; -\sum_{i=1}^{M} p_i \log_2 p_i \;+\; -\sum_{k=1}^{M} p_k \log_2 p_k = I_1 + I_2 \quad ,$$

wie man es auch erwarten würde.

Die Entropie nimmt ihr Maximum für $p_n = 1/N$ an. Zum Beweise benutzt man die Konvexität der Funktion $\varphi(x) = x\log x$. Es gilt offensichtlich $(\varphi(x_1)+\varphi(x_2))/2 \ge \varphi((x_1+x_2)/2)$ für eine konvexe Funktion, was sich auf viele Variable übertragen läßt:

$$\frac{1}{N}\sum_{n=1}^{N}\varphi(x_n) \;\ge\; \varphi\Big(\frac{1}{N}\sum_{n=1}^{N}x_n\Big) \qquad \rightarrow \qquad -\sum_{n=1}^{N} p_n\log_2 p_n \;\le\; -\log_2\frac{1}{N}$$