Anhang II zur Vorlesung Kryptologie: Entropie

von
Peter Hellekalek

Fakultät für Mathematik, Universität Wien, und
Fachbereich Mathematik, Universität Salzburg

Tel: +43-(0)662-8044-5310
Fax: +43-(0)662-8044-137
e-mail: peter.hellekalek@sbg.ac.at
web: http://random.mat.sbg.ac.at/

Wien, 24. April 2009

Inhaltsverzeichnis

Kapitel 1

Entropie

Dieses Kapitel ist eine Ausarbeitung und Adaption von Welsh[Wel88, Ch.1], unter Verwendung von Cover and Thomas[CT91, Ch.2].

▷ **Ziel**
Wir klären die Fragen nach

- einem mathematischen Maß für die Ungewissheit des Ergebnisses von Zufallsexperimenten,
- Eigenschaften dieses Maßes,
- dem Zusammenhang Entropie – Ungewissheit – Information.

1.1 Entropie = Ungewissheit

Wir möchten die Ergebnisses eines Zufallsexperiments (Münzwurf, Würfelspiel, Roulette, Pferderennen, Fußball, ...) möglichst effizient elektronisch übertragen können. *Im Durchschnitt* sollen wir möglichst wenige Bits für die Kommunikation dieser Informationen benötigen. Anders ausgedrückt: mit welcher Fragestrategie finden wir mit möglichst wenigen Ja/Nein-Fragen das Ergebnis der Zufallsexperimente heraus?

Wie können wir diese Aufgabe mathematisch fassen? Dazu hat sich Claude Shannon grundlegende Gedanken gemacht, welche die Basis der modernen Informationstheorie bilden, siehe die Arbeiten [Sha48, Sha49].[1]

Ein mathematisches Maß für die Unsicherheit muß bestimmte Voraussetzungen erfüllen. Um einige Einsichten zu gewinnen, betrachten wir die folgenden Aussagen:

1. Die Ungewissheit, wie ein Rennen zwischen zwei gleich schnellen Rennpferden ausgegangen ist, ist geringer als die Ungewissheit, wie ein Wettrennen zwischen acht gleichwertigen Pferden geendet hat.

[1] Beide Arbeiten finden sich im Internet.

2. Das Ergebnis eines Roulettespiels ist ungewisser, als das Ergebnis eines Würfelexperiments.

3. Wenn ein fairer Würfel geworfen wird, dann ist das Ergebnis ungewisser als bei einem stark verfälschten Würfel.

4. Unter der Laplace-Annahme bei der Schlüsselauswahl ist die Ungewissheit, mit welchem Schlüssel eine DES-Nachricht verschlüsselt wurde, geringer als die Ungewissheit bei einer AES-Nachricht.
(Die Schlüssellänge beträgt bei DES 56 Bit, bei AES mindestens 128 Bit.)

Allen diesen Aussagen liegt die folgende Fragestellung zugrunde.

1.1 Bemerkung (Zentrale Fragestellung)
Gegeben sei eine diskrete Zufallsvariable X mit endlichem Wertevorrat $x_1, \ldots, x_n$. Wir möchten durch Fragen ermitteln, welchen –uns unbekannten– Wert x_i die ZV X angenommen hat.

Wie groß ist die *durchschnittliche Anzahl* von Ja/Nein-Fragen, die wir zur Feststellung des Wertes benötigen?

Technischer formuliert: wie groß ist die durchschnittliche Bitanzahl, die wir benötigen, um die Werte der ZV X darzustellen?

1.2 Beispiel
Sei X die ZV, die das Ergebnis der Pferderennen zwischen den acht Pferden Nummer 0 bis 7 beschreibt, und sei P_X die Wahrscheinlichkeitsverteilung von X,

$$P_X = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right).$$

Wir wollen die Rennergebnisse möglichst Bit-sparend übertragen können. Es ist dann nicht effizient, die Nummer des siegreichen Pferdes zu übermitteln, denn dazu würden wir in jedem Fall drei Bit benötigen ($0 \equiv (0,0,0), 1 \equiv (0,0,1), \ldots, 7 \equiv (1,1,1)$). Wir werden in Kürze sehen, dass wir im Durchschnitt mit zwei Bit auskommen. Denken Sie dabei an Ja/Nein-Fragen: Pferd Nummer 0 ist als Sieger viel wahrscheinlicher als die anderen Pferde.

1.3 Beispiel
Wenn die ZV X das Ergebnis eines Münzwurfs beschreibt, $X \sim (p, 1-p)$, $0 \leq p \leq 1$, und die ZV Y ein Alternativexperiment mit der Verteilung $Y \sim (p, 1-p)$, dann sind diese beiden ZVen vom Standpunkt der Ungewissheit über das Ergebnis eines Versuches gleich. Das noch zu findende Maß für die Ungewissheit wird daher nur von der Wahrscheinlichkeitsverteilung der ZV abhängen. Es wird also für obiges X und Y den gleichen Wert annehmen.

1.4 Bemerkung (Grundprinzipien)
Wir erkennen einige Grundprinzipien, die für ein Maß für die Ungewissheit sinnvoll sind:

1. Nicht die numerische Größe der Werte $x_1, \ldots, x_n$, welche die Zufallsvariable X annehmen kann, wird in dieses Maß eingehen, sondern die Anzahl n der möglichen Werte von X.

2. Ein Maß für die Ungewissheit soll, für gegebenes n, nur von der Wahrscheinlichkeitsverteilung $P_X = (p_1, \ldots, p_n)$, $p_i = \mathsf{P}(X = x_i)$, $1 \leq i \leq n$, der ZV X abhängen

1.5 Bemerkung (Logarithmus dualis)
Im Folgenden bezeichnet ld den Logarithmus zur Basis 2. Es gilt die Beziehung

$$\log x \;=\; \log 2 \cdot \operatorname{ld} x, \quad x > 0. \tag{1.1}$$

1.6 Bemerkung (Festlegung)
Wir legen fest:

$$0 \operatorname{ld} 0 := 0.$$

Diese Festlegung ist naheliegend, da für den rechtsseitigen Grenzwert der Funktion $x \operatorname{ld} x$ gilt[2]:

$$\lim_{x \downarrow 0} x \operatorname{ld} x = 0.$$

1.7 Definition (Entropie einer ZV)
Sei X eine diskrete Zufallsvariable mit endlichem Wertevorrat $W_X = \{x_1, \ldots, x_n\}$ und Verteilung $X \sim P_X = (p_1, \ldots, p_n)$. Dann definieren wir die *Entropie der Zufallsvariablen* X als die Zahl

$$\mathrm{H}(X) \;=\; -\sum_{i=1}^{n} p_i \operatorname{ld} p_i. \tag{1.2}$$

Die Entropie eines Zufallsvektors wird analog definiert.

1.8 Definition (Entropie eines Zufallsvektors)
Sei $\mathbf{X}$ ein Zufallsvektor mit endlichem Wertevorrat $W_{\mathbf{X}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ und Wahrscheinlichkeitsverteilung $\mathbf{X} \sim P_{\mathbf{X}} = (p_1, \ldots, p_n)$. Dann definieren wir die *Entropie des Zufallsvektors* $\mathbf{X}$ als die Zahl

$$\mathrm{H}(\mathbf{X}) \;=\; -\sum_{i=1}^{n} p_i \operatorname{ld} p_i,$$

Angesichts der Definition der Entropie eines Zufallsvektors liegt es nahe, für eine endlichen Folge von ZVen folgenden Entropiebegriff einzuführen.

1.9 Definition (Gemeinsame Entropie von ZVen)
Sei $X_1, X_2, \ldots, X_m$ eine endliche Folge von diskreten ZVen, jede mit endlichem Wertevorrat. Dann definieren wir die *gemeinsame Entropie* $\mathrm{H}(X_1, \ldots, X_n)$ der ZVen $X_1, \ldots, X_n$ durch

$$\mathrm{H}(X_1, \ldots, X_n) \;=\; -\sum \operatorname{ld} p(x_1, \ldots, x_n),$$

wobei

$$p(x_1, \ldots, x_n) \;=\; \mathsf{P}(X_1 = x_1, \ldots, X_n = x_n),$$

und wobei über alle möglichen Werte des Vektors $\mathbf{X} = (X_1, \ldots, X_n)$ summiert wird.

[2] Zum Beweis können Sie die Regel von de l'Hospital verwenden, unter Beachtung der Identität (1.1).

Wenn wir die vorangegangenen Definitionen des Entropiebegriffes vergleichen, so stellen wir fest, dass es stets um die *Wahrscheinlichkeitsverteilungen* der ZVen gegangen ist und nicht um die ZVen selbst. Es liegt daher nahe, die Entropie einer Wahrscheinlichkeitsverteilung zu definieren.

1.10 Definition (Entropie einer W-Verteilung)
Sei $P = (p_1, \ldots, p_n)$ eine Wahrscheinlichkeitsverteilung.

Dann definieren wir die Entropie von P als die Zahl

$$\mathrm{H}(P) \;=\; -\sum_{i=1}^{n} p_i \,\mathrm{ld}\, p_i.$$

1.11 Satz
Sei $P = (p_1, \ldots, p_n)$ eine Wahrscheinlichkeitsverteilung. Dann gilt:

1. Für $\mathrm{H}(p_1, \ldots, p_n)$ gelten folgende Schranken:
$$0 \;\leq\; \mathrm{H}(p_1, \ldots, p_n) \;\leq\; \mathrm{ld}\, n.$$
2. Zur unteren Schranke: $\mathrm{H}(p_1, \ldots, p_n) = 0$ genau dann, wenn die W-Verteilung entartet ist, wenn also ein i existiert mit $p_i = 1$.
3. Zur oberen Schranke: $\mathrm{H}(p_1, \ldots, p_n) = \mathrm{ld}\, n$ genau dann, wenn $p_i = 1/n$, $\forall\, i,\, 1 \leq i \leq n$.

Beweis. Siehe Korollar 1.22 oder Welsh[Wel88, Ch1.2, Theorem 1, p.5]. □

1.12 Bemerkung (Extremalität der Gleichverteilung)
Der Begriff der Entropie (einer Zufallsvariablen bzw. einer Verteilung) ist das gewünschte Maß für die Ungewissheit bzw. jene Zahl, die angibt, wie viele Bits wir im Durchschnitt benötigen, um die Werte einer Zufallsvariablen zu beschreiben.

Wie uns die Aussagen von Satz 1.11 zeigen, ist die Entropie im Fall der Gleichverteilung maximal und sie ist minimal im Fall einer konstanten Zufallsvariablen bzw. einer entarteten Verteilung. Dies entspricht unserer Intuition!

1.13 Bemerkung (Entropie zur Basis b)
Wir sind an der Kodierung und Verschlüsselung von Bitströmen interessiert. Deshalb haben wir die Entropie zur Basis 2 definiert, indem wir in der Definition von H den Zweier-Logarithmus ld verwendet haben.

Wenn wir die Entropie über den Logarithmus $\log_b$ zur Basis b definieren möchten, so ergibt sich wegen $\log_b x = \log_b 2 \cdot \mathrm{ld}\, x$ die folgende Umrechnung:

$$\mathrm{H}_b(P) \;=\; \log_b 2 \cdot \mathrm{H}(P).$$

1.14 Bemerkung (Charakterisierung der Entropie einer W-Verteilung)
Interessant ist das folgende Resultat:

Sei H eine reellwertige Funktion, die für jedes n und für jede Wahrscheinlichkeitsverteilung $(p_1, \ldots, p_n)$ definiert sei. Wenn H die folgenden Eigenschaften besitzt,

1. $\mathrm{H}(p, 1-p)$ ist eine stetige Funktion von p, $p \in [0,1]$,

2. $\mathrm{H}(p_1, \ldots, p_n) = H(p_{\pi(1)}, \ldots, p_{\pi(n)})$ für jede Permutation π der Indexmenge $\{1, \ldots, n\}$,

3. Es gilt folgende Zerlegungseigenschaft:

$$\begin{aligned} \mathrm{H}(p_1, p_2, p_3, \ldots, p_n) &= \mathrm{H}(p_1 + p_2, p_3, \ldots, p_n) \\ &\quad + (p_1 + p_2)\ \mathrm{H}\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right), \end{aligned}$$

dann gilt notwendigerweise

$$\mathrm{H}(p_1, \ldots, p_n) = -\lambda \sum_{i:p_i>0}^{n} p_i \operatorname{ld} p_i,$$

mit einer positiven Konstanten λ. Einen Beweis für diese Behauptung finden Sie zum Beispiel in Welsh[Wel88, Appendix I] oder in der Originalarbeit von Shannon[Sha48, p.10; Appendix 2, p.28] (siehe auch die Anmerkungen in Cover and Thomas[CT91, Ex.4, p.42]).

1.2 I-Divergenz

Der folgende Begriff der I-Divergenz zweier Wahrscheinlichkeitsverteilungen ist aus mehreren Gründen interessant. In dieser Vorlesung wird uns dieses *Distanzmaß zwischen Verteilungen* helfen, einige Aussagen und Begriffe besser zu verstehen und Beweise zu erleichtern. Für die Bedeutung dieses Begriffes in Zusammenhang mit statistischem Testen verweisen wir auf Wegenkittl[Weg02].

1.15 Definition (I-Divergenz)
Seien $P = (p_1, \ldots, p_n)$ und $Q = (q_1, \ldots, q_n)$ zwei Wahrscheinlichkeitsverteilungen und es gelte $q_i > 0\ \forall\ i,\ 1 \le i \le n$.

Dann verstehen wir unter der *I-Divergenz von P und Q* (auch *relativen Entropie* oder *Kullback-Leibler Distanz von P und Q* genannt) die Größe

$$\mathrm{I}(P||Q) = \sum_{i=1}^{n} p_i \operatorname{ld} \frac{p_i}{q_i}.$$

1.16 Bemerkung (Verallgemeinerung)
In Definition 1.15 läßt sich die Bedingung "$q_i > 0\ \forall\ i,\ 1 \le i \le n$" zur Voraussetzung "$p_i + q_i > 0\ \forall\ i,\ 1 \le i \le n$", abschwächen, wenn wir $0 \cdot \operatorname{ld} \frac{0}{q} = 0$ und $p \cdot \operatorname{ld} \frac{p}{0} = \infty$ setzen (p und q beide größer als Null).

Im Folgenden werden wir einige Eigenschaften von $\mathrm{I}(P||Q)$ benötigen, die einen Ausflug in die elementare Analysis erforderlich machen, zur Jensenschen Ungleichung.

1.17 Definition (Konvexe und konkave Funktionen)
Sei (a, b) ein reelles Intervall und sei $f : (a, b) \to \mathbb{R}$. Die Funktion f heißt *konvex* auf (a, b), wenn gilt:

$$\begin{aligned} &\forall\, x_1, x_2 \in (a,b), \\ &\quad \forall\, \lambda \in [0,1] : \\ &\qquad f(\lambda x_1 + (1-\lambda)x_2) \;\leq\; \lambda f(x_1) + (1-\lambda) f(x_2). \end{aligned} \tag{1.3}$$

Die Funktion f heißt *streng konvex* auf (a, b), wenn in (1.3) nur für $\lambda = 0$ und $\lambda = 1$ die Gleichheit gilt.

Die Funktion f heißt *(streng) konkav* auf (a, b), wenn die Funktion $-f$ (streng) konvex auf (a, b) ist.

1.18 Lemma
Sei f in (a, b) zweimal differenzierbar. Wenn die zweite Ableitung f'' in (a, b) nichtnegativ (positiv) ist, dann ist f konvex (streng konvex) auf (a, b).

Beweis. Siehe die angegebene Literatur. □

1.19 Beispiel
Aus dem Lemma folgt:

$$\begin{aligned} f(x) &= \ln x \quad \text{ist streng konkav auf } [0, +\infty[, \\ f(x) &= x \cdot \ln x \quad \text{ist streng konvex auf } [0, +\infty[, \end{aligned}$$

Der folgende Satz ist ein wichtiges technisches Hilfsmittel für zahlreiche Beweise in der Informationstheorie.

1.20 Satz (Jensensche Ungleichung)
Sei f eine konvexe Funktion und sei X eine reellwertige Zufallsvariable. Dann gilt

1. Jensensche Ungleichung:

$$f(E[X]) \;\leq\; E[f(X)].$$

2. Wenn die Funktin f streng konvex ist, dann gilt

$$f(E[X]) = E[f(X)] \quad \Longleftrightarrow \quad \mathsf{P}(X = E[X]) = 1, \text{ d.h. } X \text{ ist konstant.}$$

Beweis. Siehe die angegebene Literatur.. □

1.21 Satz (Informationsungleichung)
Seien $P = (p_1, \ldots, p_n)$ und $Q = (q_1, \ldots, q_n)$ zwei Wahrscheinlichkeitsverteilungen und es gelte $q_i > 0\ \forall\, i,\ 1 \leq i \leq n$.

Dann gilt

1. $\mathrm{I}(P||Q) \geq 0$,
2. $\mathrm{I}(P||Q) = 0$ genau dann, wenn $P = Q$.

Beweis. Siehe die angegebene Literatur. □

1.22 Korollar (Eigenschaften von H)
Mit Hilfe von Satz 1.21 können wir die Behauptungen von Satz 1.11 beweisen:

1. Für $\mathrm{H}(p_1, \ldots, p_n)$ gelten folgende Schranken:
$$0 \leq \mathrm{H}(p_1, \ldots, p_n) \leq \mathrm{ld}\, n.$$
2. Zur unteren Schranke: $\mathrm{H}(p_1, \ldots, p_n) = 0$ genau dann, wenn die W-Verteilung entartet ist, wenn also ein i existiert mit $p_i = 1$.
3. Zur oberen Schranke: $\mathrm{H}(p_1, \ldots, p_n) = \mathrm{ld}\, n$ genau dann, wenn $p_i = 1/n$, $\forall\, i,\ 1 \leq i \leq n$.

Beweis. Sei $P = (p_1, \ldots, p_n)$ eine Wahrscheinlichkeitsverteilung und bezeichne $U_n = (1/n, \ldots, 1/n)$ die Gleichverteilung.

Wegen $-\sum_{i=1}^{n} p_i \,\mathrm{ld}\, p_i \geq 0$ ist die Behauptung $\mathrm{H}(p_1, \ldots, p_n) \geq 0$ trivial.

Für die Summanden $-p_i \,\mathrm{ld}\, p_i$, $1 \leq i \leq n$, in $\mathrm{H}(p_1, \ldots, p_n)$ gilt:

$$-p_i \,\mathrm{ld}\, p_i = \begin{cases} 0 & \text{falls } p_i \in \{0, 1\}, \\ > 0 & \text{falls } 0 < p_i < 1. \end{cases}$$

Offensichtlich gilt daher $\mathrm{H}(p_1, \ldots, p_n) = 0$ genau dann, wenn ein p_i mit $p_i = 1$ existiert. Die Verteilung P ist dann notwendig entartet.

Es gilt

$$\begin{aligned} \mathrm{I}(P||U_n) &= \sum_{i=1}^{n} p_i \,\mathrm{ld}\, \frac{p_i}{1/n} \\ &= -\sum_{i=1}^{n} p_i \,\mathrm{ld}(1/n) + \sum_{i=1}^{n} p_i \,\mathrm{ld}\, p_i \\ &= \mathrm{ld}\, n - H(P). \end{aligned}$$

Aus $\mathrm{I}(P||U_n) \geq 0$ sowie der Eigenschaft $\mathrm{I}(P||U_n) = 0$ genau dann, wenn $P = U_n$ folgt $\mathrm{H}(P) \leq \mathrm{ld}\, n$ und weiters die wichtige Extremaleigenschaft der Gleichverteilung,

$$\mathrm{H}(U_n) = \mathrm{ld}\, n.$$

□

Manchmal ist es erforderlich, zwei beliebige diskrete Wahrscheinlichkeitsverteilungen P und Q zu vergleichen, ohne Einschränkungen über Q. Dann ist zwar die I-Divergenz nicht definiert, aber wir können die folgende Ungleichung zeigen.

1.23 Lemma ("key lemma" in Welsh[Wel88, Ch.1.2, p.5])
Seien $P = (p_1, \ldots, p_n)$ und $Q = (q_1, \ldots, q_n)$ zwei Wahrscheinlichkeitsverteilungen. Dann folgt

1. Es gilt folgende fundamentale Ungleichung:
$$-\sum_{i=1}^{n} p_i \operatorname{ld} q_i \;\geq\; -\sum_{i=1}^{n} p_i \operatorname{ld} p_i, \tag{1.4}$$
2. In der Ungleichung (1.4) gilt Gleichheit genau dann, wenn $P = Q$.

Beweis. Wir unterscheiden zwei Fälle: im ersten Fall seien alle Wahrscheinlichkeiten q_i positiv (d.h. der Träger $T(Q)$ ist gleich der Menge $\{1, \ldots, n\}$, im zweiten Fall existieren Indizes i mit $q_i = 0$.

Fall 1: Sei $q_i > 0 \;\forall\, i$, $1 \leq i \leq n$.
Die Ungleichung (1.4) gilt genau dann, wenn
$$\sum_{i=1}^{n} p_i \operatorname{ld} p_i - \sum_{i=1}^{n} p_i \operatorname{ld} q_i \;=\; \underbrace{\sum_{i=1}^{n} p_i \operatorname{ld} \frac{p_i}{q_i}}_{\mathrm{I}(P||Q)} \;\geq\; 0.$$

Alles weitere, auch die Äquivalenz von $P = Q$ zur Gleichheit in (1.4) folgt aus Satz 1.21 zur I-Divergenz.

Fall 2: Es existiere ein Index i mit $q_i = 0$.
Wir zerlegen den Summationsbereich geeignet:
$$\begin{aligned} \{1, \ldots, n\} \;=\; & T(Q) \\ & \cup\; \underbrace{\{i : p_i = 0 \wedge q_i = 0\}}_{A} \\ & \cup\; \underbrace{\{i : p_i > 0 \wedge q_i = 0\}}_{B} \end{aligned}$$

Daraus folgt für die Summe auf der linken Seite der Ungleichung (1.4):
$$-\sum_{i=1}^{n} p_i \operatorname{ld} q_i \;=\; -\sum_{i \in T(Q)} p_i \operatorname{ld} q_i \;-\sum_{i \in A} p_i \operatorname{ld} q_i \;-\sum_{i \in B} p_i \operatorname{ld} q_i \tag{1.5}$$

Fall 2.1: $B \neq \emptyset$
Dann ist $-\sum_{i \in B} p_i \operatorname{ld} q_i$ gleich $+\infty$. Da auf der rechten Seite die nichtnegative Zahl $\mathrm{H}(P)$ steht, ist die Ungleichung (1.4) in diesem Fall trivial.

Fall 2.2: $B = \emptyset$
Die Aussage "$B = \emptyset$" ist logisch äquivalent zur Aussage "$p_i > 0 \implies q_i > 0$". Daher gilt die Beziehung $T(P) \subseteq T(Q)$. Damit ist aber die Verteilung $(p_i : i \in T(Q))$ eine Wahrscheinlichkeitsverteilung. Nach Teil 1 des Beweises folgt die Ungleichung
$$-\sum_{i \in T(Q)} p_i \operatorname{ld} q_i \;\geq\; -\sum_{i \in T(Q)} p_i \operatorname{ld} p_i.$$

Aus der Zerlegung (1.5) folgt wegen $\sum_{i\in A} p_i \operatorname{ld} q_i = \sum_{i\in A} p_i \operatorname{ld} p_i = 0$ die Ungleichung (1.4).

Wir untersuchen als Nächstes, wann bei Fall 2 in der Ungleichung (1.4) die Gleichheit eintritt.

Als Erstes merken wir an, dass bei Gleichheit in (1.4) der Fall $B \neq \emptyset$ nicht eintreten kann, da dann in (1.4) links $+\infty$ und rechts die reelle Zahl $\mathrm{H}(P)$ steht.

Die gleichen Überlegungen wie in Fall 2.2 führen zur Identität

$$-\sum_{i\in T(Q)} p_i \operatorname{ld} q_i \;=\; -\sum_{i\in T(Q)} p_i \operatorname{ld} p_i.$$

Da (nach Definition) für alle $i \in T(Q)$ gilt, dass $q_i > 0$ ist, sind wir in Fall 1. Dort wurde gezeigt, dass aus der Gleichheit folgt, dass gilt:

$$q_i \;=\; p_i \quad \forall\, i \in T(Q).$$

Wegen $q_i = p_i \;\forall\, i \in A$ gilt $q_i \;=\; p_i \;\forall\, i \in \{1,\ldots,n\}$. Dies ist nichts anderes als die gesuchte Aussage $P = Q$. □

1.24 Satz
Seien X und Y zwei Zufallsvariable mit endlichem Wertevorrat. Dann gilt

1. Abschätzung der gemeinsamen Entropie $\mathrm{H}(X,Y)$:

$$\mathrm{H}(X,Y) \;\leq\; \mathrm{H}(X) + \mathrm{H}(Y),$$

2. Gleichheitsbedingung:
$\mathrm{H}(X,Y) = \mathrm{H}(X) + \mathrm{H}(Y)$ genau dann, wenn X und Y unabhängig sind.

Beweis. Siehe die angegebene Literatur. □

1.25 Korollar
Die folgenden Aussagen sind direkte Verallgemeinerungen von Satz 1.24, bei gleicher Beweismethode.

1. Seien $X_1,\ldots,X_n$ Zufallsvariable, jeweils mit endlichem Wertevorrat. Dann gilt
$$\mathrm{H}(X_1,\ldots,X_n) \;\leq\; H(X_1) + \ldots + \mathrm{H}(X_n)$$
und es gilt Gleichheit genau dann, wenn $X_1,\ldots,X_n$ unabhängig sind.

2. Seien $\mathbf{X}$ und $\mathbf{Y}$ zwei Zufallsvektoren, jeweils mit endlichem Wertevorrat. Dann gilt
$$\mathrm{H}(\mathbf{X},\mathbf{Y}) \;\leq\; \mathrm{H}(\mathbf{X}) + \mathrm{H}(\mathbf{Y}),$$
mit Gleichheit genau dann, wenn $\mathbf{X}$ und $\mathbf{Y}$ unabhängig sind.

1.3 Bedingte Entropie

Die Ungewissheit, welche Werte eine Zufallsvariable angenommen hat, läßt sich durch Zusatzinformationen vermindern. Beispiele dafür sind das Abschreiben in der Schule oder ganz allgemein die Spionage. Wir fassen dies in eine präzise mathematische Form.

1.26 Bemerkung (Erinnerung)
Wir betrachten im gesamten Kapitel 1 ausschließlich diskrete Zufallsvariable und Zufallsvektoren mit *endlichem* Wertevorat.

1.27 Definition (Bedingte Entropie)
Sei $(\Omega, \mathcal{A}, \mathsf{P})$ ein Wahrscheinlichkeitsraum, sei $X : \Omega \to \mathbb{R}$ eine Zufallsvariable und sei $W_X = \{x_1, \ldots, x_n\}$ der endliche Wertevorrat von X.

Unter einem *Ereignis* A in Ω verstehen wir eine messbare Teilmenge A von Ω, also $A \in \mathcal{A}$.

Unter der *bedingten Entropie* $\mathrm{H}(X|A)$ von X unter A verstehen wir die Zahl

$$\mathrm{H}(X|A) \;=\; -\sum_{i=1}^{n} \mathsf{P}(X = x_i|A)\ \operatorname{ld} \mathsf{P}(X = x_i|A).$$

Wenn $Y : \Omega \to \mathbb{R}$ eine weitere Zufallsvariable ist, mit Wertebereich $W_Y = \{y_1, \ldots, y_m\}$, dann nennen wir die Zahl

$$\mathrm{H}(X|Y) \;=\; \sum_{j=1}^{m} \mathrm{H}(X|Y = y_j)\ \mathsf{P}(Y = y_j)$$

die *bedingte Entropie* von X unter Y.

1.28 Beispiel
Wir betrachten das Würfeln mit einem fairen Würfel. Sei

X ... das Ergebnis des Würfelns
Y ... das Ergebnis des Würfelns modulo 2 (d.h. gerade/ungerade)

Dann gilt $W_X = \{1, 2, \ldots, 6\}$, $W_Y = \{0, 1\}$, $X \sim (1/6, 1/6, 1/6, 1/6, 1/6, 1/6)$ und $Y \sim (1/2, 1/2)$.

Wir sehen (für die Details siehe die Vorlesung):

$$\begin{aligned} H(X|Y) &= \operatorname{ld} 3, \\ H(X) &= \operatorname{ld} 6 \sim 2,58. \end{aligned}$$

1.29 Bemerkung (Eigenschaften der bedingten Entropie)
Wir fassen einige Eigenschaften der bedingten Entropie zusammen.

1. $\mathrm{H}(X|A)$ gibt die Ungewissheit an welchen Wert die ZV X annimmt, wenn das Ereignis A eingetreten ist.

2. $\mathrm{H}(X|Y)$ gibt die Ungewissheit an welchen Wert die ZV X annimmt, wenn wir $\mathrm{H}(X|Y = y_j)$ über alle Werte y_j der ZV Y mitteln.
3. $\mathrm{H}(X|X) = 0$
   Dies ist leicht nachzuweisen, denn $\mathrm{H}(X|X = x_i) = 0, \forall\, i, 1 \leq i \leq n$.
4. Wenn die ZVen X und Y unabhängig sind, dann gilt

$$\mathrm{H}(X|Y) \;=\; \mathrm{H}(X).$$

   (Beweis in der Vorlesung.)

1.30 Frage
Ist die folgende Äquivalenz richtig?

$$\mathrm{H}(X|Y) \;=\; \mathrm{H}(X) \iff X, Y \text{ unabhängig}$$

Die Antwort auf diese Frage werden wir in Kürze geben.

1.31 Satz Es gilt:

$$\mathrm{H}(X|Y) = 0 \iff \exists\, g: \quad X = g(Y).$$

Beweis. Siehe die angegebene Literatur. □

1.32 Satz (Kettenregel)
Es gilt die Kettenregel:

$$\begin{aligned} \mathrm{H}(X,Y) \;&=\; \mathrm{H}(X|Y) + \mathrm{H}(Y) \qquad \text{bzw.} \\ \mathrm{H}(X|Y) \;&=\; \mathrm{H}(X,Y) - \mathrm{H}(Y). \end{aligned}$$

Beweis. Siehe die angegebene Literatur. □

1.33 Korollar (Unabhängigkeitsbedingung)
Es gelten folgende wichtige Aussagen:

1. $\mathrm{H}(X|Y) \;\leq\; \mathrm{H}(X)$
2. $\mathrm{H}(X|Y) \;=\; \mathrm{H}(X) \quad \iff \quad X, Y$ unabhängig.

Beweis. Nach der Kettenregel gilt $\mathrm{H}(X,Y) \;=\; \mathrm{H}(X|Y) + \mathrm{H}(Y)$. Nach Satz 1.24 gilt $\mathrm{H}(X,Y) \;\leq\; \mathrm{H}(X) + \mathrm{H}(Y)$, mit Gleichheit genau dann, wenn X und Y unabhängig sind. □

1.34 Satz (Kettenregel für Zufallsvektoren)
Seien $\mathbf{X}$ und $\mathbf{Y}$ zwei Zufallsvektoren. Es gilt die Kettenregel:

$$\begin{aligned} \mathrm{H}(\mathbf{X},\mathbf{Y}) \;&=\; \mathrm{H}(\mathbf{X}|\mathbf{Y}) + \mathrm{H}(\mathbf{Y}) \qquad \text{bzw.} \\ \mathrm{H}(\mathbf{X}|\mathbf{Y}) \;&=\; \mathrm{H}(\mathbf{X},\mathbf{Y}) - \mathrm{H}(\mathbf{Y}). \end{aligned}$$

Beweis. Analog zum Beweis von Satz 1.32. □

1.35 Korollar
Die Aussagen von Korollar 1.33 gelten auch für Zufallsvektoren.

1.4 Wechselseitige Information

Wir möchten den Informationsgewinn für die Zufallsvariable X messen der sich ergibt, wenn wir die Zufallsvariable Y beobachten.

1.36 Definition (Wechselseitige Information)
Seien X und Y zwei diskrete Zufallsvariable mit endlichem Wertevorrat. Unter der *wechselseitigen Information* von X und Y verstehen wir die Zahl

$$\mathrm{I}(X|Y) \;=\; \mathrm{H}(X) - \mathrm{H}(X|Y).$$

1.37 Bemerkung (Wechselseitige Information von Zufallsvektoren)
Die wechselseitige Information $\mathrm{I}(\mathbf{X}|\mathbf{Y})$ von Zufallsvektoren $\mathbf{X}$ und $\mathbf{Y}$ wird analog definiert, $\mathrm{I}(\mathbf{X}|\mathbf{Y}) \;=\; \mathrm{H}(\mathbf{X}) - \mathrm{H}(\mathbf{X}|\mathbf{Y})$.

1.38 Bemerkung (Interpretation)
Die wechselseitige Information $\mathrm{I}(X|Y)$ gibt an, wieviel Ungewissheit über X durch die Kenntnis von Y beseitigt wird. $\mathrm{I}(X|Y)$ mißt also den *Informationsgewinn* durch Y.

1.39 Bemerkung (Verwechslungsgefahr)
Die wechselseitige Information $\mathrm{I}(X|Y)$ darf nicht mit der I-Divergenz verwechselt werden. Beachten Sie in diesem Zusammenhang, dass wir die wechselseitige Information für *Zufallsvariable*, die I-Divergenz aber für *Wahrscheinlichkeitsverteilungen* definiert haben.

Es gilt folgender Zusammenhang.

1.40 Lemma
Seien X und Y zwei diskrete Zufallsvariable mit endlichem Wertevorrat und gemeinsamer Verteilung $P_{(X,Y)} = (p(x,y) : x \in T_X, y \in T_Y)$, wobei $T_X = \{x : p(x) > 0\}$ und $T_Y = \{y : p(y) > 0\}$ die Trägermengen der Randverteilungen P_X von X und P_Y von Y sind. Dann gilt:

$$\underbrace{\mathrm{I}(X|Y)}_{\text{Wechselseitige Information}} \quad = \quad \underbrace{\mathrm{I}(P_{(X,Y)}||P_X \times P_Y)}_{\text{I-Divergenz}}.$$

Beweis. Siehe die angegebene Literatur. □

1.41 Lemma (Eigenschaften von $\mathrm{I}(X|Y)$)
Es gilt:

1. $\mathrm{I}(X|Y) = \mathrm{I}(Y|X)$.
2. $\mathrm{I}(X|Y) \geq 0$,
   mit Gleichheit genau dann, wenn X und Y unabhängig sind.
3. $\mathrm{I}(X|Y) \leq \min\{\mathrm{H}(X), \mathrm{H}(Y)\}$,
   mit Gleichheit genau dann, wenn gilt:
   - Falls $\min\{\mathrm{H}(X), \mathrm{H}(Y)\} = \mathrm{H}(X)$: $X = g(Y)$,
   - Falls $\min\{\mathrm{H}(X), \mathrm{H}(Y)\} = \mathrm{H}(Y)$: $Y = f(X)$.

Beweis. Siehe die angegebene Literatur. □

Literaturverzeichnis

[CT91] Th.M. Cover and J.A. Thomas. *Elements of Information Theory.* Wiley, New York, 1991.

[Sha48] C. E. Shannon. A mathematical theory of communication. *Bell Sys. Tech. J.*, 27:623–656, 1948.

[Sha49] C. E. Shannon. Communication theory of secrecy systems. *Bell Sys. Tech. J.*, 28:657–715, 1949.

[Weg02] S. Wegenkittl. A generalized ϕ-divergence for asymptotically multivariate normal models. *Journal of Multivariate Analysis*, **83**:288–302, 2002.

[Wel88] D. Welsh. *Codes and Cryptography.* Oxford Science Publ., 1988.