1 Grundlagen

1.1 Erste Grundbegriffe

1.2 Kryptographische Systeme

1.3 Informationstheoretische Grundlagen

Die Überlegungen dieses Kapitels basieren auf der Informationstheorie von Shannon. Er beschäftigte sich Mitte letzten Jahrhunderts mit

- dem Problem des gestörten Kanals und
- dem Problem der Geheimhaltung,

also mit der Kryptologie durchaus verwandten Problemen.

Zentraler Begriff der Informationstheorie ist der Begriff der Information, bzw. der Entropie. Wir wollen diesen Begriff nun bilden. Dazu betrachten wir folgendes Modell:

Sei X eine Nachrichtenquelle, die n Nachrichten $x_1, x_2, \ldots, x_n$ verschicken kann, und dabei die Nachricht x_i mit der Wahrscheinlichkeit p_i verschickt. Sei insbesondere $\sum_{i=1}^{n} p_i = 1$ und für $i = 1, \ldots, n$ auch $p_i > 0$. Dazu kodiert sie die Nachrichten in Bitfolgen, die aneinandergereiht werden.

Ein Empfänger sollte nun bei Beginn einer Nachricht empfangsbereit sein, am Ende einer gültigen Nachricht aber empfangsbereit für die nächste Nachricht sein.

Die Bitfolge (der Code) zu einer Nachricht darf also nicht Anfangsstück des Codes einer anderen Nachricht sein, da sonst die andere Nachricht nicht empfangen werden kann.

Seien also $l_1, l_2, \cdots, l_n$ die Längen der Codes für die Nachrichten $x_1, x_2, \cdots, x_n$. Wir wollen diese Längen so bestimmen, daß die erwartete (durchschnittliche) Codelänge einer Nachricht von X minimal wird.

Für die erwartete Codelänge ergibt sich:

$$E(l) = \sum_{i=1}^{n} p_i l_i$$

Um die l_i sinnvoll so wählen zu können, daß E minimal wird, beweisen wir zunächst folgendes Lemma:

Lemma 1.3.a)

1. *Sei M eine Menge endlicher Bitfolgen derart, daß keine eine andere als Anfangsstück enthält. Die Länge einer Bitfolge m sei mit $l(m)$ bezeichnet. Dann gilt:*

$$\sum_{m \in M} 2^{-l(m)} \leq 1$$

Dabei gilt das Gleichheitszeichen genau dann, wenn von jeder unendlichen Bitfolge ein (und damit genau ein) Anfangsstück zu M gehört.

2. *Umgekehrt gibt es für jede monoton steigende Folge $L = (l_1, \dots l_n)$ natürlicher Zahlen mit*
$$\sum_{i=1}^{n} 2^{-l_i} \leq 1$$
eine Menge $M = \{m_1, \dots, m_n\}$ von Bitfolgen mit Längen $l(m_i) = l_i$ derart, daß kein Element von M ein anderes als Anfangsstück bestizt.

Beweis: Wir beweisen 1. mit Induktion nach der Länge $l(M)$ der längsten Bitfolge aus M. Falls $l(M) = 0$, so ist entweder $M = \emptyset$, die gewünschte Ungleichung gilt mit '$<$' und keine unendliche Folge hat ein Anfangsstück in M, oder M enthält genau die Bitfolge der Länge 0, die gewünschte Ungleichung gilt mit '$=$' und jede Folge enthält diese Bitfolge als Anfangsstück. Damit ist der Induktionsanfang gezeigt.

Im Schritt gehen wir davon aus, daß $l(M) \geq 1$. Damit kann M insbesondere die leere Bitfolge nicht enthalten.

Sei nun M_0 die Menge aller Elemente aus M, die mit 0 beginnen und M_1 die Menge aller Elemente aus M, die mit 1 beginnen. Weiterhin sei $M_0' = \{m' | 0m' \in M_0\}$ und $M_1' = \{m' | 1m' \in M_1\}$.

Dann gilt $l(M_0') = l(M_0) - 1 \leq l(M) - 1$ sowie $l(M_1') = l(M_1) - 1 \leq l(M) - 1$.

Weiterhin ist kein Element aus M_0' in Anfangsstück eines anderen Elements aus M_0', da ansonsten durch voranstellen einer 0 bei beiden Elementen zwei Elemente aus M entstehen, von denen eines das andere als Anfang enthält. Gleiches gilt für M_1'. Also ist die Induktionsvoraussetzung auf M_0' und M_1' anwendbar.

Wir erhalten:
$$\begin{aligned}
\sum_{m \in M} 2^{-l(m)} &= \sum_{0m' \in M_0} 2^{-l(0m')} + \sum_{1m' \in M_1} 2^{-l(1m')} \\
&= \frac{1}{2} \left(\sum_{m' \in M_0'} 2^{-l(m')} + \sum_{m' \in M_1'} 2^{-l(m')} \right) \\
&\leq \frac{1}{2}(1+1) = 1
\end{aligned}$$

Nun gilt das Gleichheitszeichen genau dann, wenn M_0' und M_1' jeweils zu jeder unendlichen Bitfolge ein Anfangsstück enthält. Dann enthält aber auch M zu jeder unendlichen Bitfolge ein Anfangsstück. Enthält umgekehrt beispielsweise M_0 zu einer unendlichen Bitfolge b kein Anfangsstück, so enthält M offenbar zu $0b$ kein Anfangsstück.

Also gilt das Gleichheitszeichen tatsächlich genau dann, wenn M zu jeder unendlichen Bitfolge ein Anfangsstück enthält. Damit ist 1. gezeigt.

Wir zeigen nun 2. mittels Induktion nach n. Für $n = 0$ ist die Aussage mit $M = \emptyset$ erfüllt.

Im Induktionsschritt ($n \geq 1$) folgt:

$$\sum_{i=1}^{n-1} 2^{-l_i} = 2^{-l_n} + \sum_{i=1}^{n} 2^{-l_i} \leq 1 - 2^{-l_n} < 1$$

Damit ist die Induktionsvoraussetzung auf die Folge $l_1, l_2, \ldots, l_{n-1}$ anwendbar und liefert eine Menge M'. Wegen $\sum\limits_{i=1}^{n-1} 2^{-l_i} < 1$ und dem schon bewiesenen Teil 1. von Lemma 1.3.a) gibt es eine unendliche Folge b, welche kein Anfangsstück in M' hat. Sei m das Anfangsstück der Länge l_n von b. Da $l_n \geq l_i$ für $i = 1, \ldots, n$ gilt, ist aber l_n auch kein Anfangsstück einer Folge aus M'. Damit gilt die zu beweisende Aussage mit $M = M' \cup \{m\}$. $\square$

Wir definieren nun:

$$H(X) = \min\{\sum_{i=1}^{n} p_i l_i | \sum_{i=1}^{n} 2^{-l_i} \leq 1\}$$

Gilt bei der Nebenbedingung '$<$', so können wir offenbar ein l_i geeignet vergrößern, ohne den Wert der zu minimierenden Summe zu verkleinern. Also haben wir sogar

$$H(X) = \min\{\sum_{i=1}^{n} p_i l_i | \sum_{i=1}^{n} 2^{-l_i} = 1\}$$

Dieses Minimierungsproblem lösen wir mit der Multiplikatorenmethode von Lagrange:

$$\begin{aligned}
L(l_1, l_2, \ldots, l_n, \lambda) &= \sum_{i=1}^{n} p_i l_i + \lambda \left(-1 + \sum_{i=1}^{n} 2^{-l_i} \right) \\
\delta_{l_i} L(l_1, l_2, \ldots, l_n, \lambda) &= p_i - \lambda 2^{-l_i} \ln(2) \\
\delta_{\lambda} L(l_1, l_2, \ldots, l_n, \lambda) &= -1 + \sum_{i=1}^{n} 2^{-l_i} \\
\lambda \ln(2) &= p_i 2^{l_i} \\
l_i &= c \cdot \mathrm{ld}(p_i) \ \text{ mit } \sum_{i=1}^{n} 2^{-l_i} = \sum_{i=1}^{n} p^{-c} = 1 \\
\text{also } c = -1 \quad &\text{und} \quad l_i = -\mathrm{ld}(p_i) = \mathrm{ld}(\frac{1}{p_i})
\end{aligned}$$

Damit erhalten wir $l_i = -\mathrm{ld}(p_i)$ als den *Informationsgehalt* einer Nachricht, die mit Wahrscheinlichkeit p_i ausgesand wird, und

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

als mittleren Informationsgehalt einer Nachricht der Quelle X. Insbesondere ist $H(X)$ untere Schranke für die mittlere Zahl der Bits einer Nachricht von X bei irgendeiner Kodierung.

Desweiteren können wir die Nachrichten von X gemäß Lemma 1.3.a)2. so kodieren, dass $l(x_i) = \lceil l_i \rceil < l_i + 1$ gilt. Damit gibt es eine zulässige Codierung der Nachrichten der Nachrichtenquelle X derart, daß die erwartete Länge des Codes einer Nachricht von X höchstens um 1 von $H(X)$ abweicht.

Wollen wir nun zwei unabhängige Nachrichten aus den Quellen X und Y zu einer Nachricht aus der Quelle $X \times Y$ zusammenfassen, so erhalten wir

$$\begin{aligned}
H(X \times Y) &= -\sum_{x \in X} \sum_{y \in Y} p(x)p(y) \mathrm{ld} p(x)p(y) \\
&= -\sum_{x \in X} p(x) \sum_{y \in Y} p(y)(\mathrm{ld}(p(x)) + \mathrm{ld}(p(y)) \\
&= -\sum_{x \in X} p(x) \sum_{y \in Y} p(y) \mathrm{ld}(p(x) - \sum_{x \in X} p(x) \sum_{y \in Y} p(y) \mathrm{ld}(p(y)) \\
&= -\sum_{x \in X} p(x) \mathrm{ld}(p(x) \sum_{y \in Y} p(y) \\
&= -\sum_{x \in X} p(x) \mathrm{ld}(p(x) + H(y) \sum_{x \in X} p(x) \\
&= H(x) + H(y)
\end{aligned}$$

Das ist wieder eine untere Schranke für den Informationsgehalt der so zusammengesetzten Nachricht.

Setzt die Quelle X nun m unabhängige Nachrichten $x_1, x_2, \ldots x_m$ im Block ab, so kann man einen solchen Block auch als eine Nachricht der Quelle X^m auffassen. Es gibt also eine Codierung für solche Blöcke mit einem Erwartungswert für die Codelänge eines zufälligen Blockes von weniger als

$$H(X^m) + 1 = m\left(H(X) + \frac{1}{m}\right)$$

Bit.

Hat X also viele Nachrichten zu versenden, so kann X durch Kombinieren der Nachrichten zu Blöcken tatsächlich mit der durchschnittlichen Länge einer Nachricht beliebig dicht an $H(X)$ herankommen.

Daher bezeichnen wir $H(X)$ als *Entropie* von X oder auch als durchschnittlichen Informationsgehalt einer Nachricht aus X.

Möchte man je zwei Nachrichten zweier abhängiger Nachrichtenquellen X und Y gemeinsam versenden, so ergibt sich der durchschnittliche Informationsgehalt der Kompositnachricht durch

$$\begin{aligned}
H(Y \times X) &= -\sum_{y\in Y}\sum_{x\in X} p(x \wedge y)\mathrm{ld}(p(x \wedge y)) \\
&= -\sum_{y\in Y} p(y) \sum_{x\in X} p(x|y)(\mathrm{ld}(p(y)) + \mathrm{ld}(p(x|y))) \\
&= -\sum_{y\in Y} p(y)\mathrm{ld}(p(y)) \sum_{x\in X} p(x|y) - \sum_{y\in Y} p(y) \sum_{x\in X} p(x|y)\mathrm{ld}(p(x|y))) \\
&= H(Y) - \sum_{y\in Y} p(y) \sum_{x\in X} p(x|y)\mathrm{ld}(p(x|y))
\end{aligned}$$

Es ergibt sich, das man, um die Kompositnachricht zu kennen, wenn man die von Y gesendete Nachricht schon kennt, noch $-\sum_{y\in Y} p(y) \sum_{x\in X} p(x|y)\mathrm{ld}(p(x|y))$ zusätzliche Bits an Information benötigt. Daher wird mit

$$H(X|Y) = -\sum_{y\in Y} p(y) \sum_{x\in X} p(x|y)\mathrm{ld}(p(x|y))$$

die Äquivokation von X unter Kenntnis von Y bezeichnet, also der zusätzliche durchschnittliche Informationsgehalt einer Nachricht von X, wenn man die Nachricht aus Y schon kennt (auch Unsicherheit von X bei Kenntnis von Y).

Die gegebene Nachricht kann dabei der Chiffretext, die gesuchte Nachricht der Klartext sein (oder der Schlüssel).

Sei nun L eine Sprache (beispielsweise Deutsch) mit k Buchstaben (beispielsweise 26), X sei eine Quelle der Nachrichten, die n Buchstaben lang sind. Dann bezeichnet

- $r = \frac{H(X)}{n}$ die Rate, bzw. Dichte von L für Nachrichten der Länge n (mittlerer Informationsgehalt eines Buchstabens)
- $R = \mathrm{ld}k$ die absolute Rate oder absolute Dichte von L (Informationsgehalt eines Buchstabens, wenn die Buchstaben gleichwahrscheinlich und unabhängig vorkommen)
- $D = R - r$ die Redundanz von L.

Redundanz spiegelt statistische Eigenschaften von Sprache wieder, Häufigkeitsverteilungen von Buchstaben, Digramme, Trigramme: Im Deutschen sind

1. E, N, I die häufigsten Einzelbuchstaben (Englisch: E, T, A),
2. ER, EN, CH die häufigsten Digramme (Englisch: TH und EN),
3. EIN, ICH, DER die häufigsten Trigramme (Englisch THE und ING).

r gibt die durchschnittliche Anzahl Bits an Information pro Buchstabe wieder, das sind für große n in Deutsch bzw. Englisch jeweils 1 bis 1.5 Bit pro Buchstabe. Beim 26 Buchstaben-Alphabet erhalten wir $R = \mathrm{ld}26 \approx 4.7$ Bits pro Buchstabe als absolute Rate. Damit liegt die Redundanz ungefähr bei 3.2 bis 3.7, also (mit $R = 100\%$) bei etwa 68%.