CODIERUNGSTHEORIE

KURS ZELL AN DER PRAM, FEBRUAR 2005

1. Das Problem

1.1. **Quellcodierung und Datenkompression.** Wir wollen eine Nachricht über einen digitalen Kanal, der nur 0 oder 1 übertragen kann, schicken. Die Nachricht ist eine Folge aus den Zeichen `A`, `B`, `C`, `D`, `E`. Eine typische Nachricht wäre etwa

```
AABAAAAAAAAACAAAAAAAABAAAAAAAAAEAAAAAAAAAAAEAAEAAEAAAAA
AAAEAACAABBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
EAAAAAAAAAAAAAAEAAAAAABAAAAABAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAEAAEAAAAAAAAAAAAAAAAABAAAAAAAAAAAAAAAAAAAB
EAAAAAAAAAAAAAAAAAAAAAAAAAAAAACAAAACAAAAAAAAAABAAAAAAB
AAAAAAAAAAAAAAAAAAAACABAAAABAAAAEAAAAAAAAAAAAABAAAAAAAA
CAAAABAAAAAAEAAAABAAAAAAAAAAAACAAAAAAAAAAAAAAAABAAAAAAA
BAAAAAACAAABAAAAAAAAAAAAAAAAAAAABAAAAAAAAAAAAAAAAAAAB
AAAAAAAAABAAAAAABAAAAAAAAAAAAAAAAAAAAAAAAAAAAAABAAAABB
AAAAA
```

Dabei wissen wir, dass an jeder Stelle der Nachricht mit Wahrscheinlichkeit 0.9 ein `A`, mit 0.06 ein `B`, mit 0.015 ein `C`, mit 0.015 ein `D` und mit Wahrscheinlichkeit 0.01 ein `E` vorkommt.

Ziel ist es, für die Übertragung diese Nachricht als Folge von 0 und 1 zu kodieren. Dabei sollen wir für jedes Zeichen im Durchschnitt nur höchstens 0.8 Bits benötigen – eine Folge von 100 Zeichen sollte im Durchschnitt also auf 80 Bits komprimiert werden können.

Ist es möglich, eine solche Nachricht so stark zu komprimieren? Versuchen Sie, ein Verfahren zur Kompression zu finden!

Date: 11. Februar 2005.

Erhard Aichinger und Peter Mayr, Institut für Algebra, Johannes Kepler Universität Linz, `erhard@algebra.uni-linz.ac.at`, `peter.mayr@algebra.uni-linz.ac.at`, `http://www.algebra.uni-linz.ac.at/`.

2. Einige Begriffe der Informationstheorie

Ein Kommunikationssystem besteht aus folgenden Teilen (s. [HQ95, Ash90]).

(1) Nachrichtenquelle
(2) Quellencodierer
(3) Kanalcodierer
(4) Kanal
(5) Rausch- oder Fehlerquelle
(6) Kanaldecodierer
(7) Quellendecodierer
(8) Nachrichtensenke

Die Begriffe “Verschlüsselung” und “Codierung”.

3. Wörter und eindeutig decodierbare Codes

Definition 3.1. Wörter über $\{0,1\}$, Zusammenhängen von Wörtern, $\{0,1\}^*$.

Definition 3.2. Sei $k \in \mathbb{N}$, und seien $p_1, p_2, \ldots, p_k \in [0,1]$ so, dass $\sum_{i=1}^{k} p_i = 1$, seien $A_1, \ldots, A_k \in \{0,1\}^*$, und sei $\mathcal{C} = (A_1, A_2, \ldots, A_k)$. Die *durchschnittliche Wortlänge* $\overline{n}$ *von* $\mathcal{C}$ bezüglich $(p_1, \ldots, p_k)$ ist gegeben durch

$$\overline{n} = p_1|A_1| + p_2|A_2| + \cdots + p_k|A_k|.$$

Definition 3.3. Sei $k \in \mathbb{N}$. Ein *Code aus* k *Codewörtern* ist eine Folge $(A_1, \ldots, A_k)$ von Wörtern in $\{0,1\}^*$, sodass für alle $i, j \in \{1, \ldots, k\}$ mit $i \neq j$ auch $A_i \neq A_j$ gilt.

Definition 3.4. Sei $k \in \mathbb{N}$, und sei $\mathcal{C} = (A_1, \ldots, A_k)$ ein Code. $\mathcal{C}$ ist *eindeutig decodierbar*, wenn für alle $m, n \in \mathbb{N}$ und für alle $D_1, \ldots, D_m$ aus $\mathcal{C}$ und $E_1, \ldots, E_n$ aus $\mathcal{C}$ folgendes gilt: Wenn

$$D_1 * D_2 * \cdots * D_m = E_1 * E_2 * \cdots * E_n,$$

dann gilt $m = n$, und für alle $i \in \{1, \ldots, m\}$ gilt $D_i = E_i$.

Wie findet man eindeutig decodierbare Codes?

Übungsaufgaben 3.5.

(1) [HQ95, S.36] Zeigen Sie, dass der Code $\{0, 10, 011, 11111\}$ eindeutig decodierbar ist.
(2) Ist der Code $\{a, c, ad, abb, bad, deb, bbcde\}$ eindeutig decodierbar?
(3) Seien $V, W \in \{0,1\}^*$, also Wörter über dem Alphabet $\{0,1\}$. Wie müssen V und W aussehen, damit $V * W = W * V$ gilt? (Dabei ist $V * W$ das Wort, das durch Zusammenhängen von V und W entsteht.

4. Präfixcodes

Definition 4.1. Seien A, B Wörter über $\{0,1\}$. A ist ein *Präfix* von B, wenn $A = B$, oder wenn es ein Wort C gibt, sodass $A * C = B$.

Definition 4.2. Seien $A_1, A_2, \ldots, A_k$ Wörter über $\{0,1\}$. $(A_1, A_2, \ldots, A_k)$ ist ein *Präfix-Code*, wenn es keine $i, j \in \{1, 2, \ldots, k\}$ gibt, sodass $i \neq j$ und A_i ein Präfix von A_j ist.

Satz 4.3. *Präfixcodes sind eindeutig decodierbar.*

Satz 4.4. *Sei $k \in \mathbb{N}$, und seien $n_1, n_2, \ldots, n_k \in \mathbb{N}$. Dann gibt es genau dann einen Präfixcode $(A_1, A_2, \ldots, A_k)$ über $\{0,1\}$, sodass $|A_1| = n_1$, …, $|A_2| = n_2$, $|A_k| = n_k$, wenn*

$$\frac{1}{2^{n_1}} + \frac{1}{2^{n_2}} + \cdots + \frac{1}{2^{n_k}} \leq 1.$$

Es gilt sogar für jeden eindeutig decodierbaren Code (nicht nur für jeden Präfixcode), dass

$$\frac{1}{2^{n_1}} + \frac{1}{2^{n_2}} + \cdots + \frac{1}{2^{n_k}} \leq 1.$$

Das ist aber schwieriger zu beweisen (cf. [Ash90, S. 35]).

Übungsaufgaben 4.5.

(1) Zeigen Sie: Präfix-Codes sind eindeutig decodierbar.
(2) Sei $k \in \mathbb{N}$, und seien $n_1, n_2, \ldots, n_k \in \mathbb{N}$. Zeigen Sie: Es gibt genau dann einen Präfixcode $(A_1, A_2, \ldots, A_k)$ über $\{0,1\}$, sodass $|A_1| = n_1, \ldots, |A_2| = n_2, |A_k| = n_k$, wenn

$$\frac{1}{2^{n_1}} + \frac{1}{2^{n_2}} + \cdots + \frac{1}{2^{n_k}} \leq 1.$$

5. Entropie

Definition 5.1. (Entropie) $H(p_1, \ldots, p_k) := \sum_{i=1}^{k} -p_i \log_2(p_i)$. Es gibt ein Problem, falls ein $p_i = 0$ ist.

Übungsaufgaben 5.2. Bestimmen Sie $H(0.5, 0.5)$, $H(0.3, 0.7)$, $H(0.1, 0.9)$, $H(.5, .25, .25)$, $H(0.6, 0.3, 0.05, 0.05)$.

Satz 5.3. *(Noiseless coding theorem) Sei $k \in \mathbb{N}$, und seien $p_1, p_2, \ldots, p_k \in (0, 1]$ so, dass $\sum_{i=1}^{k} p_i = 1$, sei $\mathcal{C} = (A_1, A_2, \ldots, A_k)$ ein Präfixcode, und sei $\overline{n}$ die durchschnittliche Wortlänge von $\mathcal{C}$ bezüglich $(p_1, \ldots, p_k)$. Dann gilt*

$$\overline{n} \geq H(p_1, \ldots, p_k).$$

Satz 5.4. *Sei $k \in \mathbb{N}$, seien $p_1, p_2, \ldots, p_k \in (0, 1]$ so, dass $\sum_{i=1}^{k} p_i = 1$, und sei $H := H(p_1, \ldots, p_k)$. Dann gibt es für diese Wahrscheinlichkeiten einen binären Präfixcode $(A_1, \ldots, A_k)$, für dessen durchschnittliche Wortlänge $\overline{n}$ bezüglich $(p_1, \ldots, p_k)$ gilt:*

$$H + 1 > \overline{n} \geq H.$$

Übungsaufgaben 5.5. In einer aus den drei Zeichen zusammengesetzten Nachricht kommen die Zeichen x_1, x_2, x_3 mit den Wahrscheinlichkeiten $0.9, 0.05, 0.05$ vor. Wir wollen diese Zeichen durch $0, 1$-Folgen übertragen.

(1) Man kann beweisen, dass wir das mit einem Code machen können, dessen durchschnittliche Codewortlänge im Intervall

$$[H(0.9, 0.05, 0.05), H(0.9, 0.05, 0.05) + 1]$$

liegt. Finden Sie einen solchen Code!

(2) Wir fassen jetzt immer zwei Nachrichtenzeichen zusammen und übertragen diese Ausgänge gemeinsam. Wieviele Zeichen brauchen Sie jetzt im Durchschnitt für die Übertragung eines Nachrichtenzeichens?

6. Konstruktion optimaler Codes

Sei $k \in \mathbb{N}$, und seien $p_1, \ldots, p_k$ Zahlen in $(0, 1]$ mit $\sum_{i=1}^{k} p_i = 1$. Ein Präfix Code $\mathcal{C} = (C_1, C_2, \ldots, C_k)$ ist *optimal* für $(p_1, \ldots, p_k)$, wenn es keinen Präfixcode mit kleinerer durchschnittlicher Codewortlänge gibt.

Satz 6.1. *Sei $k \in \mathbb{N}$, seien $p_1, \ldots, p_k \in (0, 1]$ so, dass $\sum_{i=1}^{k} p_i = 1$ und $p_1 \geq p_2 \geq p_3 \geq \cdots \geq p_{k-1} \geq p_k > 0$, und sei $\mathcal{D} = (D_1, D_2, \ldots, D_{k-1})$ ein optimaler Code für die Wahrscheinlichkeiten $(p_1, p_2, \ldots, p_{k-2}, p_{k-1} + p_k)$. Sei $\mathcal{C} = (C_1, C_2, \ldots, C_{k-1}, C_k)$ gegeben durch $C_i = D_i$ für $i \in \{1, 2, \ldots, k-2\}$, $C_{k-1} = D_{k-1} * 0$, $C_k = D_{k-1} * 1$. Dann ist $\mathcal{C}$ ein optimaler Code für die Wahrscheinlichkeiten $(p_1, \ldots, p_k)$.*

Dieser Satz liefert den Huffman-Algorithmus.

Übungsaufgaben 6.2. Konstruieren Sie – etwa mit dem Verfahren von Huffman – optimale binäre Codes für folgende Wahrscheinlichkeiten, und berechnen Sie die durchschnittliche Codewortlänge.

(1) $(0.8, 0.1, 0.06, 0.02, 0.02)$.
(2) $(0.2, 0.2, 0.2, 0.2, 0.2)$.
(3) $(.2, .18, .1, .1, .1, .061, .059, .04, .04, .04, .04, .03, .01)$.

Buchstabenhäufigkeit (in Promille) in deutschen Texten ([Pom04]):

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
60	17	27	54	180	16	32	41	81	3	13	33	23	106	27	8	0	72	69	57	46	9	15	0	0	11

Literatur

[Ash90] R. B. Ash, *Information theory*, Dover Publications Inc., New York, 1990, Corrected reprint of the 1965 original.

[HQ95] W. Heise and P. Quattrocchi, *Informations- und Codierungstheorie*, third ed., Springer-Verlag, Berlin, 1995.

[Pom04] K. Pommerening, *Kryptologie (Vorlesungsunterlagen)*, Published electronically at `http://www.uni-mainz.de/~pommeren/Kryptologie/`, 2004.