JKU Young Scientists Matheseminar

Matheseminar WS 2013/14

Codierung
und
Information

„Das grundlegende Problem der Kommunikation besteht darin, an einer Stelle entweder genau oder angenähert eine Nachricht wiederzugeben, die an einer anderen Stelle ausgewählt wurde.“

(Claude E. Shannon, Begründer der Informationstheorie)

Codierungstheorie

Ziel der Codierungstheorie ist es, Methoden anzugeben, wie man Daten **schnell** über einen **fehleranfälligen** Kanal transportieren kann. Es gibt zwei Aspekte:

- **Quellencodierung** und *Datenkompression*: Wie kann man eine Nachricht möglichst kurz darstellen?
- **Kanalcodierung** und *Fehlerkorrektur*: Wie kann man beim Senden über einen unsicheren Kanal trotzdem sichere Übertragung erreichen?

Kryptologie[1]

Kryptologie umfasst sowohl die **Kryptographie** als auch die **Kryptoanalyse**.

- **Kryptographie** ist die Wissenschaft von der Ver- und Entschlüsselung von Daten mit Hilfe mathematischer Verfahren.
- Sie befasst sich allgemein mit dem Thema **Informationssicherheit**, also der Konzeption, Definition und Konstruktion von Informationssystemen, die widerstandsfähig gegen unbefugtes Lesen und Verändern sind.
- Die **Kryptoanalyse** die Wissenschaft von der Analyse und vom Entschlüsseln verschlüsselter Daten.

Kanalcodierung - Problem

Wir sollen die erhaltene 0/1-Folge über einen Kanal übertragen.Durch Rauschen wird im Mittel jedes zehnte Bit verfälscht (d.h. nur 90% der Nachricht wird korrekt übertragen). Wir fassen die Nachricht zu Blöcken aus 5 Ziffern zusammen. Wie kann man so übertragen, dass im Durchschnitt zumindest 94% der Blöcke richtig ankommen?

- Projektwoche Angewandte Mathematik 2014

Quellencodierung - Problem

Nachricht:

BAAAABCACAAADAAAADAABCAAA
AABDBAAABABBAAAAAAAABAAAB
AAAAACAABABADCBABAABABBAA
ABAAADAAAAAAABAAABAAAAAB

- Wie kann man eine solche Nachricht so als 0/1-Folge darstellen, dass im Durchschnitt pro übertragenem Zeichen möglichst wenig Bits benötigt werden?
- Wie viele Bits/Zeichen benötigt man mindestens?
- Wie findet man ein Verfahren zur Datenkompression?

Nachrichtentechnik

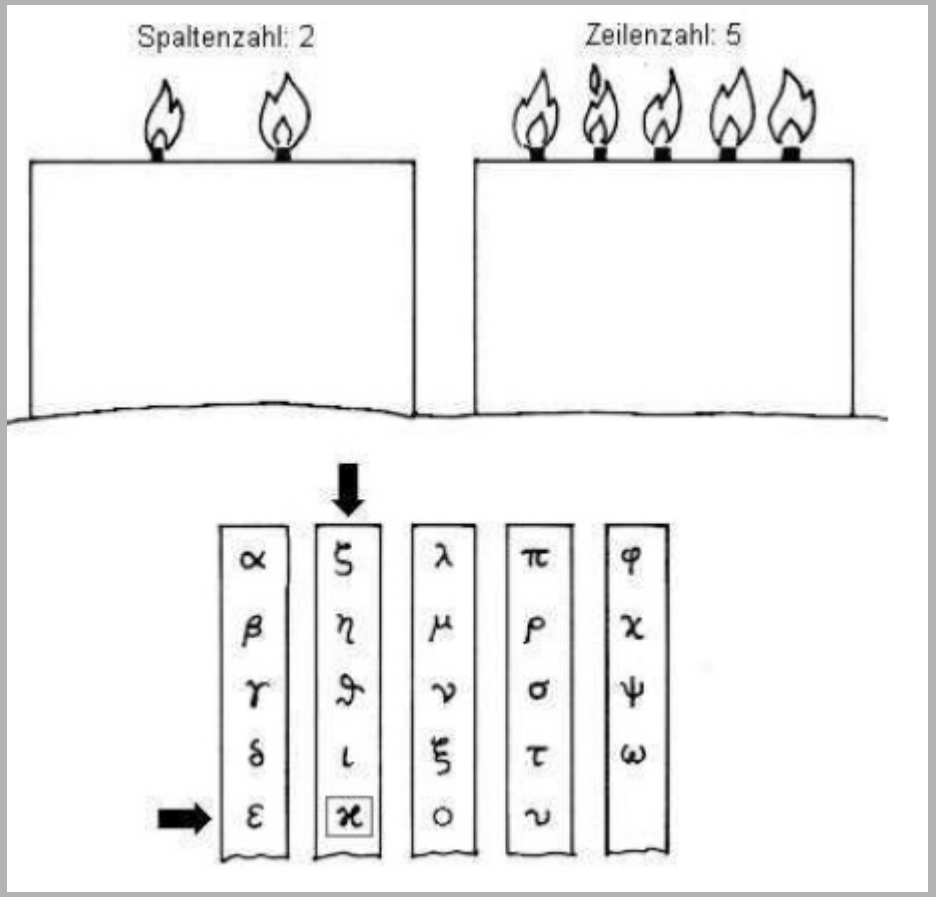
Sie befasst sich mit der

- Darstellung und Übertragung
- der Vermittlung und
- der Verarbeitung

von Nachrichten.

Die Anfänge der Nachrichtentechnik reichen weit in das Altertum zurück. Die Grundlage der Nachrichtentechnik wird durch die Erfindung der Schrift und der Zahlenzeichen ab etwa 4000 v. Chr. gelegt.

Nachrichtentechnik

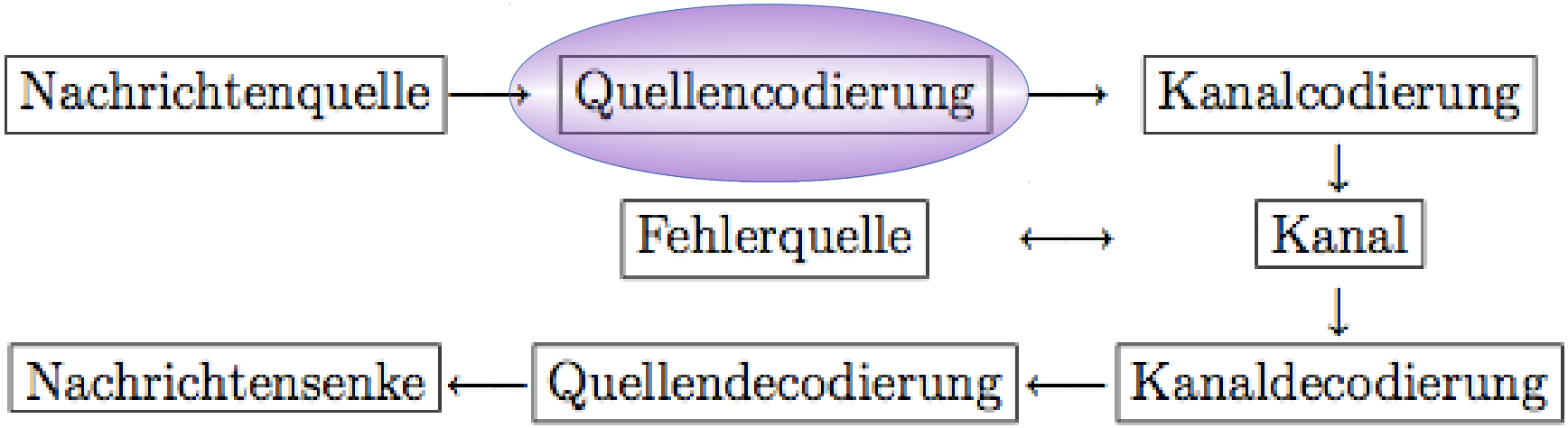
Um 180 v. Chr.	Fackeltelegraph von Polybios (optische Telegrafie) 
Anfang des 19. Jhdt.	Höhepunkt der optischen Telegrafie 1834 Strecke Berlin – Koblenz (ca. 600 km)

Nachrichtentechnik

Um 1850	Elektrische Telegrafie Nachrichtenübertragung bleibt **digital**.
Ende des 19. Jhdt.	Nachrichtentechnik wird **analog**. 1876 (Alexander G. Bell): Telefon
Anfang des 20. Jhdt.	Statistische Methoden und Vorstellungen der Wahrscheinlichkeitsrechnung werden in der Nachrichtentechnik aufgegriffen.
Mitte des 20. Jhdt.	Übergang von der analogen zur digitalen Technik 1947: Erfindung des Transistors 1970: erster Mikroprozessor auf dem Markt
Ende des 20. Jhdt.	Die in der ersten Hälfte des 20. Jhdt. gefundenen theoretischen Ansätze können aufgrund des technischen Fortschritts in bezahlbare Geräte umgesetzt werden. 1991: WWW wird erstmals eingesetzt 1993: digitaler Mobilfunk (GSM-Netz) in Österreich 1994: digitales Fernsehen in den USA

Nachrichtenübertragungssysteme

Wichtige Komponenten:



Aufgabenstellung 1

Nachrichtenquellen mit Nachrichtenvorrat N = {A, B, C, D}

Beispiel für typische Nachricht:

BAAAABCACAAADAAAADAABCAAA
AABDBAAABABBAAAAAAAABAAAB
AAAAACAABABADCBABAABABBAA
ABAAADAAAAAAABAAABAAAAAB

Übertragungswahrscheinlichkeiten:

A … 0,7 B … 0,2 C … 0,05 D … 0,05

- Findet eine Codierung der Zeichen A, B, C und D mit Hilfe von 0 und 1.
- Teilt euch in jeder Gruppe in Zweierpaare auf. Erstellt jeweils eine Nachricht und codiert diese mit dem in a) erstellen Code. Diese Nachricht soll jeweils die andere Zweiergruppe decodieren.
- Versucht, die durchschnittliche Codewortlänge für eure Codierung zu berechnen. Geht es eventuell auch noch „besser“, d. h. findet ihr einen Code der mit weniger Zeichen auskommt?

Ergebnisse:

Ausgangssituation:

- Eine Nachrichtenquelle, die verschiedene Nachrichten produziert, z.B. N = {A, B, C, D}.
- Ein Kanal, der in der Lage ist, Binärzeichen zu übertragen.

Dabei nehmen wir an, dass der Kanal störungsfrei ist (engl. ***noiseless channel***).

Ergebnisse:

Ziel der Quellencodierung:

Die Nachrichten sollen in eine für die Übertragung geeignete binäre Form gebracht werden. Dabei sollen folgende Forderungen beachtet werden:

- Die Codierung soll **effizient** sein!
  Forderung 1: Die Codewörter sollen so kurz sein wie möglich.
- Die Codierung soll **umkehrbar** sein!
  Forderung 2: Aus der Folge der übertragenen Zeichen müssen die codierten Nachrichten eindeutig rekonstruiert werden können.

Aufgabenstellung 2

1. $B^3 = \{000, 001, 010, 011, 100, 101, 110, 111\}$

2. a) $3^4 = 81$

b) (1) 2^n b) (2) q^n

3. $2^1 + 2^2 + 2^3 + 2^4 = 30$

4. a) $0{,}2^5 \cdot 0{,}15^2 \cdot 0{,}05^2 \cdot 0{,}05 \cdot 0{,}2 = 1{,}8 \cdot 10^{-10}$

4. b) $0{,}2 \cdot 0{,}15 \cdot 0{,}2 + (0{,}15 \cdot 0{,}2)^2 = 0{,}0069$

Zusatz:
Anfang In 15 Tagen beginnen die Ferien. Ende

Eindeutig decodierbare Codes

- Zur Quellencodierung werden vor allem **Codes variabler Länge** verwendet.
- Forderung: Die Codierung soll **umkehrbar** sein!
- Decodiere folgende, mit dem Morsecode codierte Nachricht (Tabelle S. 4): • • – • • • • –
- Lösungen:
  elu, usa, idea, eeba, feea, fia, …
- Wie lautete die Nachricht?

Eindeutig decodierbare Codes

- Problem bei Codes variabler Länge:
  Im Allgemeinen ist eine eindeutige Dekodierung nicht möglich.
- Ein Code heißt **eindeutig decodierbar**, wenn jede endliche Folge von Binärzeichen (Bits) höchstens einer Nachrichtensequenz entspricht.

Eindeutig decodierbare Codes

Eine Teilmenge *C* von {0,1}* (d.h. eine Menge von Wörtern über dem Binäralphabet) ist ein eindeutig decodierbarer Code, falls für alle Codewörter

$A_1, \ldots, A_n \in C$ und $B_1, \ldots, B_m \in C$ gilt:

wenn $A_1 * A_2 * \ldots * A_n = B_1 * B_2 * \ldots * B_m$

dann gilt, dass $n = m$

und für alle $i \in 1, 2, \ldots, m$ gilt,

dass $A_i = B_i$

Aufgabenstellung 3

Partnerarbeit:

Sind die folgenden Codes eindeutig decodierbar?

Beispiel 3.1. C_1 = {a, c, ad, abb, bad, deb, bbcde}

Beispiel 3.2.
a) C_{2a} = {0, 010} b) C_{2b} ={110, 1110, 1011, 1101}

Beispiel 3.3. C_3 = {0, 01, 011, 0111}

Beispiel 3.4.
C_4 = {00, 10, 010, 111, 0110, 1101, 11001}

Beispiel 3.5. C_5 = = {11, 101, 1011}

Zusatzbeispiel: C = {0, 10, 011, 11111}

Präfixcodes (präfixfreie Codes)

- Dabei handelt es sich um einen Code, bei dem kein Codewort gleichzeitig der Beginn (d. h. ein Präfix) eines anderen Codeworts darstellt. (Fano-Bedingung)
- Beispiel: C = {0, 10, 110, 111}

  0 ist kein Präfix von 10, 110, 111
  10 ist kein Präfix von 0, 110, 111
  110 ist kein Präfix von 0, 10, 111
- Präfixcodes sind **eindeutig decodierbar.**
- Alphabet A_1 = {a, b}, Alphabet A_2 = {0, 1}

  Code: f(a) = 0, f(b) = 01 … kein Präfixcode
  001000101010 = _____

Aufgabenstellung 4

Einzel-/Partnerarbeit:

Beispiel 4.1.
Ist der Code C = {1, 00, 010, 0110, 0111} ein Präfixcode?

Beispiel 4.2.
Erstelle für folgenden Nachrichtenvorrat einen Präfixcode über dem Binäralphabet.

a) N_1 = {A, B, C, D}

b) N_2 = {0, 1, 2, … , 8, 9}

Aufgabenstellung 4

Lösung 4.1. C ist ein Präfixcode

Lösung 4.2.

a) c(A) = 0; c(B) = 10; c(C) = 110; c(D) = 111
C = {0, 10, 110, 111}

b) C = {0000, 0001, 0010, …., 1000, 1001} oder

C = {0, 10, 110, 1110, 11110, 111110, 1111110, 11111110, 111111110, 111111111} oder

C = {00, 010, 0110, 0111, 100, 101, 1100, 1101, 1110, 1111}

Ungleichung von Kraft / McMillan

Satz: Sei C ein **eindeutig decodierbarer** Code über dem Binäralphabet B = {0, 1}, der aus k Wörtern c_1, c_2, …, c_k besteht, die die Längen n_1, n_2, …, n_k haben.

Dann gilt die Ungleichung von Kraft und McMillan:

$$\frac{1}{2^{n_1}} + \frac{1}{2^{n_2}} + \ldots + \frac{1}{2^{n_k}} = \sum_{i=1}^{k} \frac{1}{2^{n_i}} \leq 1$$

- Wenn ein Code die Ungleichung von Kraft und McMillan nicht erfüllt, dann ist er nicht eindeutig decodierbar.
- Ist der Code *C* = {a, c, ad, abb, bad, deb, bbcde} eindeutig decodierbar?

Ungleichung von Kraft / McMillan

Satz: Es seien $k \in \mathbf{N}$ und $n_1, n_2, \ldots, n_k \in \mathbf{N}$. Wenn die Ungleichung

$$\frac{1}{2^{n_1}} + \frac{1}{2^{n_2}} + \ldots + \frac{1}{2^{n_k}} = \sum_{i=1}^{k} \frac{1}{2^{n_i}} \leq 1$$

von Kraft und McMillan erfüllt ist, dann existiert ein Präfixcode $C = \{c_1, c_2, \ldots, c_k\}$ über dem Binäralphabet $B = \{0, 1\}$, dessen k Codewörter $c_1, c_2, \ldots, c_k$ der Reihe nach die Längen $n_1, n_2, \ldots, n_k$ haben.

Eindeutig decodierbare Codes

Jeder eindeutig decodierbare Code kann durch einen Präfixcode, der dieselbe Anzahl von Codewörtern mit denselben Codewortlängen hat, ersetzt werden.

- **Beispiel 4.4.** Gegeben sind folgende Codewortlängen: $n_1 = 2$, $n_2 = 2$, $n_3 = 3$ $n_4 = 3$, $n_5 = 4$, $n_6 = 5$. Überprüfe ob es dazu einen Präfixcode gibt und wenn ja, konstruiere einen solchen.

- **Lösung 4.4.** C = {00, 01, 100, 101, 1100, 11010}

Datenkompression

- Die Codierung soll **effizient** sein!
  Forderung: Die Codewörter sollen so kurz sein wie möglich.
- Statistische Codierung:
  Häufig zu erwartende Nachrichten werden durch kurze Codewörter, seltenere Nachrichten durch längere Codewörter beschrieben.
- Bekannte Verfahren:
  - Codierung nach Fano
  - Codierung nach Huffman
  - Arithmetische Codierung

Datenkompression

Es ist $C = \{c_1, c_2, \ldots, c_k\}$ ein Präfixcode.
Die Längen der einzelnen Codewörter sind $n_1, n_2, \ldots, n_k$ und $p_1, p_2, \ldots, p_k \in]0, 1]$ sind die Wahrscheinlichkeiten, mit denen die Codewörter ausgewählt werden.
Dabei muss gelten:

$$\sum_{i=1}^{k} p_i = 1$$

Die **durchschnittliche Wortlänge** von C ist gegeben durch

$$\bar{n} = p_1 \cdot n_1 + p_2 \cdot n_2 + \ldots + p_k \cdot n_k$$

Datenkompression

Nachricht:
AAABAACAADBBAAABAAAAABAAAAAAABAA DAAAA. . ..

70%*A*, 20%*B*, 5%*C*, 5%*D*.

Zeichen	Vorschlag 1	Vorschlag 2	Vorschlag 3
A	00	0	0
B	01	10	10
C	10	110	110
D	00	01	111
$\bar{n}$	2	---	1,4

Fragestrategien

Beispiel 5.1.

Wie viele Fragen muss Alice höchstens stellen, um herauszufinden, in welchem Monat und an welchem Tag Bob Geburtstag hat, wenn dieser auf Fragen nur mit „Ja“ oder „Nein“ antworten darf?

Beispiel 5.2.

In einer großen Schachtel befinden sich 20 Kugeln: 6 schwarze, 4 rote, 4 blaue, 3 gelbe und 3 weiße. Alice zieht eine Kugel und notiert sich deren Farbe. Bob möchte nun wissen, welche Farbe die gezogene Kugel hat; Alice antwortet auf Fragen jedoch nur mit „Ja“ oder „Nein“.

a) Erstelle eine Fragestrategie!

b) Wie viele Fragen muss Bob dann **durchschnittlich**

Beispiel 5.2. - Variante 1

1. Ist $X =$ gelb?
2. Ist $X =$ weiß?
3. Ist $X =$ rot?
4. Ist $X =$ blau?
5. Ist $X =$ schwarz?

Die Anzahl Z_1 der notwendigen Fragen bei Variante 1 ist eine Zufallsvariable, die folgende Werte annimmt:

X	Z_1 (Anzahl der Fragen)	Wahrscheinlichkeit
gelb	1	$\frac{3}{20} = 0{,}15$
weiß	2	$\frac{3}{20} = 0{,}15$
rot	3	$\frac{4}{20} = 0{,}20$
blau	4	$\frac{4}{20} = 0{,}20$
schwarz	5	$\frac{6}{20} = 0{,}30$

Die durchschnittliche Anzahl der benötigten Fragen ist

$$E(Z_1) = 1 \cdot 0{,}15 + 2 \cdot 0{,}15 + 3 \cdot 0{,}2 + 4 \cdot 0{,}2 + 5 \cdot 0{,}3 = 3{,}35$$

Beispiel 5.2. - Variante 2

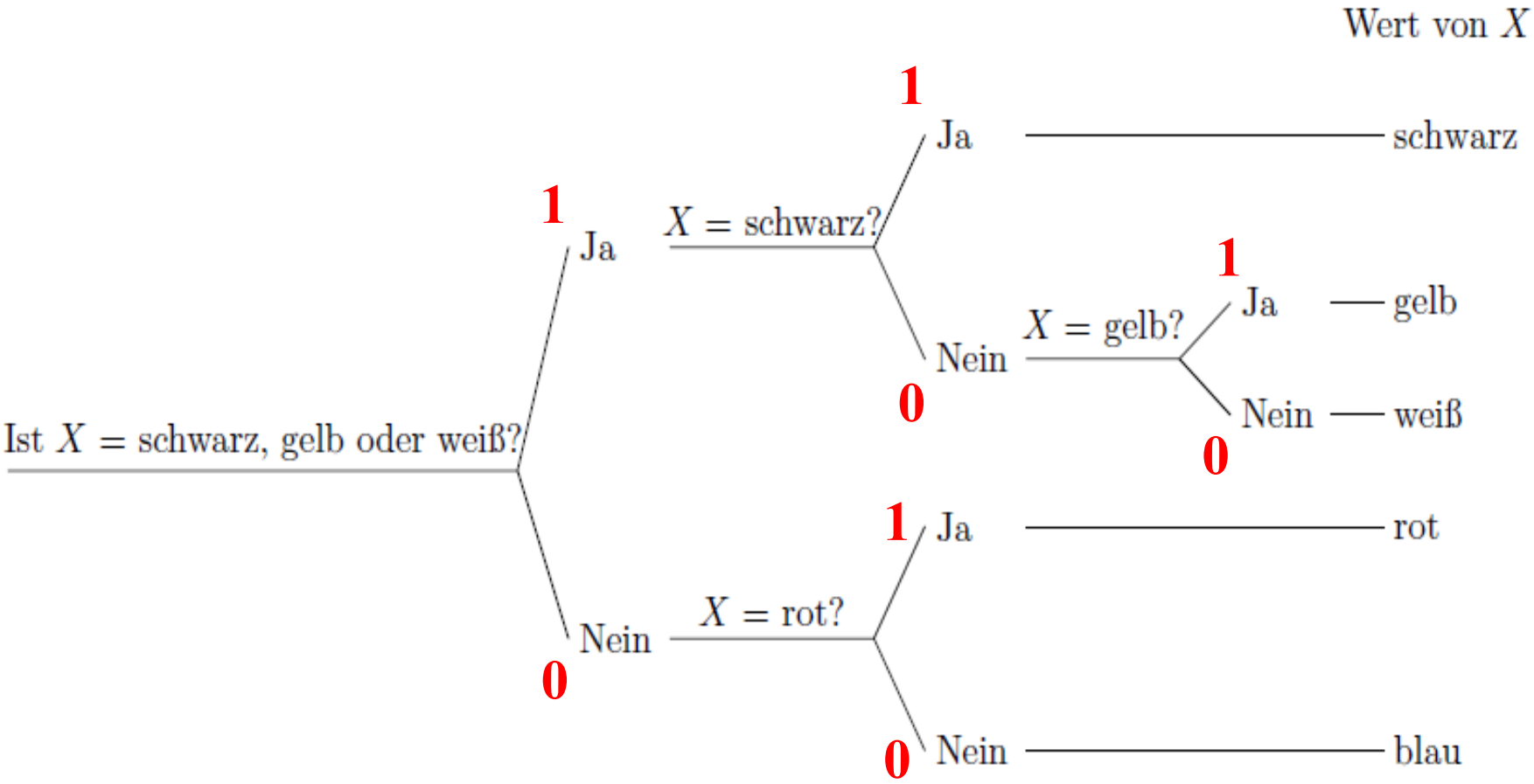
1. Ist $X =$ schwarz?
2. Ist $X =$ rot?
3. Ist $X =$ blau?
4. Ist $X =$ gelb?
5. Ist $X =$ weiß?

X	Z_2 (Anzahl der Fragen)	Wahrscheinlichkeit
schwarz	1	0,30
rot	2	0,20
blau	3	0,20
gelb	4	0,15
weiß	5	0,15

Die durchschnittliche Anzahl der benötigten Fragen ist

$$E(Z_2) = 1 \cdot 0{,}3 + 2 \cdot 0{,}2 + 3 \cdot 0{,}2 + 4 \cdot 0{,}15 + 5 \cdot 0{,}15 = 2{,}65$$

Beispiel 5.2. - Variante 3



Beispiel 5.2. - Variante 3

- Binärcodierung: jedes „JA“ durch eine „1“ und jedes „NEIN“ durch eine „0“ ersetzen
- c(S) = 11
  c(G) = 101
  c(W) = 100
  c(R) = 01
  c(B) = 00
- Durchschittliche Codewortlänge: 2,3 bit/Zeichen

Huffman-Codierung

Dieses Verfahren liefert einen **Präfixcode** mit **kleinster durchschnittlicher Codewortlänge**
(sog. optimaler Code).

1. Ordnen nach fallenden Auftrittswahrscheinlichkeiten (bzw. Häufigkeiten)
2. Zusammenfassen der beiden unwahrscheinlichsten Symbole
3. Wiederholung von Schritt 1 und 2
4. Wahrscheinlichere Nachrichtenmenge wird mit 1 und die unwahrscheinlichere mit 0 codiert.
5. Fortfahren, bis alle Nachrichten codiert sind

Huffman-Codierung

Beispiel 5.3. Konstruiere mit dem Verfahren von Huffman optimale binäre Codes für folgende Wahrscheinlichkeiten und berechne die durchschnittliche Codewortlänge.

a) (0,8; 0,1; 0,06; 0,02; 0,02)

b) (0,2; 0,2; 0,2; 0,2; 0,2)

c) (0,2; 0,18; 0,1; 0,1; 0,1; 0,061; 0,059; 0,04; 0,04; 0,04; 0,04; 0,03; 0,01)

Huffman-Codierung

Lösung 5.3.

a) C = {1, 01, 001, 0001, 0000}; n = 1,34 bit/Zeichen

b) C = {10, 111, 110, 01, 00}; n = 2,4 bit/Zeichen

c) C = {01, 111, 100, 001, 1101, 1010, 0001, 11001,
11000, 10111, 10110, 00001, 00000}
n = 3,42 bit/Zeichen

Huffman-Codierung

Der Huffman-Algorithmus liefert in seiner Grundform immer dann unbefriedigende Ergebnisse, wenn

- eine Binärquelle vorliegt, beispielsweise
  N = {A, B},
- die Wahrscheinlichkeit des häufigsten Symbols deutlich größer ist als 50%,
- es statistische Bindungen zwischen den Symbolen in der Eingangsfolge gibt

Huffman-Codierung

Beispiel 5.4.

Nachricht:

AAAAAAABCBAAAABAAAAAAABAAAAABAAAAABAAAAAC

AABAABAAAAAAABABAAAAACAAAAACBACABAAABAAAA

AAAAAAAAAABAABAAAACAAAAABAABABAAAABAABAAA

CACAAABAAABAABABAACBAACAA

Dabei wissen wir, dass die Nachrichtenquelle an jeder Stelle der Nachricht mit einer Wahrscheinlichkeit von 0,8 ein A, mit 0,15 ein B und mit 0,05 ein C ausgibt.

Dabei sollen wir für jedes Zeichen im Durchschnitt nur höchstens 0,9 Bits benötigen – eine Folge von 100 Zeichen sollte im Durchschnitt also auf 90 Bits komprimiert werden können.

- Erstelle eine Huffman-Code für diesen Nachrichtenvorrat.
- Wie kann man die Codierung verändern, so dass die durchschnittliche Codewortlänge kleiner wird?

Huffman-Codierung

Lösung 5.4.

a) C = {1, 01, 00} ... c(A) = 1; c(B) = 01; c(C) = 00; n = 1,2 bit/Nachricht

b) Man fasst jeweils zwei Symbole zu einem Paar zusammen:
AA → a, AB → b, AC → c, BA→ d, BB → e, BC → f, CA → g, CB → h, CC → i

dann lautet der zusammengefasste Nachrichtenvorrat

N' = {a, b, c, d, e, f, g, h, i} mit folgenden Wahrscheinlichkeiten:

$p_a = p_{AA} = 0{,}8^2 = 0{,}64$ $p_d = p_{BA} = 0{,}12$ $p_g = p_{CA} = 0{,}04$

$p_b = p_{AB} = 0{,}8 \cdot 0{,}15 = 0{,}12$ $p_e = p_{BB} = 0{,}0225$ $p_h = p_{CB} = 0{,}0075$

$p_c = p_{AC} = 0{,}8 \cdot 0{,}05 = 0{,}04$ $p_f = p_{BC} = 0{,}0075$ $p_i = p_{CC} = 0{,}0025$

C = {1, 011, 0011, 010, 0001, 00000, 0010, 000011, 000010}

Durchschn. Codewortlänge n = 1,8675 bit/Zweiertupel = **0,93375 bit/Zeichen**

Huffman-Codierung

Wenn die Zeichen $A_1, A_2, \ldots, A_k$ mit Wahrscheinlichkeiten $p_1, p_2, \ldots, p_k$ auftreten, so braucht jedes Quellcodierungsverfahren, das für beliebig lange Dateien funktioniert, im Mittel zumindest

$$\sum_{i=1}^{k} p_i \cdot {}^2\log\left(\frac{1}{p_i}\right) = -\sum_{i=1}^{k} p_i \cdot {}^2\log\left(p_i\right)$$

Bits pro Nachrichtenzeichen.

Durch geeignete Codierungsverfahren kann man dieser Schranke beliebig nahe kommen.

Huffman-Codierung - Anwendung

- MP3
- JPEG
- Telefax
- MPEG
- ZIP

Quellen

- [1)] http://de.wikipedia.org/wiki/Kryptographie
- http://www.Intwww.de/downloads/Informationstheorie/Aufgaben/Kapitel2/
- Quellcodierung: Edith Lindenbauer, Diplomarbeit. Informationstheorie und Quellencodierung im Schulunterricht – Didaktische Aufbereitung der *mathematischen Grundlagen*, 2005.