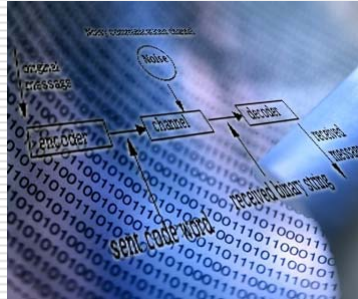


Kapitel 6: Codierung Diskreter Quellen



Ziele des Kapitels

ETH

- Die Entropie als Informationsmass für die Güte eines Codes
- Begriff der Datenkompression
- Eindeutig decodierbare Codes
- Mittlere Codelänge kann nicht kleiner als Quellenentropie sein
- Kraft'sche Ungleichung
- 1. Shannon'sches Codierungstheorem

Codierung
Diskreter Quellen

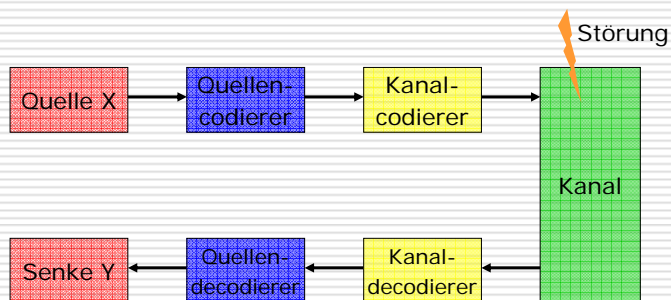
Informationstheorie
Copyright M. Gross, ETH Zürich 2005, 2006

2

Übertragungsmodell

ETH

- Wir betrachten das folgende Übertragungsmodell



Codierung
Diskreter Quellen

Informationstheorie
Copyright M. Gross, ETH Zürich 2005, 2006

3

Definitionen

ETH

- Die **Quellencodierung** ist die erste Stufe der Codierung in unserem Modell
- Hier soll die eineindeutige Codierung in einer **möglichst redundanzfreien** Form erfolgen
- Die Kanalcodierung ist die zweite Stufe der Codierung
- Hier wird oft gezielt zusätzliche Kontrollinformation zum Störungsschutz in den Code eingebaut
- In diesem Kapitel behandeln wir ausschliesslich redundanzfreie Quellencodierung

Codierung
Diskreter Quellen

Informationstheorie
Copyright M. Gross, ETH Zürich 2005, 2006

4

Definitionen

ETH

- **Definition:** Ein Code C über dem Codealphabet Δ mit $|\Delta|=D$ für eine Menge χ ist eine Abbildung von χ auf Δ^*
- Für $x \in \chi$ bezeichnet $C(x)$ das Codewort und $l_c(x)$ die **Länge** von $C(x)$
- Wenn mit $x_1 \neq x_2$ auch $C(x_1) \neq C(x_2)$ sowie $C(x)$ nie das leere Wort ist, dann heisst der Code **nicht-degeneriert**
- Oftmals bezeichnet man einfach die Menge der Codewörter als Code
- **Beispiel:** Für $\chi = \{a,b,c,d\}$ ist $C(a)=0$; $C(b)=10$, $C(c)=110$ und $C(d)=111$ ein Code

Codierung
Diskreter Quellen

Informationstheorie
Copyright M. Gross, ETH Zürich 2005, 2006

5

Definitionen

ETH

- Wir bezeichnen mit $||$ die Konkatenation einzelner Codewörter
- **Definition:** Ein Code C ist **eindeutig decodierbar**, wenn folgende Abbildung eindeutig ist

$$[x_1 || \dots || x_n] \rightarrow [C(x_1) || \dots || C(x_n)]$$

- **Definition:** Ein Code C heisst **präfixfrei**, wenn kein Codewort ein Präfix eines anderen Codewortes ist, also für zwei Codewörter c und c' , niemals $c'=c//d$ gilt mit $d \in \Delta^*$



Präfixfreiheit und eindeutige Decodierbarkeit sind Eigenschaften der Codewortmenge. Jeder präfixfreie Code ist eindeutig decodierbar.

Codierung
Diskreter Quellen

Informationstheorie
Copyright M. Gross, ETH Zürich 2005, 2006

6

Definitionen

ETH

- Im folgenden sind wir an der Codierung einer diskreten Quelle, also einer Zustandsvariablen X mit Verteilung $p_X(x)$ interessiert.
- Wir unterscheiden zwischen **average case** und **worst case** Analyse
- **Definition:** Ein Code C zur Codierung einer Zufallsvariablen heisst **optimal**, wenn die mittlere Codelänge minimal ist, also

$$E[l_C(X)] = \sum_{x \in \chi} p_X(x) l_C(x) \rightarrow \min$$

- **Definition:** Ein Code C , dessen Codewörter alle (nicht) gleich lang sind, heisst **(un)gleichmässig**



Codierung
Diskreter Quellen

Informationstheorie
Copyright M. Gross, ETH Zürich 2005, 2006

Ungleichmässige Codes

ETH

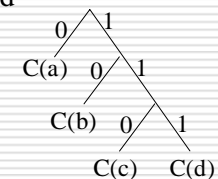
- Code 1: Gegeben $\chi = \{a,b,c,d\}$ sowie $C(a)=0$, $C(b)=10$, $C(c)=110$ und $C(d)=111$
- präfixfrei, Baumstruktur

000110110011010111

0|0|0|110|110|0|110|10|111

a a a c c a c b d

- ungleichmässig



Codierung
Diskreter Quellen

Informationstheorie
Copyright M. Gross, ETH Zürich 2005, 2006

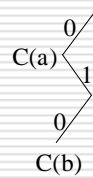
8

Ungleichmässige Codes

ETH

- Code 2: Gegeben $\chi = \{a,b\}$ sowie $C(a)=0$, $C(b)=010$
- Nicht präfixfrei, dennoch eindeutig decodierbar

00010010001000100
 0|0|010|010|0|010|0|010|0
 a a b b a b a b a



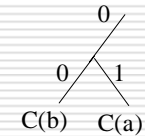
- ungleichmässig

Gleichmässige Codes

ETH

- Code 3: Gegeben $\chi = \{a,b\}$ sowie $C(a)=01$, $C(b)=00$

0001000001
 00|01|00|00|01
 b a b b a



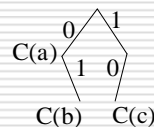
- gleichmässig

Gleichmässige Codes

ETH

- Code 4: Gegeben $\chi = \{a,b,c\}$ sowie $C(a)=0$, $C(b)=01$, $C(c)=10$
- nicht präfixfrei

010010100
 0|10|0|10|10|0
 a c a c c a
 01|0|01|0|10|0
 b a b a c a

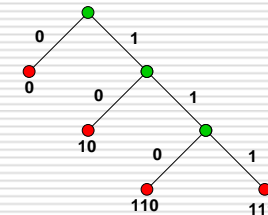


- ungleichmässig

Codebäume

ETH

- Beispiel: Für $\chi = \{a,b,c,d\}$ ist $C(a)=0$, $C(b)=10$, $C(c)=110$ und $C(d)=111$ ein Code
- Wir stellen fest, dass der Code als Baum dargestellt werden kann



Codebäume

ETH

- Jeder Code kann als Teilmenge der Knoten eines Baumes dargestellt werden.
- Jeder Knoten ist entweder ein Blatt (keine Nachfolgeknoten), oder hat höchstens D Nachfolgeknoten
- In unserem Beispiel ist $D=2$, also binärer Baum
- Ein Codebaum ist **ausgefüllt**, wenn jeder innere Knoten genau D Nachfolger hat
- Ein präfixfreier Code ist ausgefüllt, wenn der Codebaum ausgefüllt ist und jedem Blatt ein Codewort zugeordnet ist

Codierung
Diskreter Quellen

Informationstheorie
Copyright M. Gross, ETH Zürich 2005, 2006

13

Codebäume

ETH

- Sei B die Menge aller Blätter eines D -ären Codebaumes T , und sei $P(b)$ die Wahrscheinlichkeit eines Blattes $b \in B$
- Es gilt:
$$\sum_{b \in B} P(b) = 1$$
- Die **Blattentropie** von T ist definiert als

$$H_T = - \sum_{b \in B} P(b) \log P(b)$$

- Die Tiefe $t(b)$ eines Blattes ist die Distanz von der Wurzel

Codierung
Diskreter Quellen

Informationstheorie
Copyright M. Gross, ETH Zürich 2005, 2006

14

Codebäume

ETH

- Dann ist folglich die mittlere Tiefe t_T von T

$$t_T = \sum_{b \in B} P(b) t(b)$$

- Für einen ausgefüllten, präfixfreien Code C für eine Zufallsvariable X mit Codebaum T ist t_T gleich der mittleren Wortlänge,

$$p_X(x) = P(b) \quad \text{mit} \quad C(x) = b$$

- ...wenn die Blätter mit den Wahrscheinlichkeiten der codierten Symbole versehen werden

Codierung
Diskreter Quellen

Informationstheorie
Copyright M. Gross, ETH Zürich 2005, 2006

15

Kraft'sche Ungleichung

ETH

- **Theorem** (Kraft'sche Ungleichung): Ein D -ärer präfixfreier Code mit L Codewörtern der Längen l_1, \dots, l_L existiert genau dann, wenn

$$\sum_{i=1}^L D^{-l_i} \leq 1$$



Die Kraft'sche Ungleichung ist also eine hinreichende Bedingung für die Existenz eines Präfixcodes der gegebenen Struktur

Codierung
Diskreter Quellen

Informationstheorie
Copyright M. Gross, ETH Zürich 2005, 2006

16

Kraft'sche Ungleichung

ETH

- **Beweis I:** (Präfixcode- \rightarrow Ungleichung)
- Wir zeigen zunächst, dass für jeden D -ären präfixfreien Code diese Bedingung gilt:

$$\sum_{i=1}^L D^{-i} \leq 1$$

- Dies ist äquivalent mit

$$\sum_{b \in B} D^{-t(b)} \leq 1$$

Codierung
Diskreter Quellen

Informationstheorie
Copyright M. Gross, ETH Zürich 2005, 2006

17

Kraft'sche Ungleichung

ETH

- Angenommen, der Baum wächst von der Wurzel her (Induktion)
- Für die Wurzel gilt $|B|=1$:

$$\sum_{b \in B} D^{-t(b)} = D^{-t(b)} = D^{-0} = 1$$

- Wird nun die Wurzel durch D Blätter einer um 1 höheren Tiefe ersetzt, so gilt $|B|=D$ sowie

$$\sum_{b \in B} D^{-t(b)} = \sum_{i=1}^D D^{-1} = 1$$

- Die Summe bleibt also unverändert

Codierung
Diskreter Quellen

Informationstheorie
Copyright M. Gross, ETH Zürich 2005, 2006

18

Kraft'sche Ungleichung

ETH

- Ersetzen wir nun erneut ein Blatt der Tiefe 1 durch D Blätter der Tiefe 2, so gilt $|B|=D-1+D=2 \cdot D-1$ sowie

$$\sum_{b \in B} D^{-t(b)} = \sum_{i=1}^{D-1} D^{-1} + \sum_{i=1}^D D^{-2} = \sum_{i=1}^{D-1} D^{-1} + D^{-1} = 1$$

- Allgemein gilt beim Ersatz eines Blattes der Tiefe t
 $D^{-t(b)} = D \cdot D^{-t(b)-1}$

- Bei weniger neuen Blättern nimmt die Summe strikt ab
- Also gilt für einen D -ären Baum mit Blattmenge B

$$\sum_{b \in B} D^{-t(b)} \leq 1$$

- Mit Gleichheit für einen ausgefüllten Baum. Q.e.d

Codierung
Diskreter Quellen

Informationstheorie
Copyright M. Gross, ETH Zürich 2005, 2006

19

Kraft'sche Ungleichung

ETH

- **Beweis II:** (Ungleichung- \rightarrow Präfixcode)
- Wir zeigen nun, dass die Bedingung auch hinreichend ist (konstruktiv)
- Wir bezeichnen mit w_j die Anzahl der Codewörter der Länge j
- Somit gilt:

$$\sum_{j=1}^{\infty} w_j = |B|$$

- Sowie $\sum_{i=1}^L D^{-i} = \sum_{j=1}^{\infty} w_j D^{-j} \leq 1$

- Wir konstruieren nun sukzessive einen vollständigen D -ären Baum
- Wir beginnen bei Tiefe 1 und D Blättern

Codierung
Diskreter Quellen

Informationstheorie
Copyright M. Gross, ETH Zürich 2005, 2006

20

Kraft'sche Ungleichung

ETH

- Es bleiben w_1 Blätter als Codewörter stehen und $D-w_1$ Blätter werden zu insgesamt $D(w_1-D)$ Knoten der Tiefe 2.
- Wir belassen wiederum w_2 Blätter als Codewörter und ersetzen den Rest durch insgesamt $D(D(w_1-D)-w_2)$ Knoten der Tiefe 3 usw. für $m=1,2,3,\dots$
- Dies funktioniert genau dann, wenn in jedem Schritt noch genug Blätter als Codewörter vorhanden sind, also

$$w_m \leq D^m - \sum_{j=1}^{m-1} w_j D^{m-j}$$

- Division durch D^m ergibt: $\sum_{j=1}^m w_j D^{-j} \leq 1$
- Q.e.d.

Codierung
Diskreter Quellen

Informationstheorie
Copyright M. Gross, ETH Zürich 2005, 2006

21

Implikationen

ETH

- Für den Fall eines binären, präfixfreien Codes gilt entsprechend

$$\sum_{i=1}^L 2^{-l_i} \leq 1$$

- Jeder präfixfreie Code muss auch die Kraft'sche Ungleichung erfüllen
- Die Kraft'sche Ungleichung ist eine notwendige Bedingung für die Existenz eines eindeutig decodierbaren Codes mit spezifizierten Codewortlängen
- Sie ist eine notwendige Bedingung für die eindeutige Dekodierbarkeit
- Nicht jeder Code der Struktur ist eindeutig decodierbar

Codierung
Diskreter Quellen

Informationstheorie
Copyright M. Gross, ETH Zürich 2005, 2006

22

Kraft'sche Ungleichung

ETH

- Codierung einer diskreten Quelle X mit $L=6$ Zeichen und Codes mit $l_{\max} = 4$
- Wir analysieren 4 Varianten:

X	K1	K2	K3	K4
a	00	00	00	0
b	01	01	01	100
c	10	10	10	101
d	110	110	110	110
e	111	1110	1110	1110
f	1101	1101	1111	1111

Codierung
Diskreter Quellen

Informationstheorie
Copyright M. Gross, ETH Zürich 2005, 2006

23

Kraft'sche Ungleichung

ETH

- Wir prüfen:
 - Präfixbedingung
 - Kraft'sche Ungleichung
- K1
 - nicht erfüllt: d und f haben den gleichen Präfix
 - nicht erfüllt
$$3 \cdot \frac{1}{4} + 2 \cdot \frac{1}{8} + \frac{1}{16} > 1$$
- K2
 - nicht erfüllt: d und f haben den gleichen Präfix
 - erfüllt
$$3 \cdot \frac{1}{4} + \frac{1}{8} + 2 \cdot \frac{1}{16} = 1$$

Codierung
Diskreter Quellen

Informationstheorie
Copyright M. Gross, ETH Zürich 2005, 2006

24

Kraft'sche Ungleichung

ETH

- K3
1) erfüllt
2) erfüllt
- K4
1) erfüllt
2) erfüllt
- Eindeutig dekodierbar sind nur K3 und K4
- Kraft'sche Ungleichung impliziert Existenz eines decodierbaren Codes

1. Shannon'sches Codierungstheorem

ETH

- Wir suchen Schranken für die mittlere Codelängen optimaler, präfixfreier Codes
- **Theorem (Shannon):** Die mittlere Codewortlänge $E[l_C(X)]$ eines optimalen präfixfreien Codes C über einem Codealphabet Δ mit $|\Delta|=D$ für eine Zufallsvariable X erfüllt

$$\frac{H(X)}{\log D} \leq E[l_C(X)] < \frac{H(X)}{\log D} + 1$$

- Für $D=2$ gilt

$$H(X) \leq E[l_C(X)] < H(X) + 1$$



Absolut fundamental!

1. Shannon'sches Codierungstheorem

ETH

- **Beweis (untere Schranke):** Wir betrachten den entsprechenden Codebaum und erinnern uns, dass

$$t_T = E[l_C(X)] \quad \text{sowie} \quad H_T = H(X)$$

- Die untere Schranke entspricht somit
- $$H_T \leq t_T \log D$$
- Wir beweisen die Ungleichung mittels Induktion über Teilbäume
 - Für einen leeren Baum gilt die Beziehung trivialerweise
 - Jeder D -äre Baum kann als Wurzel mit bis zu D disjunkten Teilbäumen aufgefasst werden, T_1, \dots, T_D mit disjunkten Blattmengen B_1, \dots, B_D

1. Shannon'sches Codierungstheorem

ETH

- Sei die Summe der Wahrscheinlichkeiten der Blätter in Teilbaum T_i mit Blattmenge B_i
- $$q_i = \sum_{b \in B_i} P(b)$$
- Wir normieren die Verteilung im Teilbaum T_i durch Division, so dass

$$1 = \sum_{b \in B_i} \frac{P(b)}{q_i}$$

- Die mittlere Blatttiefe t_i im Teilbaum ist

$$t_i = \sum_{b \in B_i} \frac{P(b)}{q_i} (t(b) - 1)$$

1. Shannon'sches Codierungstheorem **ETH**

- Die Blattentropie ist entsprechend

$$H_i = - \sum_{b \in B_i} \frac{P(b)}{q_i} \log \frac{P(b)}{q_i}$$

- Als Induktionsannahme gilt

$$H_i \leq t_i \log D \quad \forall i$$

- Die mittlere Blatttiefe des ganzen Baum ist dann

$$t_T = \sum_{i=1}^D q_i (t_i + 1) = 1 + \sum_{i=1}^D q_i t_i$$

- Die Entropie ist

$$H_T = - \sum_{i=1}^D q_i \left(\sum_{b \in B_i} \frac{P(b)}{q_i} \log \frac{P(b)}{q_i} \right)$$

1. Shannon'sches Codierungstheorem **ETH**

- Entsprechend umgeformt

$$H_T = - \sum_{i=1}^D \left(\sum_{b \in B_i} P(b) \log P(b) \right)$$

$$H_T = - \sum_{i=1}^D \left(\sum_{b \in B_i} P(b) \left(\log \frac{P(b)}{q_i} + \log q_i \right) \right)$$

$$H_T = - \sum_{i=1}^D \left(q_i \log q_i + \sum_{b \in B_i} P(b) \log \frac{P(b)}{q_i} \right)$$

$$H_T = \sum_{i=1}^D \left(-q_i \log q_i - q_i \sum_{b \in B_i} \frac{P(b)}{q_i} \log \frac{P(b)}{q_i} \right)$$

1. Shannon'sches Codierungstheorem **ETH**

- Entsprechend umgeformt

$$H_T = \sum_{i=1}^D (-q_i \log q_i + q_i H_i)$$

$$H_T = H([q_1 \dots q_D]) + \sum_{i=1}^D q_i H_i$$

- Aufgrund unserer Annahmen sowie den Grenzen der Entropiefunktion gilt

$$H_T \leq \log D + \sum_{i=1}^D q_i t_i \log D$$

$$H_T \leq \log D (1 + \sum_{i=1}^D q_i t_i) = t_T \log D$$

1. Shannon'sches Codierungstheorem **ETH**

- Beweis (obere Schranke):** Eine intuitiv sinnvolle Wahl der Codewortlängen wäre

$$l_c(x) = -\log_D p_X(x) \quad \forall x \in X$$

- Die Codewortlänge sollte also dem negativen Logarithmus der entsprechenden Symbolwahrscheinlichkeiten entsprechen
- Die Kraft'sche Ungleichung wäre nun mit Gleichheit erfüllt, da

$$\sum_{x \in X} D^{-l_c(x)} = \sum_{x \in X} D^{-(-\log_D p_X(x))} = \sum_{x \in X} p_X(x) = 1$$

- Ein solcher Code existiert allerdings nur, wenn alle Codelängen ganzzahlig sind

1. Shannon'sches Codierungstheorem **ETH**

- Also wählen wir statt dessen

$$l_C(x) = \lceil -\log_D p_X(x) \rceil \quad \forall x \in X$$

- Die Kraft'sche Ungleichung gilt immer noch, da

$$\sum_{x \in X} D^{-l_C(x)} = \sum_{x \in X} D^{-\lceil -\log_D p_X(x) \rceil} \leq \sum_{x \in X} D^{\log_D p_X(x)} = \sum_{x \in X} p_X(x) = 1$$

- Somit existiert ein Code für diese Codelängen.
Seine mittlere Codelänge ist

$$E[l_C(x)] = -\sum_{x \in X} p_X(x) \lceil \log_D p_X(x) \rceil \leq 1 - \sum_{x \in X} p_X(x) \log_D p_X(x) = 1 + \frac{H(X)}{\log D}$$

- Wir nutzen hierbei, das

$$\lceil u \rceil \leq u + 1$$

Implikationen **ETH**

- Wir haben zum ersten Mal die Nützlichkeit der Entropie als Informationsmass gerechtfertigt
- Wir erhalten einen Zusammenhang zwischen der Entropie einer Zufallsvariablen und der Güte eines optimalen Codes dafür
- Obere Schranke ist um 1 grösser als untere Schranke
- Untere Schranke wird im allgemeinen nicht erreicht
- Der Einfluss von Fehlern bei zu starker Kompression ist noch nicht bekannt