ETH Zürich
Dept. Informatik

Prof. Markus Gross
WS 05/06

Informationstheorie

Lösung 6

6.1 Das "Perpetuum mobile" der Datenkompression

a) Dies ist nicht möglich, wenn die Entropie der Daten > 200 GB beträgt; Kompression würde in diesem Fall zwangsläufig zu Datenverlust führen. Im Allgemeinen ist es nicht möglich, ein Kompressionverfahren zu definieren, das jedes gegebene Datenpaket um einen fixen Prozentsatz verkleinert. Der Erfinder gibt vor, ein solches gefunden zu haben.

b) Die Wahrscheinlichkeit, dass die (zufälligen) Daten wirklich so speziell sind, dass sie sich kurz beschreiben lassen, ist extrem klein. Wahrscheinlich ist das Kompressionsprogramm über Nacht den Testdaten angepasst worden. Möglicherweise ist auch Information über das spezifische Datenpaket im Kompressionsprogramm selbst enthalten. Ein praktisch brauchbarer Kompressionsalgorithmus sollte aber universell sein, also nicht von den Daten abhängen. Beim zweiten Test sollten die Testdaten erst unmittelbar vor der Codierung erzeugt werden.

Kommentar: Diese Geschichte beruht auf Tatsachen. Der durchaus ehrliche Erfinder, welcher zwar nichts von Informationstheorie verstand, dafür aber von seinen eigenen Fähigkeiten umso mehr überzeugt war, glaubte sogar, sein Verfahren könne iterativ eingesetzt werden, um die Daten immer stärker zu komprimieren. Er bot sein Verfahren einem der bedeutendsten Hersteller mathematischer Software an.

6.2 Eindeutig decodierbare und präfixfreie Codes

a) $\{\epsilon, 01, 11\}$: Dieser Code ist offensichtlich degeneriert, da er das leere Wort enthält, und somit nicht eindeutig decodierbar.

b) $\{0, 010\}$: Nicht präfixfrei, jedoch suffixfrei, und damit eindeutig decodierbar.

c) $\{0, 01, 011, 111\}$: ebenso wie der Code aus b): Nicht präfixfrei, jedoch suffixfrei, und damit eindeutig decodierbar.

d) $\{110, 1110, 1011, 1101\}$: Nicht eindeutig decodierbar, da $1101 \parallel 110 = 110 \parallel 1110$.

e) $\{00, 10, 010, 111, 0110, 1101, 11001\}$: Präfixfrei und eindeutig decodierbar. Der Baum ist nicht ausgefüllt, denn z.B. 0110 tritt als Codewort auf, aber kein Wort mit $0111\ldots$.

f) $\{11, 101, 1011\}$: Obwohl weder präfix- noch suffixfrei, ist dieser Code eindeutig decodierbar: Eine ungerade Anzahl Einsen zwischen zwei Nullen in einer Codewortfolge identifiziert das Codewort "1011" eindeutig, eine gerade Anzahl "101". Mehr als drei Einsen zwischen zwei Nullen identifizieren eines oder mehrere Vorkommen von "11".

6.3 Richtig oder falsch?

a) Falsch; Die Entropie H(X) ist nur eine Untergrenze für die mittlere Codewortlänge eines optimalen Codes.

b) Richtig. Beweis: siehe Vorlesungsunterlagen / Skript.

c) Falsch; Für einen nicht-binären Code kann mit der gegebenen Anzahl Codewörter möglicherweise kein ausgefüllter Baum erzeugt werden.

d) Falsch. Gegenbeispiel: Der Code mit den Codewörtern $\{0, 010\}$.

6.4 Kurzaufgaben

a) Dieser Code ist eindeutig decodierbar. Ein möglicher Decodieralgorithmus: Falls nach einem möglichen Codewort eine 0 folgt, ist das Codewort zu Ende, falls eine 1 folgt, noch nicht.

b) Eine Möglichkeit, die ganzen Zahlen mit dem Alphabet $\{0, 1, 2\}$ zu codieren, ist die folgende: Die Zahlen werden binär codiert, wobei für die negativen Zahlen jeweils das Zweierkomplement verwendet wird. Damit dieses erkannt werden kann, werden positive Zahlen mit einer führenden Null geschrieben. Die Zwei wird als Stopsymbol verwendet, womit garantiert wird, dass der Code präfixfrei ist. Will man das Zweierkomplement umgehen, kann eine einfache Zwei als Stopsymbol vor positiven Zahlen verwendet werden, vor negativen Zahlen hingegen wird eine doppelte Zwei geschrieben.

c)

(i) Dieser Code ist offensichtlich einfach schlecht gestaltet. Die Kraft'sche Ungleichung erfüllt er nämlich, auch wenn eindeutige Decodierung nicht möglich ist: $\sum_{i=1}^{L} D^{-l_i} = 1/4+1/8+1/16 = 7/16$. Im ersten und/oder dritten Codewort eins bis zwei Nullen durch eine Eins zu ersetzen löst das Problem.

(ii) Ein Code über $\mathcal{D} = \{0, 1\}$ mit den Codewortlängen 1, 3, 3, 3, 3, 3, 3 ist nie eindeutig decodierbar, da (Kraft'sche Ungleichung) $\sum_{i=1}^{L} D^{-l_i} = 1/2+6*1/8 > 1$. Macht man aus dem ersten Codewort "1"das Codewort "10", ist die Kraft'sche Ungleichung erfüllt und der Code präfixfrei (ein Codewort hat entweder Länge drei oder aber ($\hat{=}$ XOR) beginnt mit "10").

(iii) Die Kraft'sche Ungleichung ist erfüllt ($\sum_{i=1}^{L} D^{-l_i} = 4 * 1/9 + 8 * 1/27 = 20/27$), der Code ist aber nicht eindeutig decodierbar. Das bisher nicht verwendete "2" kann als Stoppsymbol verwendet werden und hinter (oder vor, oder zwischen) alle Codewörter der Länge 2 gestellt werden.