Vorlesung 15a

Quellencodieren und Entropie

S sei eine endliche (oder abzählbar unendliche) Menge (ein “Alphabet”).

Die Elemente von S nennen wir *Buchstaben*.

Wir wollen die Buchstaben $a, b, \ldots$ durch (möglichst kurze) `01`-Folgen $k(a), k(b), \ldots$ codieren.

Dabei soll so etwas ausgeschlossen sein:

$$k(a) = \texttt{001},\ k(b) = \texttt{00101}.$$

Definition:

Eine Abbildung

$$k : S \to \bigcup_{l \geq 1} \{0,1\}^l$$

heißt **(binärer) Präfixcode**,

wenn kein $k(a)$ Anfangsstück irgendeines $k(b)$, $a \neq b$, ist.

Ist $k(a) = k_1(a) \ldots k_l(a)$, dann nennt man

$$\ell(a) = l$$

die *Länge* (oder auch die Anzahl der Bits)

des *Codeworts* $k(a)$.

Binäre Präfixcodes kann man sich auch als

(verwurzelte, planare (d.h. “in die Ebene gezeichnete”))

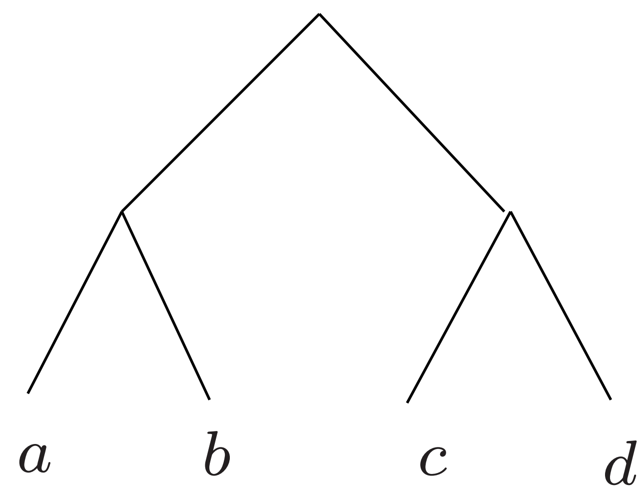
binäre Bäume vorstellen,

deren Blätter bijektiv

mit den Buchstaben des Alphabets beschriftet sind.

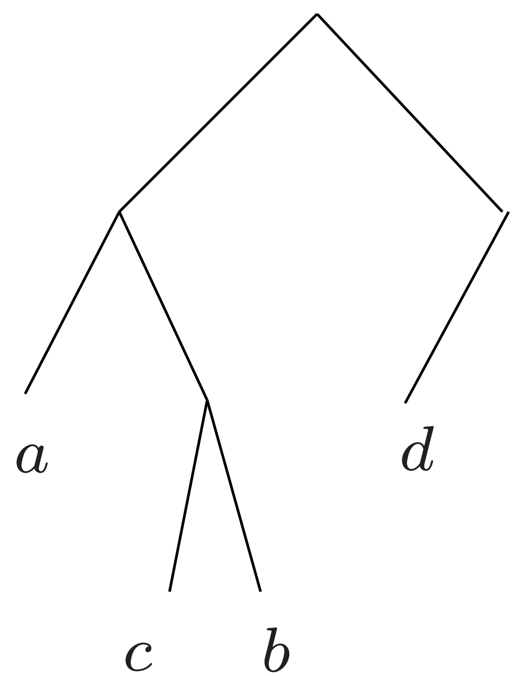
Beispiel: $S = \{a, b, c, d\}$

$$k(a) := 00, \;\; k(b) := 01, \;\; k(c) := 10, \;\; k(d) := 11.$$



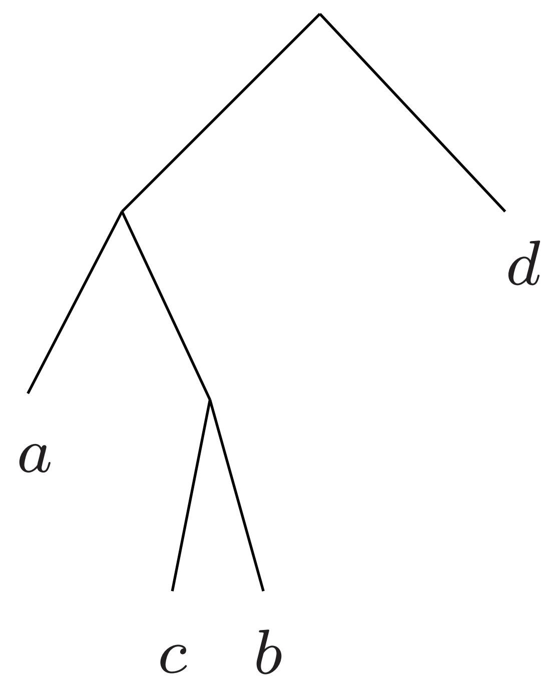
Beispiel: $S = \{a, b, c, d\}$

$$k(a) := 00, \;\; k(b) := 011, \;\; k(c) := 010, \;\; k(d) := 10.$$



Beispiel: $S = \{a, b, c, d\}$

$$k(a) := 00, \quad k(b) := 011, \quad k(c) := 010, \quad k(d) := 1.$$



Merkmale eines solchen binären Baumes:

Ein Knoten ist ausgezeichnet als *Wurzel*.

Jeder Knoten (außer der Wurzel) hat genau einen Vorgänger.

Jeder Knoten hat zwei, einen oder keinen Nachfolger.

Die Knoten ohne Nachfolger heißen *Blätter*,
die mit Nachfolger heißen *innere Knoten*.

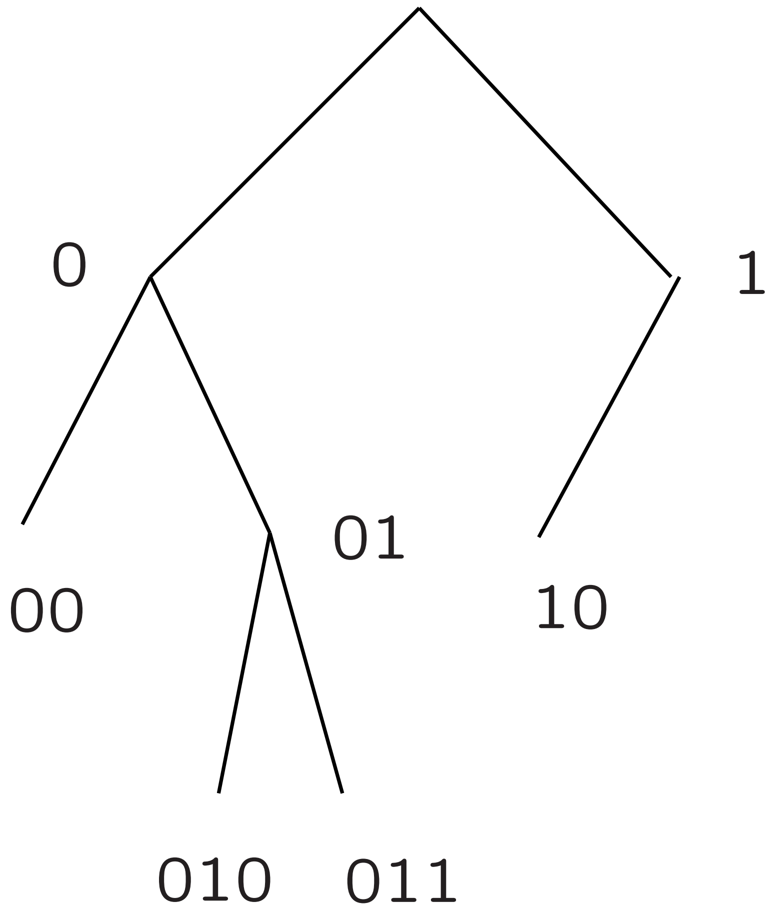
Regeln für die Beschriftung der Knoten mit 01-Wörtern:

Die Wurzel trägt das Wort der Länge 0 (das “leere Wort”).

Das Wort an einem Knoten (außer der Wurzel)
setzt das Wort am Vorgängerknoten um ein Bit fort
(die Wörter in der Tiefe l haben somit die Länge l).

Verschiedene Knoten tragen verschiedene Wörter.

Beispiel:



Die Fano-Kraft-Ungleichung (Teil 1)

Für jeden binären Präfixcode k gilt:

$$\sum_{a \in S} 2^{-\ell(a)} \leq 1 \ .$$

Beweis durch ein Gedankenexperiment:

Erzeuge per Münzwurf 01-Folgen und stoppe,

sobald eines der Codewörter vollendet ist.

Wegen der Präfixeigenschaft

schließen sich diese Ereignisse paarweise aus. Also gilt:

$$1 \geq \mathbf{P}(\text{man trifft schließlich auf ein Codewort})$$

$$= \sum_{a \in S} 2^{-\ell(a)}. \quad \square$$

Die Fano-Kraft-Ungleichung (Teil 2)

Ist $\ell : S \to \mathbb{N}$ eine Abbildung mit $\sum_{a \in S} 2^{-\ell(a)} \leq 1$,

dann gibt es einen binären Präfixcode,

für den $\ell(a)$ die Länge von $k(a)$ ist für alle $a \in S$.

Beweis: Wir ordnen die $a_1, a_2, \ldots$ so, dass

$$\ell(a_1) \leq \ell(a_2) \leq \cdots.$$

Angenommen $a_1, \ldots, a_m$ ist schon codiert,

aber a_{m+1} noch nicht. Dann ist $\sum_{i=1}^{m} 2^{-\ell(a_i)} < 1$, also

$$1 - \sum_{i=1}^{m} 2^{-\ell(a_i)} \geq 2^{-\ell(a_m)} \geq 2^{-\ell(a_{m+1})}.$$

Daraus folgt: Bei einem Münzwurf der Länge $\ell(a_{m+1})$

wirft man mit W'keit $2^{-\ell(a_{m+1})}$ ein Wort w,

das keines der $k(a_1), \ldots, k(a_m)$ als Anfangsstück enthält.

Setze dann $k(a_{m+1}) := w$. $\square$

Die Ungleichung von Fano und Kraft

formuliert für binäre Bäume:

S sei eine abzählbare Menge

und $\ell(a),\ a \in S$, seien natürliche Zahlen.

Genau dann gibt es einen binären Baum, dessen Blätter

bijektiv mit den Elementen $a \in S$ beschriftet sind

und Tiefen $\ell(a)$ haben,

wenn $\quad \sum_{a \in S} 2^{-\ell(a)} \leq 1 \quad$ gilt.

Sparsames Codieren zufälliger Buchstaben

Sei X eine S-wertige Zufallsvariable

(ein “zufälliger Buchstabe”)

mit Verteilungsgewichten $\rho(a) = \mathbf{P}(X = a)$.

Gefragt ist nach einem binären Präfixcode,

dessen erwartete Codelänge

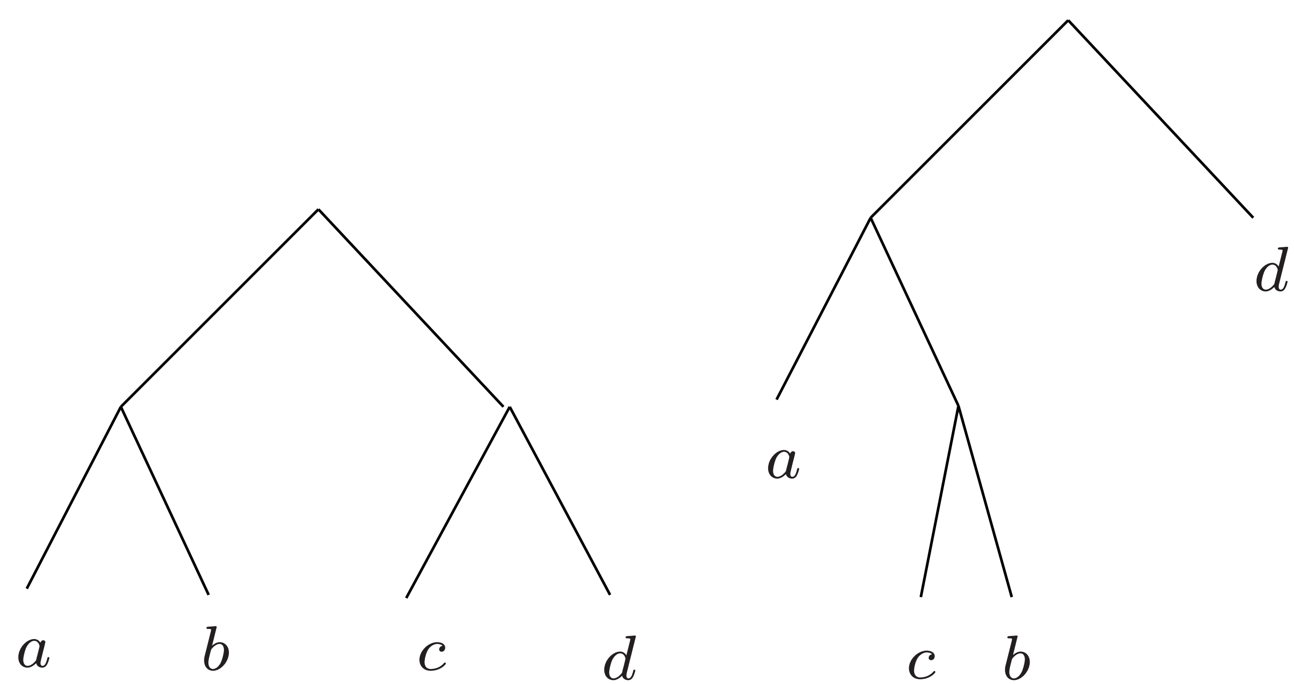
$$\mathbf{E}[\ell(X)] = \sum_{a \in S} \ell(a)\rho(a)$$

möglichst klein ist.

Beispiel: $S = \{a, b, c, d\}$.

Sind die vier Ausgänge gleich wahrscheinlich,

dann ist der Code links günstiger als der rechts.



Allgemeiner:

Gilt $\#S = 2^l$ mit $\rho(a) = 2^{-l},\ a \in S$,

dann ist es am besten, alle 2^l Buchstaben

mit den 01-Folgen der Länge l zu codieren.

In diesem Fall gilt

$$\ell(a) = l = -\log_2 \rho(a) \quad \text{für alle } a \in S.$$

Ein *Shannon-Code* ist ein Präfixcode,

bei dem jeder Buchstabe a mit einer 01-Folge codiert wird,

deren Länge $\ell(a)$ durch Aufrunden von $-\log_2 \rho(a)$

auf die nächste ganze Zahl entsteht, also

$$-\log_2 \rho(a) \leq \ell(a) < -\log_2 \rho(a) + 1 \ .$$

Solche Codes gibt es immer,

denn für die so festgelegten $\ell(a)$ folgt

$$\sum_{a \in S} 2^{-\ell(a)} \leq \sum_{a \in S} \rho(a) = 1 \ ,$$

die Fano-Kraft Ungleichung ist also erfüllt.

Wir werden zeigen:

Shannon-Codes

verfehlen die minimal mögliche erwartete Codelänge

um höchstens ein Bit.

Sei dazu

$$\mathbf{H}_2[X] := -\sum_{a \in S} \rho(a) \log_2 \rho(a),$$

mit $0 \log 0 := 0$.

Satz

(Quellencodierungssatz von Shannon)

a) Für jeden binären Präfixcode gilt

$$\mathbf{E}[\ell(X)] \geq \mathbf{H}_2[X].$$

b) Für binäre Shannon-Codes gilt außerdem

$$\mathbf{E}[\ell(X)] < \mathbf{H}_2[X] + 1.$$

Beweis:

Für Shannon-Codes gilt

$$\ell(a) < -\log_2 \rho(a) + 1$$

und folglich

$$\mathbf{E}[\ell(X)] < \sum_a \rho(a)(-\log_2 \rho(a) + 1) = \mathbf{H}_2[X] + 1.$$

Dies beweist erst einmal Teil b) des Satzes.

Jetzt zu Teil a):

Für beliebige Codes gilt

$$\mathbf{H}_2[X] - \mathbf{E}[\ell(X)] = \sum_{a:\rho(a)>0} \rho(a) \log_2 \frac{2^{-\ell(a)}}{\rho(a)} .$$

Nun ist die Logarithmusfunktion konkav

und liegt unterhalb ihrer Tangente im Punkt 1.

Folglich gilt $\log_2 x \le c \cdot (x - 1)$ mit geeignetem $c > 0$, und

$$\mathbf{H}_2[X] - \mathbf{E}[\ell(X)] \le c \sum_{a:\rho(a)>0} \rho(a) \Big(\frac{2^{-\ell(a)}}{\rho(a)} - 1 \Big)$$

$$\le c \cdot \Big(\sum_a 2^{-\ell(a)} - 1 \Big).$$

Nach Fano-Kraft ist die rechte Seite ≤ 0, also

$$\mathbf{H}_2[X] - \mathbf{E}[\ell(X)] \le 0. \quad \square$$

Die im Quellencodierungssatz auftretende Größe

$$\mathbf{H}_2[X] := -\sum_{a \in S} \rho(a) \log_2 \rho(a)$$

heißt die **Entropie** von X (zur Basis 2).

Es ist sinnvoll, auch andere Basen zu erlauben.

Wir schreiben

$$\mathbf{H}[X] := -\sum_{a \in S} \rho(a) \log \rho(a),$$

wobei log zu einer beliebigen, festen Basis genommen wird:

$$\mathbf{H}_b[X] := -\sum_{a \in S} \rho(a) \log_b \rho(a).$$

Die Entropie ist *der* fundamentale Begriff
der Informationstheorie.

Nach dem Quellenkodierungssatz gibt die Entropie
fast genau die mittlere Anzahl von Ja-Nein Fragen an,
die notwendig und
- bei guter Wahl des Codes - auch hinreichend ist,
um den unbekannten Wert von X von jemandem zu erfragen,
der X beobachten kann.

Dies ist gemeint, wenn man die Entropie beschreibt als den
Grad von Unbestimmtheit oder Ungewissheit
über den Wert, den X annimmt.