

Eindeutig entschlüsselbar  $\Rightarrow \sum_{j=1}^n 2^{-\ell_j} \leq 1$

- Sei  $C$  eindeutig entschlüsselbar mit  $C(a_j) = \ell_j$ ,  $\ell = \max_j \{\ell_j\}$ .
- Wählen  $r \in \mathbb{N}$  beliebig. Betrachten

$$\left( \sum_{j=1}^n 2^{-\ell_j} \right)^r = \sum_{i=1}^{r\ell} n_i 2^{-i}$$

- Analog zum Beweis zuvor:  $n_i =$  Anzahl Strings aus  $\{0, 1\}^i$ , die sich als Folge von  $r$  Codeworten schreiben lässt.
- $C$  eindeutig entschlüsselbar: Jeder String aus  $\{0, 1\}^i$  lässt sich als höchstens eine Folge von Codeworten schreiben, d.h.  $n_i \leq 2^i$ .
- Damit gilt  $\sum_{i=1}^{r\ell} n_i 2^{-i} \leq r\ell \Rightarrow \sum_{j=1}^n 2^{-\ell_j} \leq (r\ell)^{\frac{1}{r}}$
- Für  $r \rightarrow \infty$  folgt  $\sum_{j=1}^n 2^{-\ell_j} \leq 1$ .

# Huffman Kodierung

**Szenario:** Quelle  $Q$  mit Symbole  $\{a_1, \dots, a_n\}$

- $a_i$  sortiert nach absteigenden Quellws.  $p_1 \geq p_2 \geq \dots \geq p_n$ .

## Algorithmus Huffman-Kodierung

**Eingabe:** Symbole  $a_i$  mit absteigend sortierten  $p_i$ ,  $i = 1, \dots, n$ .

- 1 IF ( $n=2$ ), Ausgabe  $C(a_1) = 0$ ,  $C(a_2) = 1$ .
- 2 ELSE
  - 1 Bestimme  $k \in \mathbb{Z}_{n-1}$  mit  $p_k \geq p_{n-1} + p_n \geq p_{k+1}$ .
  - 2  $(p_1, \dots, p_k, p_{k+1}, p_{k+2}, \dots, p_{n-1}) \leftarrow (p_1, \dots, p_k, p_{n-1} + p_n, p_{k+1}, \dots, p_{n-2})$
  - 3  $(C(a_1), \dots, C(a_{k-1}), C(a_{k+1}), \dots, C(a_{n-2}), C(a_k)0, C(a_k)1) \leftarrow$   
Huffmann-Kodierung( $a_1, \dots, a_{n-1}, p_1, \dots, p_{n-1}$ )

**Ausgabe:** kompakter Präfixcode für  $Q$

**Laufzeit:**  $\mathcal{O}(n^2)$

( $\mathcal{O}(n \log n)$  mit Hilfe von Heap-Datenstruktur)

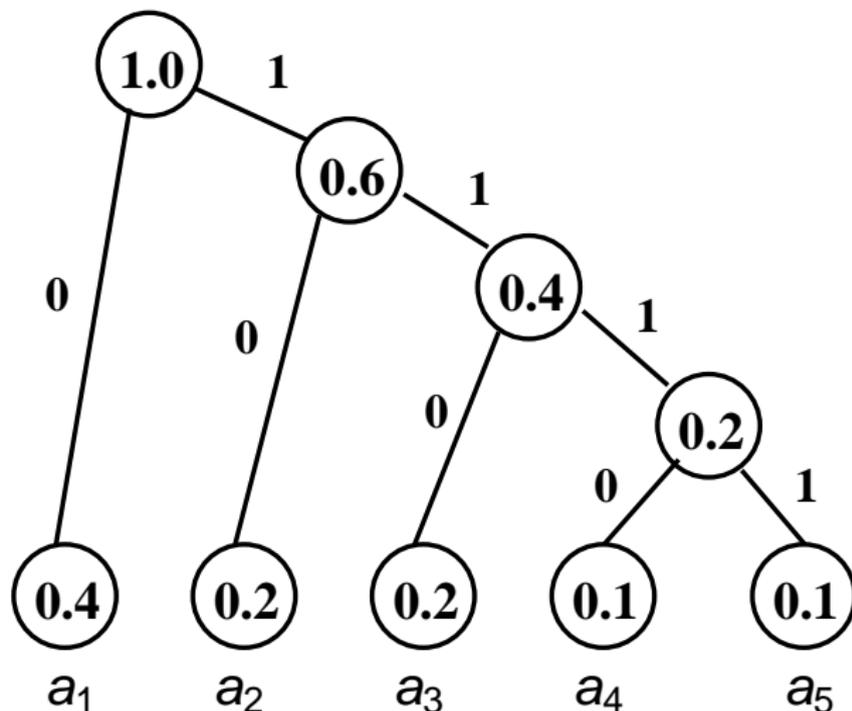
# Beispiel Huffman-Kodierung

**Beispiel:**  $p_1 = 0.4$ ,  $p_2 = p_3 = 0.2$ ,  $p_4 = p_5 = 0.1$

$a_i$	$p_i$	$C(a_i)$	$p_i$	$C(a_i)$	$p_i$	$C(a_i)$	$p_i$	$C(a_i)$
$a_1$	0.4	00	0.4	00	<b>0.4</b>	1	<b>0.6</b>	0
$a_2$	0.2	01	0.2	01	0.4	00	0.4	1
$a_3$	0.2	11	<b>0.2</b>	10	0.2	01		
$a_4$	0.1	100	0.2	11				
$a_5$	0.1	101						

- **Fett** gedruckt: Stelle  $k$  Man beachte:  $k$  ist nicht eindeutig, d.h.  $C$  ist nicht eindeutig.
- $E(C) = (0.4 + 0.2 + 0.2) * 2 + 2 * 0.1 * 3 = 2.2$
- Huffman-Tabelle: Spalten 1 und 3. Mittels Huffman-Tabelle kann jeder String  $m \in A^*$  in Zeit  $\mathcal{O}(|C(m)|)$  kodiert werden.

## Wahl eines anderen $k$



$$E(C') = 0.4 * 1 + 0.2 * (2 + 3) + 0.1 * 2 * 4 = 2.2$$

# Eigenschaften kompakter Codes

Sei  $l_i := |C(a_i)|$ .

## Lemma: Eigenschaften kompakter Codes

Sei  $C$  ein kompakter Code, oBdA ist  $C$  ein Präfixcode.

- 1 Falls  $p_i > p_j$ , dann ist  $l_i \leq l_j$
- 2 Es gibt mindestens zwei Codeworte in  $C$  mit maximaler Länge.
- 3 Unter den Worten mit maximaler Länge existieren zwei Worte, die sich nur in der letzten Stelle unterscheiden.

# Beweis der Eigenschaften

## Beweis:

- ① Sei  $l_i > l_j$ . Dann gilt

$$\begin{aligned} p_i l_i + p_j l_j &= p_i(l_i - l_j + l_j) + p_j(l_j - l_i + l_i) \\ &= p_i l_j + p_j l_i + (l_i - l_j)(p_i - p_j) > p_i l_j + p_j l_i \end{aligned}$$

D.h. vertauschen der Kodierungen von  $a_i$  und  $a_j$  verkürzt den Code.

- ② Sei  $c = c_1 \dots c_n \in C$  das einzige Codewort mit maximaler Länge. Streichen von  $c_n$  führt zu einem Präfixcode mit kürzerer erwarteter Codewortlänge.
- ③ Annahme: Alle Paar von Codeworten maximaler Länge unterscheiden sich nicht nur in der letzten Komponente.
- ▶ Entferne die letzte Komponente eines beliebigen Codewortes maximaler Länge.
  - ▶ Wir erhalten einen Präfixcode mit kürzerer Länge.

# Optimalität der Huffman-Kodierung

## Satz

Die Huffman-Kodierung liefert einen kompakten Code.

### Beweis per Induktion über $n$ .

- **IA:**  $n = 2$ : Für  $\{a_1, a_2\}$  ist die Codierung  $\{0, 1\}$  kompakt.
- **IS:**  $n - 1 \rightarrow n$ : Sei  $C'$  kompakt für  $\{a_1, \dots, a_n\}$ .
  - ▶ Lemma,2:  $C'$  enthält zwei Codeworte maximaler Länge.
  - ▶ Lemma,3: Unter den Codeworten maximaler Länge gibt es zwei Codeworte  $c_0, c_1 \in C'$  mit  $c \in \{0, 1\}^*$ , die sich nur in der letzten Stelle unterscheiden.
  - ▶ Lemma,1: Die beiden Symbole  $a_{n-1}, a_n$  mit kleinster Quellws besitzen maximale Codewortlänge. Vertausche die Kodierungen dieser Symbole mit  $c_0, c_1$ .
  - ▶  $a_{n-1}$  oder  $a_n$  tauchen mit Ws  $p_{n-1} + p_n$  auf.
  - ▶ **IA:** Huffman-Kodierung liefert kompakten Präfixcode  $C$  für  $a_1, \dots, a_{n-2}, a'$  mit Quellws  $p_1, \dots, p_{n-2}, p_{n-1} + p_n$
  - ▶  $C(a_1), \dots, C(a_{n-2}), C(a')0 = c_0, C(a')1 = c_1$  ist Präfixcode mit erwarteter Codewortlänge  $E(C')$ , d.h. die Huffman-Kodierung liefert einen kompakten Präfixcode.

# Informationsgehalt einer Nachricht

Betrachten folgendes Spiel

- Gegeben: Quelle  $Q$  mit unbekanntem Symbolen  $\{a_1, a_2\}$  und  $p_1 = 0.9, p_2 = 0.1$ .
- Zwei Spieler erhalten rundenweise je ein Symbol.
- Gewinner ist, wer zuerst beide Symbole erhält.

Szenario:

- Spieler 1 erhält in der ersten Runde  $a_1$  und Spieler 2 erhält  $a_2$ .
- **Frage:** Wer gewinnt mit höherer Ws? Offenbar Spieler 2.

**Intuitiv:** Je kleiner die Quellws, desto höher der Informationsgehalt.

# Eigenschaft von Information

## Forderungen für eine Informationsfunktion

- 1  $I(p) \geq 0$ : Der Informationsgehalt soll positiv sein.
- 2  $I(p)$  ist stetig in  $p$ : Kleine Änderungen in der Ws  $p$  sollen nur kleine Änderungen von  $I(p)$  bewirken.
- 3  $I(p_i) + I(p_j) = I(p_i p_j)$ :
  - ▶  $X$  = Ereignis, dass  $a_i$  und  $a_j$  nacheinander übertragen werden.
  - ▶ Informationsgehalt von  $X$ :  $I(p_i) + I(p_j)$ ,  $W_s(X) = p_i p_j$

## Satz zur Struktur von $I(p)$

Jede Funktion  $I(p)$  für  $0 < p \leq 1$ , die obige drei Bedingungen erfüllt, ist von der Form

$$I(p) = C \log \frac{1}{p}$$

für eine positive Konstante  $C$ .

## Beweis: Form von $I(p)$

- Forderung 3 liefert  $I(p^2) = I(p) + I(p) = 2I(p)$ .
- Induktiv folgt:  $I(p^n) = nI(p)$  für alle  $n \in \mathbb{N}$  und alle  $0 < p \leq 1$ .
- Substitution  $p \rightarrow p^{\frac{1}{n}}$  liefert:  $I(p) = nI(p^{\frac{1}{n}})$  bzw.  $I(p^{\frac{1}{n}}) = \frac{1}{n}I(p)$
- Damit gilt für alle  $q \in \mathbb{Q}$ :  $I(p^q) = qI(p)$ .
- Für jedes  $r \in \mathbb{R}$  gibt es eine Sequenz  $q_i$  mit  $\lim_{n \rightarrow \infty} q_n = r$ . Aus der Stetigkeit von  $I(p)$  folgt

$$I(p^r) = I\left(\lim_{n \rightarrow \infty} p^{q_n}\right) = \lim_{n \rightarrow \infty} I(p^{q_n}) = \lim_{n \rightarrow \infty} q_n I(p) = rI(p)$$

- Fixiere  $0 < q < 1$ . Für jedes  $0 < p \leq 1$  gilt

$$\begin{aligned} I(p) &= I(q^{\log_q p}) = I(q) \log_q p = -I(q) \log_q \left(\frac{1}{p}\right) = -I(q) \frac{\log_2 \frac{1}{p}}{\log_2 q} \\ &= C \log_2 \frac{1}{p} \quad \text{mit } C = -I(q) \cdot \frac{1}{\log_2(q)} > 0. \end{aligned}$$