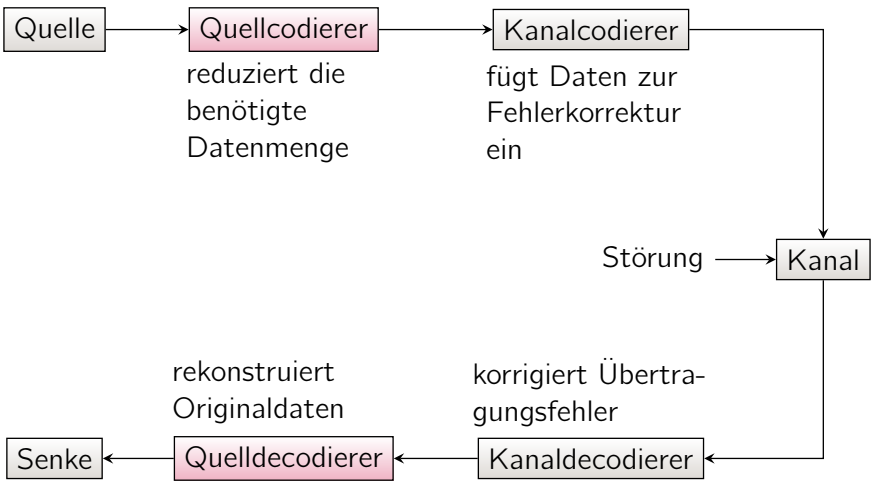
3 Codierung diskreter Quellen

3 Codierung diskreter Quellen



3.1 Einführung

- Die Quellcodierung wandelt Sequenzen von Quellsymbolen in Bitsequenzen.
- Aus Effizienzgründen soll die erzeugte Bitsequenz möglichst kurz sein.
- Selbstverständlich soll sich aus der Bitsequenz die Symbolsequenz eindeutig wiederherstellen lassen.

Definition 3.1

Eine *(Quell)codierung* $\mathcal{C}$ bildet jedes Symbol $x \in X$ einer Quelle $\mathcal{X}$ auf ein *Codewort* $\mathcal{C}(x) \in \{0,1\}^+ = \bigcup_{l=1}^{\infty} \{0,1\}^l$ ab, wobei $x \neq x' \Rightarrow \mathcal{C}(x) \neq \mathcal{C}(x')$. Eine Sequenz von Symbolen $x_1 \dots x_N$ wird auf die Sequenz $\mathcal{C}(x_1) \dots \mathcal{C}(x_N)$ abgebildet. Die Menge C aller Codewörter $\mathcal{C}(x)$ wird *Code* genannt.

Beispiel 3.1

Die Zahlen $\{1, \dots, 6\}$ lassen sich durch die binäre Darstellung $1 \mapsto 001, 2 \mapsto 010, 3 \mapsto 011, 4 \mapsto 100, 5 \mapsto 101, 6 \mapsto 110$ codieren. Die Sequenz 123 würde damit zu 001010011 codiert und könnte eindeutig decodiert werden.

Beispiel 3.2

Eine alternative Codierung der Zahlen $\{1, \dots, 6\}$ wäre $1 \mapsto 1, 2 \mapsto 10, 3 \mapsto 11, 4 \mapsto 100, 5 \mapsto 101, 6 \mapsto 110$. Die Sequenz 123 würde damit zu 11011 codiert, was als 123, 1211, 151, 611 oder 63 decodiert werden könnte.

Definition 3.2

Eine Codierung ist *eindeutig decodierbar*, wenn jedes Paar unterschiedlicher Sequenzen $x_1 \dots x_N \neq x'_1 \dots x'_M$ zu unterschiedlichen Sequenzen $\mathcal{C}(x_1) \dots \mathcal{C}(x_N) \neq \mathcal{C}(x'_1) \dots \mathcal{C}(x'_M)$ codiert wird.

Definition 3.3

Ein Code (eine Codierung) heißt *gleichmäßig*, wenn alle Codewörter die gleiche Länge haben.

Satz 3.1

Jede gleichmäßige Codierung ist eindeutig decodierbar.

Beweis.

Dank der einheitlichen Länge der Codewörter ist die Zerlegung jeder Sequenz $\mathcal{C}(x_1) \dots \mathcal{C}(x_N)$ in die Codewörter $\mathcal{C}(x_i)$ eindeutig. □

3.2 Ungleichmäßige Codierung

- Für eine höhere Code-Effizienz (weniger benötigte Bits) sind ungleichmäßige Codierungen unverzichtbar.
- Seien $l_1, \ldots, l_N$ die Längen der Codewörter eines Codes C für ein Quellalphabet mit N Symbolen.
- Welche Bedingung müssen die Längen $l_1, \ldots, l_N$ erfüllen, damit $\mathcal{C}$ eindeutig decodierbar ist?

Satz 3.2 (Satz von McMillan)

Seien $l_1, \ldots, l_N$ die Codewortlängen einer eindeutig decodierbaren Codierung, so gilt

$$K = \sum_{i=1}^{N} 2^{-l_i} \leq 1. \tag{3.1}$$

Zu jeder Menge von Codewortlängen, die diese Bedingung erfüllen, gibt es eine eindeutig decodierbare Codierung.

Aber nicht jede Codierung, die die Bedingung erfüllt, ist eindeutig decodierbar!

Beweis.

Wir beweisen zunächst nur den ersten Teil; die Existenz einer eindeutig decodierbaren Codierung bei erfüllter Bedingung wird später für eine spezielle Klasse von Codierungen gezeigt.

Wir betrachten

$$K^n = \left(\sum_{i=1}^{N} 2^{-l_i} \right)^n = \left(2^{-l_1} + 2^{-l_2} + \cdots + 2^{-l_N} \right)^n, \quad n \geq 1 \tag{3.2}$$

was sich in eine Summe von N^n Termen der Form

$$2^{-l_{i_1} - l_{i_2} - \cdots - l_{i_n}} = 2^{-k_i}, \qquad i \in 1, \ldots, N^n \tag{3.3}$$

mit $k_i = l_{i_1} + l_{i_2} + \cdots + l_{i_n}$ expandieren lässt. Sei $m = \max(l_1, l_2, \ldots, l_N)$ die maximale Codewortlänge, so gilt offenbar $n \leq k_i \leq nm$. Daher lässt sich

$$K^n = \sum_{k=n}^{nm} N_k 2^{-k} \tag{3.4}$$

schreiben, wobei N_k die Anzahl der Terme mit $k_i = k$ angibt, was gleichzeitig die Anzahl der Codewortsequenzen der Gesamtlänge k ist. Aufgrund der eindeutigen Decodierbarkeit muss $N_k \leq 2^k$, also

$$K^n \leq \sum_{k=n}^{nm} 2^k 2^{-k} = \sum_{k=n}^{nm} 1 = nm - n + 1 \leq nm \tag{3.5}$$

für beliebig große n, und daher $K \leq 1$. □

3.3 Präfix-Codes

Definition 3.4

Ein Code wird Präfix-Code genannt, wenn kein Codewort $\mathcal{C}(x)$ Präfix eines anderen Codeworts $\mathcal{C}(x')$ ist, also kein $\boldsymbol{c} \in \{0,1\}^*$ exististiert, sodass $\mathcal{C}(x)\boldsymbol{c} = \mathcal{C}(x')$ für $x \neq x'$.

Beispiel 3.3

Der Code $\{000, 001, 010, 011, 10, 11\}$ ist ein Präfix-Code.

Beispiel 3.4

Der Code $\{1, 10, 11, 100, 101, 110\}$ aus Beispiel 3.2 ist *kein* Präfix-Code.

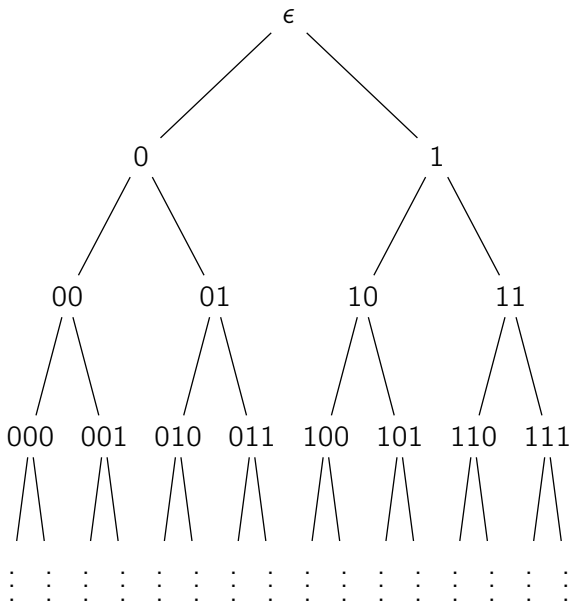
Satz 3.3

Eine auf einem Präfix-Code aufbauende Codierung ist eindeutig decodierbar.

Beweis.

Seien die ersten l Bits einer Codesequenz $\mathcal{C}(x)$. Aufgrund der Präfix-Eigenschaft gibt es kein Codewort $\mathcal{C}(x')$, dass ebenfalls den Anfang der Codesequenz bilden kann. Damit kann das erste Zeichen eindeutig zu x decodiert werden und außerdem die verbleibende Sequenz nach demselben Schema ebenfalls eindeutig decodiert werden. □

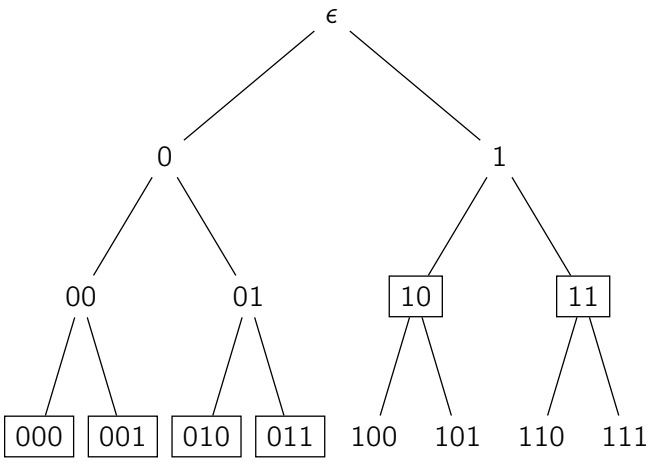
3.3.1 Codebäume



- Jedes Codewort wird als Knoten eines binären Baumes repräsentiert.
- Alle Präfixe eines Codewortes finden sich im Codebaum oberhalb des Codeworts.
- Bei einem Präfix-Code befindet sich zwischen Wurzel und einem Blatt des Codebaums maximal ein Codewort.

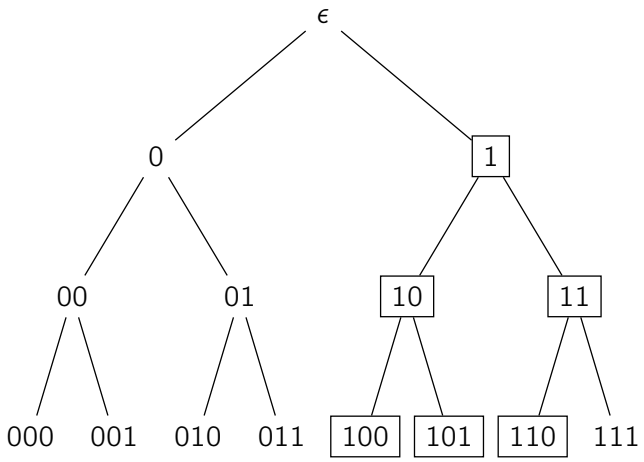
Beispiel 3.5

Der Präfix-Code $\{000, 001, 010, 011, 10, 11\}$ besitzt folgenden Code-Baum:



Beispiel 3.6

Der Code $\{1, 10, 11, 100, 101, 110\}$ (*kein* Präfix-Code) besitzt folgenden Code-Baum:



3.3.2 Kraftsche Ungleichung

Satz 3.4 (Kraftsche Ungleichung)

Seien $l_1, \ldots, l_N$ die Codewortlängen eines Präfix-Codes, so gilt

$$K = \sum_{i=1}^{N} 2^{-l_i} \leq 1. \tag{3.6}$$

Zu jeder Menge von Codewortlängen, die diese Bedingung erfüllen, gibt es einen Präfix-Code.

Beweis zum ersten Teil.

Sei $m = \max(l_1, \ldots, l_N)$. Es wird der vollständige binäre Baum der Höhe m betrachtet; er besitzt 2^m Blätter. Unterhalb eines Codeworts der Länge l liegen 2^{m-l} Blätter. Jedes Blatt liegt unter höchstens einem Codewort, deshalb

$$\sum_{i=1}^{N} 2^{m-l_i} \leq 2^m, \quad \text{also} \quad \sum_{i=1}^{N} 2^{-l_i} \leq 1. \tag{3.7}$$

□

Beweis zum zweiten Teil.

O.B.d.A. gelte $l_1 \leq l_2 \leq \cdots \leq l_N$. $\mathcal{C}(x_1)$ sei ein beliebiger Knoten auf Höhe l_1; da

$$2^{m-l_1} < \sum_{i=1}^{N} 2^{m-l_i} \leq 2^m \quad (\text{für } N > 1) \tag{3.8}$$

gibt es mindestens ein Blatt, das nicht unter $\mathcal{C}(x_1)$ liegt. Der Knoten auf Höhe l_2 über diesem Blatt besitzt daher $\mathcal{C}(x_1)$ *nicht* als Präfix und wird als $\mathcal{C}(x_2)$ verwendet. Sind auf diese Weise die ersten $M < N$ Codewörter zugeordnet, liegen

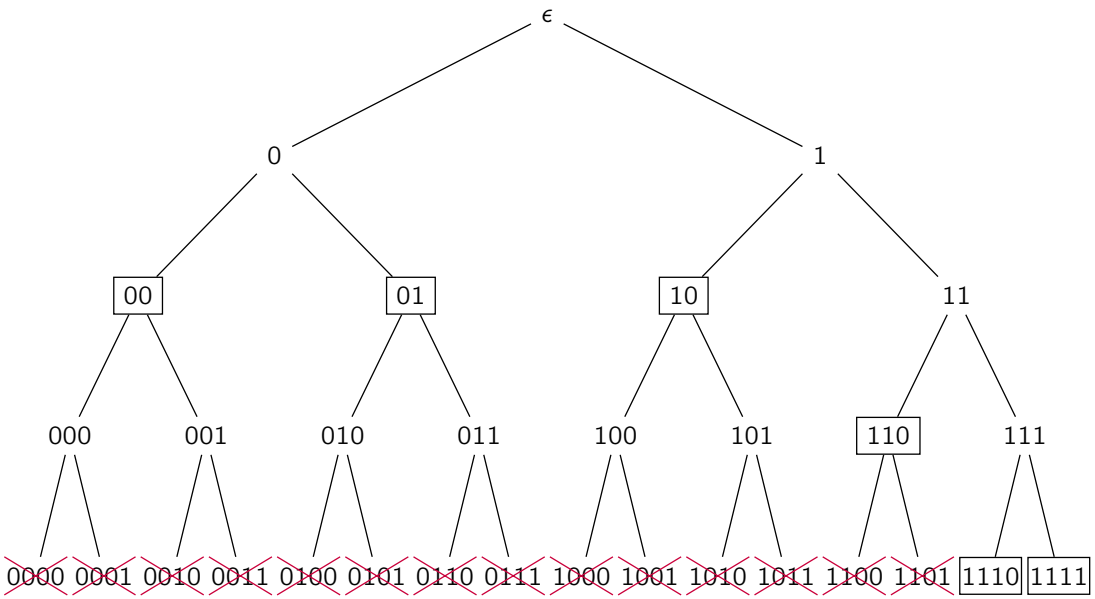
$$\sum_{i=1}^{M} 2^{m-l_i} < \sum_{i=1}^{N} 2^{m-l_i} \leq 2^m \tag{3.9}$$

Blätter unter Codewörtern, und es gibt weiterhin mindestens ein Blatt, das nicht unter einem Codewort liegt und für die Zuordnung von $\mathcal{C}(x_{M+1})$ verwendet werden kann. Wiederholung dieses Schemas bis N erzeugt einen Präfix-Code. □

Da Präfix-Codes eindeutig decodierbar sind, vervollständigt dies den Beweis zu Satz 3.2 (Satz von McMillan).

Beispiel 3.7

Es soll ein Präfix-Code mit den Längen $l_1 = l_2 = l_3 = 2, l_4 = 3, l_5 = l_6 = 4$ konstruiert werden.



Nach den Sätzen von McMillan und Kraft gelten dieselben Bedingungen in Bezug auf die Codewortlängen für die Existenz von eindeutig decodierbaren und Präfix-Codes. Dies bedeutet

- *nicht*, dass jeder eindeutig decodierbare Code ein Präfix-Code ist (z.B. $\{1, 10, 100, 1000\}$)
- aber, dass es zu jedem eindeutig decodierbaren Code einen Präfix-Code mit denselben Codewortlängen gibt (z.B. $\{1, 01, 001, 0001\}$).

Häufig ist es daher ausreichend, nur Präfix-Codes zu betrachten.

3.4 Grenzen der Code-Effizienz

Definition 3.5

Die Codierung $\mathcal{C}$ einer Quelle $\mathcal{X}$ hat die *mittlere Codewortlänge*

$$L(\mathcal{C}) = \sum_{x \in X} p_X(x) l_x, \tag{3.10}$$

wobei l_x die Länge des Codeworts $\mathcal{C}(x)$ bezeichnet.

Beispiel 3.8

Die Codierung $1 \mapsto 001, 2 \mapsto 010, 3 \mapsto 011, 4 \mapsto 100, 5 \mapsto 101, 6 \mapsto 110$ aus Beispiel 3.1 für den gezinkten Würfel aus Beispiel 2.5 ($p_X(1) = \frac{1}{21}$, $p_X(2) = \frac{2}{21}$, $p_X(3) = \frac{3}{21}$, $p_X(4) = \frac{4}{21}$, $p_X(5) = \frac{5}{21}$, $p_X(6) = \frac{6}{21}$) hat offensichtlich die mittlere Codewortlänge $L(\mathcal{C}) = 3$.

Beispiel 3.9

Die Codierung $1 \mapsto 0000, 2 \mapsto 0001, 3 \mapsto 001, 4 \mapsto 01, 5 \mapsto 10, 6 \mapsto 11$ führt für diesen Würfel auf die mittlere Codewortlänge
$L(\mathcal{C}) = (\frac{1}{21} + \frac{2}{21}) \cdot 4 + \frac{3}{21} \cdot 3 + (\frac{4}{21} + \frac{5}{21} + \frac{6}{21}) \cdot 2 = \frac{51}{21} \approx 2{,}429$.

Satz 3.5

Sei $\mathcal{C}$ eine eindeutig decodierbare Codierung der Quelle $\mathcal{X}$, so gilt

$$L(\mathcal{C}) \geq H(\mathcal{X}). \tag{3.11}$$

Beweis.

Unter Ausnutzung von Lemma 2.1 und mit $K = \sum_{x \in X} 2^{-l_x} \leq 1$ folgt

$$\begin{aligned} H(\mathcal{X}) &= -\sum_{x \in X} p_X(x) \operatorname{ld}\left(p_X(x)\right) && (3.12)\\ &\leq -\sum_{x \in X} p_X(x) \operatorname{ld}\left(\frac{2^{-l_x}}{K}\right) && (3.13)\\ &= -\sum_{x \in X} p_X(x)\big(-l_x - \operatorname{ld}(K)\big) = \sum_{x \in X} p_X(x) l_x + \operatorname{ld}(K) \sum_{x \in X} p_X(x) && (3.14)\\ &= L(\mathcal{C}) + \operatorname{ld}(K) \leq L(\mathcal{C}). && (3.15) \end{aligned}$$

□

Dabei wird $L(\mathcal{C}) = H(\mathcal{X})$ erreicht, wenn $K = 1$ und $p_X(x) = 2^{-l_x}$ bzw. $l_x = -\operatorname{ld}\left(p_X(x)\right)$, was nur in Spezialfällen auf ganzzahlige Längen führt.

Satz 3.6

Zu jeder Quelle $\mathcal{X}$ existiert eine Präfix-Codierung $\mathcal{C}$, sodass

$$H(\mathcal{X}) \leq L(\mathcal{C}) \leq H(\mathcal{X}) + 1. \tag{3.16}$$

Beweis.

Eine solche Codierung $\mathcal{C}$ ist gegeben durch die *Shannon-Codierung*, bei der

$$-\operatorname{ld}\left(p_X(x)\right) \leq l_x \leq -\operatorname{ld}\left(p_X(x)\right) + 1. \tag{3.17}$$

Damit gilt $K = \sum_{x \in X} 2^{-l_x} \leq \sum_{x \in X} 2^{\operatorname{ld}(p_X(x))} = \sum_{x \in X} p_X(x) = 1$, womit die Kraftsche Ungleichung erfüllt ist. Multiplikation von Gleichung (3.17) mit $p_X(x)$ und Summation über alle x führt auf

$$\underbrace{-\sum_{x \in X} p_X(x) \operatorname{ld}\left(p_X(x)\right)}_{H(\mathcal{X})} \leq \underbrace{\sum_{x \in X} p_X(x) l_x}_{L(\mathcal{C})} \leq \underbrace{-\sum_{x \in X} p_X(x) \operatorname{ld}\left(p_X(x)\right)}_{H(\mathcal{X})} + \underbrace{\sum_{x \in X} p_X(x)}_{1}. \tag{3.18}$$

□

Satz 3.7 (Erstes Shannonsches Codierungstheorem)

Durch Codierung der erweiterten Quelle $\mathcal{X}^n$ kann für hinreichend große n eine Codierung mit durchschnittlicher Codewortlänge beliebig nah an der Entropie erreicht werden.

Beweis.
Sei $\mathcal{C}'$ eine Codierung von $\mathcal{X}^n$ mit

$$H(\mathcal{X}^n) \leq L(\mathcal{C}') \leq H(\mathcal{X}^n) + 1, \tag{3.19}$$

die nach Satz 3.6 existiert. Nach Korollar 2.1 gilt

$$n \cdot H(\mathcal{X}) \leq L(\mathcal{C}') \leq n \cdot H(\mathcal{X}) + 1. \tag{3.20}$$

Für die auf ein Symbol der ursprünglichen Quelle $\mathcal{X}$ bezogene mittlere Codewortlänge $L_1 = L(\mathcal{C}')/n$ ergibt sich damit

$$H(\mathcal{X}) \leq L_1 \leq H(\mathcal{X}) + 1/n, \tag{3.21}$$

und offenbar $\lim_{n\to\infty} H(\mathcal{X}) + 1/n = H(\mathcal{X})$. □

3.5 Optimal-Codierung

Wir haben gesehen, dass

- die Shannon-Codierung maximal ein Bit mehr pro Symbol zur Codierung benötigt, als durch die Entropie als untere Schranke gegeben ist
- dies ausreicht, um bei Betrachtung von erweiterten Quellen asymptotisch ideal zu codieren.

Wie codiert man die praktisch relevanten endlichen Erweiterungen oder nicht erweiterten Quellen?

Definition 3.6

Ein eindeutig decodierbare Codierung $\mathcal{C}$ ist *optimal*, wenn für jede andere eindeutig decodierbare Codierung $\mathcal{C}'$ gilt, dass $L(\mathcal{C}) \leq L(\mathcal{C}')$.

3.5.1 Huffman-Codierung

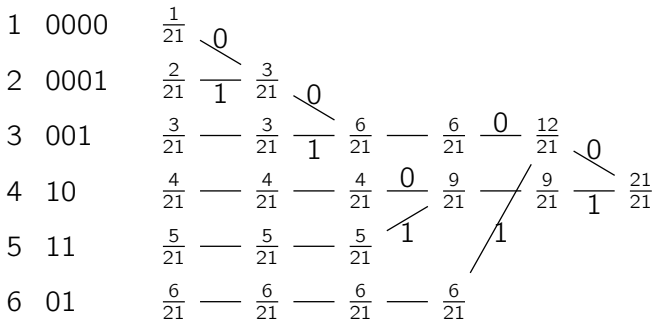
Algorithmus der Huffman-Codierung für eine Quelle $\mathcal{X}$:

1. Bestimme die beiden Symbole $x_1, x_2 \in X$ mit der kleinsten Auftrittswahrscheinlichkeit[2].
2. Konstruiere eine neue Quelle $\mathcal{X}'$ mit $X' = X \setminus \{x_1, x_2\} \cup \{x'_{1,2}\}$ und $p_{X'}(x') = p_X(x')$ für $x' \neq x'_{1,2}$, $p_{X'}(x'_{1,2}) = p_X(x_1) + p_X(x_2)$. (Es werden also x_1 und x_2 zu einem Symbol $x'_{1,2}$ zusammengefasst.)
3. Bestimme die Huffman-Codierung $\mathcal{C}'$ zu $\mathcal{X}'$ mit $x'_1 \mapsto 0, x'_2 \mapsto 1$, falls X' nur zwei Symbole enthält.
4. Wähle $\mathcal{C}(x) = \mathcal{C}'(x)$ für $x \neq x_1, x_2$, $\mathcal{C}(x_1) = \mathcal{C}'(x'_{1,2})0$, $\mathcal{C}(x_2) = \mathcal{C}'(x'_{1,2})1$.

[2] Auswahl bei Nicht-Eindeutigkeit beliebig

Beispiel 3.10

Es sei wieder der gezinkte Würfel aus Beispiel 2.5 ($p_X(x) = \frac{x}{21}$) betrachtet.



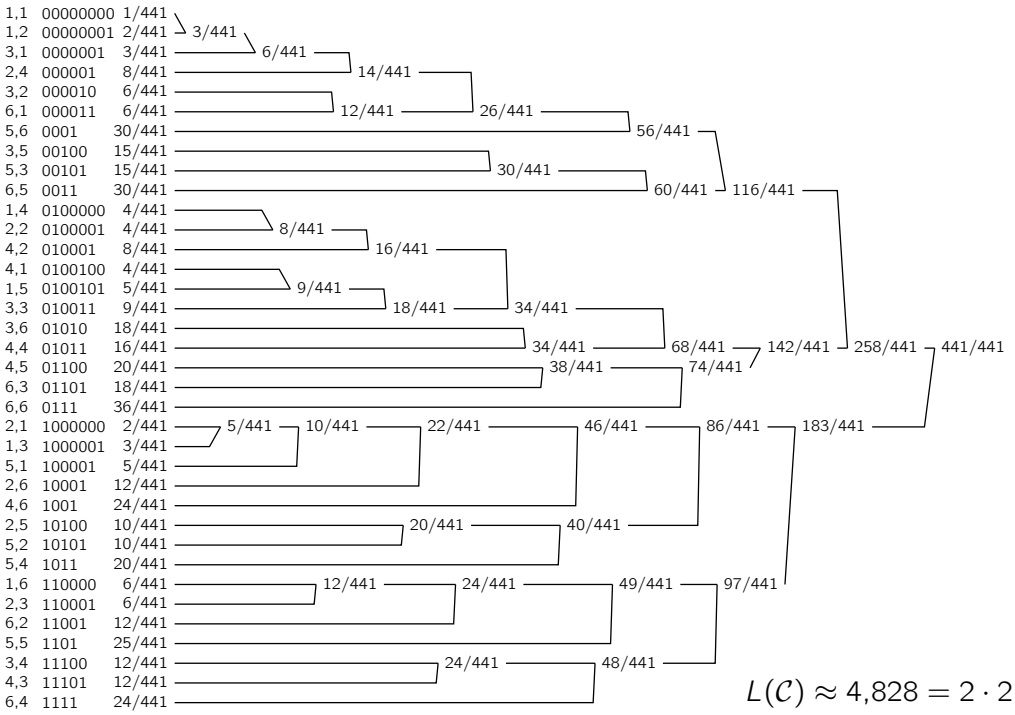
Die mittlere Codewortlänge beträgt

$$L(\mathcal{C}) = \left(\frac{1}{21} + \frac{2}{21}\right) \cdot 4 + \frac{3}{21} \cdot 3 + \left(\frac{4}{21} + \frac{5}{21} + \frac{6}{21}\right) \cdot 2 = \frac{51}{21} \approx 2{,}429$$

Zum Vergleich: $H(\mathcal{X}) \approx 2{,}398$.

Beispiel 3.11

Es sei die Erweiterung $\mathcal{X}^2$ des gezinkten Würfels ($p_{X,X}(x_1, x_2) = \frac{x_1 x_2}{441}$) betrachtet.



$L(\mathcal{C}) \approx 4{,}828 = 2 \cdot 2{,}414$

3.6 Zusammenfassung

- Aus den Sätzen von Kraft und McMillan folgt, dass Präfix-Codierungen genauso effizient sein können wie allgemeine, eindeutig decodierbare Codierungen.
- Die Entropie der zu codierenden Quelle stellt eine untere Schranke der mittleren Codewortlänge dar.
- Erstes Shannonsches Codierungstheorem: Durch blockweise Codierung (Codierung der erweiterten Quelle) lässt sich diese Schranke asymptotisch erreichen.
- Die Huffman-Codierung lässt sich durch sukzessives Zusammenfassen der beiden unwahrscheinlichsten Symbole bilden.