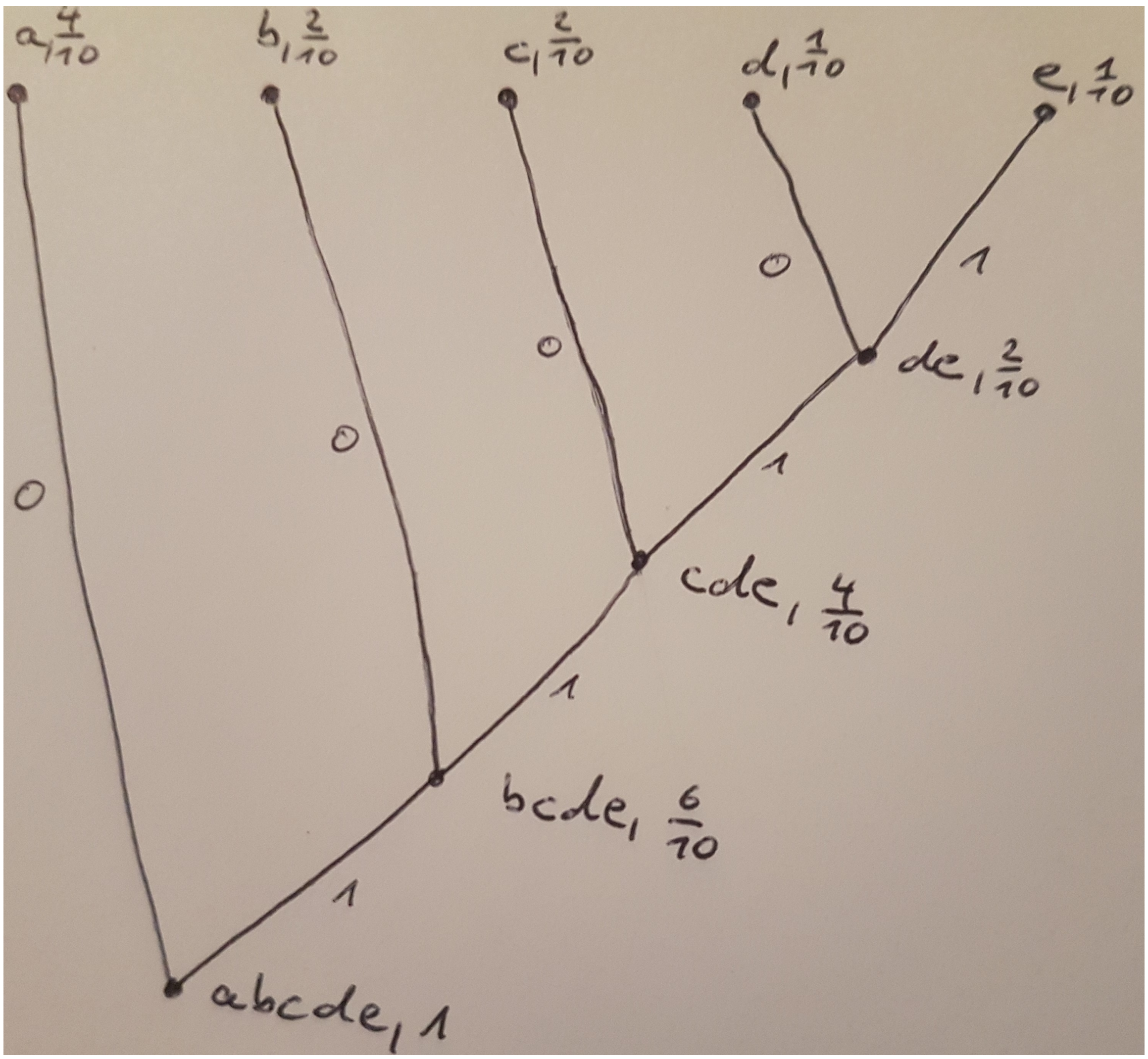
Übungsblatt 2

Auf dem gesamten Übungszettel behandeln wir die Begriffe (binärer) Wurzelbaum und Präfixcode als Synonyme. Dies ist durch Blatt 1 Aufgabe 3 gerechtfertigt.

Aufgabe 1. Kodieren Sie mit Huffman eine Quelle mit Wahrscheinlichkeitsverteilung $p = (0.4, 0.2, 0.2, 0.1, 0.1)$.

Sei $X = \{a, b, c, d, e\}$ eine Quelle mit $p(a) = 0.4$, $p(b) = p(c) = 0.2$ und $p(d) = p(e) = 0.1$. Eine Huffman Kodierung dieser lautet $a \mapsto 0$, $b \mapsto 10$, $c \mapsto 110$, $d \mapsto 1110$ und $e \mapsto 1111$, wobei der dazugehörige gewichtete Wurzelbaum wie folgt aussieht.



Aufgabe 2. Zeigen Sie $\overline{L}(C_{i-1}) = \overline{L}(C_i) + h_{i-1}$, wobei h_{i-1} die Summe zweier minimaler Wahrscheinlichkeiten der Quelle X_{i-1} ist.

Beweis. Seien, wie auf dem Aufgabenzettel beschrieben, T_i die Wurzelbäume assoziiert zu den Codes C_i für $1 \leq i \leq n$. Diese lassen sich auf natürliche Weise wie folgt ineinander einbetten. Es gilt

$$T_n \hookrightarrow T_{n-1} \hookrightarrow \ldots \hookrightarrow T_1,$$

wobei T_n die Wurzel der T_i für $1 \leq i \leq n$ und T_1 der Wurzelbaum einer Huffman Kodierung ist. Bezeichne mit B_i die Menge der Blätter des Wurzelbaumes T_i für $1 \leq i \leq n$. Beachte, dass die obige Wahrscheinlichkeitsverteilung p von der Quelle X_i abhängt. Aus Gründen der Übersichtlichkeit ignoriert unsere Notation diesen Fakt.
Nach Konstruktion gibt es eindeutige $c \in B_i$ und $c_1, c_2 \in B_{i-1}$ mit $p(c) = p(c_1) + p(c_2)$ und $B_i \setminus \{c\} = B_{i-1} \setminus \{c_1, c_2\}$ für alle $2 \leq i \leq n$. Bemerke, dass die Summe zweier minimaler Wahrscheinlichkeiten $h_{i-1} = p(c) = p(c_1) + p(c_2)$ ist. Wir erhalten somit

$$\begin{aligned}
\overline{L}(C_{i-1}) &= \sum_{x \in B_{i-1}} p(x)\ell(x) \\
&= \sum_{x \in B_i \setminus \{c\}} p(x)\ell(x) + p(c_1)\ell(c_1) + p(c_2)\ell(c_2) \\
&= \sum_{x \in B_i \setminus \{c\}} p(x)\ell(x) + p(c_1)(\ell(c) + 1) + p(c_2)(\ell(c) + 1) \\
&= \sum_{x \in B_i} p(x)\ell(x) + p(c_1) + p(c_2) \\
&= \overline{L}(C_i) + h_{i-1}.
\end{aligned}$$

□

Bemerkung. *Wir bezeichnen den Schritt, beschrieben auf dem Aufgabenzettel, der aus dem Code C_i den Code C_{i-1} erzeugt, mit* Huffman-Schritt*.*

Lemma. *Seien $n \geq 2$ und $X_1 = \{x_1, \ldots, x_n\}$ eine n-elementige Quelle mit Wahrscheinlichkeitsverteilung p und nehme o.E. an, dass $p(x_1) \geq p(x_2) \geq \ldots \geq p(x_n)$. Sei nun C_2 ein optimaler Code der Quelle $X_2 = \{x_1, \ldots, x_{n-2}, x_{n-1}x_n\}$ mit Wahrscheinlichkeitsverteilung $p'(x_i) = p(x_i)$ für $1 \leq i \leq n-2$ und $p'(x_{n-1}x_n) = p(x_{n-1}) + p(x_n)$. So ist der Code C_1 der Quelle X_1 mit Wahrscheinlichkeitsverteilung p, der durch einen Huffman-Schritt entsteht, optimal.*

Beweis. Sei A_1 ein optimaler Code der Quelle X_1 mit Wahrscheinlichkeitsverteilung p. Wegen der Optimalität gilt folgende Aussage. Für je zwei Blätter x, y von A_1 mit $l(x) \geq l(y)$ gilt $p(x) \leq p(y)$. (Wenn dem nicht so ist, so tausche einfach im entsprechenden Baum x und y. Dies liefert einen kleineren Wert für $\overline{L}$). Mit anderen Worten, Blätter kurzer Pfade haben hohe Wahrscheinlichkeiten, wobei Blätter langer Pfade niedrige Wahrscheinlichkeiten haben.
Sei $c_1 \in X_1$ ein Blatt von A_1 maximaler Länge. So existiert wegen der Optimalität von A_1 ein weiteres Blatt c_2 mit demselben Vorgänger. Da die Längen von c_1 und c_2 maximal sind, sind ihre Wahrscheinlichkeiten minimal, d.h. $p(c_1) = p(x_n)$ und $p(c_2) = p(x_{n-1})$. Wir können ohne Einschränkung annehmen, dass $c_1 = x_n$ und $c_2 = x_{n-1}$, denn Blätter derselben Wahrscheinlichkeit dürfen im Baum getauscht werden und dabei bleibt die Optimalität des Codes erhalten. Sei nun A_2 der Code auf X_2 bzgl. p', der aus A_1 mittels eines Huffman-Schritts, der x_{n-1} und x_n zusammenlegt, entstanden ist. So gilt nach Aufgabe 2

$$\overline{L}(A_1) - \overline{L}(A_2) = p(x_{n-1}) + p(x_n) = \overline{L}(C_1) - \overline{L}(C_2),$$

was zu

$$\overline{L}(A_1) - \overline{L}(C_1) = \overline{L}(A_2) - \overline{L}(C_2)$$

äquivalent ist.
Da A_2 und C_2 Codes auf der Quelle X_2 bzgl. p' sind, folgt aus der Optimalität von C_2, dass $\overline{L}(C_2) \leq \overline{L}(A_2)$. Somit erhalten wir

$$\overline{L}(A_1) - \overline{L}(C_1) = \overline{L}(A_2) - \overline{L}(C_2) \geq 0.$$

Also gilt $\overline{L}(A_1) \geq \overline{L}(C_1)$, d.h. C_1 ist optimal. □

Aufgabe 3. Zeigen Sie, dass im Huffman-Code, die im i-ten Schritt entstandene Quelle X_i, durch C_i optimal kodiert ist.

Beweis. Wir beweisen die Aussage per vollständiger Induktion nach i. Sei zunächst $i = n = |X| = |X_1|$, so besteht der Code C_i nur aus einem Wort und somit ist $\overline{L}(C_i) = 0$ minimal. Sei nun $1 < i < n$, so ist nach Induktionsvoraussetzung C_i ein optimaler Code. Da C_{i-1} durch einen Huffman-Schritt aus C_i entstanden ist, ist C_{i-1} nach obigen Lemma optimal. □

Aufgabe 4. Es sei p die Wahrscheinlichkeitsverteilung auf den Quellen A und es sei $| A |= r$. Zeigen Sie

(a) $H(p) \leq \log_2(r)$ und

(b) $H(p)$ ist maximal, falls p die Gleichverteilung ist.

Beweis. (a)

Lemma (Vorlesung: Lemma (2.3))**.** *Seien $s_1, \ldots, s_r, y_1, \ldots, y_r$ positive reelle Zahlen, so gilt für $q \geq 2$*

$$0 \leq \sum_{i=1}^{r} y_i \log_q(\frac{y_i}{s_i})$$

und genau dann Gleichheit, wenn $s_i = y_i$ für alle $1 \leq i \leq r$.

Seien ohne Einschränkung $p_i = 0$ für alle $k+1 \leq i \leq r$ für ein $1 \leq k \leq r$. So gilt

$$\sum_{i=1}^{k} p_1 = 1 \geq \sum_{i=1}^{k} \frac{1}{r}$$

und damit nach obigen Lemma

$$0 \leq \sum_{i=1}^{k} p_i \log_2(p_i r) = \sum_{i=1}^{k} p_i \log_2(p_i) + \sum_{i=1}^{k} p_i \log_2(r) = -H(p) + \sum_{i=1}^{r} p_i \log_2(r) = -H(p) + \log_2(r)$$

(b) Ist nun p gleichverteilt, so ist in der Situation von Teilaufgabe (a) $k = r$ und $p_i = \frac{1}{r}$ für alle $1 \leq i \leq r$. Das obige Lemma liefert somit $H(p) = \log_2(r)$.

□