Zeichenkette endliche Folge von Symbolen aus $\Sigma$

Kode Funktion $C: \Sigma \to \{0,1\}^*$

Kodierung $s = s_1 s_2 \dots s_\ell$

$\to C(s) = C(s_1)\, C(s_2) \dots C(s_\ell)$

Wollen Kode, der (i) eindeutig ist

(ii) möglichst kurz

- Kode heißt präfixfrei / Präfixcode wenn kein Codewort Präfix eines anderen Kodeworts ist.
- Präfixfreie Kodes sind eindeutig dekodierbar
- Präfixfreie Kodes ensprechen Binärbaum

Blätter: Zeichen aus $\Sigma$

Pfade Wurzel $\to$ Blatt: Kodewörter

Problem Gegeben Alphabet $\Sigma = \{\sigma_1, \dots, \sigma_k\}$ mit Häufigkeiten $h_1, h_2, \dots, h_k$

(= # Vorkommen von $\sigma_i$ im Strings $s$)

finde optimalen präfixfreien Code $C: \Sigma \to \{0,1\}^*$, d.h.

präfixfreier Code $C$, so dass Gesamtlänge $C(s) = \sum_{\sigma \in \Sigma} \underbrace{|C(\sigma)|}_{\text{Länge von } C(s)} h_\sigma$

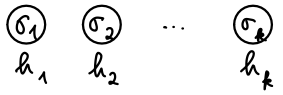
minimal ist

Idee: Konstruiere den Baum für $C$ gierig

Lege Knoten für jedes Symbol an.

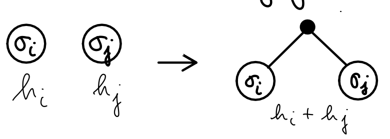
Annotiere jeden Knoten mit entsprechenden Häufigkeiten

Annotiere Knoten mit Häufigkeiten



Wähle zwei Knoten mit kleinsten Häufigkeiten, vereinige diese zu

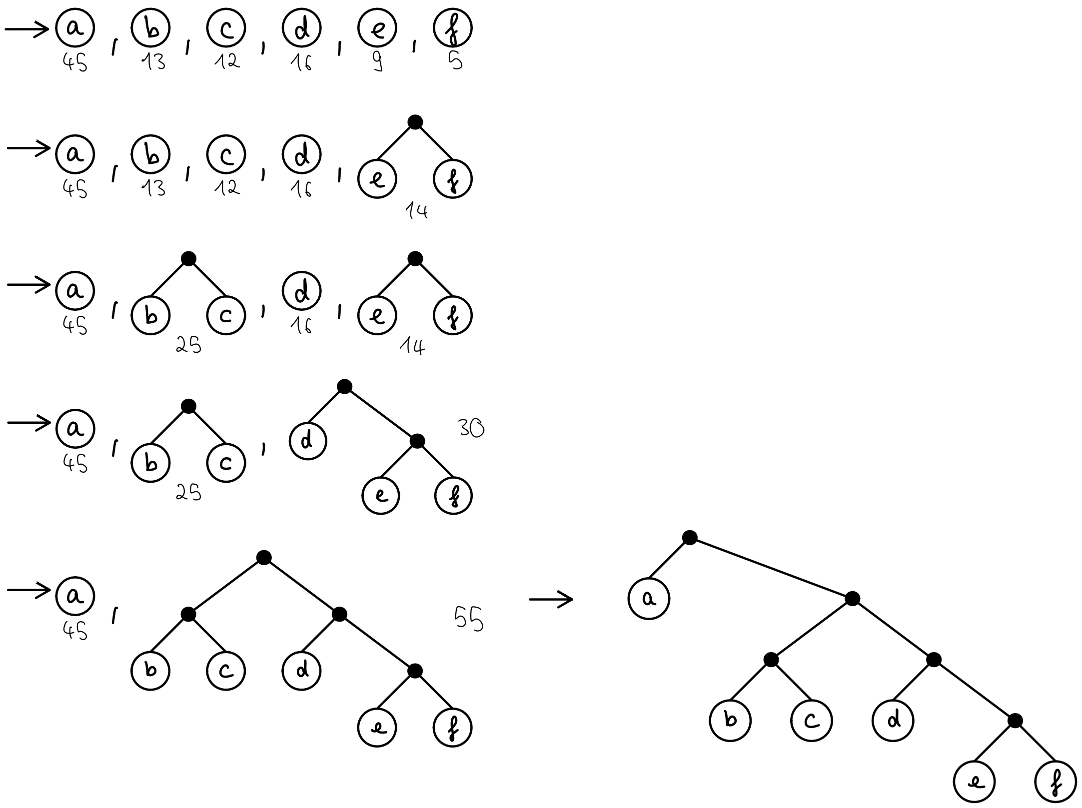
Teilbaum mit Häufigkeit $h_i + h_j$



Allgemein: Wähle Teilbäume mit kleinsten Häufigkeiten, vereinige diese, addiere Häufigkeiten

Wiederhole, bis nur ein Baum übrig ist. Dieser ist der Präfixcode.

<u>Bsp</u> a:45 b:13 c:12 d:16 e:9 f:5




$a \to 0,\ b \to 100,\ c \to 101,\ d \to 110,\ e \to 1110,\ f \to 1111$

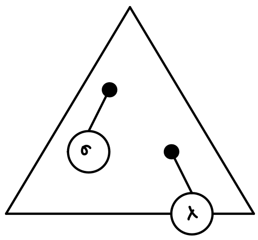
Dieser Kode heißt Huffman-Kode.

<u>Satz</u> Huffman-Kodes sind optimale Präfixcodes. D.h. für ein Alphabet $\Sigma = \{\sigma_1, \dots, \sigma_k\}$ und Häufigkeiten $h_{\sigma_1}, \dots, h_{\sigma_k}$ liefert der Algorithmus einen Kode $C$, so dass $\sum_{\sigma \in \Sigma} |C(\sigma)| h_\sigma$ minimal ist.

<u>Lemma 1</u> Sei $\sigma \in \Sigma$, so dass $h_\sigma$ minimal ist, und $\tau \in \Sigma$, so dass $h_\tau$ minimal in $\Sigma \setminus \{\sigma\}$ ist. Dann existiert ein optimaler Präfixcode für $\Sigma$ mit den gegebenen Häufigkeiten, so dass die Blätter für $\sigma$ und $\tau$ Geschwister sind und maximale Tiefe haben (kein Blatt ist tiefer).

Beweis Sei $C^*$ ein optimaler Präfixcode.

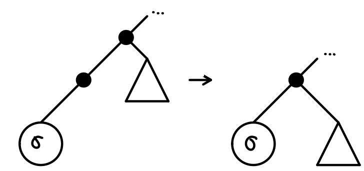
(i) Es existiert ein optimaler Präfixcode $C^{**}$, so dass das Blatt für $\sigma$ maximale Tiefe hat.




- Gilt für $C^* \to$ fertig ($C^{**} = C^*$)
- Sei $\lambda$ ein Zeichen mit maximaler Tiefe. Tausche $\sigma$ mit $\lambda$. Nach Wahl von $\sigma$ wird dadurch die Summe $\sum_{\sigma \in \Sigma} |C^*(\sigma)| h_\sigma$ nicht größer

(ii) Es existiert ein optimaler Präfixcode $C^{***}$, so dass die Blätter für $\sigma$ und $\tau$ Geschwister sind und maximale Tiefe haben.

- Gilt in $C^{**} \to$ fertig ($C^{***} = C^{**}$).
- Sonst muss $\sigma$ in $C^{**}$ Geschwister haben, sonst nicht optimal




Tausche Geschwister von $\sigma$ mit $\tau$. Nach Wahl von $\tau$ kann die Gesamtlänge nicht wachsen. $\square$

Beweis des Satzes

Induktion nach $k = |\Sigma|$

Basis: $k = 2$ ✓ (0, 1 genügt)

Schritt IA: HK ist optimal für alle Alphabete der Größe $k-1$ und alle Häufigkeiten

z.z. HK ist optimal für alle Alphabete der Größe $k$ und alle Häufigkeiten

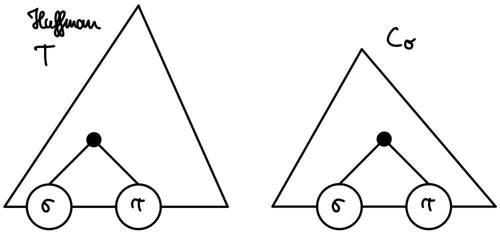
Nimm an: Es existieren $\Sigma = \{\sigma_1, \dots, \sigma_k\}$ und Häufigkeiten $h_{\sigma_1}, \dots, h_{\sigma_k}$, so dass HK nicht optimal ist für $h_{\sigma_1}, \dots, h_{\sigma_k}$.

Seien $\sigma$ und $\tau$ die Symbole, die HK zuerst vereinigt.

Nach L1 existiert ein optimaler Kode $C_\sigma$, in dem

$\sigma$ und $\tau$ Geschwister sind und maximale Tiefe haben.

<u>Also</u>




, so dass

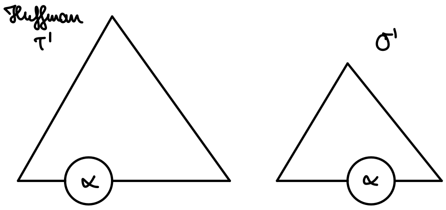
$$\sum_{\sigma\in\Sigma} |C_H(\sigma)| h_\sigma > \sum_{\sigma\in\Sigma} |C_\sigma(\sigma)| h_\sigma \quad (\text{HK nicht optimal})$$

$$\Leftrightarrow \sum_{\sigma\in\Sigma} (|T(\sigma)| - |O(\sigma)|) h_\sigma > 0$$

Sei $\alpha$ ein neues Symbol. Definiere Alphabet

$\Sigma' = \Sigma \setminus \{\sigma, \tau\} \cup \{\alpha\}$ mit $h_\alpha = h_\sigma + h_\tau$

Betrachte:




<u>Beh</u> $O'$ ist besser als $T'$

$$\sum_{\sigma\in\Sigma'} (|T'(\sigma)| - |O'(\sigma)|) h_\sigma = \sum_{\substack{\sigma\in\Sigma \\ \sigma\neq\sigma,\tau}} (|T(\sigma)| - |O(\sigma)|) h_\sigma + (|T'(\alpha)| - |O'(\alpha)|) h_\alpha$$

Es ist: $|T'(\alpha)| - |O'(\alpha)| = |T(\sigma)| - |O(\sigma)| = |T(\tau)| - |O(\tau)|$, da in den jeweiligen Bäumen $\sigma$ und $\tau$ immer auf der gleichen Ebene liegen und $\alpha$ auf der Elternebene.

Also gilt

$$(|T'(\alpha)| - |O'(\alpha)|) \underbrace{h_\alpha}_{h_\sigma + h_\tau} = (|T(\sigma)| - |O'(\sigma)|) h_\sigma + (|T(\tau)| - |O'(\tau)|) h_\tau$$

Also ist $\sum_{\sigma\in\Sigma'} (|T'(\sigma)| - |O'(\sigma)|) h_\sigma = \sum_{\sigma\in\Sigma} (|T(\sigma)| - |O(\sigma)|) h_\sigma > 0$, also ist $O'$ besser als $T'$ im Widerspruch zur I.A. ↯ $\square$