Codierungstheorie

MATTHIAS KRIESELL

Technische Universität Ilmenau

November 2017 — April 2020

Inhaltsverzeichnis

Kapitel 1

Präfixfreie Codes

1.1 Codes

Ein *Alphabet* A ist eine nichtleere endliche Menge von *Buchstaben*. Ein *Wort* der *Länge* ℓ über A ist eine Folge $w = (a_1, \ldots, a_\ell)$ mit $a_i \in A$ für alle $i \in \{1, \ldots, \ell\}$, wobei die Klammern und Kommata der Einfachheit halber fortgelassen werden, also $w = a_1 \ldots a_\ell$. Der Fall $\ell = 0$ ist zugelassen, das *leere Wort* wird meist mit ε bezeichnet. Die Menge aller Wörter der Länge n über A ist also gleich A^n, die Menge *aller* Wörter über A wird mit A^* bezeichnet, und ist $w \in A^*$ irgendein Wort, so bezeichnen wir mit $\ell(w)$ seine Länge.

Die *Konkatenation* (oder: Verkettung, Hintereinanderschreibung) zweier Wörter $u = a_1 \ldots a_\ell$ und $w = b_1 \ldots b_m$ ist das Wort $uw := a_1 \ldots a_\ell b_1 \ldots b_m$. In diesem Fall ist $\ell(uw) = \ell(u) + \ell(w)$. Betrachtet man die Konkatenation als eine Operation auf A^*, so ist sie nach Definition assoziativ, und das leere Wort ε ist ihr neutrales Element, also $\varepsilon w = w = w\varepsilon$ für alle $w \in A^*$.

Ein Wort $u \in A^*$ heißt *Präfix* des Wortes $w \in A^*$, falls es ein Wort $v \in A^*$ mit $uv = w$ gibt. Die Präfixe von $w = a_1 \ldots a_\ell$ sind also genau die Wörter $a_1 \ldots a_i$ mit $i \in \{0, \ldots, \ell\}$. Denkt man sich die Elemente aus A^* als Ecken eines gerichteten Graphen und zieht eine Kante von u nach w, wenn u das Präfix der Länge $\ell(w) - 1$ von w ist, so entsteht ein (unendlicher, gerichteter) Baum T mit Wurzel ε. Der Außengrad an jeder Ecke ist $|A|$, der Innengrad von ε ist 0, jedes andere Wort hat Innengrad 1. Die Präfixe des Wortes w sind genau die Ecken des ε, w-Weges in T. Man nennt T auch den *Baum aller Wörter über* A.

Ein *Code über* A ist eine endliche Teilmenge C von A^*. Der von C induzierte Baum ist der durch C und alle Präfixe aller Wörter aus C im Baum aller Wörter über A induzierte Teilbaum. C heißt *präfixfrei*, falls kein Wort in C Präfix eines anderen Wortes in C ist. Offensichtlich ist ein Code genau dann präfixfrei, wenn die Wörter aus C genau die *Blätter* des von C induzierten Baumes sind

(das heißt die Ecken mit Außengrad 0). Die Blätter eines endlichen Teilbaumes des Baumes aller Wörter über A, der ε enthält, bilden einen präfixfreien Code.

Satz 1 *Sei C ein präfixfreier Code über dem Alphabet A. Dann gilt* $\sum_{w \in C} \frac{1}{|A|^{\ell(w)}} \leq 1$.

Beweis. Für $C = \emptyset$ ist die Behauptung klar. Für alle nichtleeren Codes C ist $n(C) := \max\{\ell(w) : w \in C\}$ definiert und wir beweisen die Behauptung induktiv über $n(C)$. Im Fall $n(C) = 0$ ist $C = \{\varepsilon\}$ und die Behauptung richtig. Für $n(C) > 0$ entstehe C' aus C durch Ersetzen aller Wörter der Länge $n(C)$ durch ihr Präfix der Länge $n(C) - 1$. Die Menge D all dieser Präfixe enthält keine Wörter aus C, und C entsteht aus C' durch Ersetzen jedes Wortes w' aus D durch bis zu $|A|$ Wörter aus C der Länge $n(C) = \ell(w') + 1$. Also gilt mit Induktionsvoraussetzung

$$\sum_{w \in C} \frac{1}{|A|^{\ell(w)}} \leq \sum_{w' \in C \cap C'} \frac{1}{|A|^{\ell(w')}} + \sum_{w' \in D} |A| \cdot \frac{1}{|A|^{\ell(w')+1}} = \sum_{w' \in C'} \frac{1}{|A|^{\ell(w')}} \leq 1.$$

□

Sei $p : \mathbb{N} \to \mathbb{N}$. Der *Träger* von p ist die Menge $\{j \in \mathbb{N} : p(j) \neq 0\}$. Für p mit endlichem Träger heißt

$$\sum_{j \in \mathbb{N}} \frac{p(j)}{|A|^j} \leq 1$$

die *Kraftsche Ungleichung*. Ist C ein Code, so heißt die durch $p(j) := |\{w \in C : \ell(w) = j\}|$ definierte Funktion $p : \mathbb{N} \to \mathbb{N}$ das *Längenprofil* von C. Satz 1 besagt demnach, daß das Längenprofil eines präfixfreien Codes die Kraftsche Ungleichung erfüllt. Der folgende Satz besagt, daß die Kraftsche Ungleichung auch hinreichend für die Existenz eines Codes mit vorgegebenem Längenprofil ist. Sie charakterisiert also die möglichen Längenprofile präfixfreier Codes unter allen natürlichen Folgen.

Satz 2 *Sei A ein Alphabet und $p : \mathbb{N} \to \mathbb{N}$ mit endlichem Träger erfülle die Kraftsche Ungleichung. Dann gibt es einen präfixfreien Code über A mit Längenprofil p.*

Beweis. Wir bauen den Code C zu gegebenem p schrittweise auf, indem wir nacheinander jeweils alle Wörter der Länge $0, 1, 2, \ldots$ festlegen. Zu jedem Zeitschritt t sei a_t die Anzahl der Wörter der Länge t, die kein Präfix unter den bisher gewählten Wörtern haben, also $a_0 = 1$. Wir wählen unter diesen Wörtern $p(t)$ viele aus. Die übrigen $a_t - p(t)$ Wörter verlängern wir um einen Buchstaben, wodurch sich genau $a_{t+1} := |A| \cdot (a_t - p(t))$ Wörter der Länge $t+1$ ergeben, die kein Präfix unter den bisher gewählten Wörtern haben. Zunächst erhalten wir eine Rekursionsgleichung für die a_t in ganzen Zahlen, aus der wir durch Induktion $a_t - p(t) = |A|^t \cdot (1 - \sum_{j=0}^{t} \frac{p(j)}{|A|^j})$ gewinnen. Da p die Kraftsche Ungleichung erfüllt, ist die rechte Seite stets größer oder gleich Null, so daß die Auswahl von $p(t)$ Wörtern tatsächlich in jedem Schritt möglich ist. □

Aus dem vorangegangenenen Beweis ergibt sich zugleich eine Methode, einen präfixfreien Code zu vorgegebenem Längenprofil zu konstruieren.

Ein Code C über A heißt *eindeutig entzifferbar*, wenn für alle Wörter $u_1, \ldots, u_r$, $w_1, \ldots, w_s \in C$ aus $u_1 \ldots u_r = w_1 \ldots w_s$ stets $r = s$ und $u_i = w_i$ für alle $i \in \{1, \ldots, r\}$ folgt.

Satz 3 *Jeder präfixfreie Code $C \neq \{\varepsilon\}$ ist eindeutig entzifferbar.*

Beweis. Wir halten uns an die Definition und führen Induktion über $\ell(u_1 \ldots u_r)$. Weil $C \neq \{\varepsilon\}$ ist, folgt $\varepsilon \notin C$, also impliziert $\ell(u_1 \ldots u_r) = 0$ stets $r = 0$, was den Induktionsanfang sichert. Es gilt $\ell(u_1) = \ell(w_1)$, denn andernfalls sind u_1, w_1 verschieden und u_1 Präfix von w_1 oder w_1 Präfix von u_1, was der Voraussetzung an C widerspricht. Hieraus folgt $u_1 = w_1$, also auch $u_2 \ldots u_r = w_2 \ldots w_s$, und durch Induktion folgt $r - 1 = s - 1$ und $u_i = w_i$ für alle $i \in \{2, \ldots, r\}$. □

Die Umkehrung von Satz 3 gilt nicht: Es gibt eindeutig entzifferbare Codes, die nicht präfixfrei sind (Übung), jedoch erfüllen ihre Längenprofile ebenfalls die Kraftsche Ungleichung:

Satz 4 *Das Längenprofil p eines eindeutig entzifferbaren Codes C über dem Alphabet A erfüllt die Kraftsche Ungleichung.*

Beweis. Weil C eindeutig entzifferbar ist, sind aus verschiedenen Codewortfolgen zusammengesetzte Wörter aus A^* verschieden. Es gibt daher genau $p(\ell_1) \cdot \cdots \cdot p(\ell_k)$ verschiedene Wörter aus A^* (nicht weniger) der Form $w = u_1 \ldots u_k$ mit $u_i \in C$ und $\ell(u_i) = \ell_i$ für alle $i \in \{1, \ldots, k\}$. Somit ist die Zahl

$$d(k, \ell) := \sum_{\substack{(\ell_1, \ldots, \ell_k) \in \mathbb{N}^k \\ \ell_1 + \cdots + \ell_k = \ell}} p(\ell_1) \cdot \cdots \cdot p(\ell_k)$$

die Anzahl aller Wörter der Länge ℓ über A, die sich als Konkatenation von k Wörtern aus C darstellen lassen. Natürlich gilt $d(k, \ell) \leq |A^\ell| = |A|^\ell$, und so

folgt mit $L := \max\{j : p(j) \neq 0\}$

$$\begin{aligned}
(\sum_{\ell\in\mathbb{N}} \frac{p(\ell)}{|A|^\ell})^k &= (\sum_{\ell=0}^{L} \frac{p(\ell)}{|A|^\ell})^k \\
&= \sum_{\substack{(\ell_1,\ldots,\ell_k)\in\mathbb{N}^k \\ 0\le\ell_1,\ldots,\ell_k\le L}} \prod_{i=1}^{k} \frac{p(\ell_i)}{|A|^{\ell_i}} \\
&= \sum_{(\ell_1,\ldots,\ell_k)\in\mathbb{N}^k} \prod_{i=1}^{k} \frac{p(\ell_i)}{|A|^{\ell_i}} \\
&= \sum_{\ell=0}^{kL} \sum_{\substack{(\ell_1,\ldots,\ell_k)\in\mathbb{N}^k \\ \ell_1+\cdots+\ell_k=\ell}} \prod_{i=1}^{k} \frac{p(\ell_i)}{|A|^{\ell_i}} \\
&= \sum_{\ell=0}^{kL} \frac{d(k,\ell)}{|A|^\ell} \\
&\le \sum_{\ell=0}^{kL} 1 \\
&= kL.
\end{aligned}$$

(Für die zweite Identität rufe man sich in Erinnerung, wie man *allgemein ausmultipliziert*, also ein Produkt von Summen als Summe von Produkten darstellt.) Also gilt $\sum_{\ell\in\mathbb{N}} \frac{p(\ell)}{|A|^\ell} \le \sqrt[k]{kL}$ für alle $k \in \mathbb{N}$ und darum $\sum_{\ell\in\mathbb{N}} \frac{p(\ell)}{|A|^\ell} \le \lim_{k\to\infty} \sqrt[k]{kL} = 1$, also $\sum_{\ell\in\mathbb{N}} \frac{p(\ell)}{|A|^\ell} \le 1$. □

Aus den vorangegangenen Sätzen gewinnen wir:

Satz 5 *Sei A ein Alphabet und $p : \mathbb{N} \to \mathbb{N}$ mit endlichem nichtleeren Träger, $p \neq (1, 0, 0, \ldots)$. Dann sind äquivalent:*

(i) Es gibt einen präfixfreien Code über A mit Längenprofil p,

(ii) Es gibt einen eindeutig entzifferbaren Code über A mit Längenprofil p,

(iii) p erfüllt die Kraftsche Ungleichung.

Beweis. Der Ringschluß gelingt mit Satz 3, Satz 4 und Satz 1. □

1.2 Codierungen

Bislang haben wir nur „innere" Eigenschaften von Codes betrachtet und den Vorgang des Codierens außen vor gelassen. Um die Zielsetzung der Quellencodierung formulieren zu können, werden wir den Horizont weiten: Eine *Codierung* eines Alphabets Q über dem Alphabet A ist eine Injektion $f : Q \to A^*$. Sie heißt *eindeutig entzifferbar*, wenn *der von f erzeugte Code* $f(Q) = \{f(x) : x \in Q\}$ eindeutig entzifferbar ist. Da es immer darum geht, nicht nur einzelne Buchstaben, sondern ganze Texte zu codieren, setzen wir f zu einer Abbildung $f^* : Q^* \to A^*$ fort, indem wir $f^*(x_1 \dots x_\ell) := f(x_1) \dots f(x_\ell)$ für $x_1, \dots, x_\ell \in Q$, $\ell \geq 0$ definieren. Erwartungsgemäß kommt:

Satz 6 *Eine Codierung f ist genau dann eindeutig entzifferbar, wenn f^* injektiv ist.*

Beweis. Sei f eine Codierung von Q über A. Ist f eindeutig entzifferbar und gilt $f(x_1 \dots x_r) = f(y_1 \dots y_s)$, also $f(x_1) \dots f(x_r) = f(y_1) \dots f(y_s)$, so folgt $r = s$ und $f(x_i) = f(y_i)$ für alle $i \in \{1, \dots, r\}$. Aus der Injektivität von f folgt $x_i = y_i$ für alle $i \in \{1, \dots, r\}$, also $x_1 \dots x_r = y_1 \dots y_s$. Dies zeigt, daß f^* injektiv ist. Sei nun umgekehrt f^* injektiv und $u_1 \dots u_r = w_1 \dots w_s$ für gewisse u_i, w_j aus $f(Q)$. Seien $x_1, \dots, x_r$ und $y_1, \dots, y_s$ die Urbilder der $u_1, \dots, u_r$ bzw. $w_1, \dots, w_s$ in Q. Dann ist $f(x_1 \dots x_r) = f(x_1) \dots f(x_r) = u_1 \dots u_r = w_1 \dots w_s = f(y_1) \dots f(y_s) = f(y_1 \dots y_s)$. Wegen der Injektivität von f^* sind die Wörter $x_1 \dots x_r = y_1 \dots y_s$ gleich, also sind sie gleichlang und stimmen buchstabenweise überein, d. h. $r = s$ und $x_i = y_i$ für $i \in \{1, \dots, r\}$, also auch $f(x_i) = f(y_i)$ und somit $u_i = w_i$ für alle $i \in \{1, \dots, r\}$. Dies zeigt, daß $f(Q)$ und damit auch f eindeutig entzifferbar ist. □

Fortan wollen wir uns das „Quellalphabet" Q mit einem Wahrscheinlichkeitsmaß p versehen denken (eine sogenannte „diskrete Quelle ohne Gedächtnis"). Wir sprechen dann von einer *Quelle Q mit p*, im Fall $|Q| = 1$ von einer *trivialen* Quelle. Für eine Codierung f von Q über A nennen wir den Erwartungswert der Zufallsvariablen $\ell \circ f$, also $L(f) = \sum_{x \in Q} p(x)\ell(f(x))$, die *mittlere Codewortlänge*. Unser Ziel wird es sein, die mittlere Codewortlänge möglichst zu minimieren, und zwar bei festem A (für $|A| \geq |Q|$ ist stets $L(f) = 1$ zu erreichen). Wie gut man hier sein kann, hängt natürlich zuallererst von der Größe von Q ab, aber ganz erheblich auch vom Maß p.

Unabhängig von einer konkreten Codierung $f : Q \to A$ wollen wir zunächst den *Informationsgehalt* bzw. *Überraschungswert* $I_A(x)$ eines Zeichens $x \in Q$ über dem Alphabet A beziffern. Da die Elemente in Q nur durch ihre Wahrscheinlichkeit abgegrenzt sind, hängt $I_A(x)$ nur von $p(x)$ ab, und wir können uns einfach die Frage stellen, welche Wortlänge wir zum Ausdrücken von $p \in [0, 1]$ über dem Alphabet A benötigen: Im Fall $p = 1/n$ ist dies $\log_{|A|}(n) = -\log_{|A|}(p)$, und so lassen wir uns inspirieren und definieren

$$I_A(x) := -\log_{|A|}(p(x)).$$

Tritt x sicher (mit Wahrscheinlichkeit 1) ein, so birgt dies keine Überraschung, der Informationsgehalt ist dann Null. Mit kleiner werdender Wahrscheinlichkeit überrascht das Eintreten von x mehr und mehr, und das Eintreten des unmöglichen Ereignisses ist überraschend ohnegleichen.

Der erwartete (mittlere) Informationsgehalt heißt *Entropie*, definiert durch

$$H_A(Q) = \sum_{x \in Q} p(x) I_A(x).$$

Dabei werden die Terme mit $p(x) = 0$ einfach fortgelassen. (Dies ist auch wegen $\lim\limits_{p \to 0} p \cdot (-\log_{|A|}(p)) = 0$ sinnvoll.)

Satz 7 *Sei Q mit p Quelle, A ein Alphabet und f eine eindeutig entzifferbare Codierung von Q über A. Dann gilt $H_A(Q) \leq L_A(Q, f)$.*

Beweis. In der Analysis zeigt man: $\ln x \leq x - 1$ für $x > 0$, also $\log_{|A|}(x) = \frac{\ln x}{\ln |A|} \leq \frac{1}{\ln |A|}(x-1)$. Hieraus folgt

$$\begin{aligned}
H_A(Q) - L_A(Q,f) &= \sum_{x \in Q} p(x)(-\log_{|A|}(p(x)) - \ell(f(x))) \\
&= \sum_{x \in Q} p(x)\big(\log_{|A|}\big(\frac{1}{p(x)}\big) - \log_{|A|}(|A|^{\ell(f(x))})\big) \\
&= \sum_{x \in Q} p(x) \log_{|A|}\big(\frac{1}{p(x)|A|^{\ell(f(x))}}\big) \\
&\leq \sum_{x \in Q} p(x) \frac{1}{\ln |A|}\big(\frac{1}{p(x)|A|^{\ell(f(x))}} - 1\big) \\
&= \frac{1}{\ln |A|} \sum_{x \in Q} \big(\frac{1}{|A|^{\ell(f(x))}} - p(x)\big) \\
&= \frac{1}{\ln |A|}\big(\sum_{x \in Q} \frac{1}{|A|^{\ell(f(x))}} - \sum_{x \in Q} p(x)\big) \\
&= \frac{1}{\ln |A|}\big(\sum_{x \in Q} \frac{1}{|A|^{\ell(f(x))}} - 1\big).
\end{aligned}$$

Die Kraftsche Ungleichung liefert $\sum\limits_{x \in Q} \frac{1}{|A|^{\ell(f(x))}} \leq 1$ und damit die Behauptung.

□

Satz 8 *Sei Q mit p nichttriviale Quelle und A ein Alphabet. Dann gibt es eine eindeutig entzifferbare Codierung f von Q über A mit $L_A(Q, f) < H_A(Q) + 1$.*

Beweis. Für $x \in Q$ mit $p(x) > 0$ sei $\ell(x)$ die ganze Zahl aus dem halboffenen Intervall $[\, I_A(x), I_A(x) + 1\,)$. Wegen $\ell(x) \geq I_A(x) = -\log_{|A|} p(x)$ kommt

$|A|^{-\ell(x)} \leq |A|^{\log_{|A|} p(x)} = p(x)$, also $\sum_{x\in Q} \frac{1}{|A|^{\ell(x)}} \leq \sum_{x\in Q} p(x) = 1$. Definiert man $k(\ell) := |\{x \in Q : \ell(x) = \ell\}|$, so erfüllt folglich $k : \mathbb{N} \to \mathbb{N}$ die Kraftsche Ungleichung. Daher gibt es einen eindeutig entzifferbaren Code C über A mit Längenprofil k und wir können daraus sofort eine Bijektion $f : Q \to C$ mit $\ell(f(x)) = \ell(x)$ gewinnen. $f : Q \to A^*$ ist dann eine eindeutig entzifferbare Codierung, und es gilt $L_A(Q, f) = \sum_{x\in Q} p(x)\ell(f(x)) = \sum_{x\in Q} p(x)\ell(x) < \sum_{x\in Q} p(x)(I_A(x) + 1) = H_A(Q) + 1$. □

Seien wieder Q mit p eine Quelle und A ein Alphabet. Wir behaupten, daß es stets eine eindeutig entzifferbare Codierung von Q über A mit *minimaler* mittlerer Codewortlänge gibt. Grundsätzlich gibt es natürlich unendlich viele eindeutig entzifferbare Codierungen von Q über A mit beliebig großer mittlerer Codewortlänge und denkbar wäre es, daß die Menge aller realisierbaren mittleren Codewortlängen zwar ein Infimum hat, dieses jedoch selber nicht durch eine („optimale") Codierung realisierbar wäre. Wir werden aber jetzt zeigen, daß es zu jeder eindeutig entzifferbaren Codierung f nur endlich viele eindeutig entzifferbare Codierungen gibt, die eine höchstens ebenso große mittlere Codewortlänge wie f besitzt, woraus die Behauptung folgt. Sei dazu $q := \min\{p(x) : x \in Q, p(x) > 0\}$. Sei weiter f irgendeine eindeutig entzifferbare Codierung von Q über A und $m := L_A(Q, f)$. Für eine eindeutig entzifferbare Codierung g von Q über A mit $\ell(g(y)) > m/q$ für ein $y \in Q$ folgt $L_A(Q, g) = \sum_{x\in Q} p(x)\ell(g(x)) \geq p(y)\ell(g(y)) > m = L_A(Q, f)$. Folglich sind die Wortlängen der Codes all derjenigen eindeutig entzifferbaren Codierungen g von Q über A mit $L_A(Q, g) \leq L_A(Q, f)$ durch m/q beschränkt. Da es nur endlich viele Teilmengen von A^* mit Wortlängen höchstens m/q gibt und jeweils nur endlich viele injektive Abbildungen der endlichen Menge Q in eine solche Menge, gibt es nur endlich viele eindeutig entzifferbare Codierungen g von Q über A mit $L_A(Q, g) \leq L_A(Q, f)$. Hieraus folgt, daß die Menge

$$\{L_A(Q, f) : f \text{ eindeutig entzifferbare Codierung von } Q \text{ über } A\} \qquad (1)$$

ein Minium besitzt.

Eine Codierung f von Q über A heißt *präfixfrei*, wenn $f(Q)$ präfixfrei ist; außer im Fall $f(Q) = \{\varepsilon\}$ ist eine solche präfixfreie Codierung wegen Satz 3 auch eindeutig entzifferbar. Ist umgekehrt f eindeutig entzifferbar so gibt es wegen Satz 5 einen präfixfreien Code $C \subseteq A^*$ mit demselben Längenprofil wie $f(Q)$. Es gibt dann eine Bijektion $\phi : f(Q) \to C$ mit $\ell(\phi(w)) = \ell(w)$ für alle $w \in f(Q)$, so daß $g := \phi \circ f : Q \to A^*$ eine präfixfreie Codierung von Q über A mit $\ell(g(x)) = \ell(\phi(f(x))) = \ell(f(x))$ ist. Infolgedessen kommt $L_A(Q, g) = L_A(Q, f)$, die Menge aus (1) ist also — außer im Fall einer trivialen Quelle — gleich

$$\{L_A(Q, f) : f \text{ präfixfreie Codierung von } Q \text{ über } A\}. \qquad (2)$$

Ihr Minimum wird mit $L_A(Q)$ bezeichnet und heißt *minimale mittlere Codewortlänge* (einer präfixfreien Codierung von Q über A).

Satz 9 *Sei Q mit p nichttriviale Quelle und A ein Alphabet. Dann gilt $H_A(Q) \leq L_A(Q) < H_A(Q) + 1$.*

Beweis. Die beiden Ungleichungen folgen aus Satz 7 bzw. Satz 8. □

1.3 Optimale Codierungen

Eine präfixfreie Codierung $f : Q \to A^*$ mit $L_A(Q, f) = L_A(Q)$ wollen wir *optimal* nennen.

Lemma 1 *Sei A ein Alphabet und Q mit p eine Quelle der Größe $|Q| = 1+n\cdot(|A|-1)$ und S eine $|A|$-elementige Menge mit minimalem $p(S)$. Dann gibt es eine optimale Codierung f von Q über A derart, daß alle Ecken des von $f(Q)$ induzierten Baumes 0 oder $|A|$ Außennachbarn haben und alle Wörter aus $f(S)$ die Länge $m := \max\{\ell(w) : w \in f(Q)\}$ und dasselbe Präfix der Länge $m-1$.*

Beweis. Sei f irgendeine optimale Codierung von Q über A, wobei zusätzlich $N(f) := \sum_{x\in Q,\, p(x)=0} \ell(f(x))$ möglichst klein sei. Sei $m := \max\{\ell(w) : w \in f(Q)\}$ die maximale Codewortlänge.

Sei T der von $f(Q)$ induzierte Teilbaum im Baum aller Wörter über A. Jede Ecke der Länge m ist ein Blatt und hat daher Außengrad 0. Solange es zwei (oder mehr) Ecken u, w der Länge $m-1$ mit von 0 und $|A|$ verschiedenem Außengrad r bzw. s gibt, können wir $d := \min\{|A| - r, s\}$ verschiedene Außennachbarn von w wählen — etwa mit Urbild $q_1, \ldots, q_d$ unter f — und durch d noch nicht in C enthaltene Wörter $ua_1, \ldots, ua_q$ ersetzen, also $f(q_j) := ua_j$ redefinieren. Auch T wird redefiniert, die Anzahl der Ecken der Länge $m-1$ mit von 0 und $|A|$ verschiedenem Außengrad sinkt, doch bleiben die $\ell(f(q))$ und somit $L_A(Q, f)$ unverändert. So dürfen wir annehmen, daß alle bis auf höchstens eine Ecke der Länge $m-1$ Außengrad 0 oder k haben. Jede Ecke u der Länge $k < m-1$ hat den Außengrad 0 oder $|A|$, denn sonst wäre u kein Blatt und Präfix eines nicht in T enthaltenen Wortes ua der Länge $k+1$, und wir können ein Wort w maximaler Länge, etwa mit Urbild q unter f, durch ua ersetzen, also $f(q) := ua$ redefinieren; dadurch verändert sich $L_A(Q, f)$ um $-p(w)m + p(w)(k+1)$, bleibt also unverändert oder sinkt. Ersteres impliziert $p(w) = 0$, was der Minimalität von $N(f)$ widerspricht, letzteres steht im Widerspruch zur Optimalität von f.

Hätte nun T eine Ecke w des Außengrades $r \notin \{0, |A|\}$, so hat w die Länge $m-1$ und alle Außennachbarn von w sind Blätter. Der Baum, der durch Löschung dieser r Blätter aus T entsteht, hat bei jeder Ecke den Außengrad 0 oder $|A|$; induktiv zeigt man, daß die Anzahl der Blätter eines solchen Baumes kongruent 1 modulo $|A|-1$ und die Anzahl der Blätter von T darum kongruent $1 + r - 1 = r$ modulo $|A|-1$ ist, woraus $r = 1$ mit Voraussetzung folgt. Ist nun u der einzige Außennachbar von w, etwa mit Urbild q unter f, so können wir u durch w ersetzen, also $f(q) := w$ redefinieren; dadurch verändert sich $L_A(Q, f)$ um $-p(w)m + p(w)(m-1)$, bleibt also unverändert oder sinkt, was abermals widersprüchlich ist.

Also hat *jede* Ecke von T Außengrad 0 oder $|A|$; insbesondere gibt es eine $|A|$-elementige Menge $R_f \subseteq Q$ von Buchstaben deren Bilder unter f die Länge m haben und die alle dasselbe Präfix der Länge $m-1$ haben.

Wir verändern jetzt f so, daß S möglichst viele Elemente aus R_f enthält. Gäbe es ein $x \in S \setminus R_f$ mit $\ell(u := f(x)) =: k \leq m$, so gibt es ein $z \in R_f \setminus S$, und für dieses gilt $p(z) \geq p(x)$, weil sonst $S' := (S \setminus \{x\}) \cup \{z\}$ der Minimalität von $p(S)$ widerspricht. Dann aber verändert sich durch $L_A(Q, f)$ durch Redefinition $f(x) := w$, $f(z) := u$ um $-p(x)k - p(z)m + p(x)m + p(z)k = (p(x) - p(z))(m-k) \leq 0$: Die Optimalität des ursprünglichen f impliziert also die Optimalität des redefinierten f und $(p(x)-p(z))(m-k) = 0$. Im Falle $m = k$ bleibt $N(f)$ unverändert, da die Längen von $f(x), f(z)$ unverändert sind. Andernfalls ist $p(x) = p(z)$ und die Summe $\ell(f(x)) + \ell(f(z))$ trägt unverändert zu $N(f)$ bei (oder nicht bei). Also bleibt $N(f)$ in jedem Fall unverändert. Zum redefinierten R_f ist ein weiteres Element aus S hinzugekommen, nämlich x. Diesen Schritt können wir wiederholen, bis schließlich alle Elemente aus S in R_f enthalten sind. □

1.4 Huffman-Codierung

Sei A ein Alphabet. Zu allen Quellen Q mit p der Größe $|Q| = 1 + n \cdot (|A| - 1)$ beschreiben wir rekursiv eine Codierung $h : Q \to A^*$ wie folgt.

Für $n = 0$ ist $Q = \{x\}$ und wir setzen wir $h(x) = \varepsilon$.

Für $n > 0$ ist $|Q| \geq |A|$ und wir wählen eine $|A|$-elementige Menge S mit minimalem $p(S)$. Q' entsteht aus $Q \setminus S$ durch Hinzufügen eines neuen Elements s, p' aus $p|Q'$ durch die Erweiterung $p'(s) := p(S)$. Rekursiv erhalten wir $h' : Q' \to A^*$, und h entsteht aus $h'|Q\setminus\{s\}$ durch die Erweiterung $h(x) := h'(s)\alpha(x)$ für alle $x \in S$, wobei $\alpha : S \to A$ irgendeine Bijektion sei.

Eine auf diese (nicht vollständig deterministische) Weise hergestellte Codierung h heißt *Huffman*-Codierung.

Man überlegt sich leicht, daß h präfixfrei ist (Übung). Praktisch führt man das Verfahren durch, indem man die Wahrscheinlichkeiten absteigend ordnet, die $|A|$ letzten Einträge der Liste ausstreicht, deren Summe jedoch als neuen Eintrag in die Restliste einordnet nebst $|A|$ vielen Kanten zu den gerade Ausgestrichenen, und iteriert bis zur 1. Der „Zusammenfassungsbaum" kann dann (ab der Wurzel beliebig) in den Baum aller Wörter eingebettet werden und liefert den Code bzw. die Codierung.

Satz 10 *Sei Q eine Quelle mit p und $|Q| = 1 + n \cdot (|A| - 1)$. Dann ist jede Huffman-Codierung von Q über A optimal.*

Beweis. Die Behauptung ist offensichtlich richtig für $n \leq 1$. Für $n > 1$ entstehe

die Huffman-Codierung h durch Wahl einer $|A|$-elementigen Menge S mit minimalem $p(S)$: Q' entsteht aus $Q \setminus S$ durch Hinzufügen eines neuen Elements s, p' aus $p|Q'$ durch die Erweiterung $p'(s) := p(S)$. Rekursiv erhalten wir die Huffman-Codierung $h' : Q' \to A^*$, und h entsteht aus $h'|(Q \setminus \{s\})$ durch die Erweiterung $h(x) := h'(s)\alpha(x)$ für alle $x \in S$, wobei $\alpha : S \to A$ irgendeine Bijektion sei. Dann ist $L_A(Q,h) = L_A(Q,h') + \delta$ mit $\delta := -p'(s)\ell(h'(s)) + \sum_{x\in S} p(x)\ell(h(x)) = -p'(s)\ell(h'(s)) + \sum_{x\in S} p(x)(\ell(h'(s)) + 1) = p(S)$.

Wir wollen h mit einer optimalen Codierung f wie im Lemma vergleichen. Sei dazu u das gemeinsame Präfix der Länge $m-1$ der Wörter aus $p(S)$. Wir definieren $f' : Q' \to A^*$ durch $f'(x) := f(x)$ für alle $x \in Q \setminus S$ und $f'(s) := u$. Dann ist auch $L_A(Q,f) = L_A(Q,f') + \delta$ (gleiche Rechnung wie oben für f, f' statt h, h'). Wegen der durch die Induktionsvoraussetzung gegebenen Optimalität von h' folgt $L_A(Q,f) \le L_A(Q,h) = L_A(Q',h') + \delta \le L_A(Q',f') + \delta = L_A(Q,f)$, also $L_A(Q,h) = L_A(Q,f)$, und somit ist auch h optimal. □

In Satz 10 wird verlangt, daß das Quellalphabet Q mit p die Größe $1 + n(|A|-1)$ für ein geeignetes n hat, wobei A das Codealphabet ist. Man kann dies für ein Q beliebiger Größe sicherstellen, indem man Q mit (höchstens $|A|-2$) Dummyzeichen auffüllt; diese treten mit Wahrscheinlichkeit 0 auf, die Wahrscheinlichkeiten der Zeichen aus Q bleiben unverändert. So entstehe die Quelle Q' mit p'. Sei h' eine Huffman-Codierung von Q' mit p' über A. Die Einschränkung von h' auf Q, $h := h'|Q$, ist dann eine optimale Codierung von Q über A (Übung).

1.5 Übungen

1. Sei A ein Alphabet ungerader Größe und T ein endlicher Teilbaum des Baumes aller Wörter über A mit $\varepsilon \in V(T)$, in dem jede Ecke außer den Blättern Außengrad $|A|$ hat. Man zeige:

   (i) $|T|$ besitzt ungeradzahlig viele Blätter.

   (ii) Das Längenprofil des von den Blättern von T gebildeten präfixfreien Codes erfüllt die Kraftsche Ungleichung mit Gleichheit.

2. Man beschreibe die präfixfreien Codes für deren Längenprofil Gleichheit in der Kraftschen Ungleichung gilt (zum Beispiel mit Hilfe der von ihnen induzierten Bäume oder durch eine Maximalitätseigenschaft).

3. Man zeige: Der Code $\{11, 1110\}$ ist eindeutig entzifferbar. Ist er präfixfrei?

4. Gibt es einen präfixfreien Code über $A = \{0,1\}$ mit sechs Codewörtern der Länge (a) $1,3,3,3,3,3$ bzw. (b) $2,3,3,3,3,3$? Wenn ja, gebe man einen solchen Code an.

5. Sei $r \ge 2$. Zur Codierung eines 26-buchstabigen Alphabets X soll ein präfixfreier Code C über einem r-buchstabigen Alphabet A verwendet

werden, wobei alle Codewörter aus C die gleiche Länge haben sollen. (a) Geht dies mit $r = 2$ und $\ell = 4$? (b) In Abhängigkeit von r bestimme man das kleinste ℓ, für welches ein solcher Code C existiert; geben Sie $\ell(2)$ und $\ell(3)$ an.

6. Zu gegebenem n sei $p : \mathbb{N} \to \mathbb{N}$ definiert durch $p(j) = 1$ für $j \in \{1, \ldots, n\}$ und $p(j) = 0$ sonst. Man konstruiere einen präfixfreien Code C über $A = \{0, 1\}$ mit Längenprofil p.

7. Zeigen Sie, daß $L_A(Q)$ und $H_A(Q)$ nur von Q mit p und $|A|$ abhängen.

8. Sei Q mit p eine Quelle mit 65535 Buchstaben für alle $x \in Q$ sowie A ein Alphabet mit 7 Buchstaben. Kann p so beschaffen sein, daß $L_A(Q) = H_A(Q)$ gilt?

9. Man bestimme die Entropie einer gleichverteilten Quelle Q über A.

10. Zeigen Sie, daß eine Huffman-Codierung präfixfrei ist.

11. Man zeige: Für alle reellen Zahlen $x > 0$ ist $\ln x \leq x - 1$ mit Gleichheit nur für $x = 1$.

12. Sei Q mit p eine Quelle und q eine weitere Wahrscheinlichkeitsverteilung auf Q. Sei A ein Alphabet. Man zeige $H_A(Q) \leq -\sum_{q \in Q} p(x) \log_{|A|} q(x)$ mit Gleichheit genau dann, wenn $p = q$ gilt.

13. Seien p eine Wahrscheinlichkeitsverteilung auf der endlichen Menge Q und A ein Alphabet. Man zeige $H_A(Q) \leq \log_{|A|} |Q|$ mit Gleichheit genau dann, wenn p die Gleichverteilung auf Q ist.

14. Wie groß kann die Entropie eines 27-buchstabigen Quelle Q über dem Alphabet $\{0, 1\}$ höchstens sein?

15. Für eine Quelle Q mit p betrachte man die *Produktquelle* $Q^2 = Q \times Q$, versehen mit dem Produktmaß q, also $q(xy) = p(x)p(y)$ für alle $x, y \in Q$. Man zeige $L_A(Q \times Q) \leq 2L_A(Q)$.

16. Man beweise, daß die im Anschluß an Satz 10 konstruierte Codierung einer beliebig großen Quelle Q mit p über A optimal ist.

17. Sei $Q = \{A, B, C, D, E, F, G, H\}$ eine Quelle mit p, wobei p durch folgende Tabelle gegeben ist:

x	A	B	C	D	E	F	G	H
$p(x)$	30%	25%	10%	10%	10%	5%	5%	5%

Man konstruiere eine optimale Codierung von Q über $A = \{0, 1\}$ und bestimme $L_A(Q)$.

18. Sei $A = \{0, 1\}$. Man betrachte folgendes 22-buchstabiges Wort w über dem Alphabet $Q = \{A, C, G, T\}$:

ACAACTTCGTCGCGCACATCCA

Sei p die durch die Häufigkeit des Auftretens der Buchstaben aus Q in diesem konkreten w gegebene Verteilung.

(i) Man bestimme p und konstruiere eine optimale Codierung von Q mit p über dem Alphabet A.

(ii) Man betrachte die Produktquelle $Q \times Q$ mit dem Produktmaß q und konstruiere eine optimale Codierung von $Q \times Q$ mit q über dem Alphabet A.

(iii) Man fasse w als 11-buchstabiges Wort über dem Alphabet $Q \times Q$ auf. Man bestimme die Menge $R \subseteq Q \times Q$ der Buchstaben, die in w auftreten sowie die durch Häufigkeiten des Auftretens gegebene Verteilung r auf R. Man konstruiere eine optimale Codierung von R mit q über dem Alphabet A.

(iv) Man bestimme die optimale mittlere Codewortlänge in allen drei genannten Szenarien und vergleiche.