4.3 Universelles Hashing

Definition 29

Eine Klasse $\mathcal{H}$ von Hashfunktionen von U nach $[0..n-1]$ heißt universell, falls für alle $x, y \in U$ mit $x \neq y$ gilt

$$\frac{|\{h \in \mathcal{H};\ h(x) = h(y)\}|}{|\mathcal{H}|} \leq \frac{1}{n},$$

wenn $h \in \mathcal{H}$ gleichverteilt gewählt wird.

Satz 30

Sei $\mathcal{H}$ eine universelle Klasse von Hashfunktionen für eine Hashtabelle der Größe n und sei $h \in \mathcal{H}$ zufällig gleichverteilt gewählt. Für eine Menge S von $m \leq n$ Schlüsseln ist dann die erwartete Anzahl von Kollisionen eines festen Schlüssels $x \in S$ mit anderen Elementen aus S kleiner als 1.

Beweis:

Sei x fest. Setze

$$C_x(y) =_{\text{def}} \begin{cases} 1 & \text{falls } h(x) = h(y)\,; \\ 0 & \text{sonst.} \end{cases}$$

Dann gilt

$$\begin{aligned} \mathbb{E}[C_x(y)] &= 0 \cdot \Pr[h(x) \neq h(y)] + 1 \cdot \Pr[h(x) = h(y)] \\ &= \Pr[h(x) = h(y)] \leq \frac{1}{n}\,. \end{aligned}$$

Für $C_x =_{\text{def}} \sum C_x(y)$ folgt damit

$$\mathbb{E}[C_x] = \sum_{y \in S \setminus \{x\}} C_x(y) \leq \frac{m-1}{n} < 1\,.$$

□

Sei $U = \{0, 1, \ldots, n-1\}^{r+1}$, für eine Primzahl n. Definiere

$$\mathcal{H} =_{\text{def}} \{h_\alpha;\ \alpha \in U\},$$

wobei

$$h_\alpha : U \ni (x_0, x_1, \ldots, x_r) \mapsto \sum_{i=0}^{r} \alpha_i x_i \bmod n \in \{0, 1, \ldots, n-1\}.$$

Lemma 31

$\mathcal{H}$ ist universell.

Beweis:

Seien $x, y \in U$ mit $x \neq y$. Wir nehmen o.B.d.A. an, dass $x_0 \neq y_0$. Ist $h_\alpha(x) = h_\alpha(y)$ für ein $\alpha \in U$, so gilt

$$\alpha_0(y_0 - x_0) = \sum_{i=1}^{r} \alpha_i(x_i - y_i) \bmod n\,.$$

Da n prim ist, ist $\mathbb{Z}_n$ ein Körper, und es gibt, bei vorgegebenen x, y und $\alpha_1, \ldots, \alpha_r$, genau ein α, so dass $h_\alpha(x) = h_\alpha(y)$.

Für festes x und y gibt es damit genau n^r Möglichkeiten, α zu wählen, so dass $h_\alpha(x) = h_\alpha(y)$.

Damit:

$$\frac{|\{h_\alpha \in \mathcal{H};\ h_\alpha(x) = h_\alpha(y)\}|}{|\mathcal{H}|} = \frac{n^r}{n^{r+1}} = \frac{1}{n}\,.$$

□

Wie groß müssen universelle Klassen von Hashfunktionen sein?

- Aus dem Beispiel:

$$|\mathcal{H}| = n^{r+1} = |U|\,.$$

- Es gibt Konstruktionen für Klassen der Größe $n^{\log(|U|)}$ bzw. $|U|^{\log n}$.

Satz 32

Sei $\mathcal{H}$ eine universelle Klasse von Hashfunktionen $h : U \to \{0, 1, \ldots, n-1\}$. Dann gilt

$$|\mathcal{H}| \geq n \left\lfloor \frac{\log(|U|) - 1}{\log n} \right\rfloor .$$

Beweis:

Sei $\mathcal{H} = \{h_1, h_2, \ldots, h_t\}$. Betrachte die Folge $U = U_0 \supseteq U_1 \supseteq U_2 \supseteq \cdots \supseteq U_t$, die definiert ist durch

$$U_i =_{\text{def}} U_{i-1} \cap h_i^{-1}(y_i)\,,$$

wobei $y_i \in \{0, 1, \ldots, n-1\}$ so gewählt ist, dass $|U_i|$ maximiert wird. Damit gilt

- h_i ist auf U_i konstant
- $|U_i| \geq \frac{|U_{i-1}|}{n}$, d.h. $|U_i| \geq \frac{|U|}{n^i}$.

Sei nun $\bar{t} = \left\lfloor \frac{\log(|U|)-1}{\log n} \right\rfloor$. Dann folgt

$$\log |U_{\bar{t}}| \geq \log |U| - \bar{t} \log n \geq \log |U| - \left(\frac{\log(|U|) - 1}{\log n} \right) \cdot \log n = 1\,.$$

Beweis:

Sei $\mathcal{H} = \{h_1, h_2, \ldots, h_t\}$. Betrachte die Folge $U = U_0 \supseteq U_1 \supseteq U_2 \supseteq \cdots \supseteq U_t$, die definiert ist durch

$$U_i =_{\text{def}} U_{i-1} \cap h_i^{-1}(y_i)\,,$$

wobei $y_i \in \{0, 1, \ldots, n-1\}$ so gewählt ist, dass $|U_i|$ maximiert wird. Damit gilt

- h_i ist auf U_i konstant
- $|U_i| \geq \frac{|U_{i-1}|}{n}$, d.h. $|U_i| \geq \frac{|U|}{n^i}$.

Seien $x, y \in U_{\bar{t}}$, $x \neq y$. Dann ist

$$\bar{t} \leq |\{h \in \mathcal{H};\ h(x) = h(y)\}| \leq |\mathcal{H}|/n$$

und damit

$$|\mathcal{H}| \geq n\bar{t} = n \left\lfloor \frac{\log(|U|) - 1}{\log n} \right\rfloor .$$

□

4.4 Perfektes Hashing

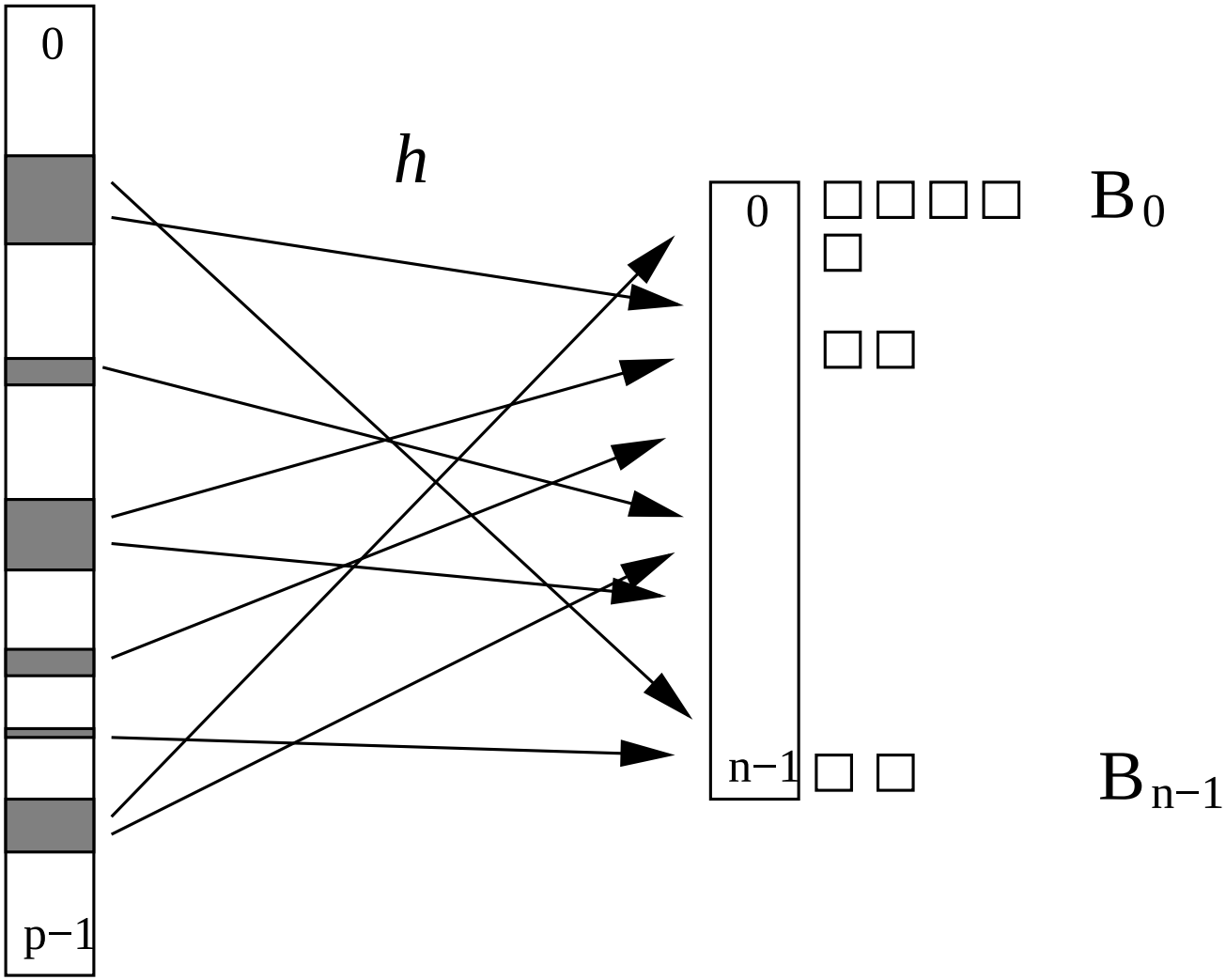
Das Ziel des perfekten Hashings ist es, für eine Schlüsselmenge eine Hashfunktion zu finden, so dass keine Kollisionen auftreten. Die Größe der Hashtabelle soll dabei natürlich möglichst klein sein.

4.4.1 Statisches perfektes Hashing

Sei $U = \{0, 1, \ldots, p-1\}$, p prim, das Universum, $n \in \mathbb{N}$ die Größe des Bildbereichs $\{0, 1, \ldots, n-1\}$ der Hashfunktionen und $S \subseteq U$, $|S| = m \leq n$, eine Menge von Schlüsseln.

Eine Hashfunktion $h : U \to \{0, 1, \ldots, n-1\}$ partitioniert S in „Buckets“

$$B_i = \{x \in S;\ h(x) = i\},\ \text{für } i = 0, 1, \ldots, n-1\,.$$



Hashfunktion h mit Buckets B_i

Definition 33

$\mathcal{H} = \mathcal{H}_{2,n}$ bezeichne die Klasse aller Funktionen

$$h_{a,b} : U \to \{0, 1, \ldots, n-1\}$$

mit

$$h_{a,b}(x) = ((a \cdot x + b) \bmod p) \bmod n \text{ für alle } x \in U\,,$$

wobei $0 < a < p$ und $0 \leq b < p$.

Lemma 34

$\mathcal{H}$ ist universell, d.h. für alle $x, y \in U$ mit $x \neq y$ gilt

$$\Pr[h(x) = h(y)] \leq \frac{1}{n}\,,$$

wenn h zufällig und gleichverteilt aus $\mathcal{H}$ gewählt wird.

Beweis:

Sei $h_{a,b}(x) = h_{a,b}(y) = i$. Dann ist

$$i = \underbrace{(ax + b) \bmod p}_{\alpha} = \underbrace{(ay + b) \bmod p}_{\beta} \pmod{n}$$

Sei $\alpha \in \{0, \ldots, p-1\}$ fest. Dann gibt es in der obigen Kongruenz $\lceil p/n \rceil - 1$ Möglichkeiten für β, nämlich

$$\beta \in \{i, i+n, i+2n, \ldots\} \setminus \{\alpha\},$$

da $\alpha \neq \beta$ und $x \neq y$ gilt.

Beweis:

Also gibt es höchstens

$$p \cdot \left(\left\lceil \frac{p}{n} \right\rceil - 1 \right) = p \cdot \left(\left(\left\lfloor \frac{p-1}{n} \right\rfloor + 1 \right) - 1 \right) \leq \frac{p(p-1)}{n}$$

Möglichkeiten für das Paar (α, β). Jedes Paar (α, β) bestimmt aber genau ein Paar (a, b), da $\mathbb{Z}_p$ ein Körper ist.

Weil es insgesamt $p(p-1)$ Paare (a, b) gibt und h uniform zufällig aus $\mathcal{H}$ ausgewählt wird, folgt

$$\Pr[h(x) = h(y)] \leq \frac{p(p-1)/n}{p(p-1)} = \frac{1}{n}$$

für jedes Paar $x, y \in U$ mit $x \neq y$. □

Lemma 35

Sei $S \subseteq U$, $|S| = m$. Dann gilt:

1. $$\mathbb{E}\left[\sum_{i=0}^{n-1} \binom{|B_i|}{2}\right] \leq \frac{m(m-1)}{2n}$$

2. $$\mathbb{E}[\sum_{i=0}^{n-1} |B_i|^2] \leq \frac{m(m-1)}{n} + m$$

3. $$\Pr[h_{a,b} \textit{ ist injektiv auf } S] \geq 1 - \frac{m(m-1)}{2n}$$

4. $$\Pr[\sum_{i=0}^{n-1} |B_i|^2 < 4m] > \frac{1}{2}, \textit{ falls } m \leq n$$

Beweis:

Definiere die Zufallsvariablen $X_{\{x,y\}}$ für alle $\{x,y\} \subseteq S$ gemäß

$$X_{\{x,y\}} = \begin{cases} 1 & \text{falls } h(x) = h(y)\,, \\ 0 & \text{sonst}\,. \end{cases}$$

Wegen Lemma 34 gilt $\mathbb{E}[X_{\{x,y\}}] = \Pr[h(x) = h(y)] \leq 1/n$ für alle Paare $\{x,y\} \subseteq S$. Weiter ist

$$\mathbb{E}\left[\sum_{i=0}^{n-1} \binom{|B_i|}{2}\right] = |\{\{x,y\} \subseteq S;\ h(x) = h(y)\}|$$
$$\leq \binom{m}{2} \cdot \frac{1}{n}\,.$$

Beweis (Forts.):

Da $x^2 = 2 \cdot \binom{x}{2} + x$ für alle $x \in \mathbb{N}$, folgt

$$\mathbb{E}[\sum_{i=0}^{n-1} |B_i|^2] = \mathbb{E}\left[\sum_{i=0}^{n-1} \left(2 \cdot \binom{|B_i|}{2} + |B_i|\right)\right]$$
$$\stackrel{(1)}{\leq} 2 \cdot \frac{m(m-1)}{2n} + m\,.$$

Aus der Markov-Ungleichung ($\Pr[X \geq t] \leq \frac{\mathbb{E}[X]}{t}$ für alle $t > 0$) folgt

$$\Pr[h_{a,b} \text{ nicht injektiv auf } S] = \Pr\left[\sum_{i=0}^{n-1} \binom{|B_i|}{2} \geq 1\right]$$
$$\stackrel{(1)}{\leq} \frac{m(m-1)}{2n}\,.$$

Beweis (Forts.):

Für $m \leq n$ folgt aus (2), dass

$$\mathbb{E}[\sum_{i=0}^{n-1} |B_i|^2] \leq m + m = 2m\,.$$

Also folgt, wiederum mit Hilfe der Markov-Ungleichung, dass

$$\Pr\left[\sum_{i=0}^{n-1} |B_i|^2 > 4m\right] \leq \frac{1}{4m} \cdot 2m = \frac{1}{2}\,.$$

□

Die Struktur der perfekten Hashtabelle nach

Michael L. Fredman, János Komlós, Endre Szemerédi:
Storing a sparse table with $\mathcal{O}(1)$ worst case access time,
Journal of the ACM **31**(3), p. 538–544 (1984)

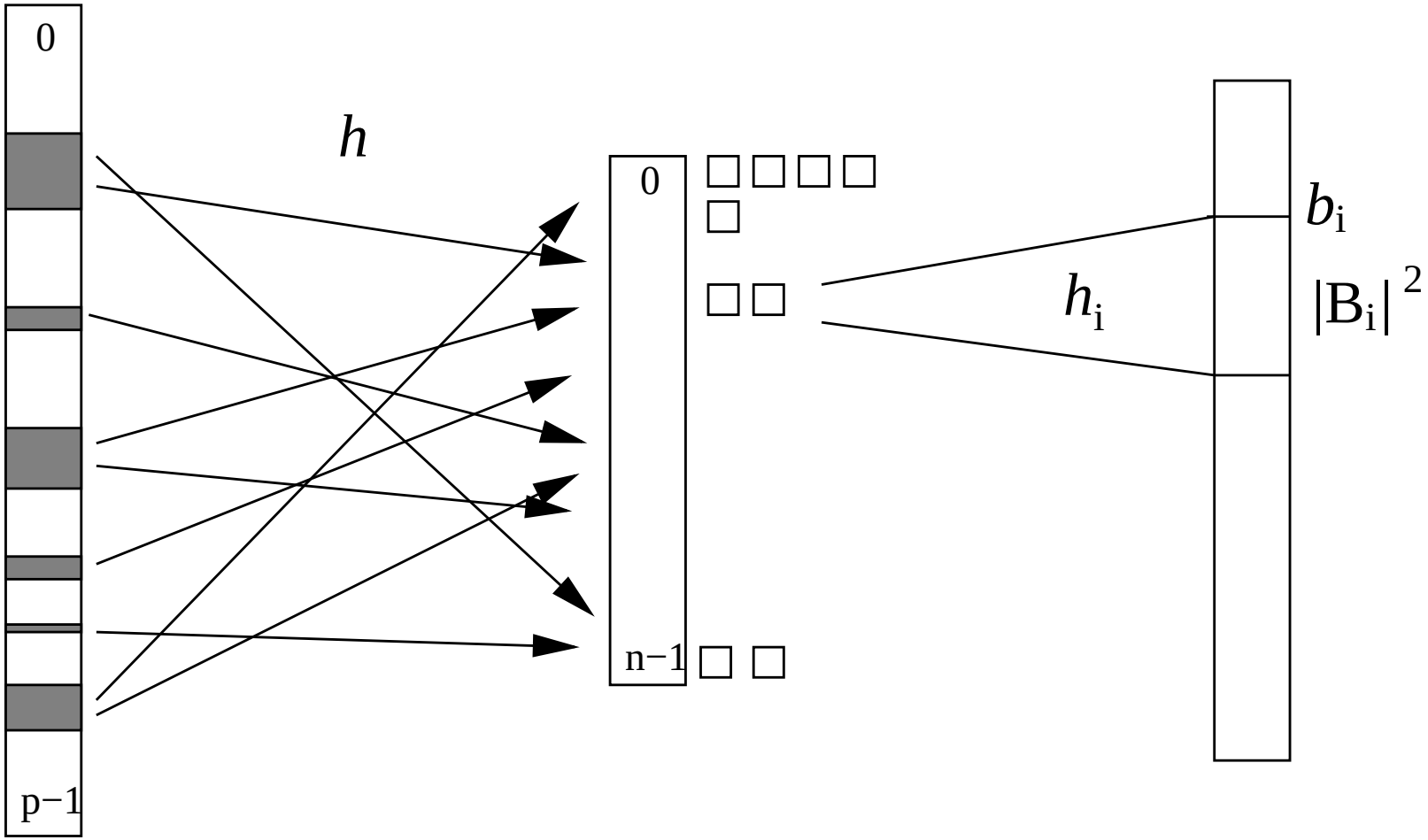
verwendet ein zweistufiges Hashverfahren.

Für einen gegebenen Schlüssel x wird zunächst $i = h(x)$ berechnet, um über den Tabellenplatz $T[i]$, b_i, $|B_i|$ und $h_i \in \mathcal{H}_{2,|B_i|^2}$ zu ermitteln. Dann wird im Tabellenplatz $T'[b_i + h_i(x)]$ nachgeschaut, ob x da abgespeichert ist. Falls ja, wird **true** ausgegeben und sonst **false**.

Falls

$$\sum_{i=0}^{n-1} |B_i|^2 < 4n$$

ist, so wird nur $\mathcal{O}(n)$ Platz verwendet.



Zweistufige Hashtabelle nach Fredman, Komlós und Szemerédi

Algorithmus für Hashtabelle nach FKS:

Eingabe: $S \subseteq U$, $|S| = m \leq n$
Ausgabe: Hashtabelle nach FKS

1. Wähle $h \in \mathcal{H}_{2,n}$ zufällig. Berechne $h(x)$ für alle $x \in S$.
2. Falls $\sum_i |B_i|^2 \geq 4m$, dann wiederhole 1.
3. Konstruiere die Mengen B_i für alle $0 \leq i < n$.
4. **for** $i = 0$ **to** $n - 1$ **do**
   (a) wähle $h_i \in \mathcal{H}_{2,|B_i|^2}$ zufällig
   (b) falls h_i auf B_i nicht injektiv ist, wiederhole (a)

Ein Durchlauf der Schleife bestehend aus den Schritten 1. und 2. benötigt Zeit $\mathcal{O}(n)$. Gemäß Lemma 35 ist die Wahrscheinlichkeit, dass Schritt 1. wiederholt werden muss, $\leq 1/2$ für jedes neue h.

Die Anzahl der Schleifendurchläufe ist also geometrisch verteilt mit Erfolgswahrscheinlichkeit $\geq 1/2$, und es ergibt sich

$$\mathbb{E}[\#\text{ Schleifendurchläufe}] \leq 2\,.$$

Also ist der Zeitaufwand für diese Schleife $\mathcal{O}(n)$. Schritt 3. kostet offensichtlich ebenfalls Zeit $\mathcal{O}(n)$.

Für jedes $i \in \{0, \ldots, n-1\}$ gilt, ebenfalls gemäß Lemma 35, dass

$$\Pr[h_i \text{ ist auf } B_i \text{ injektiv}] \geq 1 - \frac{|B_i|(|B_i|-1)}{2|B_i|^2} > \frac{1}{2}\,.$$

Damit ist auch hier die erwartete Anzahl der Schleifendurchläufe ≤ 2 und damit der erwartete Zeitaufwand

$$\mathcal{O}(|B_i|^2)\,.$$

Insgesamt ergibt sich damit für Schritt 4. wie auch für den gesamten Algorithmus ein Zeitaufwand von

$$\mathcal{O}(n)\,.$$