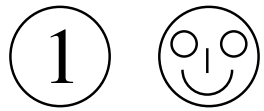
Etwas Wahrscheinlichkeitstheorie für den Hausgebrauch



Hash-Beispiel

Elementarereignisse Ω — Hash-Funktionen $\{0..m-1\}^{\text{Key}}$

Ereignisse: Teilmengen von Ω — $\mathscr{E}_{42} = \{h \in \Omega : h(4) = h(2)\}$

p_x =Wahrscheinlichkeit von $x \in \Omega$. $\sum_x p_x = 1$!

Gleichverteilung: $p_x = \frac{1}{|\Omega|}$ — $p_h = m^{-|\text{Key}|}$

$\mathbb{P}[\mathscr{E}] = \sum_{x \in \mathscr{E}} p_x$ — $\mathbb{P}[\mathscr{E}_{42}] = \frac{1}{m}$

Zufallsvariable (ZV) $X_0 : \Omega \to \mathbb{R}$ — $X = |\{e \in M : h(e) = 0\}|$

0-1-Zufallsvariable (Indikator-ZV) $I : \Omega \to \{0,1\}$

Erwartungswert $\mathrm{E}[X] = \sum_{y \in \Omega} p_y X(y)$ — $\mathrm{E}[X] = \frac{|M|}{m}$

Linearität des Erwartungswerts: $\mathrm{E}[X+Y] = \mathrm{E}[X] + \mathrm{E}[Y]$

Beispiel: Variante des Geburtstagsparadoxon

Wieviele Gäste muss eine Geburtstagsparty "im Mittel" haben, damit mindestens zwei Gäste den gleichen Geburtstag haben?
Gäste (Keys) $1..n$.
Elementarereignisse: $h \in \Omega = \{0..364\}^{\{1..n\}}$.
Definiere Indikator-ZV $I_{ij} = 1$ gdw $h(i) = h(j)$.
Anzahl Paare mit gleichem Geburtstag: $X = \sum_{i=1}^{n} \sum_{j=i+1}^{n} I_{ij}$.

$$\begin{aligned}
\mathrm{E}[X] &= \mathrm{E}[\sum_{i=1}^{n} \sum_{j=i+1}^{n} I_{ij}] = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathrm{E}[I_{ij}] \\
&= \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathbb{P}[I_{ij} = 1] = \frac{n(n-1)}{2} \cdot \frac{1}{365} \\
&\stackrel{!}{=} 1 \Leftrightarrow n = \frac{1}{2} + \sqrt{\frac{1}{2^2} + 730} \approx 27.52
\end{aligned}$$

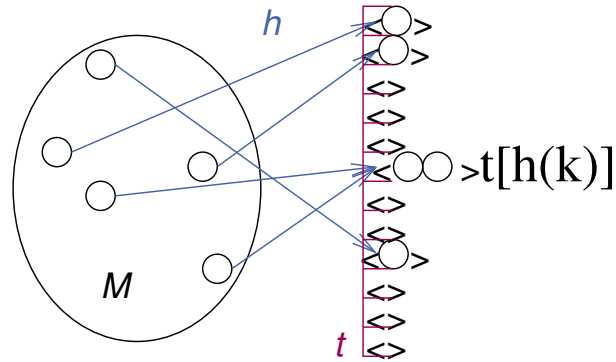
Mehr zum Geburtstagsparadoxon

Standardfomulierung:
Ab wann lohnt es sich zu wetten, dass es zwei Gäste mit gleichem Geburtstag gibt? Etwas komplizierter. Antwort: $n \geq 23$
Verallgemeinerung: Jahreslänge m = Hashtabelle der Größe m:
eine zufällige Hashfunktion $h : 1..n \to 0..m-1$ ist nur dann mit vernünftiger Wahrscheinlichkeit perfekt wenn $m = \Omega(n^2)$.
Riesige Platzverschwendung.

Analyse für zufällige Hash-Funktionen

Theorem 1

$\forall k$: *die erwartete Anzahl kollidierender Elemente ist* $\mathrm{O}(1)$ *falls* $|M| \in \mathrm{O}(m)$.



Beweis.

Für festen Schlüssel k definiere Kollisionslänge X

$X := |t[h(k)]| = |\{e \in M' : h(e) = h(k)\}|$ mit

$M' = \{e \in M : \mathrm{key}(e) \neq k\}$.

Betrachte die 0-1 ZV $X_e = 1$ für $h(e) = h(k)$, $e \in M'$ und $X_e = 0$ sonst.

$$\mathrm{E}[X] = \mathrm{E}[\sum_{e \in M'} X_e] = \sum_{e \in M'} \mathrm{E}[X_e] = \sum_{e \in M'} \mathbb{P}[X_e = 1] = \frac{|M'|}{m}$$

$$\in \mathrm{O}(1)$$

Das gilt unabhängig von der Eingabe M. □

Zufällige Hash-Funktionen?

Naive Implementierung: ein Tabelleneintrag pro Schlüssel.

$\rightsquigarrow$ meist zu teuer

Weniger naive Lösungen: kompliziert, immer noch viel Platz.

$\rightsquigarrow$ meist unsinnig

$\rightsquigarrow$ unrealistisch

Universelles Hashing

Idee: nutze nur bestimmte "einfache" Hash-Funktionen

Definition 2

$\mathscr{H} \subseteq \{0..m-1\}^{\text{Key}}$ ist *universell*
falls für alle x, y in Key mit $x \neq y$ und zufälligem $h \in \mathscr{H}$,

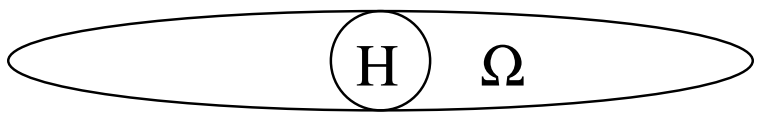
$$\mathbb{P}[h(x) = h(y)] = \frac{1}{m} \quad .$$

Theorem 3

Theorem 1 gilt auch für universelle Familien von Hash-Funktionen.

Beweis.

Für $\Omega = \mathscr{H}$ haben wir immer noch $\mathbb{P}[X_e = 1] = \frac{1}{m}$.
Der Rest geht wie vorher. □



Eine einfache universelle Familie

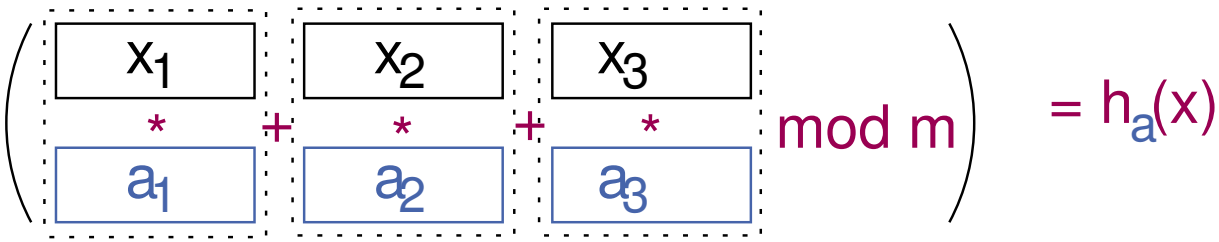
m sei eine Primzahl, Key $\subseteq \{0,\ldots,m-1\}^k$

Theorem 4

Für $\mathbf{a} = (a_1,\ldots,a_k) \in \{0,\ldots,m-1\}^k$ *definiere*

$h_{\mathbf{a}}(\mathbf{x}) = \mathbf{a}\cdot\mathbf{x}$ **mod** m, $H^{\cdot} = \left\{ h_{\mathbf{a}} : \mathbf{a} \in \{0..m-1\}^k \right\}$.

$H^{\cdot}$ *ist eine universelle Familie von Hash-Funktionen*



Beispiel für $H^{\cdot}$

Für $\mathbf{a} = (a_1, \ldots, a_k) \in \{0, \ldots, m-1\}^k$ definiere
$h_{\mathbf{a}}(\mathbf{x}) = \mathbf{a} \cdot \mathbf{x} \bmod m$, $H^{\cdot} = \left\{ h_{\mathbf{a}} : \mathbf{a} \in \{0..m-1\}^k \right\}$.
$k = 3$, $m = 11$
wähle $\mathbf{a} = (\mathbf{8}, \mathbf{1}, \mathbf{5})$.
$h_{\mathbf{a}}((1,1,2)) = (8,1,5) \cdot (1,1,2) = 8 \cdot 1 + 1 \cdot 1 + 5 \cdot 2 = 19 \equiv 8 \bmod 11$

Beweis.

Betrachte $\mathbf{x} = (x_1, \ldots, x_k)$, $\mathbf{y} = (y_1, \ldots, y_k)$ mit $x_j \neq y_j$
zähle $\mathbf{a}$ mit $h_{\mathbf{a}}(\mathbf{x}) = h_{\mathbf{a}}(\mathbf{y})$.
Für jede Wahl der a_i, $i \neq j$, $\exists$ genau ein a_j mit $h_{\mathbf{a}}(\mathbf{x}) = h_{\mathbf{a}}(\mathbf{y})$:

$$\sum_{1 \leq i \leq k} a_i x_i \equiv \sum_{1 \leq i \leq k} a_i y_i (\mathbf{mod}\ m)$$

$$\Leftrightarrow a_j(x_j - y_j) \equiv \sum_{i \neq j, 1 \leq i \leq k} a_i(y_i - x_i) (\mathbf{mod}\ m)$$

$$\Leftrightarrow a_j \equiv (x_j - y_j)^{-1} \sum_{i \neq j, 1 \leq i \leq k} a_i(y_i - x_i) (\mathbf{mod}\ m)$$

m^{k-1} Möglichkeiten die a_i (mit $i \neq j$) auszuwählen.
m^k ist die Gesamtzahl der $\mathbf{a}$, d. h.,

$$\mathbb{P}[h_{\mathbf{a}}(x) = h_{\mathbf{a}}(\mathbf{y})] = \frac{m^{k-1}}{m^k} = \frac{1}{m}.$$

Bit-basierte Universelle Familien

Sei $m = 2^w$, Key $= \{0,1\}^k$

Bit-Matrix Multiplikation: $H^{\oplus} = \left\{ h_{\mathbf{M}} : \mathbf{M} \in \{0,1\}^{w \times k} \right\}$

wobei $h_{\mathbf{M}}(\mathbf{x}) = \mathbf{M}x$ (Arithmetik mod 2, d. h., xor, and)

Tabellenzugriff: $H^{\oplus[]} = \left\{ h^{\oplus[]}_{(t_1,\ldots,t_b)} : t_i \in \{0..m-1\}^{\{0..2^a-1\}} \right\}$

wobei $h^{\oplus[]}_{(t_1,\ldots,t_b)}((x_0,x_1,\ldots,x_b)) = x_0 \oplus \bigoplus_{i=1}^{b} t_i[x_i]$