Hashing

Philip Bille   Inge Li Gørtz   Eva Rotenberg

Hashing

- Universe U,
- Range $[m] = \{0, 1, 2, \dots, m-1\}$,
- The class of all functions $U \to [m]$,
- A hash function is a random variable in $\uparrow$ that class of functions.
- Example: The truly random hash function assigns each $x \in U$ to a uniformly random value in $[m]$, in a way that is independent of all other values $y_1, \dots, y_i \in U$, $y_1 \neq x, \dots, y_i \neq x$.
- Question to you: is this the same as choosing uniformly at random from the class of all functions $U \to [m]$?
- Truly random hash function – not very practical. Also much more powerful than usually necessary. Let's consider hash functions that are just good enough. Universal hashing.

Universal Hashing

- Universe U, range $[m] = \{0, 1, 2, \ldots, m-1\}$,
- Random variable h in the class of all functions $U \to [m]$,
- Universal means: $P[h(x) = h(y)] \leq 1/m$ for $x \neq y$, $x, y \in U$.
- In words: the pairwise collision probability is as low as fully random.
- c-approximately universal means $P[h(x) = h(y)] \leq c/m$ for $x \neq y$.
- E.g: hashing with chaining. Works with full (utopian) randomness. Works with universal? Works with $O(1)$-approximate universal?

Strong Universality

- Universe U, range $[m] = \{0, 1, 2, \ldots, m-1\}$,
- Random variable h in the class of all functions $U \to [m]$,
- Strongly universal means bounded probability of pairwise events:
  - for $x \neq y \in U$ and any $q, r \in [m]$, $P[h(x) = q \wedge h(y) = r] = 1/m^2$
- In words: given different values x and y from the universe, all m^2 possible outcomes of the pair $(h(x), h(y))$ are equally likely.
- Questions: can a deterministic function be universal? Strongly?
- Observation: being strongly universal is equivalent to being:
  - uniform: $h(x)$ takes each value in $[m]$ with probability $1/m$
  - 2-independent: $h(x_1)$ is independent of $h(x_2)$ for $x_2 \neq x_1$.
- c-approximately strongly universal:
  - c-approximately uniform (probability $\leq c/m$)
  - 2-independent (like above).

Example function: Multiply mod prime [warmup]

- Warmup: consider $[m] = [p]$ with $p \geq |U|$.
- Let a, b be random numbers in $[p] = \{0, 1, \ldots, p-1\}$.
- Consider the function $\tilde{h}_{a,b}(x) = ax + b \mod p$.
- What is the probability $\tilde{h}_{a,b}(x) = q \wedge \tilde{h}_{a,b}(y) = r$? $(x \neq y.)$
- $ax + b = q$ and $ay + b = r$, so $a(x - y) = q - r$.
  Since $\mathbb{Z}/p$ is a field, unique $a \in [p]$ solves $\uparrow$. And then, b unique.
- So: Given x, y, every value pair (q, r) corresponds uniquely to a pair a, b, such that $\tilde{h}_{a,b}(x) = q \wedge \tilde{h}_{a,b}(y) = r$. Since each pair (a, b) is equally likely, all value pairs q, r are equally likely.
- Question: We may sometimes choose $a = 0$. Is this good or bad?

Example function: Multiply mod prime

- We have that $\tilde{h}_{a,b}(x) : U \to [p]$ is strongly universal.
- If, on the other hand, we restrict to $a \neq 0$, we have no collisions.
- Now, for any $m \leq [p]$, consider $h(x) = \tilde{h}_{a\neq 0,b}(x) \mod m$.
- When do we have a collision $h(x) = h(y)$ for $x \neq y$?
- Let q denote $\tilde{h}_{a,b}(x)$ and r denote $\tilde{h}_{a,b}(y)$, then the collision happens when $q \equiv r \mod m$.
- For a given q, there are at most $\lceil p/m \rceil$ such values r.
- But if $a \neq 0$, only $\leq \lceil p/m \rceil - 1$ of them can be the value $\tilde{h}_{a,b}(y)$.
- So, we get $\sum_{q\in[p]} P[h(x) = h(y) | h(x) = q]$ and we found this was $\leq \sum_{q\in[p]} \lceil p/m \rceil - 1$; all in all $\leq p \cdot (\lceil p/m \rceil - 1) \leq p(p-1)/m$.
- That $\uparrow$ many collision pairs out of $(p-1) \times p$ choices for a, b gives collision probability $\leq \frac{p(p-1)/m}{p(p-1)} = 1/m$.