COS226 Precept 6 - Advanced Fall '25

Precept Outline

- Uniform hashing assumption
- Universal hashing and pairwise uniformity

A. Uniform Hashing Assumption

In our discussion of hash functions, there is an elephant in the room that we never really addressed: we adopt the uniform hashing assumption, haven't shown you any examples of functions that satisfy it.

Let's start by proving this is impossible: show that there is no function $h\colon K \to [m]$ (where K is the set of keys, e.g., all 4-byte integers or 8-byte doubles) such that, for all $x \in K$ and $i \in [m]$, we have $\mathbb{P}[h(x) = i] = 1/m$. *Hint: don't overthink it!*

B. Universal Hashing

The (formal) to this conundrum is to pick a *random function* instead;[1] more precisely, given a family of hash functions $\mathcal{H}$ (with domain K and codomain $[m]$), we sample a uniformly random $h \leftarrow \mathcal{H}$ before any operations and use it throughout the lifetime of the hash table. (Concretely, we would add an instance variable `hashFunction` and assign it `hashFamily[StdRandom.uniformInt(hashFamily.length)]` in the constructor.)

A family of hash functions is called *universal* if, for all keys $x, y \in K$ with $x \neq y$, we have

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

(Note that only h is random, whereas x and y are not.)

For each of the families of functions below, either prove its universality or prove that it is not (by finding a pair of distinct keys that collide with probability $\neq 1/m$). For simplicity, assume that m is a prime number (so you can use the fact that every $y \in [m-1]$ is invertible modulo m); and that $K = [m]$.

- $\mathcal{H}_1 = \{h_b : b \in [m]\}$, where $h_b(x) = x + b \pmod{m}$.
- $\mathcal{H}_2 = \{h_a : a \in [m-1]\}$, where $h_a(x) = a \cdot x \pmod{m}$.
- $\mathcal{H}_3 = \{h_a : a \in [m]\}$, where $h_a(x) = a \cdot x \pmod{m}$.
- $\mathcal{H}_4 = \{h_c : c \in [m]\}$, where $h_c(x) = (x + c)^2 \pmod{m}$.

[1] A solution widely used in practice is to forgo the uniform hashing assumption entirely and make an assumption on the "unpredictability" of the output; see, e.g., the AES cipher.

- $\mathcal{H}_5 = \{h_a : a \in [m]\}$, where $h_a(x) = a \cdot x^2 \pmod{m}$.
- $\mathcal{H}_6 = \{h_{a,b} : a \in [m], b \in [m]\}$, where $h_{a,b}(x) = a \cdot x + b \pmod{m}$.
  (Notice that sampling from $\mathcal{H}_6$ is equivalent to sampling uniform a and b independently.)

C. Pairwise Independence

A stronger property that hash functions is pairwise independence, also known as *strong universality*: for any pair $x, y \in K$ of keys and $z, w \in [m]$ of hash values,

$$\mathbb{P}[h(x) = z \text{ and } h(y) = w] = \frac{1}{m^2}.$$

(Equivalently, the hash values of two distinct keys are uniform and independent.)

For each family of functions above, either prove that is satisfies pairwise independence or give a counterexample (a pair of keys and values such that the probability is $\neq 1/m^2$).

D. Boolean Hash Functions

As a warm-up, consider the function $h\colon \{0,1\} \to \{0,1\}$ generated as follows: sample two uniformly random and independent bits $a, b \leftarrow \{0,1\}$ and set $h(x) = (a \cdot x) \oplus b$ (where $\oplus$ is the XOR operation, given by $0 \oplus x = x$ and $1 \oplus x = 1 - x$ for $x \in \{0,1\}$).

What is the probability that $(h(0), h(1)) = (x, y)$ for each $x, y \in \{0,1\}^2$?

Now let $m < n$ and consider the hash function $h\colon \{0,1\}^n \to \{0,1\}^m$, where for each i, the i^{th} bit of $h(x)$ is generated by sampling $a_i, b_i \leftarrow \{0,1\}$ and setting $h_i(x) = (a_i \cdot x_i) \oplus b_i$. (All a_i and b_i are uniform and independent.)

Show this family of functions does not satisfy pairwise independence, namely, that there are four bit strings $x \neq x' \in \{0,1\}^n$ and $y, y' \in \{0,1\}^m$ such that the probability that $h(x) = y$ and $h(x') = y'$ is not $1/2^{m+1}$.

Consider the following alternative definition of $h\colon \{0,1\}^n \to \{0,1\}^m$: for each i, the i^{th} bit of $h(x)$ is generated by sampling $a_{i1}, a_{i2}, \ldots, a_{in}, b_i \leftarrow \{0,1\}$ and setting $h_i(x) = (a_{i1} \cdot x_1) \oplus \cdots \oplus a_{in} \cdot x_n \oplus b_i$. (All a_{ij} and b_i are uniform and independent).

Show that this does satisfy pairwise independence.

Finally, show that h does not satisify 4-wise independence: find four strings $x, y, z, w \in \{0, 1\}^n$ such that, *for every* h, the value of $h(w)$ is completely determined by $h(x)$, $h(y)$ and $h(z)$.