CSE6512 Lecture 9 Notes

Ruofan Jin
September 27, 2011

1 Hashing

Definition A family H of hash functions is *2-universal* if for any $x, y \in M$ with $x \neq y$,

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

where $h \in H, h : M \to N, M = \{0, 1, \ldots, m-1\}, N = \{0, 1, \ldots n-1\}$.

Construction:

Pick a prime $p \geq m$, use the field $\mathbb{Z}_p = \{0, 1, \cdots p-1\}$. Let $f_{a,b}(x) = ax+b \bmod p$, for $a, b \in \mathbb{Z}_P, a \neq 0$. Let $g(x) = x \bmod n$, and let $h_{a,b}(x) = g\left(f_{a,b}(x)\right) = (ax+b) \bmod p \bmod n$. Then

$$H = \{h_{a,b} : a, b \in \mathbb{Z}_p, a \neq 0\},$$

$$|H| = p(p-1).$$

Note. *∃ a prime number between m and $2m$ for any integer m ⇒ Any member of H can be specified with $O(\log m)$ bits.*

Definition $\delta(x, y, h) = \begin{cases} 1 & \text{if } h(x) = h(y), x \neq y \\ 0 & \text{otherwise} \end{cases}$

$\delta(X, y, h), \delta(x, y, H)$, etc. can be defined likewise.

Fact. *If $H$is 2-univeral, then $\forall x, y \in M$, $\delta(x, y, H) \leq \frac{|H|}{n}$.*

Fact. *If H is 2-univeral and $S \subseteq M$, then $\forall x \in M$ and a randomly picked $h \in H$, $E\left[\delta(x, S, h)\right] \leq \frac{|S|}{n}$.*

Lemma. *If $H = \{h_{a,b} : a, b \in \mathbb{Z}_p, a \neq 0\}$, then $\forall x, y \in \mathbb{Z}_p, x \neq y$,*

$$\delta(x, y, H) = \delta(\mathbb{Z}_p, \mathbb{Z}_p, g).$$

Proof. The above statement says that the number of functions in H under which x and y collide is the same as the number of pairs in $\mathbb{Z}_P$ that collide under g. Fix x and y and let $r = (ax+b) \bmod p$ and $s = (ay+b) \bmod p$. Note that if $x \neq y$, then $r \neq s$. Let $F(a, b) = ((ax+b) \bmod p, (ay+b) \bmod p)$.

Fact. *F is one-to-one and onto.*

Consider two pairs (a_1, b_1) and (a_2, b_2). Let $F(a_1, b_1) = (r_1, s_1)$ and $F(a_2, b_2) = (r_2, s_2)$.

If $(a_1, b_1) \neq (a_2, b_2)$, can $(r_1, s_1) = (r_2, s_2)$? Note that $r_1 = (a_1x + b_1) \bmod p$ and $r_2 = (a_2x + b_2) \bmod p$. If $r_1 = r_2$, then, $x = (b_2 - b_1)(a_1 - a_2)^{-1} \bmod p$. Similarly, if $s_1 = s_2$, we can see that $y = (b_2 - b_1)(a_1 - a_2)^{-1} \bmod p$. As a result, if $(r_1, s_1) = (r_2, s_2)$, it will imply that $x = y$ which is a contradiction. Thus F is one-to-one.

Let (r, s) be any pair from $\mathbb{Z}_P$ with $r \neq s$. We can solve $ax + b \bmod p = r$ and $ay + b \bmod p = s$ to get a unique pair (a, b) with $a \neq 0$. Thus F is onto.

We realize that the function $f_{a,b}$ cannot make x and y to collide if $x \neq y$. For a given x and y (with $x \neq y$), when we change a and b, $F(a, b)$ ranges over all pairs (r, s) (with $r, s \in \mathbb{Z}_p$ and $r \neq s$). Collisions happen only because of the function g.

For a given x and y, the number of hash functions that make x and y to collide is the same as the number of pairs (a, b) (with $a \neq 0$) for which $h_{a,b}(x) = h_{a,b}(y)$. This number is the same as the number of pairs (a, b) for which $g(f_{a,b}(x)) = g(f_{a,b}(y))$. In turn, this number is the same as the number of pairs (r, s) (with $r, s \in \mathbb{Z}_p$ and $r \neq s$) for which $g(r) = g(s)$.

□

Lemma. *H is* 2-universal. *i.e.* $\delta(x, y, H) \le \frac{|H|}{n}$.

Proof. Let $A_Z = \{x \in \mathbb{Z}_P : g(x) = Z\}, Z = 0, 1, \ldots, n-1$.
Note that $A_Z \le \lceil \frac{p}{n} \rceil$ for any $Z \in N$.
$\Rightarrow \delta(\mathbb{Z}_P, \mathbb{Z}_P, g) \le p\left(\lceil \frac{p}{n} \rceil - 1\right) \le p\frac{(p-1)}{n} = \frac{|H|}{n}$. □

2 Searching in $O(1)$ time (M. Ajtai, J. Komlós & E. Szemerédi, 1985)

Let $M = \{0, 1, \ldots, m-1\}$, $N = \{0, 1, \ldots, n-1\}$. W.L.O.G., let $p = m+1$ be a prime number.
For any $1 \le k \le m$, let $h_k(x) = kx \bmod p \bmod n$.
Let $V \subseteq M$ be the input set where $|V| = v$. Let $B_i(k, n, V)$ be the set of elements of V that are hashed into i, i.e.,

$$B_i(k, n, V) = \{x \in V : h_k(x) = i\}, i = 0, 1, \ldots, n-1.$$

Let $b_i(k, n, V) = |B_i(k, n, V)|$.

Lemma. $\sum_{k=1}^{m} \sum_{i=0}^{n-1} \binom{b_i(k,n,V)}{2} < \frac{mv^2}{n}$ *for all $V \subseteq M$ and $n > v$.*

Proof. $\binom{b_i(k,n,V)}{2}$ is the number of sets $\{x, y\}$, s.t. x and y collide under h_k and $h_k(x) = i$. And $\sum_{i=0}^{n-1} \binom{b_i(k,n,V)}{2}$ is the number of sets $\{x, y\}$, s.t., x and y collide under h_k. Likewise, $\sum_{k=1}^{m} \sum_{i=0}^{n-1} \binom{b_i(k,n,V)}{2}$ is the number of tuples $(k, \{x, y\})$, s.t. x and y collide under h_k.
$\Rightarrow$ It is the number of tuples $(k, \{x, y\})$, s.t., $kx \bmod p \bmod n = ky \bmod p \bmod n$.
$\Rightarrow k(x-y) \bmod p \in \left\{\pm n, \pm 2n, \ldots, \pm \lfloor \frac{p-1}{n} \rfloor n\right\}$
Note that $k(x-y) \bmod p = jn$ has a unique solution for k if we fix x and y, for any $j = 1, \ldots, \lfloor \frac{p-1}{n} \rfloor$.
$\Rightarrow$ For any $x, y \in \mathbb{Z}_P, \exists \le 2\left(\frac{p-1}{n}\right)$ functions h_k under which x and y collide.
$\Rightarrow \sum_{k=1}^{m} \sum_{i=0}^{n-1} \binom{b_i(k,n,V)}{2} \le \binom{v}{2} \frac{2(p-1)}{n} < \frac{(p-1)v^2}{n} = \frac{mv^2}{n}$. □

Corollary. $\exists k$,*s.t.*, $\sum_{i=0}^{n-1} \binom{b_i(k,n,V)}{2} < \frac{v^2}{n}$.