Universal families of hash functions

Scribe: Muhan Li

In this lecture, we will define several common families of 2-universal hash functions.

Scenario 1: hash numbers to numbers

In this scenario, we will hash $u = \{0, \ldots, N\}$, to a hash domain of size n: $\{0, \ldots, n-1\}$, where n is the number of hash buckets.

Hash function definition

Pick a prime number $p > N$, (p is deterministic here, randomness is in a and b), $N < p \leq 2N$. We assume $p > N > n$.

By Bertrand's postulate, We can always find such a prime number.

Define hash family:

$$\mathcal{H} = \{h_{ab} \mid h_{ab} : X \to ((ax + b) \; mod \; p) \; mod \; n\}$$
$$s.t.: \begin{cases} 1 \leq a < p \\ 0 \leq b < p \\ \mathbb{Z}_p = \{0, \ldots, p-1\} \end{cases}$$

where $\mathbb{Z}_p$ are fields defined with the following operations:

$$\begin{cases} a +_p b = a + b \; mod \; p & \text{(addition)} \\ a -_p b = (p + a - b) \; mod \; p & \text{(subtraction)} \\ a *_p b = a * b \; mod \; p & \text{(multiplication)} \\ a /_p b = q, \; where \; a = b * q \; mod \; p, q \in \mathbb{Z}_p & \text{(division)} \end{cases}$$

2-universality proof

Proof:

Suppose $x \in u$ and $y \in u$, $x \neq y$

Then we can represent collision of x and y as (represented by a linear system):

$$\begin{bmatrix} ax + b = u \; mod \; p \\ ay + b = v \; mod \; p \\ u = v \; mod \; n \end{bmatrix}, \text{also equivalent to} \begin{cases} \begin{pmatrix} x & 1 \\ y & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} u \\ v \end{pmatrix} \; mod \; p \\ \\ u = v \; mod \; n \end{cases}$$

We can draw a conclusion about the uniqueness of solution (a, b)

> For any x, y, there exists a one to one correspondence between pairs (a, b), that cause collisions and pairs (u, v)
>
> $s.t.: \; u = v \; mod \; n, \quad u, v \in \mathbb{Z}_p$

This is because $x \neq y$, the coefficient matrix $\begin{pmatrix} x & 1 \\ y & 1 \end{pmatrix}$ is invertible, and since $a \geq 1$, for any pairs of $(u, v), u = v,$ there exists and only exists one (a, b) which solves the system.

Now we need to prove the probability definition in 2-universality, $u = v \, mod \, p$ requires following conditions:

$$\begin{cases} v \in \{u - ln, \ldots, u - n, u, u + n, \ldots, u + kn\} \\ u + kn \leq p - 1 \\ u - ln \geq 0 \end{cases}$$

Therefore the upper bound of the number of v is $\lceil (p-1)/n \rceil$, and number of (u, v) pairs is bounded by $p * ((p-1)/n + 1)$

and:

$$\begin{aligned} & P\{h_{ab}(x) = h_{ab}(y)\} \\ & \leq \frac{(p(\frac{p-1}{n} + 1))}{\text{Number of (a, b)}} \\ & = \frac{(p(\frac{p-1}{n} + 1))}{p(p-1)} \\ & = \frac{1}{n} + o(1) \\ & \simeq \frac{1}{n} \end{aligned}$$

Scenario 2: hash fixed length binary strings to numbers

Hash function definition

In this scenario, we will hash a domain made up of fixed length binary strings of length m: $u = \{0,1\}^m$, to a hash domain of size 2^d: $\{0, \ldots, 2^d - 1\}$, with the number of hash buckets $n = 2^d$, suppose $m \gg d$.

Pick a random matrix A as:

$$\text{d rows}\left\{ \underbrace{\begin{pmatrix} & A & \end{pmatrix}}_{\text{m columns}} \right. \in \{0,1\}^{d \times m}$$

And compute hash value as:

$$h_A = Ax \, mod \, 2 \qquad \text{(x is a binary vector of length m, } x \in \{0,1\}^m\text{)}$$

2-universality proof

Proof:

Suppose $Ax = Ay \mod 2, x \neq y$ creates a collision, then $A(x-y) = 0 \mod 2, x \neq y$

let $z = x - y$, then $Az = 0 \mod 2, z \neq 0$:

$$Az = \begin{pmatrix} \overrightarrow{a_1} \\ \ldots \\ \overrightarrow{a_d} \end{pmatrix} * \begin{pmatrix} z_1 \\ \ldots \\ z_m \end{pmatrix} = \begin{pmatrix} 0 \\ \ldots \\ 0 \end{pmatrix} \quad \mod 2$$

The goal $\forall z \neq 0, \ P\{Az = 0\} \leq \frac{1}{2^d}$ stands if:

1. $P\{< a_i, z >= 0 \mod 2\} = 1/2$

2. All products $< a_1, z > \mod 2, \dots, < a_d, z > \mod 2$ are independent

For the second condition:

> All products $< a_1, z > \mod 2, \dots, < a_d, z >$ mod 2 are independent, if a_i are chosen independently.
>
> This is easy to prove using the independence definition: X and Y are independent **iff** $P\{X \in S_X, Y \in S_Y\} = P\{X \in S_X\}P\{Y \in S_Y\}$, a detailed proof can be seen at here.

For the first condition:
Proof:

> $< a_i, z > \mod 2 = \sum_{j=1}^{m} a_{ij} * z_j \mod 2$
>
> Suppose j_0 is the index of a non-zero element in vector z:
>
> $$\begin{aligned} z = (0, \dots, &1, \dots) \\ &\uparrow j_0 \end{aligned}$$
>
> $$\begin{aligned} &\sum_{j=1}^{m} a_{ij} * z_j \mod 2 \\ &= \underbrace{\sum_{j=1, j!=j_0}^{m} a_{ij} * z_j}_{L} + \underbrace{a_{ij_0} * z_{j_0}}_{R} \mod 2 \\ &= \underbrace{\sum_{j=1, j!=j_0}^{m} a_{ij} * z_j}_{L} + \underbrace{a_{ij_0}}_{R} \mod 2 \end{aligned}$$
>
> Now we need to prove $P\{L + R = 0 \mod 2\} = \frac{1}{2}$, and it is clear to see that $P\{R = 0\} = P\{R = 1\} = \frac{1}{2}$
>
> Since L is not dependent on R, for any given L, their is a R that makes $L + R = 0 \mod 2$, and probability of this specific R is $\frac{1}{2}$ as shown above, we have conditional probability:
>
> $$P\{L + R = 0 \mod 2 | L\} = \frac{1}{2}$$
>
> Therefore:
>
> $$P\{L + R = 0 \mod 2\} = \mathbb{E}_L\{P\{L + R = 0 \mod 2 | L\}\} = \frac{1}{2}$$

This concludes our 2-universality proof.

Perfect hashing: never collides

In lecture 1 we have concluded that a perfect hash function requires $|u| < n$

Why not use identity as the hash function? because we want universal, fixed size hash values: eg: an html tag (string) into uint64.

Algorithm used to create a perfect hashing function

Definition

Let $\mathcal{H}$ be a universal hash family $\mathcal{H} = \{h | h : u \rightarrow \{0, \ldots, n-1\}\}$

Let $n = c * |u|^2$, where c is a constant

Repeatedly pick a random $h \in \mathcal{H}$, test h on the input set s until h doesn't have any collision.

Performance analysis

We can prove that the probability of picking a bad h with more than 1 collision after k steps decreases **exponentially**.

$$\begin{aligned}
&\mathbb{E}_h\{\text{numbr of pairwise collisions}\} \\
&= \mathbb{E}_h\{\sum_{x,y\in u, x\neq y} \mathbb{I}\{h(x) = h(y)\}\} \\
&= \sum_{x,y\in u, x\neq y} P_{h\in H}\{h(x) = h(y)\} \\
&\text{First we select a pair of x and y, and then select a collision value from 0 to n-1:} \\
&= \binom{|u|}{2} * \frac{1}{n} \\
&= \frac{|u|(|u|-1)}{2n}
\end{aligned}$$

Because $n = |u|^2$, $\mathbb{E}_h\{\text{numbr of pairwise collisions}\} \leq \frac{1}{2}$

By [Markov's inequality](): $P\{|x| \geq t\} \leq \mathbb{E}\{|X|\}/t, t > 0$, we have
$P\{\text{number of pairwise collisions} > t\} \leq \mathbb{E}_h\{\text{numbr of pairwise collisions}\}/t$

So pairwise collisions > 1, probability is less or equal to 1/2, and by repeatedly choosing and testing, the intersection of bad probability deceases as fast as $\frac{1}{2^n}$

Dealing the remaining 1 collision

In order to deal with the remaining 1 collision in the chosen function, we might use a second level of hash table to further split up collisions in that bucket.

Suppose in first level buckets: $\{0, \ldots, n-1\}$, one has k collisions.

Then we will create a second level of buckets for this bucket, ranging: $\{0, \ldots, k^2 - 1\}$

Performance analysis:

Claim: $\mathbb{E}_h\{\sum_i load(i)^2\} \leq O(|u|)$, h is the hash function.

Proof:

$$
\begin{aligned}
&\mathbb{E}_h\{\sum_i load(i)^2\} \\
&= \mathbb{E}_h\{\sum_i(\sum_x \mathbb{I}\{h(x)=i\})^2\} \\
&= \mathbb{E}_h\{\sum_i(\sum_{x,y\in u} \mathbb{I}\{h(x)=i\}\mathbb{I}\{h(y)=i\})\} \\
&= \sum_{x,y\in u} \mathbb{E}_h\{\sum_i \mathbb{I}\{h(x)=i\}\mathbb{I}\{h(y)=i\}\} \\
&= \sum_{x,y\in u} P\{h(x)=h(y)\} \\
&\le \sum_{x,y\in u} \begin{cases} 1, & x=y \\ \frac{1}{n}, & x\neq y \end{cases} \\
&\text{because } 1*|u|+1/n*|u|*|u| = O(|u|), |u|<|n| \\
&= O(|u|)
\end{aligned}
$$

Applications

gperf: a perfect hash function generator