1. For every $a \in [p]^+$ there exists a unique integer $x \in [p]^+$ such that $ax \bmod p = 1$

Pf: • Since $a \in [p]^+$, $1 \le a \le p-1$, then the greatest common divisor of a and p is 1.

By Bézout's identity, if $\gcd(a,p)=1$, there exists $x, y \in \mathbb{Z}$ such that $ax+py=1$.

Reducing module p, we get $ax \equiv 1 \pmod p$

• To prove the uniqueness of x, we can use contradiction. suppose there is another integer $k \in [p]^+$ that $ak \equiv 1 \pmod p$, then $ax \equiv ak \pmod p$

By definition of congruence, $p \mid |ax-ak|$, and 0 is the only non-negative number less than p that is also divisible by p. Thus, $|ax-ak|=0$

$$\Rightarrow ax = ak$$

$$x = k$$

We thus prove uniquess that there is no other integer than x for $a \in [p]^+$ and $ax \bmod p = 1$.

2. To have a hash function that assign inputs uniformly and maximize the number of collision, we can design a hash function that maps all inputs to a single output. For example, let our hash function family as

$H(x) = 0$.

This function maps all inputs uniformly to value 0, and all inputs will collide with each other, which results in a maximum number of collisions.

3. Let p be a large prime, ; $p > |U|$
For any integers $a \in \{1, \dots, p-1\} = [p]^+$, $b \in \{0, 1, \dots, p-1\} = [p]$.
define $h_{a,b}(x) = (ax+b) \bmod p \bmod m$.
Let $\mathcal{H} = \{h_{a,b} \mid a \in [p]^+, b \in [p]\}$ be set of all $p(p-1)$ such functions
Prove: $\mathcal{H}$ is 2-universal.

Fix four integers $t_1, t_2, x_1, x_2 \in [p]$ such that $x_1 \neq x_2$,
and $t_1 \neq t_2$.

The linear system $a x_1 + b \equiv t_1 \pmod p$ ...... ①
$a x_2 + b \equiv t_2 \pmod p$ ...... ②

①−②: $a(x_1 - x_2) \equiv t_1 - t_2 \pmod p$

The linear system has a unique solution $a, b \in [p]$ with $a \neq 0$,
where $a = (t_1 - t_2)(x_1 - x_2)^{-1} \bmod p$
$b = (t_2 x_1 - t_1 x_2)(x_1 - x_2)^{-1} \bmod p$.
(notice that b is in a similar expression of a).
Since $x_1 \neq x_2$, a is non-zero if and only if $t_1 \neq t_2$.
By what we prove in problem 1, which implies to here that there is exactly one possible pair of (a, b) that gives us $a x_1 + b = t_1$ and $a x_2 + b = t_2$.
Therefore,

$$\Pr_{a,b}[(a x_1 + b) \bmod p = t_1 \text{ and } (a x_2 + b) \bmod p = t_2] = \frac{1}{p(p-1)}$$

and $\Pr_{a,b}[h_{a,b}(x_1) = h_{a,b}(x_2)] = \sum_{i=1}^{N} \frac{1}{p(p-1)} = \frac{N}{p(p-1)}$.
where N is the number of ordered pairs $(t_1, t_2) \in [p]^2$ such that $t_1 \neq t_2$, but $t_1 \bmod m = t_2 \bmod m$.
For each fixed $t_1 \in [p]$, there are at most $\lfloor \frac{p}{m} \rfloor$ integers $t_2 \in [p]$ such that $t_1 \neq t_2$ but $t_1 \bmod m = t_2 \bmod m$.
Also, since p is prime, $\lfloor p/m \rfloor \leq (p-1)/m$,
So the number of such ordered pairs: $N \leq \frac{p(p-1)}{m}$.

Therefore, $Pr(h_{a,b}(x_1) = h_{a,b}(x_2)) \leq \frac{1}{p(p-1)} \cdot \frac{p(p-1)}{m} = \frac{1}{m}$
which is the condition for 2-universality.

4. Suppose we hash n items into a table of size m.
Let C_{ij} be the indicator variable that equals 1 if and only if $i \neq j$ and $h_{a,b}(i) = h_{a,b}(j)$
Let $C = \sum_{i<j} C_{ij}$ be the total number of pairwise collision.

Since hash values are assigned uniformly at random, the probability of a collision is exactly $\frac{1}{m}$.
Linearity of expectation implies:

$$E(C) = \sum_{i<j} Pr(h(i) = h(j)) \leq \binom{n}{2} \frac{2}{m} = \frac{n(n-1)}{m}$$

By Markov's inequality,

$$Pr(C \geq 1) \leq \frac{E(C)}{1} = 1$$

$$Pr(C \geq 1) - 1 \leq 1 - 1$$

$$1 - Pr(C \geq 1) > 0$$

$$P(C = 0) > 0$$

Therefore, there exists a hash function $h \in \mathcal{H}$ that achieves 0 collisions.

5. To prove Pr(no slot receives more than 2 log n hashed keys) $\geq 1-\frac{1}{n}$ is equivalent to prove

P(slots receive more than 2 log n hashed keys) $\leq \frac{1}{n}$

Let X_i be the number of keys hashed to slot i.

$E(X_i) = n/n = 1$

Now, considering the prob that slot i contains at least k keys.

There are $\binom{n}{k}$ choices for k keys.

The prob of any particular subset of k keys being hashed in slot i is $\frac{1}{n^k}$, so the union bound ($P(A \vee B) \leq P(A) + P(B)$) implies $P(X_i \geq k) \leq \binom{n}{k}\left(\frac{1}{n}\right)^k \leq \frac{n^k}{k!}\left(\frac{1}{n}\right)^k = \frac{1}{k!}$

In this case, we consider slots that have at least 2 log n keys, so set $k = 2\log n$, we have

$k! \geq 2^k = 2^{2\log n} = n^2$

which implies $P(X_i \geq 2\log n) \leq \frac{1}{n^2}$

This probability bound holds for every slot i.

Thus, by the union bound,

P(slots receive more than 2 log n hashed keys)

$\leq \sum_{i=1}^{n} P(X_i \geq 2\log n) \leq n \cdot \frac{1}{n^2} = \frac{1}{n}$

Therefore,

$= 1 - P$(slots receive more than 2 log n hashed keys) $\geq 1 - \frac{1}{n}$

which means

Pr(no slot receives more than 2 log n hashed keys) $\geq 1 - \frac{1}{n}$

6. To estimate F_2 accurately, we need to compute the sum of products of the frequencies of all 4 elements in the dataset so that we can find a bound for Variance of the output.

Analytically, As we output Z^2 for estimation of F_2, we need to find $E(Z^2)$ and $Var(Z^2)$, where we expect variance of Z^2 to be tight enough for the estimations.

By linearity of expectation, $Var(Z^2) \leq E(Z^4)$, which thus requires 4-wise independence.

To generate hash function for integers, for example, we can design polynomial hash function as

$$h(x) = ((ax^3 + bx^2 + cx + d) \bmod p) \bmod m,$$

where a, b, c, d are random coefficients chosen uniformly from the set of integers in $[p]^+$. This hash function will return a hash value in the range $[0, m-1]$

- Since storing one integers use 32 bits,
  To store a 4 polynomial hash functions with three degrees, we need $4 \times 32 = 128$ bits.
- Alternatively, if we have 64-bits computer where an integer is stored by 64 bits, we need $4 \times 64 = 256$ bits to store the hash function.

7. **Theoretical guarantees (proof on blackboard)**

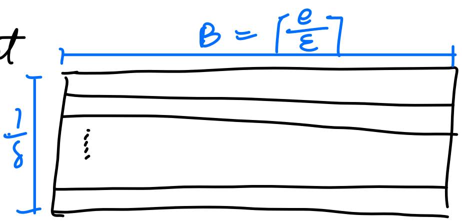
Suppose m is the length of the stream. Let the number of buckets B (#cols) be equal to $\lceil \frac{e}{\epsilon} \rceil$ and the number of repetitions (rows) set to **log(1/δ)**. Then the estimated frequency $\tilde{f}_i$ of the true frequency f_i satisfies the following guarantees:

1. $f_i \le \tilde{f}_i$
2. $\tilde{f}_i \le f_i + \epsilon m$ With probability at least 1-δ.

$B = \lceil \frac{e}{\varepsilon} \rceil$

$r = O(\log(\frac{1}{\delta}))$

Claim 1: $f_x \le \tilde{f}_x$ where x is the element



let $\hat{f}_{x,1}, \cdots, \hat{f}_{x,r}$ represent the estimated frequency of element x at row r.

m = # of items in stream

Therefore, the estimated frequency

$$\hat{f}_x = \sum_{h(y)=h(x)} f_y = f_x + \sum_{\substack{y \ne x \\ \text{but} \\ \text{hash}(x)=\text{hash}(y)}} f_y \geq f_x$$

The collision makes the estimated frequency being always larger than the actual frequency

claim 2: if $B = \lceil \frac{e}{\varepsilon} \rceil$, $r = \log\frac{1}{\delta}$,

then, $P_r(\tilde{f}_x \le f_x + \varepsilon m) \geq 1 - \delta$

($\tilde{f}_x$ is probability not to far from f_x)

proof: Pick item x and define r.vs $\{Z_1, Z_2, \cdots, Z_r\}$ s.t.

$Z_i = \underbrace{C_{i,h_i(x)}}_{\text{count in row } i} - f_x$ as the over-count in row i due to collisions

Let $X_{i,y}$ be the indicator for collision for $i \in \{1, 2, \cdots, r\}$, $y \in \{\text{distinct items}\} \setminus \{x\}$

$$X_{i,y} = \begin{cases} 1 & \text{if } h_i(y) = h_i(x) \\ 0 & \text{otherwise} \end{cases}$$

Therefore, $Z_i = \sum_{y \ne x} (f_y \cdot X_{i,y})$

$$E(Z_i) = E\left[\sum_{y \neq x} f_y \cdot X_{i,y}\right] \xrightarrow[\text{expectations}]{\text{linearity of}} \sum_{y \neq x} f_y \cdot E(X_{i,y})$$

$$= \sum_{x \neq y} f_y \cdot Pr(h_i(y) = h_i(x))$$

Since hash function family H holds universality,

$$\sum_{x \neq y} f_y \cdot Pr(h_i(y) = h_i(x)) \leq \sum_{x \neq y} f_y \cdot \frac{1}{B} \leq \frac{m}{B}$$

Therefore, expected per-row excess $E(Z_i)$ is at most $\frac{m}{B}$

By Markov's inequality:

$$Pr(Z_i \geq b \cdot E(Z_i)) \leq \frac{1}{b}, \quad \text{combine with } E(Z_i) \leq \frac{m}{B}$$

$$\underline{Pr(Z_i \geq b \cdot \tfrac{m}{B})} \leq Pr(Z_i \geq b \cdot E(Z_i)) \underline{\leq \tfrac{1}{b}}$$

$$\hookrightarrow Pr(Z_i \geq \frac{bm}{B}) \leq \frac{1}{b}$$

let $b = B\varepsilon$

$$Pr(Z_i \geq \varepsilon m) \leq \frac{1}{B\varepsilon}$$

Since $B = \lceil \frac{e}{\varepsilon} \rceil$

$$Pr(Z_i \geq \varepsilon m) \leq \frac{1}{e}$$

Therefore, $Pr(Z_i \geq \varepsilon m) = Pr(Z_i + f_x \geq f_x + \varepsilon m) \leq \frac{1}{e}$

If we repeat for r rows and take minimum, and for each row i we perform independently

$$Pr(\forall_{1 \leq i \leq r}\ Z_i \geq \varepsilon m) \leq \left(\frac{1}{e}\right)^r$$

Since $r = \log \frac{1}{\delta}$, $\left(\frac{1}{e}\right)^r = \left(\frac{1}{e}\right)^{\log \frac{1}{\delta}} = e^{\log \delta} = \delta$

Therefore, $Pr(\forall_{1 \leq i \leq r}\ Z_i \geq \varepsilon m) \leq \delta$

probability of bad estimate is $\leq \delta$

$$Pr(\exists_{1 \leq i \leq r}\ Z_i < \varepsilon m) \geq 1 - \delta$$

probability of good estimate is $\geq 1 - \delta$

so $\tilde{f}_x = \min(C_{1,h_1(x)}, C_{2,h_2(x)}, \cdots, C_{r,h_r(x)})$

$$= \min(f_x + Z_1, f_x + Z_2, \cdots, f_x + Z_r) \leq f_x + \varepsilon m \geq 1 - \delta$$