Instructions

- The homework is due on **Friday 4/7 at 5pm ET**.
- No extension will be provided, unless for serious documented reasons.
- Start early!
- Study the material taught in class, and feel free to do so in small groups, but the solutions should be a product of your own work.
- This is not a multiple choice homework; reasoning, and mathematical proofs are required before giving your final answer.

1 Theory problems [70 pts, 10 each]

In the following let p be a prime. For any integer m, define $[m] = \{0, \ldots, m-1\}$ and $[m]^+ = \{1, \ldots, m-1\}$.

1. Prove that for every $a \in [p]^+$ there exists a unique integer $x \in [p]^+$ such that

$$ax \bmod p = 1.$$

2. Answer question 1 on slide 5. Specifically, give a family of hash functions that satisfies the uniformity property but maximizes the number of collisions. Your answer should formally prove why the specific family has the two latter properties.

3. Let $h_{ab} = (ax + b) \mod p \mod m$ where $a \in [p]^+, b \in [p]$ and p is a prime such that $p \geq m$. Prove that $\mathcal{H} = \{h_{ab}\}$ is 2-universal.

4. Consider a 2-universal family of hash functions $\mathcal{H}$ that hash the universe U to $[m]$. Assume you have n keys $m > \binom{n}{2}$. Prove that there exists a hash function $h \in \mathcal{H}$ that achieves 0 collisions.

$$E[X^2] = E\left[\sum_{j=1}^{n} f_j^2 Y_j^2 + \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} f_i f_j Y_i Y_j\right] = F_2$$

   Hint: Let C be the RV of number of collisions. Prove that $Pr_{h \in \mathcal{H}}(C = 0) > 0$.

5. Suppose we hash n keys to n slots. Prove that with probability at least $1 - \frac{1}{n}$ there is no slot that receives more than $2 \log n$ hashed keys.

6. Explain why estimating F_2 requires 4-wise independence. Describe how you can generate such as hash function for integers and explain how many bits are needed to store it?

7. In class, we went over the theoretical guarantees (slide 60) of Count-Min sketch when $B = \lceil \frac{3}{\epsilon} \rceil$ and $r = O(\log(\frac{1}{\delta}))$ where $\epsilon, \delta > 0$ are the accuracy and confidence parameters. Your task is the following:

   - Write a formal proof of both guarantees 1. and 2. on slide (slide 60). Set the number of buckets $B = \lceil \frac{e}{\epsilon} \rceil$.

2 Coding [30 points]

Check the Jupyter notebook on our Git repo.

In the following let p be a prime. For any integer m, define $[m] = \{0, \ldots, m-1\}$ and $[m]^+ = \{1, \ldots, m-1\}$.

1. Prove that for every $a \in [p]^+$ there exists a unique integer $x \in [p]^+$ such that

$$ax \bmod p = 1.$$

Proof by contradiction:

Assume there exists more than one integer x satisfiny $ax \bmod p = 1$.

Suppose $ax_1 \bmod p = ax_2 \bmod p$, $x_1, x_2 \in [p]^+$, $x_1 \neq x_2$

$\Rightarrow$ $a(x_1 - x_2) \bmod p = 0$

$\because$ p is a prime, $a \in [1, p-1]$

$\therefore$ p, a are coprime, which means $(x_1 - x_2)$ can be divided by p.

$x_1 - x_2 = k \cdot p$, k is an integer.

Yet, we know that $x \in [1, \ldots, p-1]$, so $x_1 - x_2 < p$

$\frac{x_1 - x_2}{p} = k < 1$. k can't be an integer which is a contradiction.

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

2. Answer question 1 on slide 5. Specifically, give a family of hash functions that satisfies the uniformity property but maximizes the number of collisions. Your answer should formally prove why the specific family has the two latter properties.

Proof:

$H: \{h_0, h_2, \cdots, h_{m-1}\}$, $h_0(x) = 0$, $h_1(x) = 1$, $h_2(x) = 2, \cdots$ $h_{m-1}(X) = m-1$

- To show it's uniform, we need to show

$$\boxed{\Pr_{h \in H}(h(x) = i) = \frac{1}{m} \text{ for all } i, x.}$$

By definition of H above,

for all x, $h_i(x) = i$, $\Pr(h(x) = i \mid h(x) = h_i(x)) = 1$

Since there are m hash functions in H,

$$\Pr_{h \in H}(h(x) = h_i(x)) = \frac{1}{m}$$

$$\Pr(h(x) = h_i(x) \mid h(x) = i) = 1$$

* since there's no other hash function hashing x to i but $h_i(x)$.

$$\Pr_{h \in H}(h(x) = i) = \frac{\Pr(h(x) = i \mid h(x) = h_i(x)) \cdot \Pr(h(x) = h_i(x))}{\Pr(h(x) = h_i(x) \mid h(x) = i)}$$

$$= \frac{1 \cdot \frac{1}{m}}{1} = \frac{1}{m}$$

- Collision: items collide if their hash values are equal.

$\because$ for each item in $h_i(x)$ their hash values are equal.

which is $h_i(a) = h_i(b) = h_i(c) = \cdots = h_i(n) = i$

$\therefore$ The number of collisions = The number of items in the universal set.

∎

3. Let $h_{ab} = (ax + b) \mod p \mod m$ where $a \in [p]^+, b \in [p]$ and p is a prime such that $p \geq m$. Prove that $\mathcal{H} = \{h_{ab}\}$ is 2-universal.

Proof:

To prove 2-universal, we need to show that $\Pr_{h \sim H}(h(x) = h(y)) \leq \frac{1}{m}$ $x \neq y$.

$a \in \{1, \cdots, p-1\}$, $b \in \{0, \cdots, p-1\}$, $\Rightarrow a, p$ are coprime, b, p are coprime.

$r \equiv (ax+b) \mod p$, $s \equiv (ay+b) \mod p$, $r \neq s$.

$\because a \neq 0$, $\therefore$ solution (a, b) is unique.

$$ax \equiv r - b \mod p$$
$$ay \equiv s - b \mod p$$
$$ax - ay \equiv (r-b) - (s-b) \mod p$$
$$a(x-y) \equiv (r-s) \mod p$$
$$a \equiv (r-s)(x-y)^{-1} \mod p \quad \because x \neq y$$

$$ax \equiv r - b \mod p$$
$$b \equiv r - ax \mod p$$

$\Rightarrow$ $h_{ab}(x) \equiv r \mod m$, $h_{ab}(y) \equiv s \mod m$ $\Rightarrow$ $h_{ab}(x) = h_{ab}(y)$ so $h_{ab}(x) \equiv h_{ab}(y) \mod m$ $\Rightarrow$ $r \equiv s \mod m$, but $r \neq s$

$|H| = p \cdot (p-1)$,

$$\Pr_{h \sim H}(h_{ab}(x) = h_{ab}(y)) = \frac{|\overset{\text{collision}}{h_{ab}(x) = h_{ab}(y)}|}{|H|} \leqslant \frac{\frac{p(p-1)}{m}}{p(p-1)} = \frac{1}{m}$$

∎

4. Consider a 2-universal family of hash functions $\mathcal{H}$ that hash the universe U to $[m]$. Assume you have n keys $m > \binom{n}{2}$. Prove that there exists a hash function $h \in \mathcal{H}$ that achieves 0 collisions.

Hint: Let C be the RV of number of collisions. Prove that $Pr_{h \in \mathcal{H}}(C = 0) > 0$.

Proof

$\Pr(h_1(x) = h_1(y), x \neq y) = \frac{1}{m}$, and for n keys, there're $\binom{n}{2}$ pairs of keys that can collide. Thus, $\Pr(\text{at least a collision}) \leq \binom{n}{2} \cdot \frac{1}{m} < m \frac{1}{m} = 1$.

$\Pr(\text{no collision}) = 1 - \Pr(\text{at least a collision}) > 0$

$\therefore$ There exists a hash function that achieves 0 collision.

∎

5. Suppose we hash n keys to n slots. Prove that with probability at least $1 - \frac{1}{n}$ there is no slot that receives more than $2 \log n$ hashed keys.

Proof:

$$\Pr(\text{ slot } j \text{ has } k \text{ keys}) = \binom{n}{k}\left(\frac{1}{n}\right)^k$$

$$= \frac{n!}{(n-k)!\,k!} \cdot \frac{1}{n^k} = \frac{n^k}{k!\,n^k} = \frac{1}{k!} \le \frac{1}{e^k}$$

$$\Pr(\text{ slot } j \text{ has } 2\log n \text{ keys}) = \frac{1}{e^{2\log n}} = \frac{1}{n^2}$$

$$\Pr(\text{no slot} > 2\log n) = \Pr(\text{all slots} \le 2\log n) = 1 - \Pr(\text{all slots} > 2\log n)$$

$$= 1 - n \cdot \Pr(\text{slot } j > 2\log n) \le 1 - \frac{n}{n^2} = 1 - \frac{1}{n}$$

$\blacksquare$

6. Explain why estimating F_2 requires 4-wise independence. Describe how you can generate such as hash function for integers and explain how many bits are needed to store it?

$Y_j = h(j) = 1$ or -1

Since $F_2 = E(X^2) = E[\sum_{j=1}^{n} f_j^2 Y_j^2 + \sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} f_i f_j Y_i Y_j]$ requires 2-wise independence, in order to make $E(X^2)$ more accurate, we need to minimize its variance. Thus, 4-wise independence is needed.

$$Var[X^2] = E[X^4] - (E[X^2])^2 = E[X^4] - F_2^2$$

$$E[X^4] = \sum_{j=1}^{n} f_j^4 E[Y_j^4] + \binom{4}{2} \sum_{i=1}^{n}\sum_{j=i+1}^{n} E[Y_i^2 Y_j^2] f_i f_j$$

$$= \sum_{j=1}^{n} f_j^4 E(Y_j^4] + 3\left(\left(\sum_{j=1}^{n} f_i\right)^2 - \sum_{j=1}^{n} f_j^4\right)$$

$$= F_4 + 3F_2^2 - 3F_4 = 3F_2^2 - 2F_4$$

$\because E[Y_i Y_j] = \sum Y_i Y_j \Pr(Y_j) = \begin{cases} 1, & Y_i = Y_j \\ 0, & o.w. \end{cases}$

$\because 1^2 = 1,\ -1^2 = 1$

$\therefore E[Y_i^2 Y_j^2] = 1$

$\therefore E(X^2)^2 = E[Y_i^2 Y_j^2] \cdot f_i f_j = \begin{cases} 1 & f_i = f_j \\ 0 & o.w \end{cases}$

$$Var[X^2] \le 3F_2^2 - 2F_4 - F_2^2 = 2F_2^2 - 2F_4 \le 2F_2^2 = 2 \cdot (E[X^2])^2$$

$\Rightarrow 4$ bits.

7. In class, we went over the theoretical guarantees (slide 60) of Count-Min sketch when $B = \lceil \frac{3}{\epsilon} \rceil$ and $r = O(\log(\frac{1}{\delta}))$ where $\epsilon, \delta > 0$ are the accuracy and confidence parameters. Your task is the following:

   - Write a formal proof of both guarantees 1. and 2. on slide (slide 60). Set the number of buckets $B = \lceil \frac{e}{\epsilon} \rceil$.

Proof:

① $f_x \le \hat{f}_x$

Since $\hat{f}_x = f_x +$ collision,

we have $f_x \le \hat{f}_x$.

② $\hat{f}_x \le f_x + \varepsilon m$ w.p $\ge 1-\delta$

$E(\text{collision}) \le \frac{m}{B}$

$\because \hat{f}_x = f_x + \text{collision}$, $\therefore \hat{f}_x \le f_x + \varepsilon m \Rightarrow Pr(\text{min collision} \ge \varepsilon m) \le \delta$

$Pr(\text{collision}_i \ge \varepsilon m) = \frac{E(\text{collision})}{\varepsilon m} \le \frac{\frac{m}{B}}{\varepsilon m} = \frac{1}{\varepsilon B} \le \frac{1}{\varepsilon} \cdot \frac{\varepsilon}{e} = \frac{1}{e}$ for each row

$Pr(\min(\text{collision}_1, \dots, \text{collision}_r) \ge \varepsilon m) = Pr(\text{Collision}_1 \ge \varepsilon m) \cdot \ldots \cdot Pr(\text{collision}_r \ge \varepsilon m)$
$\le \left(\frac{1}{e}\right)^r = \delta$

Thus, $\hat{f}_x \le f_x + \varepsilon m$ w.p $\ge 1-\delta$.

$\blacksquare$