24 Feb 2025 — Streaming Algorithms

Announcements.

		Regrade Deadline
(1)	Quiz 3 is graded.	Sun 3/2
(2)	PSet 1 is graded.	Sun 3/9
(3)	Solution Set 1 on Canvas.	
(4)	Quiz 4 on Wed, 2/26, will cover 2/12, 2/19, 2/24.	

RECAP. A family of hash functions $\mathcal{H} = \{h : \mathcal{X} \to \mathcal{B}\}$ is 2-universal ("pairwise independent") if

$\forall x \neq y \in \mathcal{X}$ $(h(x), h(y)) \in \mathcal{B}^2$ is uniformly distributed.

Ex. $\mathcal{X} = \mathcal{B} = \mathbb{F}_p = \{0, 1, \dots, p-1 \pmod p\}$ p prime.

$$h_{a,b}(x) = ax + b \pmod p.$$

Why is $\mathcal{H} = \{ h_{a,b} \mid a, b \in \mathbb{F}_p \}$ 2-univ.?

For any fixed $x \neq y$ the pair $(h(x), h(y))$ can take any value in $\mathbb{F}_p^2$...

because $h(x)$ can be made to take any value by varying b,

and $h(y) - h(x) = ay + b - (ax + b) = a \cdot (y - x)$

can be made to take any value by varying a_i

Look at $0 \cdot (y-x) \mod p$

$1 \cdot (y-x) \mod p$

$\vdots$

$(p-1) \cdot (y-x) \mod p.$

If any two of these coincide it means $\exists\, j<k$ in $\{0, \ldots, p-1\}$ s.t.

$$j \cdot (y-x) \equiv k \cdot (y-x) \pmod p$$

$\Rightarrow$ $\underline{(j-k)} \cdot \underline{(y-x)}$ is divisible by p.

both factors strictly between $\emptyset$ and p.

$\Rightarrow$ contradiction. (p prime)

Inner Product Hashing.

$$\mathcal{X} = \mathbb{F}_p^d \qquad d > 0$$

$$\mathcal{B} = \mathbb{F}_p$$

Def. If $a = (a_1, \dots, a_d)$, $x = (x_1, \dots, x_d) \in \mathbb{F}_p^d$

let $\langle a, x \rangle = \sum_{i=1}^{d} a_i x_i \pmod p$

$$h_{\vec{a},b}(\vec{x}) = \langle \vec{a}, \vec{x} \rangle + b$$

$$h_{\vec{a},b} : \mathcal{X} \to \mathcal{B}$$

$$\mathcal{H}_p^d = \{ h_{\vec{a},b} \mid \vec{a} \in \mathbb{F}_p^d,\ b \in \mathbb{F}_p \}$$

Space complexity of representing $h \in \mathcal{H}_p^d$: $O(d)$ space to store $d+1$ values in $\mathbb{F}_p$

(assume p is small enough that its binary representation fits in $O(1)$ memory.)

time complexity of evaluating $h_{\vec{a},b}$: $O(d)$

d mult and d add in $\mathbb{F}_p$.

$\mathcal{H}_p^d$ is 2-universal by a variant of the arguments already presented.

Consider $\vec{x} \neq \vec{y} \in \mathbb{F}_p^d$.

WLOG assume $x_1 \neq y_1$.

For any fixed list of coefficients $a_2, \ldots, a_d$ we have p^2 ways to choose a_1 and b.

By varying b can make $h(x)$ take any value in $\mathbb{F}_p$.

By varying a_1 can make $h(y) - h(x)$ take any value in $\mathbb{F}_p$.

$$h(y) - h(x) = \langle \vec{a}, \vec{y} - \vec{x} \rangle$$

$$= \boxed{a_1 \cdot (y_1 - x_1)} + \boxed{\sum_{i=2}^{d} a_i (y_i - x_i)}$$

ranges over all of $\mathbb{F}_p$ as a_1 varies.

constant as $a_2, \ldots, a_d, \vec{x}, \vec{y}$ are fixed.

Streaming algorithm for distinct elements

(ε, δ) – PAC algorithm:

for all input sequences / streams

with probability $\geq 1-\delta$

the algorithm's answer is correct

within ε relative error.

Flajolet – Martin algorithm for distinct elements.....

[imprecise version, $\varepsilon = 5$]

Suppose elements have b-bit identifiers so that $\{\text{potential elements}\} \subseteq [2^b]$.

Let p be a prime $\geq 2^b$.

Sample $h = h_{a,b} \in \mathcal{H}_p^1$ uniformly at random.

initialize $z = p-1$

for $i = 1, \ldots, n$:

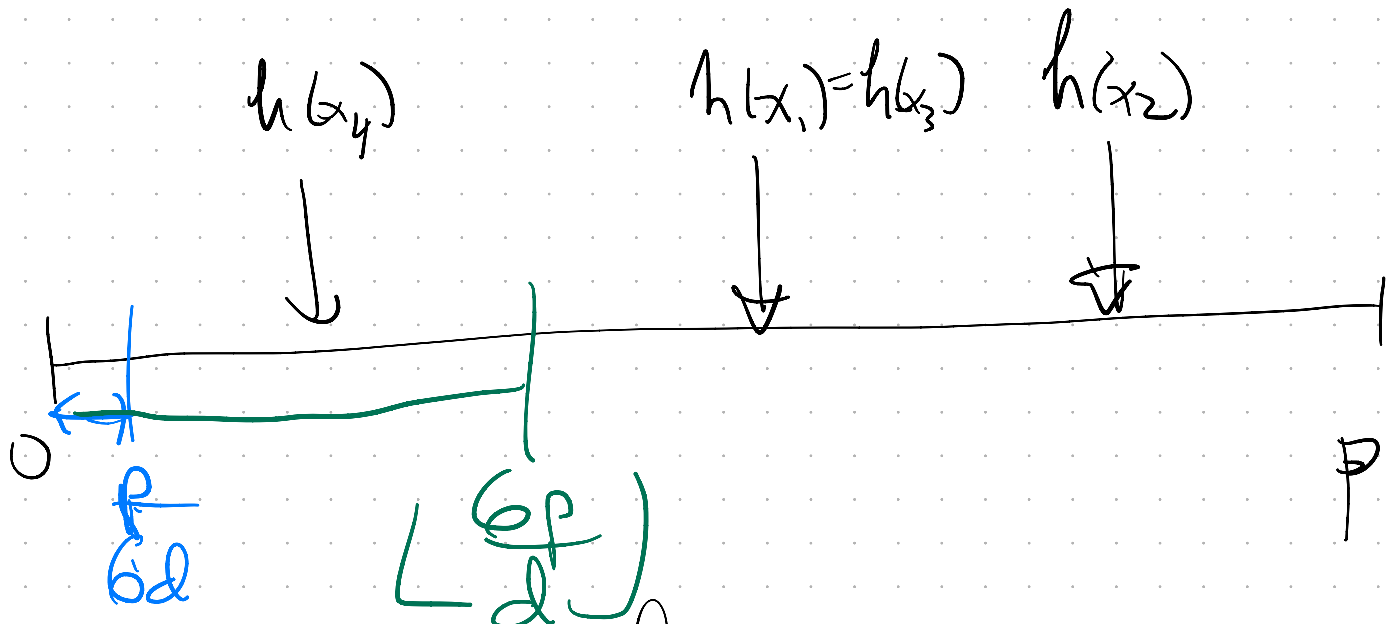
observe x_i from data stream

compute $h(x_i)$

update $z = \min\{h(x_1), \ldots, h(x_i)\}$.

After observing the stream:

output $\dfrac{p}{z+1}$.



If there are d distinct values among $x_1, \ldots, x_n$ then you expect $\{h(x_1), \ldots, h(x_n)\}$ to be d distinct numbers spread uniformly in the interval $[0, p]$.

If they were <u>exactly</u> evenly spaced then z would be at $\frac{p}{d}$.
(Maybe $\frac{p}{d+1}$) ✓

So $\dfrac{p}{z}$ ought (?) to be a good estimate of d or $d+1$

<u>Analysis.</u> I aim to prove

(A) $\Pr\left(\dfrac{p}{z+1} > 6d\right) \leq \dfrac{1}{6}$

(B) $\Pr\left(\frac{p}{Z+1} < \frac{d}{6}\right) \leq \frac{1}{6}$.

So then $\Pr\left(\frac{d}{6} \leq \frac{p}{Z+1} \leq 6d\right) \geq \frac{2}{3}$.

... i.e., the algorithm is $(\varepsilon = 5, \delta = 1/3)$-PAC

Proof of (A). $\frac{p}{Z+1} > 6d$

$$\iff Z+1 < \frac{p}{6d}.$$

Each individual hash value $h(x_i)$ is unif distrib over $\{0, \ldots, p-1\}$

$$\Rightarrow \Pr\left(h(x_i) < \frac{p}{6d}\right) \leq \frac{1}{6d}.$$

There are d distinct identifiers in the stream

$$\Rightarrow E\left[\#\text{ identifiers that hash to } < \frac{p}{6d}\right]$$

$$\leq \left(\frac{1}{6d}\right) \cdot d \leq \frac{1}{6}.$$

(linearity of expectation)

$$\Pr\left(\exists \text{ an identifier hashing to } < \frac{p}{6d}\right) \leq \frac{1}{6}$$

(Markov's inequality)

Proof of (B). Assume WLOG that $x_1, \ldots, x_d$ are the d distinct elements in the stream.

Random variables

$$X_{ik} = \begin{cases} 1 & \text{if } h(x_i) \leq k \\ 0 & \text{o.w.} \end{cases}$$

$$Y_k = \sum_{i=1}^{d} X_{ik}$$

Event (B): $\frac{p}{Z+1} < \frac{d}{6} \iff Z > \frac{6p}{d} - 1$

Set $k = \left\lfloor \frac{6p}{d} \right\rfloor$.

$$Z > \frac{6p}{6} - 1 \iff Y_k = 0.$$

$\forall i \quad \mathbb{E}[X_{ik}] \geq \frac{6}{d}$.

$\implies \mathbb{E}[Y_k] \geq 6$ (lin of exp.)

$\boxed{\text{FACT}}$ For a Bernoulli $\{0,1\}$-valued random variable X

$$Var(X) = E[X] \cdot (1 - E[X]).$$

$$\leq E[X].$$

$$Var(X_{ik}) \leq E[X_{ik}]$$

$$\sum_i Var(X_{ik}) \leq E\left[\sum X_{ik}\right]$$

$$\| \qquad\qquad\qquad \|$$

pairwise independence

$$Var(Y_k) \qquad\qquad E[Y_{1k}]$$

__Lemma.__ If $X_1, \ldots, X_n$ are __pairwise__ independent and

$$Y = X_1 + \cdots + X_n,$$

$$Var(Y) = \sum_{i=1}^{n} Var(X_i)$$

__Proof.__ $\mathrm{Var}(Y) = \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2$

$$\mathbb{E}[Y] = \sum_{i=1}^{n} \mathbb{E}[X_i].$$

$$(\mathbb{E}[Y])^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{E}[X_i] \cdot \mathbb{E}[X_j].$$

$$Y^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} X_i X_j$$

$$\mathbb{E}[Y^2] = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{E}[X_i X_j].$$

$$\mathrm{Var}(Y) = \sum_{i=1}^{n} \sum_{j=1}^{n} \underbrace{\left(\mathbb{E}[X_i X_j] - \mathbb{E}[X_i] \cdot \mathbb{E}[X_j]\right)}_{\substack{=0 \text{ when} \\ X_i, X_j \text{ independent.}}}$$

$$= \sum_{i=1}^{n} \mathbb{E}[X_i^2] - (\mathbb{E}[X_i])^2$$

$$= \sum_{i=1}^{n} \mathrm{Var}(X_i).$$