Linear Sketching

CountMin Sketch

Count Sketch

EECS 477

Seth Pettie

Streaming & Sketching

- You need to maintain a vector $\boldsymbol{x} \in \mathbb{Z}^u$ of integers. Initially $\boldsymbol{x} = 0$.
  - Update(i, Δ) : Set $x(i) = x(i) + \Delta$.
  - Query(i) : Return $x(i)$.
  - *Look at other types of queries later…*

- "Incremental" : all Δs are positive.
- "Strict Turnstyle" : positive and negative Δ, but $\forall i.\, x(i) \geq 0$ at all times.

- $\Theta(u)$ space is necessary and sufficient. Problem solved!

Streaming & Sketching

- You need to maintain a vector $\boldsymbol{x} \in \mathbb{Z}^u$ of integers. Initially $\boldsymbol{x} = 0$.
  - Update(i, Δ) : Set $x(i) = x(i) + \Delta$.
  - Query(i) : Return $\tilde{x}(i) = x(i) \pm Err(x)$.
  - *Look at other types of queries later…*

- "Incremental" : all Δs are positive.
- "Strict Turnstyle" : positive and negative Δ, but $\forall i.\, x(i) \geq 0$ at all times.

- Now how much space do you need? Depends a lot on what "$Err(x)$" is…

CountMin Sketch

- **<u>Theorem.</u>** A data structure of size $O(\epsilon^{-1} \log u)$ can handle **Update/Query** operations in $O(\log u)$ time. **Query** returns $\tilde{x}(i) = x(i) \pm \|x\|_1$.

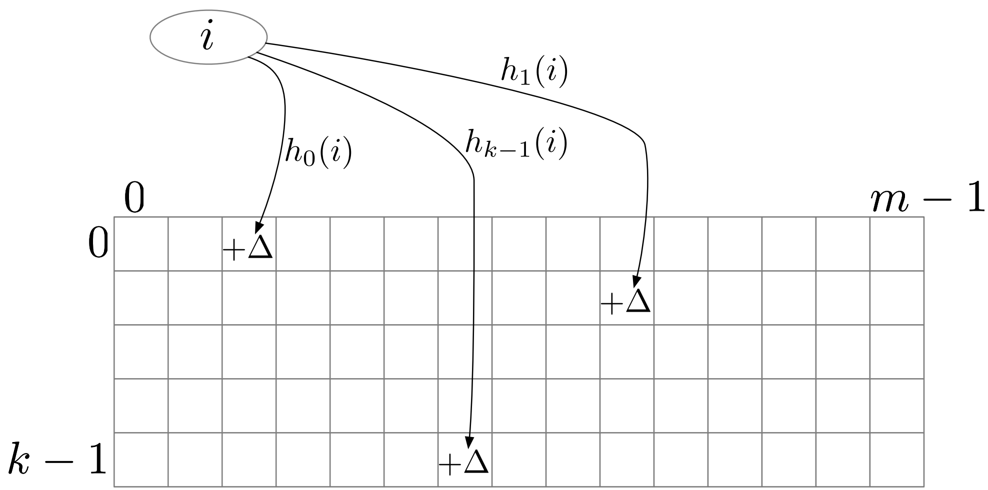
- "ℓ_1-norm" $\|x\|_1 = \sum_{i=0}^{u-1} x(i)$. Manhattan distance

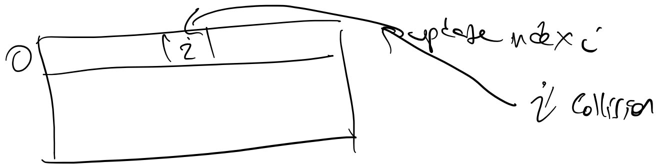
- "ℓ_2-norm" (aka Euclidean norm) $\|x\|_2 = \sqrt{\sum_{i=0}^{u-1} x(i)^2}$.

- $F_2 = 2^{\text{nd}}$ moment $= \|x\|_2^2 = \sum_{i=0}^{u-1} x(i)^2$.

CountMin Sketch

- Choose k hash functions $h_1, \ldots, h_k : [u] \rightarrow [m]$ from a 2-independent family.
- Allocate an $k \times m$ array A, initially all 0.
- Update(i, Δ) :
  - For$(j = 0;\ j < k;\ j++)$
    - $A[j, h_j(i)] \mathrel{+}= \Delta;$





$\frac{4}{m}$ collisions in expectation

$I_{i'} : h_0(i) = h_0(i')$

indicator for event

$$E(I_{i'}) = \begin{cases} 1 & i' = i \\ \frac{1}{m} & i' \neq i \end{cases}$$

$$\text{Query}(i) = \min_j \left(A[j, h_j(i)] \right)$$

$= O(k)$

$$E(A[0, h_0(i)])$$

$$= E\left(\sum_{i'=0}^{m-1} I_i \cdot x(i') \right)$$

constant

$$= \sum_{i'=0}^{m-1} E(I_{i'}) \cdot x(i')$$

x norm except x(i)

$$= x(i) + \left(\frac{\|x\|_1 - x(i)}{m} \right)$$

positive noise

$$\left(\frac{1}{m}\right)\left(x(0) + \dots + x(m-1) \right)$$

except x(i)

- What is $E(A[0, h_0(i)]) = ?$
  - Hint: indicators!
- $I_{i'}$: an indicator for the event that $h_0(i) = h_0(i')$.
- $E(A[0, h_0(i)]) = \sum_{i'=0}^{u-1} E(I_{i'}) \cdot x(i')$
- $= x(i) + (\|x\|_1 - x(i))/m$
- How should we implement Query(i)?
- Query(i)
  - Return $\min\{A[0, h_0(i)], \ldots, A[k-1, h_{k-1}(i)]\}$.

$$P(A[0,h_0(i)] - x(i) \geq err) \underset{\text{Markov's}}{\leq} \frac{E(A[0,h_0(i)] - x(i))}{err}$$

$$\leq \frac{\|x\|_1/m}{err} \overset{\text{want this}}{=} \frac{1}{2}$$

$$err = \left(\frac{2}{M}\right)\|x\|_1$$

At 12:46

$$P(\forall_j\, A[0,h_0(j)] - x(j) \geq err) \leq \left(\frac{1}{2}\right)^k = \frac{1}{\mu^2} \quad k = 2\log\mu$$

- Define Err $= \left(\frac{2}{m}\right)\|x\|_1$
- What is $\Pr(\text{Query}(i) \notin [x(i), x(i) + \text{Err}])$?

Markov's Ineq.

- $\Pr\left(A[j, h_j(i)] - x(i) > 2 \cdot E\left(A[j, h_j(i)] - x(i)\right)\right) < \frac{1}{2}$.
  - $E\left(A[j, h_j(i)] - x(i)\right) < \frac{\|x\|_1}{m} = \frac{\text{Err}}{2}$.
- Thus, $\Pr\left(A[j, h_j(i)] > x(i) + \text{Err}\right) < \frac{1}{2}$.
- Finally, $\Pr\left(\forall j. A[j, h_j(i)] > x(i) + \text{Err}\right) < \left(\frac{1}{2}\right)^k$.
- Set $k = 2\log u, m = 2\epsilon^{-1}$, then
  - $\forall i. \text{Query}(i) \in [x(i), x(i) + \epsilon\|x\|_1]$ with prob. $1 - 1/u$.

Heavy Hitters

- Want to implement

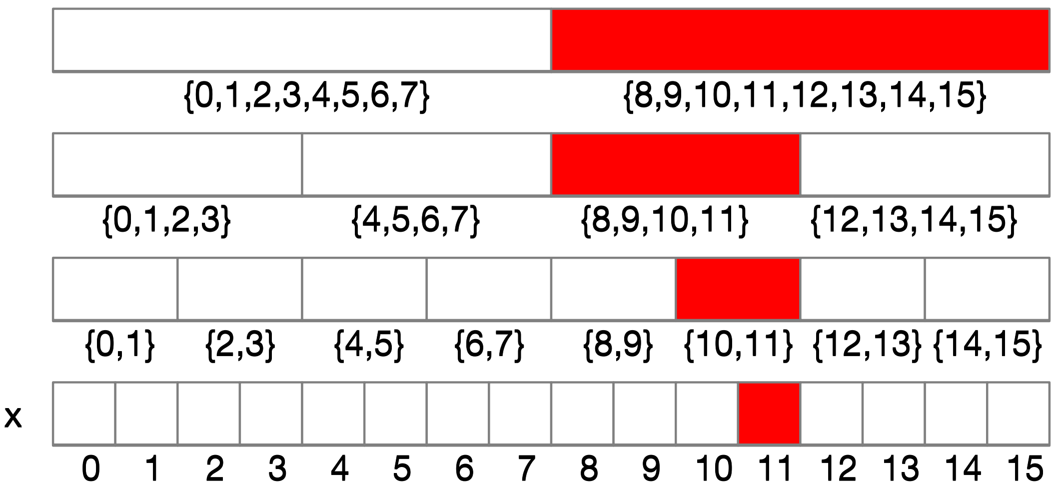
  $\text{Err} = \varepsilon \|x\|_1 \quad m = 2\varepsilon^{-1}$

  – HeavyHitters() : return a short list L that includes all (2ϵ)-heavy hitters: i s.t. $x(i) \geq 2\epsilon \|x\|_1$.

- For i from 0 to $u - 1$:
  – If $\text{Query}(i) \geq 2\epsilon \|x\|_1$, add i to L.
- **Claim:** $|L| \leq \epsilon^{-1}$.
- **Proof:** Any $i \in L$ has $x(i) \geq \epsilon \|x\|_1$.

Heavy Hitters

- Want to implement
  - HeavyHitters() : return a short list L that includes all (2ϵ)-heavy hitters: i s.t. $x(i) \geq 2\epsilon\|x\|_1$.
- At most ϵ^{-1} indices i at each level with $x(i) \geq 2\epsilon\|x\|_1$.
- $O(\epsilon^{-1}\log^2 u)$ time to find them all.



Build CountMin structure for each of $\log u$ vectors

If an element is a 2ϵ-HH, its parent is also a 2ϵ-HH.

If an element is a 2ϵ-HH, 0, 1, or 2 of its children Could be 2ϵ-HHs.

Count Sketch

- CountMin Sketch: ℓ_1 error; good for finding **2ϵ-heavy hitters.** (i s.t. $x(i) \geq 2\epsilon\|x\|_1$)
- Count Sketch: ℓ_2 error; good for finding ℓ_2 **2ϵ-heavy hitters**: i s.t $x(i) \geq 2\epsilon\|x\|_2$.
- Pick hash functions $h_0, \dots, h_{k-1}: [u] \to [m]$ as before.
- Pick hash functions $g_0, \dots, g_{k-1}: [u] \to \{-1,1\}$ from a 2-wise independent family.
- Update(i, Δ) :
  - For($j = 0;\ j < k;\ j++$)
    - $A[j, h_j(i)] += g_j(i) \cdot \Delta;$

$x = (\sqrt{n}, \underbrace{1, 1, 1, 1, 1, \_\_\_\_, 1}_{n\ 1's}, 0 \_\_)$

$\varepsilon = 0.01$

$\|x\|_2 = \sqrt{2n}$ ℓ_2 norm picks up different heavy hitters

$$E(g_0(i) \cdot A[0, h_0(i)]) = E\Big(g_0(i) \sum_{i'} I_{i'} \cdot g_0(i') \cdot x(i')\Big)$$

$$= g_0(i)^2 x(i) + \sum_{i' \neq i} \underbrace{E(I_{i'} \cdot g_0(i'))}_{0} \cdot x(i')$$

$$= g_0(i) x(i)$$

$$= x(i)$$

$$Var(\quad) = E\big((g_0(i) \cdot A[0, h_0(i)] - x(i))^2\big)$$

$$= E\Bigg(\Big(\sum_{i' \neq i} g_0(i) \cdot g_0(i') \cdot x(i') \cdot I_{i'}\Big)^2\Bigg)$$

when $i' = i''$

$$= E\Big(\sum_{i'=i''} x(i')^2 I_{i'} + \underbrace{\sum_{i' \neq i''} \overset{=1}{g_0(i)^2}\, \overset{=}{g_0(i')\, g_0(i'')}\, \overset{=}{x(i')\, x(i'')}\, \overset{=}{I_{i'}\, I_{i''}}}_{0}\Big)$$

$$= E\Big(\sum_{i'=i''} x(i')^2 \cdot I_{i'}\Big) = \sum_{i'=i''} x(i')^2 \cdot E(I_{i'}) \leq \frac{\|x\|_2^2}{m} = \frac{F_2}{m}$$

$$P\big(|g_0(i) \cdot A[0, h_0(i)] - x(i)|^2 > t^2\big) \underset{\text{Markov's}}{\leq} \frac{E(g_0(i) \cdot A[0, h_0(i)] - x(i))^2}{t^2}$$

$$\leq \frac{\|x\|_2^2}{m t^2} = \frac{1}{3}$$

Query(i): $g_0(i) \cdot A[0, h_0(i)]$

$\vdots$

get the median of this $\longrightarrow$ $g_{k-1}(i) \cdot A[k-1, h_{k-1}(i)]$

$t = \varepsilon \|x\|_2 \qquad m = 3 \cdot \varepsilon^{-2}$

$K = 10 \log n$

Count Sketch

- What is $E(A[0, h_0(i)]) = ?$
- $= E\left(\sum_{i'} I_{i'} \cdot g(i')x(i')\right)$
- $= g(i)x(i) + \sum_{i'} E\Big(I_{i'} \cdot g(i')\Big)x(i') = g(i)x(i).$
- $\mathrm{Var}(g(i)A[0, h_0(i)]) =$ $E\left(\left(\sum_{i' \neq i} g(i)g(i')I_{i'} \cdot x(i')\right)^2\right)$
- $= \sum_{i' \neq i} E(I_{i'}) \cdot x(i')^2 +$ $\sum_{i' \neq i''} g(i')g(i'')I_{i'}I_{i''}x(i')x(x'')$
- $\leq \|x\|_2^2/m = F_2/m.$

$$\Pr(|g_0(i)A[0,h_0(i)] - x(i)| > t) < \frac{E((A[0,h_0(i)] - x(i))^2)}{t^2}$$

square markov
Chebyshev

$$= t = \varepsilon\|x\|_2, \quad m = \varepsilon^2 \cdot 3$$

$$\Rightarrow \frac{\|x\|_2^2/m}{t^2} = \frac{1}{3}$$

↳ now we get good approx with prob 2/3

- Query (i)
  - Return $\text{median}_{0 \le j < k} \{g_j(i)\, A[j, h_j(i)]\}$

<u>Claim</u>: if $k \ge 60 \ln n$, Query returns $x(i) \pm \varepsilon\|x\|_2$
with prob $1 - 1/n^3$

Chernoff Bounds

No pf in slides

- $X = X_1 + \cdots + X_n$, $X_i \in \{0,1\}$, and $\Pr(X_i = 1) = p$.
- $\Pr(X > (1+\delta)np) < \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{np} < e^{-\frac{\delta^2}{2+\delta}np}$.
- $\Pr(X < (1-\delta)np) < \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{np} < e^{-\frac{\delta^2}{2}np}$.

Back to Count Sketch

- $X = X_1 + \cdots + X_k$ defined so that
  - $X_j = \begin{cases} 0 & \text{if } g_j(i)A[j,h_j(i)] \text{ is good approx of } x(i) \\ 1 & \text{if } g_j(i)A[j,h_j(i)] \text{ is bad approx of } x(i) \end{cases}$
  - $\Pr(X_j = 1) = \frac{1}{3} = p.\ \ E(X) = k/3.$
- Query(i) *could* return a bad approx. if $X \geq k/2$.
- $\Pr(X \geq (1.5)E(X)) \leq \left(\frac{e^{\frac{1}{2}}}{1.5^{1.5}}\right)^{E(X)} < e^{-.1\left(\frac{k}{3}\right)} < \frac{1}{u^2}$
- The last inequality holds for $k = 60 \ln u$.