CIS 700: “algorithms for Big Data”

Lecture 3: Streaming

Slides at http://grigory.us/big-data-class.html

Grigory Yaroslavtsev

http://grigory.us

Penn
UNIVERSITY of PENNSYLVANIA

Count-Min Sketch

- https://sites.google.com/site/countminsketch/
- Stream: $\boldsymbol{m}$ elements from universe $[\boldsymbol{n}] = \{1, 2, \dots, \boldsymbol{n}\}$, e.g.
  $$\langle x_1, x_2, \dots, x_{\boldsymbol{m}} \rangle = \langle 5, 8, 1, 1, 1, 4, 3, 5, \dots, 10 \rangle$$
- f_i = frequency of i in the stream = # of occurrences of value i, $f = \langle f_1, \dots, f_{\boldsymbol{n}} \rangle$
- Problems:
  - Point Query: For $i \in [n]$ estimate f_i
  - Range Query: For $i, j \in [n]$ estimate $f_i + \cdots + f_j$
  - Quantile Query: For $\phi \in [0,1]$ find j with $f_1 + \cdots + f_j \approx \phi m$
  - Heavy Hitters: For $\phi \in [0,1]$ find all i with $f_i \geq \phi m$

Count-Min Sketch: Construction

- Let $H_1, \dots, H_d : [n] \to [w]$ be 2-wise independent hash functions
- We maintain $d \cdot w$ counters with values:

  $c_{i,j} =$ # elements e in the stream with $H_i(e) = j$
- For every x the value $c_{i,H_i(x)} \geq f_x$ and so:

$$f_x \leq \widetilde{f_x} = \min(c_{1,H_1(x)}, \dots, c_{d,H_d(x)})$$

- If $w = \frac{2}{\epsilon}$ and $d = \log_2 \frac{1}{\delta}$ then:

$$\Pr\left[f_x \leq \widetilde{f_x} \leq f_x + \epsilon m\right] \geq 1 - \delta.$$

Count-Min Sketch: Analysis

- Define random variables $\boldsymbol{Z}_1 \dots, \boldsymbol{Z}_d$ such that $c_{i,H_i(x)} = f_x + \boldsymbol{Z}_i$

$$\boldsymbol{Z}_i = \sum_{y \neq x, H_i(y) = H_i(x)} f_y$$

- Define $\boldsymbol{X}_{i,y} = 1$ if $H_i(y) = H_i(x)$ and 0 otherwise:

$$\boldsymbol{Z}_i = \sum_{y \neq x} f_y \boldsymbol{X}_{i,y}$$

- By 2-wise independence:

$$\mathbb{E}[\boldsymbol{Z}_i] = \sum_{y \neq x} f_y \, \mathbb{E}\left[\boldsymbol{X}_{i,y}\right] = \sum_{y \neq x} f_y \Pr[H_i(y) = H_i(x)] \leq \frac{m}{w}$$

- By Markov inequality,

$$\Pr[\boldsymbol{Z}_i \geq \epsilon m] \leq \frac{1}{w\,\epsilon} = \frac{1}{2}$$

Count-Min Sketch: Analysis

- All Z_i are independent

$$\Pr[Z_i \geq \epsilon m \quad for\ all \quad 1 \leq i \leq d\ ] \leq \left(\frac{1}{2}\right)^d = \delta$$

- With prob. $1\ - \delta$ there exists j such that $Z_j \leq \epsilon m$

$$\widetilde{f_x} = \min\left(c_{1,H_1(x)}, \ldots, c_{d,H_d(x)}\right) =$$
$$= \min(f_x, +Z_1\ \ldots, f_x + Z_d) \leq f_x + \epsilon m$$

- CountMin estimates values f_x up to $\pm\epsilon m$ with total memory $O\left(\frac{\log m \log\frac{1}{\delta}}{\epsilon}\right)$`

Dyadic Intervals

- Define $\log n$ partitions of $[n]$:

$I_0 = \{1,2,3, \dots n\}$
$I_1 = \{\{1,2\}, \{3,4\}, \dots, \{n-1, n\}\}$
$I_2 = \{\{1,2,3,4\}, \{5,6,7,8\}, \dots, \{n-3, n-2, n-1, n\}\}$
…
$\mathrm{I}_{\mathrm{log\ n}} = \{\{1, 2,3, \dots, n\}\}$

- Exercise: Any interval (i, j) can be written as a disjoint union of at most $2 \log n$ such intervals.
- Example: For $n = 256$: $[48,107] = [48,48] \cup [49,64] \cup [65,96] \cup [97,104] \cup [105,106] \cup [107,107]$

Count-Min: Range Queries and Quantiles

- Range Query: For $i, j \in [n]$ estimate $f_i + \cdots f_j$
- Approximate median: Find j such that:

$$f_1 + \cdots + f_j \geq \frac{m}{2} + \epsilon m \text{ and}$$

$$f_1 + \cdots + f_{j-1} \leq \frac{m}{2} - \epsilon m$$

Count-Min: Range Queries and Quantiles

- Algorithm: Construct $\log n$ Count-Min sketches, one for each I_i such that for any $I \in I_i$ we have an estimate $\tilde{f}_l$ for f_l such that:
$$\Pr\left[f_l \le \widetilde{f_l} \le f_l + \epsilon m\right] \ge 1 - \delta$$
- To estimate $[i, j]$, let $I_1 \ldots, I_k$ be decomposition:
$$\widetilde{f_{[i,j]}} = \widetilde{f_{I_1}} + \cdots + \widetilde{f_{I_k}}$$
- Hence,
$$\Pr\left[f_{[i,j]} \le \widetilde{f_{[i,j]}} \le 2\,\epsilon m \log n\right] \ge 1 \ - 2\delta \log n$$

Count-Min: Heavy Hitters

- Heavy Hitters: For $\phi \in [0,1]$ find all i with $f_i \geq \phi m$ but no elements with $f_i \leq (\phi - \epsilon)m$
- Algorithm:
  - Consider binary tree whose leaves are [n] and associate internal nodes with intervals corresponding to descendant leaves
  - Compute Count-Min sketches for each I_i
  - Level-by-level from root, mark children I of marked nodes if $\widetilde{f_I} \geq \phi m$
  - Return all marked leaves
- Finds heavy-hitters in $O(\phi^{-1} \log n)$ steps

More about Count-Min

- Authors: Graham Cormode, S. Muthukrishnan [LATIN'04]
- Count-Min is linear:

  Count-Min(S1 + S2) = Count-Min(S1) + Count-Min(S2)

- Deterministic version: CR-Precis
- Count-Min vs. Bloom filters
  - Allows to approximate values, not just 0/1 (set membership)
  - Doesn't require mutual independence (only 2-wise)
- FAQ and Applications:
  - https://sites.google.com/site/countminsketch/home/
  - https://sites.google.com/site/countminsketch/home/faq

Fully Dynamic Streams

- Stream: $\boldsymbol{m}$ updates $(x_i, \Delta_i) \in [n] \times \mathbb{R}$ that define vector f where $f_j = \sum_{i:x_i=j} \Delta_i$.
- Example: For $n = 4$

$$\langle (1,3), (3, 0.5), (1,2), (2, -2), (2,1), (1, -1), (4,1) \rangle$$
$$f = (4, -1, 0.5, 1)$$

- Count-Min Sketch:

$$\Pr\left[|\widetilde{f_x} - f_x| + \epsilon \left|| f |\right|_1 \right] \geq 1 - \delta$$

- Count Sketch: Count-Min with **random signs** and **median** instead of min:

$$\Pr\left[|\widetilde{f_x} - f_x| + \epsilon \left|| f |\right|_2 \right] \geq 1 - \delta$$

Count Sketch

- In addition to $H_i : [n] \rightarrow [w]$ use random signs $r[i] \rightarrow \{-1,1\}$

$$c_{i,j} = \sum_{x: H_i(x) = j} r_i(x) f_x$$

- Estimate:

$$\hat{f}_x = median(r_1(x) c_{1,H_1(x)}, \dots, r_d(x) c_{d,H_d(x)})$$

- Parameters: $d = O\left(\log \frac{1}{\delta}\right), w = \frac{3}{\epsilon^{\mathbf{2}}}$

$$\Pr\left[|\widetilde{f_x} - f_x| + \epsilon ||f||_2\right] \geq 1 - \delta$$

ℓ_p-Sampling

- Stream: $\boldsymbol{m}$ updates $(x_i, \Delta_i) \in [n] \times \mathbb{R}$ that define vector f where $f_j = \sum_{i:x_i=j} \Delta_i$.
- ℓ_p-Sampling: Return random $I \in [n]$ and $R \in \mathbb{R}$:

$$\Pr[I = i] = (1 \pm \epsilon)\frac{|f_i|^p}{||f||_p^p} + n^{-c}$$

$$R = (1 \pm \epsilon) f_I$$

Application: Social Networks

- Each of n people in a social network is friends with some arbitrary set of other $n - 1$ people
- Each person knows only about their friends
- With no communication in the network, each person sends a postcard to Mark Z.
- If Mark wants to know if the graph is connected, how long should the postcards be?

Optimal F_k estimation

- Last time: (ϵ, δ)-approximate F_k
  - $\tilde{O}(n^{1-1/k})$ space for $F_k = \sum_i |f_i|^k$
  - $\tilde{O}(\log n)$ space for F_2
- New algorithm: Let (I, R) be an ℓ_2-sample.
  Return $T = \widehat{F_2} R^{k-2}$, where $\widehat{F_2}$ is an $e^{\pm\epsilon}$ estimate of F_2
- Expectation:

$$\mathbb{E}[T] = \widehat{F_2} \sum_i Pr[I = i]\,(e^{\pm\epsilon} f_i)^{k-2}$$

$$= e^{\pm\epsilon k} F_2 \sum_{i \in [n]} \frac{f_i^2}{F_2} f_i^{k-2} = e^{\pm\epsilon k} F_k$$

Optimal F_k estimation

- New algorithm: Let (I, R) be an ℓ_2-sample.
  Return $T = \widehat{F_2} R^{k-2}$, where $\widehat{F_2}$ is an $e^{\pm\epsilon}$ estimate of F_2
- Variance:

$$Var[T] \leq \mathbb{E}[T^2] = \sum_i Pr[I = i]\mathbb{E}[T^2 | I = i]$$

$$= e^{\pm 2\epsilon k} \sum_{i \in [n]} \frac{f_i^2}{F_2} F_2^2 f_i^{2(k-2)} = e^{\pm 2\epsilon k} F_2 F_{2k-2} \leq e^{\pm 2\,\epsilon k}\, n^{1-\frac{2}{k}} F_k^2$$

- Exercise: Show that $F_2 F_{2k-2} \leq n^{1-\frac{2}{k}} F_k^2$
- Overall: $\mathbb{E}[T] = e^{\pm\epsilon k} F_k, Var[T] \leq e^{\pm 2\,\epsilon k}\, n^{1-\frac{2}{k}} F_k^2$
  - Apply average + median to $O\left(n^{1-\frac{2}{k}} \epsilon^{-2} \log \delta^{-1}\right)$ copies

ℓ_2-Sampling: Basic Overview

- Assume $F_2(f) = 1$. Weight f_i by $\sqrt{w_i} = \sqrt{\frac{1}{u_i}}$, where $u_i \in_R [0,1]$:

$$f = (f_1, f_2, \dots, f_n)$$

$$g = (g_1, g_2, \dots, g_n) \text{ where } g_i = \sqrt{w_i} f_i$$

- For some value t, return (i, f_i) if there is a unique i such that $g_i^2 \geq t$

- Probability (i, f_i) is returned if t is large enough:

$$\Pr\left[g_i^2 \geq t \;\; and \; \forall j \neq i, g_j^2 < t\right] = \Pr\left[g_i^2 \geq t\right] \prod_{j \neq i} Pr\left[g_j^2 < t\right]$$

$$= \Pr\left[u_i \leq \frac{f_i^2}{t}\right] \prod_{j \neq i} \Pr\left[u_j > \frac{f_j^2}{t}\right] \approx \frac{f_i^2}{t}$$

- Probability some value is returned $\sum_i \frac{f_i^2}{t} = \frac{1}{t}$, repeat $O\left(t \log \frac{1}{\delta}\right)$ times.

ℓ_2-Sampling: Part 1

- Use Count-Sketch with parameters (m, d) to sketch g
- To estimate f_i^2:

$$g_i^2 = median_j\left(c_{j,h_j(i)}^2\right) \text{ and } \widehat{f_i^2} = \frac{\widehat{g_i^2}}{w_i}$$

- Lemma: With high probability if $d = O(\log n)$

$$\widehat{g_i^2} = g_i^2 e^{\pm\epsilon} \pm O\left(\frac{F_2(g)}{\epsilon m}\right)$$

- Corollary: With high probability if $d = O(\log n)$ and $m \gg \frac{F_2(g)}{\epsilon}$,

$$\widehat{f_i^2} = f_i^2 e^{\pm\epsilon} \pm \frac{1}{w_i}$$

- Exercise: $\Pr[F_2(g) \leq c \log n] \leq \frac{99}{100}$ for large $c > 0$.

Proof of Lemma

- Let $c_j = r_j(i) g_i + Z_j$
- By the analysis of Count Sketch $\mathbb{E}\left[Z_j^2\right] \le \frac{F_2(g)}{m}$ and by Markov:

$$\Pr\left[Z_j^2 \le \frac{3F_2(g)}{m}\right] \ge \frac{2}{3}$$

- If $|g_i| \ge \frac{2}{\epsilon}\left|Z_j\right|$, then $\left|c_{j,h_j(i)}\right|^2 = e^{\pm\epsilon}|g_i|^2$
- If $|g_i| \le \frac{2}{\epsilon}\left|Z_j\right|$, then

$$\left|c_{j,h_j(i)}^2\right| \le \left(|g_i| + \left|Z_j\right|\right)^2 - |g_i|^2 = \left|Z_j\right|^2 + 2\left|g_i Z_j\right| \le \frac{6\left|Z_j\right|^2}{\epsilon} \le 18\frac{F_2(g)}{\epsilon m}$$

where the last inequality holds with probability $2/3$

- Take median over $d = O(\log n)$ repetitions $\Rightarrow$ high probability

ℓ_2-Sampling: Part 2

- Let $s_i = 1$ if $\widehat{f_i}^2 w_i \geq \frac{4}{\epsilon}$ and $s_i = 0$ otherwise
- If there is a unique i with $s_i = 1$ then return $\left(i, \widehat{f_i}^2\right)$.
- Note that if $\widehat{f_i}^2 w_i \geq \frac{4}{\epsilon}$ then $\frac{1}{w_i} \leq \frac{\epsilon \widehat{f_i}^2}{4}$ and so

$$\widehat{f_i}^2 = f_i^2 e^{\pm\epsilon} \pm \frac{1}{w_i} = f_i^2\, e^{\pm\epsilon} \pm \frac{\epsilon \widehat{f_i}^2}{4},$$

therefore $f_i^2 = e^{\pm 4\epsilon}\, \widehat{f_i}^2$

- Lemma: With probability $\Omega(\epsilon)$ there is a unique i such that $s_i = 1$. If so then $\Pr[i = i^*] = e^{\pm 8\,\epsilon} f_{i^*}^2$
- Thm: Repeat $\Omega(\epsilon^{-1} \log n)$ times. Space: $O(\epsilon^{-2} polylog\, n)$

Proof of Lemma

- Let $\mathrm{t} = \frac{4}{\epsilon}$. We can upper-bound $\Pr[s_i = 1]$:

$$\Pr[s_i = 1] = \Pr\left[\widehat{f_i}^{\,2} w_i \geq t\right] \leq \Pr\left[\frac{e^{4\epsilon} f_i^2}{t} \geq u_i\right] \leq \frac{e^{4\epsilon} f_i^2}{t}$$

Similarly, $\Pr[s_i = 1] \geq \frac{e^{-4\epsilon} f_i^2}{t}$.

- Using independence of w_i, probability of unique i with $s_i = 1$:

$$\sum_i \Pr\left[s_i = 1, \sum_{j \neq i} s_j = 0\right] \geq \sum_i \Pr[s_i = 1]\left(1 - \sum_{j \neq i} \Pr\left[s_j = 1\right]\right)$$

$$\geq \sum_i \frac{e^{-4\epsilon} f_i^2}{t}\left(1 - \frac{\sum_{j \neq i} e^{4\epsilon} f_i^2}{t}\right)$$

$$\geq \frac{e^{-4\epsilon}\left(1 - \frac{e^{4\epsilon}}{t}\right)}{t} \approx 1/t$$

Proof of Lemma

- Let t $= \frac{4}{\epsilon}$. We can upper-bound $\Pr[s_i = 1]$:

$$\Pr[s_i = 1] = \Pr\left[\widehat{f_i}^2 w_i \geq t\right] \leq \Pr\left[\frac{e^{4\epsilon} f_i^2}{t} \geq u_i\right] \leq \frac{e^{4\epsilon} f_i^2}{t}$$

Similarly, $\Pr[s_i = 1] \geq \frac{e^{-4\epsilon} f_i^2}{t}$.

- We just showed:

$$\sum_i \Pr\left[s_i = 1, \sum_{j \neq i} s_j = 0\right] \approx 1/t$$

- If there is a unique i, probability $i = i^*$ is:

$$\frac{\Pr[s_{i^*} = 1, \sum_{j \neq i} s_j = 0]}{\sum_i \Pr\left[s_i = 1, \sum_{j \neq i} s_j = 0\right]} = e^{\pm 8\epsilon} f_{i^*}^2$$

ℓ_0-sampling

- Maintain $\widetilde{F_0}$, and (1 ± 0.1)-approximation to F_0.
- Hash items using $h_j : [n] \to \left[0, 2^j - 1\right]$ for $j \in [\log n]$
- For each j, maintain:

$$D_j = (1 \pm 0.1)|\{t | h_j(t) = 0\}|$$

$$S_j = \sum_{t, h_{j(t)}=0} f_t i_t$$

$$C_j = \sum_{t, h_j(t)=0} f_t$$

- Lemma: At level $j = 2 + \lceil \log \widetilde{F_0} \rceil$ there is a unique element in the streams that maps to 0 (with constant probability)
- Uniqueness is verified if $D_j = 1 \pm 0.1$. If so, then output S_j / C_j as the index and C_j as the count.

Proof of Lemma

- Let $j = \lceil \log \widetilde{F_0} \rceil$ and note that $2F_0 < 2^j < 12\, F_0$
- For any i, $\Pr\left[h_j(i) = 0\right] = \frac{1}{2^j}$
- Probability there exists a unique i such that $h_j(i) = 0$,

$$\sum_i \Pr\left[h_j(i) = 0 \; and \; \forall k \neq i, h_j(k) \neq 0\right]$$

$$= \sum_i \Pr\left[h_j(i) = 0\right] \Pr\left[\forall k \neq i, h_j(k) \neq 0 \middle| \, h_{j(i)} = 0\right]$$

$$\geq \sum_i \Pr\left[h_j(i) = 0\right] \left(1 - \sum_{k \neq i} \Pr\left[h_j(k) = 0 \middle| h_j(i) = 0\right]\right)$$

$$= \sum_i \Pr\left[h_j(i) = 0\right] \left(1 \; - \sum_{k \neq i} \Pr\left[h_j(k) = 0\right]\right) \geq \sum_i \frac{1}{2^j} \left(1 \; - \frac{F_0}{2^j}\right) \geq \frac{1}{24}$$

- Holds even if h_j are only 2-wise independent