Sublinear Algorihms for Big Data

Lecture 3

Grigory Yaroslavtsev

http://grigory.us

Penn
UNIVERSITY of PENNSYLVANIA

SOFSEM 2015

- URL: http://www.sofsem.cz
- 41st International Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM'15)
- When and where?
  - January 24-29, 2015. Czech Republic, Pec pod Snezkou
- Deadlines:
  - August 1st (tomorrow!): Abstract submission
  - August 15th: Full papers (proceedings in LNCS)
- I am on the Program Committee ;)

Today

- Count-Min (continued), Count Sketch
- Sampling methods
  - ℓ_2-sampling
  - ℓ_0-sampling
- Sparse recovery
  - ℓ_1-sparse recovery
- Graph sketching

Recap

- Stream: elements from universe $[\quad]$ $= \{1, 2, \ldots, \quad\}$, e.g. $\langle \ _1, \ _2, \ldots, \quad \rangle = \langle 5, 8, 1, 1, 1, 4, 3, 5, \ldots, 10 \rangle$
- $\quad$ = frequency of $\quad$ in the stream = # of occurrences of value $\ ,\quad = \langle \ _1, \ldots, \quad \rangle$

Count-Min

- $_1, \ldots,$ $:[\ \ ] \to [\ \ ]$ are 2-wise independent hash functions
- Maintain $\cdot$ counters with values:

  $_{,} = \#$ elements in the stream with $(\ \ ) =$

- For every the value $_{,\ (\ )} \geq$ and so:

$$\leq \ \tilde{}\ = \min\left(_{1,\ 1}(\ ), \ldots, \ _{,\ 1}(\ ) \right)$$

- If $= \frac{2}{}$ and $= \log_2 \frac{1}{}$ then:

$$\Pr\left[\ \ \tilde{}\ \ \ \ \| \ \| \ \right] \ 1$$

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: updates $\left(\;, \Delta \right) \in [\;] \times \mathbb{R}$ that define vector where $= \sum_{: \;=} \Delta$.
- Example: For $= 4$

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

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

Count Sketch

- In addition to $:[\ ]\to[\ ]$ use random signs $[\ ]\to\{-1,1\}$

$$, = \sum_{:\ (\ )=} (\ )$$

- Estimate:

$$\hat{\ } = (\ _1(\ )\ _{1,}\ _1(\ ), \dots, (\ )\ _{,}\ (\ ))$$

- Parameters: $= \left(\log\frac{1}{\ }\right), \quad = \frac{3}{\ }$

$$\Pr[|\tilde{\ } - \ | \pm \ ||\ ||\ ] \geq 1 -$$

ℓ -Sampling

- Stream: updates $(\ , \Delta) \in [\ ] \times \mathbb{R}$ that define vector where $= \sum_{: \ =} \Delta$.
- ℓ -Sampling: Return random $\in [\ ]$ and $\in \mathbb{R}$:

$$\Pr[\ = \ ] = (1 \pm \ )\frac{| \ |}{|| \ ||} + \ ^{-}$$

$$= (1 \pm \ )$$

Application: Social Networks

- Each of people in a social network is friends with some arbitrary set of other -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 estimation

- Yesterday: $(\; , \;)$-approximate
  - $\tilde{} (\; ^{1-1/} \;)$ space for $= \sum | \quad |$
  - $\tilde{} (\log \;)$ space for $_2$
- New algorithm: Let $(\; , \;)$ be an ℓ_2-sample.

Return $= \hat{}_2 \quad ^{-2}$, where $\hat{}_2$ is an $\pm$ estimate of $_2$

- Expectation:

Optimal estimation

- New algorithm: Let $(\;,\;)$ be an ℓ_2-sample.

  Return $= \hat{\ }_2 \ {}^{-2}$, where $\hat{\ }_2$ is an $\pm$ estimate of $_2$

- Variance:

$$[\ ] \le [\ {}^2] = \sum [\ =\ ]\,[\ {}^2/\ =\ ]$$

$$= {}^{\pm 2} \sum_{\in [\ ]} \frac{{}^2}{{}^2}\ \frac{2}{2}\ 2(\ -2) = {}^{\pm 2}\ {}_2\ {}_2^{\ -2} \le {}^{\pm 2}$$

- Exercise: Show that ${}_2\ {}_2^{\ -2} \le 1 - \frac{2}{\ }\ 2$

ℓ_2-Sampling: Basic Overview

- Assume $_2(\ )=1$. Weight by $\sqrt{\ }=\sqrt{\frac{1}{\ }}$, where

  $\in [0,1]$:

  $$=(\ _1,\ _2,\ldots,\ )$$

  $$=(\ _1,\ _2,\ldots,\ )$$ where $=\sqrt{\ }$

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

- Probability $(\ ,\ )$ is returned if is large enough:

ℓ_2-Sampling: Part 1

- Use Count-Sketch with parameters $(\quad , \quad)$ to sketch
- To estimate $\quad ^2$:

$$^2 = \qquad \left(\; ^2_{,h(\;)} \right) \text{ and } \hat{\;}^2 = \frac{\hat{\;}^2}{\quad}$$

- Lemma: With high probability if $\quad = \quad (\log \quad)$

$$\hat{\;}^2 = \quad ^2 \pm \quad \pm \quad \left(\frac{\;^2(\;)}{\quad} \right)$$

- Corollary: With high probability if $\quad = \quad (\log \quad)$ and $\quad \gg \frac{\;^2(\;)}{\quad}$,

Proof of Lemma

- Let $= \left(\right) +$
- By the analysis of Count Sketch $\left[\ {}^2 \right] \le \frac{{}^2\left(\ \right)}{\ }$ and by Markov:

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

- If $\left| \ \right| \ge \frac{2}{\ } \left| \ \right|$, then $\left| \ {}_{, h}\left(\ \right) \right|^2 = \pm \left| \ \right|^2$
- If $\left| \ \right| \le \frac{2}{\ } \left| \ \right|$, then

ℓ_2-Sampling: Part 2

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

$$\hat{}\ {}^{2} = {}^{2} \pm \pm \frac{1}{\ } = {}^{2} \pm \pm \frac{\hat{}\ {}^{2}}{4},$$

therefore ${}^{2} = \pm 4 \hat{}\ {}^{2}$

Proof of Lemma

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

$$\Pr[\quad = 1] = \Pr[\hat{\ }\ ^2 \quad \geq \quad]$$

$$\leq \Pr\left[\frac{4 \quad 2}{} \geq \quad \right] \leq \frac{4 \quad 2}{}$$

Similarly, $\Pr[\quad = 1] \geq \frac{-4 \quad 2}{}$.

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

$$\sum \Pr\left[\quad = 1 \sum \quad = 0\right] \geq \sum \Pr[\quad = 1]\left(1 - \sum \Pr\right.$$

Proof of Lemma

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

$$\Pr\left[\quad = 1\right] = \Pr\left[\hat{\ }\ ^{2} \quad \geq \quad\right]$$

$$\leq \Pr\left[\frac{4 \quad\quad 2}{\quad\quad\quad\quad} \geq \quad\right] \leq \frac{4 \quad\quad 2}{\quad\quad\quad\quad}$$

Similarly, $\Pr\left[\quad = 1\right] \geq \frac{-4 \quad\quad 2}{\quad\quad\quad\quad}$.

- We just showed:

$$\sum \Pr\left[\quad = 1, \sum \quad = 0\right] \approx 1 /$$

ℓ_0-sampling

- Maintain $\tilde{\ }_0$ and (1 ± 0.1)-approximation to $\ _0$
- Hash items using $h\ : [\ \ ] \to [0, 2\ \ - 1]$ for $\ \in [\ \log\ \ ]$
- For each $\ $, maintain:

$$= (1 \pm 0.1) |\{\ \ | h\ (\ \ ) = 0\}|$$

$$= \sum_{, h\ (\ ) = 0}$$

$$= \sum_{, h\ (\ ) = 0}$$

Proof of Lemma

- Let $= \lceil \log \tilde{}_{0} \rceil$ and note that $2 \;_{0} < 2 < 12 \;_{0}$
- For any , $\Pr\left[h\left(\;\right) = 0\right] = \dfrac{1}{2}$
- Probability there exists a unique such that $h\left(\;\right) = 0,$

$$\sum \Pr\left[h\left(\;\right) = 0 \qquad \forall \quad \neq \;, h\left(\;\right) \neq 0\right]$$

$$= \sum \Pr\left[h\left(\;\right) = 0\right] \Pr\left[\forall \quad \neq \;, h\left(\;\right) \neq 0 \middle| h_{\left(\;\right)} = 0\right]$$

Sparse Recovery

- Goal: Find such that $|| \quad - \quad ||_1$ is minimized among $'$s with at most non-zero entries.
- Definition: $(\quad) = \min_{g: ||g||_0 \leq} || \quad - \quad ||_1$
- Exercise: $(\quad) = \sum_{\notin} | \quad |$ where are indices of largest
- Using $(\quad ^{-1} \quad \log \quad)$ space we can find $\tilde{}$ such

Count-Min Revisited

- Use Count-Min with $= \left(\log\ \right), \quad = 4\ /$
- For $\in [\ \ ]$, let $\tilde{}\ = {}_{,h(\ )}$ for some row $\in [\ \ ]$
- Let $= \{ {}_{1}, \ldots, \ \}$ be the indices with max. frequencies. Let be the event there doesn't exist $\in \ /$ with $h(\ ) = h(\ )$
- Then for $\in [\ \ ]$:

$$\Pr\left[\left| \ - \tilde{}\ \right| \geq \frac{(\ )}{}\right] =$$

$$\Pr\left[not\ \ \right] \times \Pr\left[\left| \ - \tilde{}\ \right| \geq \frac{(\ )}{} \middle| \ \right]$$

$$+$$

Sparse Recovery Algorithm

- Use Count-Min with $= \left(\log \right), = 4 /$
- Let $' = (\tilde{}_{1}, \tilde{}_{2}, \ldots, \tilde{})$ be frequency estimates:

$$\left| - \tilde{} \right| \le \frac{(\ )}{}$$

- Let $\tilde{}$ be $'$ with all but the k-th largest entries replaced by 0.
- Lemma: $\left|\left|\tilde{} - \right|\right|_1 \le \left(1 + 3 \right) (\ )$

$$\|\tilde{\ } - \ \|_1 \le \left(1 + 3\ \right) \qquad (\ )$$

- Let $\ ,\ \subseteq [\ ]$ be indices corresponding to largest value of and $\tilde{\ }$.
- For a vector $\in \mathbb{R}$ and $\subseteq [\ ]$ denote as the vector formed by zeroing out all entries of except for those in .

$$\|\ -\ '\ \|_1 \le \|\ -\ \|_1 + \|\ -\ '\ \|_1$$

$$= \|\ \|_1 - \|\ \|_1 + \|\ -\ '\ \|_1$$

$$= \|\ \|_1 - \|\ '\ \|_1 + \left(\|\ '\ \|_1 - \|\ \|_1\right)$$

$$+ \|\ -\ '\ \|_1$$

$$\le \|\ \|\ - \|\ '\ \|\ +2\ \|\ -\ '\ \|$$

Thank you!

- Questions?