Improved Concentration Bounds for Count-Sketch

Gregory T. Minton[1]   Eric Price[2]

[1]MIT → **MSR New England**
[2]MIT → **IBM Almaden** → UT Austin

2014-01-06

Count-Sketch: a classic streaming algorithm

Charikar, Chen, Farach-Colton 2002

- Solves "heavy hitters" problem
- Estimate a vector $x \in \mathbb{R}^n$ from low dimensional sketch $Ax \in \mathbb{R}^m$.
- Nice algorithm
  - Simple
  - Used in Google's MapReduce standard library
- [CCF02] bounds the *maximum* error over all coordinates.
- We show, for the same algorithm,
  - *Most* coordinates have asymptotically better estimation accuracy.
  - The *average* accuracy over many coordinates will be asymptotically better *with high probability*.
  - Experiments show our asymptotics are correct.
- Caveat: we assume fully independent hash functions.

Outline

1. Robust Estimation of Symmetric Variables
   - Lemma
   - Relevance to Count-Sketch
2. Electoral Colleges and Direct Elections
   - Lemma
   - Relevance to Count-Sketch
3. Experiments!

Outline

1. Robust Estimation of Symmetric Variables
   - Lemma
   - Relevance to Count-Sketch
2. Electoral Colleges and Direct Elections
   - Lemma
   - Relevance to Count-Sketch
3. Experiments!

Estimating a symmetric random variable's mean

$\mathcal{X}$

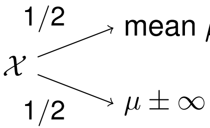
- Unknown distribution $\mathcal{X}$ over $\mathbb{R}$, symmetric about unknown μ.
  - Given samples $x_1, \ldots, x_R \sim X$.
  - How to estimate μ?

Estimating a symmetric random variable's mean

$\mathcal{X}$ → mean μ, standard deviation σ

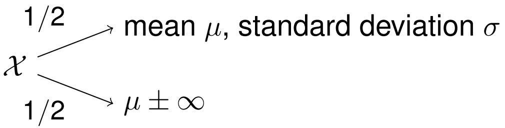
- Unknown distribution $\mathcal{X}$ over $\mathbb{R}$, symmetric about unknown μ.
  - Given samples $x_1, \ldots, x_R \sim X$.
  - How to estimate μ?
- Mean:
  - Converges to μ as $\sigma/\sqrt{R}$.

Estimating a symmetric random variable's mean



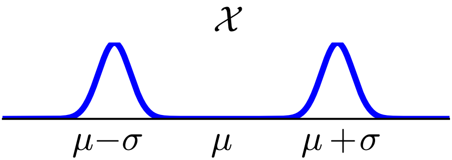
- Unknown distribution $\mathcal{X}$ over $\mathbb{R}$, symmetric about unknown μ.
  - Given samples $x_1, \ldots, x_R \sim X$.
  - How to estimate μ?
- Mean:
  - Converges to μ as $\sigma/\sqrt{R}$.
  - No robustness to outliers

Estimating a symmetric random variable's mean



- Unknown distribution $\mathcal{X}$ over $\mathbb{R}$, symmetric about unknown μ.
  - Given samples $x_1, \ldots, x_R \sim X$.
  - How to estimate μ?
- Mean:
  - Converges to μ as $\sigma/\sqrt{R}$.
  - No robustness to outliers
- Median:
  - Extremely robust
  - Doesn't necessarily converge to μ.

Estimating a symmetric random variable's mean



- Median doesn't converge
- Consider: median of pairwise means

$$\widehat{\mu} = \underset{i \in \{1, 3, 5, \ldots\}}{\text{median}} \frac{x_i + x_{i+1}}{2}$$

  - Converges as $O(\sigma/\sqrt{R})$, even with outliers.
- That is: median of $(\mathcal{X} + \mathcal{X})$ converges.

[See also: Hodges-Lehmann estimator.]

Why does median converge for $\mathcal{X} + \mathcal{X}$?

- WLOG $\mu = 0$.
- Define the Fourier transform $\mathcal{F}_{\mathcal{X}}$ of $\mathcal{X}$:

$2\pi \approx 6.28$

$$\mathcal{F}_{\mathcal{X}}(t) = \mathop{\mathbb{E}}_{x \sim \mathcal{X}}[\cos(\tau x t)]$$

(standard Fourier transform of PDF, specialized to symmetric $\mathcal{X}$.)

- Convolution $\iff$ multiplication
  - $\mathcal{F}_{\mathcal{X}+\mathcal{X}}(t) = (F_{\mathcal{X}}(t))^2 \geq 0$ for all t.

Theorem

Let $\mathcal{Y}$ be symmetric about 0 with $\mathcal{F}_{\mathcal{Y}}(t) \geq 0$ for all t and $\mathbb{E}[Y^2] = \sigma^2$. Then for all $\epsilon \leq 1$,

$$\Pr[|y| \leq \epsilon\sigma] \gtrsim \epsilon$$

Standard Chernoff bounds: median $y_1, \ldots, y_R$ converges as $\sigma/\sqrt{R}$.

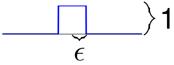
Proof

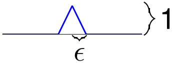
Theorem

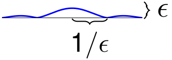
Let $\mathcal{F}_{\mathcal{Y}}(t) \geq 0$ for all t and $\mathbb{E}[Y^2] = 1$. Then for all $\epsilon \leq 1$,

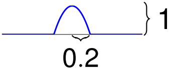
$$\Pr[|y| \leq \epsilon] \gtrsim \epsilon.$$

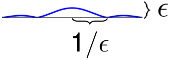
$$\mathcal{F}_{\mathcal{Y}}(t) = \mathbb{E}[\cos(\tau y t)] \geq 1 - \frac{\tau^2}{2} t^2$$

$\Pr[|y| \leq \epsilon] = \mathcal{Y} \cdot$ 

$\geq \mathcal{Y} \cdot$ 

$= \mathcal{F}_{\mathcal{Y}} \cdot$ 

$\geq$ 

$\cdot$ 

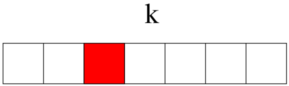
$\gtrsim \epsilon.$ ■

Outline

Count-Sketch

- Want to estimate $x \in \mathbb{R}^n$ from small "sketch."
- Hash to k buckets and sum up with random signs
  Choose random $h : [n] \to [k], s : [n] \to \{\pm 1\}$. Store

$$y_j = \sum_{i:\, h(i)=j} s(i) x_i$$

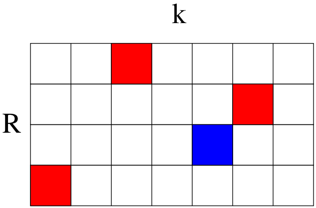


- Can estimate x_i by $\tilde{x}_i = y_{h(i)} s(i)$.

Count-Sketch

- Want to estimate $x \in \mathbb{R}^n$ from small "sketch."
- Hash to k buckets and sum up with random signs
  Choose random $h : [n] \to [k], s : [n] \to \{\pm 1\}$. Store

$$y_j = \sum_{i:\, h(i)=j} s(i)x_i$$



- Can estimate x_i by $\tilde{x}_i = y_{h(i)}s(i)$.
- Repeat R times, take the median.
- For each row,

$$\tilde{x}_i - x_i = \sum_{j \neq i} \begin{cases} \pm x_j & \text{with probability } 1/k \\ 0 & \text{otherwise} \end{cases}$$

- Symmetric, non-negative Fourier transform.

Count-Sketch Analysis

Let

$$\sigma^2 = \frac{1}{k} \min_{k\text{-sparse } x_{[k]}} \|x - x_{[k]}\|_2^2$$

be the "typical" error for a single row of Count-Sketch with k columns.

Theorem

For the any coordinate i, we have for all $t \leq R$ that

$$\Pr[|\widehat{x}_i - x_i| > \sqrt{\frac{t}{R}}\sigma] \leq e^{-\Omega(t)}.$$

(CCF02: $t = R = O(\log n)$ case; $\|\widehat{x} - x\|_\infty \lesssim \sigma$ w.h.p.)

Corollary

Excluding $e^{-\Omega(R)}$ probability events, we have for each i that

$$\mathbb{E}[(\widehat{x}_i - x_i)^2] = \sigma^2 / R$$

Estimation of *multiple* coordinates?

- What about the average error on a *set* S of k coordinates?
- Linearity of expectation: $\mathbb{E}[\|\widehat{x}_S - x_S\|_2^2] = \frac{O(1)}{R} k\sigma^2$.
- Does it concentrate?

$$\Pr[\|\widehat{x}_S - x_S\|_2^2 > \frac{O(1)}{R} k\sigma^2] < p = ???$$

- By expectation: $p = \Theta(1)$.
- If independent: $p = e^{-\Omega(k)}$.
- Sum of many variables, but not independent...
- Chebyshev's inequality, bounding covariance of error:
  - Feasible to analyze (though kind of nasty).
  - Ideally get: $p = 1/\sqrt{k}$.
  - We can get $p = 1/k^{1/14}$.
- Can we at least get "high probability," i.e. $1/k^c$ for arbitrary constant c?

Boosting the error probability

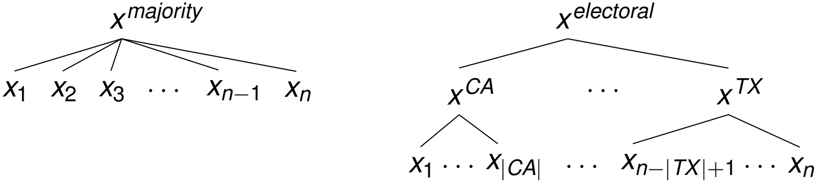
in a black box manner

- We know that $\|\widehat{x}_S - x_S\|_2$ is "small" with all but $k^{-1/14}$ probability.
- Way to get all but k^{-c} probability: repeat $100c$ times and take the median of results.
  - With all but k^{-c} probability, $> 75c$ of the $\widehat{x}_S^{(i)}$ will have "small" error.
  - Median of results has at most $3\times$ "small" total error.
- But resulting algorithm is stupid:
  - Run count-sketch with $R' = O(cR)$.
  - Arbitrarily partition into blocks of R rows.
  - Estimate is median (over blocks) of median (within block) of individual estimates.
- Can we show that the direct median is as good as the median-of-medians?

Outline

Electoral Colleges

- Suppose you have a two-party election for k offices.
  - Voters come from a distribution $\mathcal{X}$ over $\{0, 1\}^k$.
  - "True" majority slate of candidates $\overline{x} \in \{0, 1\}^k$.
  - Election day, receive ballots $x_1, \ldots, x_n \sim \mathcal{X}$.
- How to best estimate $\overline{x}$? For each office,



- Is $x^{majority}$ better than $x^{electoral}$ in every way? Is

$$\Pr[\|x^{majority} - \overline{x}\| > \alpha] \leq \Pr[\|x^{electoral} - \overline{x}\| > \alpha]$$

for all α, $\|\cdot\|$?

Electoral Colleges

- Is $x^{majority}$ better than $x^{electoral}$ in every way, so

$$\Pr[\|x^{majority} - \overline{x}\| > \alpha] \leq \Pr[\|x^{electoral} - \overline{x}\| > \alpha]$$

  for all α, $\|\cdot\|$?
- Don't know, but

Theorem

$$\Pr[\|x^{majority} - \overline{x}\| > 3\alpha] \leq 4 \cdot \Pr[\|x^{electoral} - \overline{x}\| > \alpha]$$

for all p-norms $\|\cdot\|$*.*

Proof

Theorem

$$\Pr[\|x^{majority} - \overline{x}\| > 3\alpha] \leq 4 \cdot \Pr[\|x^{electoral} - \overline{x}\| > \alpha]$$

for all p-norms $\|\cdot\|$.

Follows easily from:

Lemma (median^3)

For any $x_1, \ldots, x_n \in \mathbb{R}^k$, *we have*

$$\underset{\textit{partitions into states}}{\text{median}} \; \underset{\textit{states}}{\text{median}} \; \underset{\textit{within state}}{\text{median}} \; x_i = \underset{\textit{populace}}{\text{median}} \, x_i$$

(With $4p$ failure probability, $3/4$ of partitions have error at most α; then their median has error 3α.)

Outline

Concentration for sets

- We know that a "median-of-medians" variant of Count-Sketch would give good estimation of sets with high probability.
- Therefore the standard Count-Sketch would as well.

Theorem

For any constant c, we have for any set S of coordinates that

$$\Pr[\|\widehat{x}_S - x_S\|_2 > O(\sqrt{\frac{|S|}{R}}\sigma)] \lesssim |S|^{-c}.$$

Outline

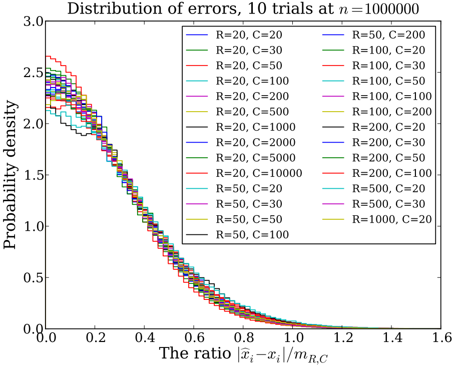
Experiments

- Claims
  1. Individual coordinates have error that concentrates like a Gaussian with standard deviation $\sigma/\sqrt{R}$.
  2. Sets of coordinates have error $O(\sigma\sqrt{k/R})$ with high probability.
- Evaluate on power-law distribution with typical parameters.

Experiments

1. Individual coordinates have error that concentrates like a Gaussian with standard deviation $\sigma/\sqrt{R}$.

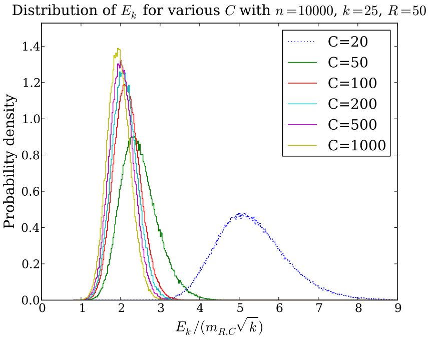
- Compare observed error to expected error for various R, C.

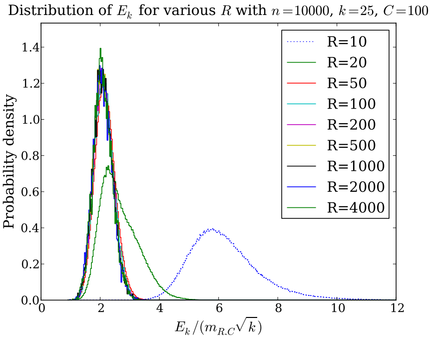


Experiments

2. Sets of coordinates have error $O(\sigma\sqrt{k/R})$ with high probability. (for large enough R, C)

- Compare observed error to expected error for various R, C.





Conclusions

- We present an improved analysis of Count-Sketch, a classic algorithm used in practice.
- Experiments show it gives the right asymptotics
- More applications of our lemmas?
- Independence?

Thank You