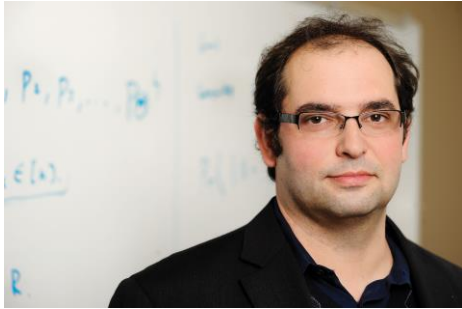


Private Data Stream Analysis for Universal Symmetric Norm Estimation



Vladimir Braverman



Joel Manning

Steven Wu

Samson Zhou



Symmetric Norms

- ❖ A norm is symmetric if it is invariant under permutations and sign flips on an input frequency vector

$$a = [1, 3, -2, 0, 0, 5, -2, 4]$$

$$b = [1, 3, 2, 0, 0, 5, 2, 4]$$

$$c = [0, 0, 1, 2, 2, 3, 4, 5]$$

$$\|a\| = \|b\| = \|c\|$$

$$L(a) = L(b) = L(c)$$

L_p Norms

- ❖ Let F_p be the frequency moment of the vector $f \in R^n$:

$$F_p = f_1^p + f_2^p + \dots + f_n^p$$

- ❖ Then the L_p norm of the frequency vector f is:

$$L_p(f) = \left(F_p(f) \right)^{1/p}$$

- ❖ **Goal:** Given an accuracy parameter α , output a $(1 + \alpha)$ -approximation to L_p
- ❖ **Motivation:** Entropy estimation, linear regression

Differential Privacy

- ❖ [DworkMcSherryNissimSmith06] Given $\varepsilon > 0$ and $\delta \in (0,1)$, a randomized algorithm $A: U^* \rightarrow Y$ is (ε, δ) -differentially private if, for every neighboring frequency vectors f and f' and for all $E \subseteq Y$,

$$\Pr[A(f) \in E] \leq e^\varepsilon \Pr[A(f') \in E] + \delta$$

Multiple Privately Queries

- ❖ Privately query $f \in R^n$ multiple times?
- ❖ Add noise to each query with scale parameter depending on the number Q of queries
- ❖ Accuracy degrades as the number Q of queries increases

Can we answer multiple queries without
sacrificing accuracy?

“Beating advanced composition”
“Avoid privacy analysis per algorithm”

Streaming Model

- ❖ **Input:** Elements of an underlying data set S , which arrives sequentially
- ❖ **Output:** Evaluation (or approximation) of a given function
- ❖ **Goal:** Use space *sublinear* in the size m of the input S

Symmetric Norms in the Streaming Model

- ❖ Given a stream S of m elements from $[n]$, let f_i be the frequency of element i . (How often it appears)

$$1\ 1\ 2\ 1\ 2\ 1\ 1\ 2\ 3 \rightarrow [5, 3, 1, 0] := f$$

- ❖ **Goal:** Given a stream S of length m that defines a frequency vector $f \in R^n$ and an accuracy parameter α , output a $(1 + \alpha)$ -approximation to $L(f)$, using space sublinear in n and m

Our Result

There exists an (ϵ, δ) -differentially private algorithm such that:

- ❖ **Input:** on a stream S of length m that defines a frequency vector $f \in R^n$ that
- ❖ **Output:** a set C , from which the $(1 + \alpha)$ -approximation to any symmetric norm with maximum modulus of concentration M can be computed with probability $1 - \delta$.
- ❖ The algorithm uses $M^2 \cdot \text{poly}\left(\frac{1}{\alpha}, \frac{1}{\epsilon}, \log(n, m), \log \frac{1}{\delta}\right)$ space

Applications

- ❖ For L_p norms, $M(\ell) = O(\log m)$ for $p \in [1,2]$ and $M(\ell) = O(n^{1/2-1/p})$ for $p > 2$ [MilmanSchectman86, KlartagVershynin07]
- ❖ Our algorithm achieves space $\text{poly log}(m)$ for $p \in [1,2]$ and $\tilde{O}(n^{1-2/p})$ for $p > 2$ in the constant α, ε and $\delta = \frac{1}{\text{poly}(m)}$ regime
- ❖ Matches known lower bounds up to log factors [Bar-YossefJayramKumarSivakumar04]
- ❖ For top k norms, $M(\ell) = \tilde{O}\left(\sqrt{\frac{n}{k}}\right)$ [BlasiokBravermanChestnutKrauthgamerYang17]

Maximum Modulus of Concentration

- ❖ Maximum modulus of concentration [MilmanSchectman86] of a norm measures the worst-case ratio of the **maximum value** to the **median value** of a norm on the L_2 -unit sphere for any restriction of the coordinates

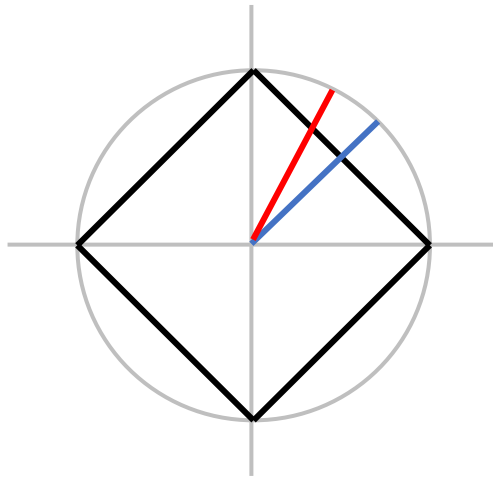
- ❖ Intuitively, quantifies the “difficulty” of computing a norm

Modulus of Concentration

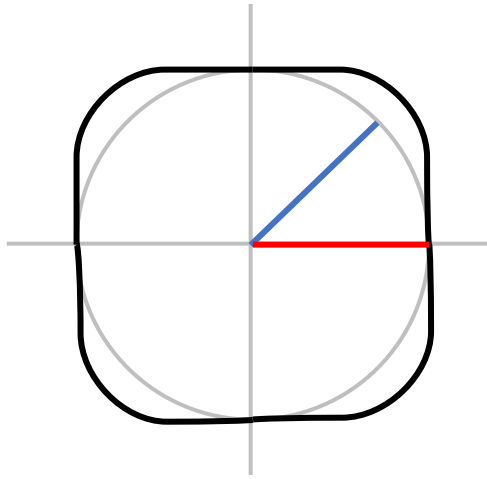
- ❖ Let $f \in R^n$ be a random vector drawn from the uniform distribution on the L_2 -unit sphere S^{n-1}
- ❖ Let b_L denote the maximum value of $L(f)$ over S^{n-1} and let M_L denote the median of $L(f)$, i.e., the unique value such that $\Pr[L(f) \geq M_L] \geq \frac{1}{2}$ and $\Pr[L(f) \leq M_L] \geq \frac{1}{2}$
- ❖ The ratio $\text{mc}(L) = \frac{b_L}{M_L}$ is the modulus of concentration of L

Modulus of Concentration

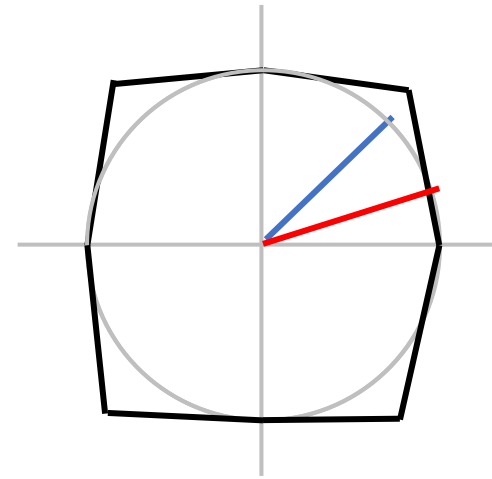
b_L is the maximum value of $L(f)$ over S^{n-1}
 M_L is the median of $L(f)$



$$b_L = \sqrt{n}$$
$$M_L \approx \sqrt{n}$$



$$b_L = 1$$
$$M_L \approx n^{-1/6}$$



Maximum Modulus of Concentration

- ❖ Maximum modulus of concentration of a norm is the maximum of the modulus of concentration of the norm restricted to sub-coordinates of R^n
- ❖ Definition is robust to “average” norms that “hide” challenging behavior embedded in lower-dimensional space
- ❖ $L(x) = \max\left(\frac{L_1(x)}{\sqrt{n}}, L_\infty(x)\right)$

Symmetric Norms

- ❖ A norm is symmetric if it is invariant under permutations and sign flips on an input frequency vector

$$a = [1, 3, -2, 0, 0, 5, -2, 4]$$

$$b = [1, 3, 2, 0, 0, 5, 2, 4]$$

$$c = [0, 0, 1, 2, 2, 3, 4, 5]$$

$$L(a) = L(b) = L(c)$$

Approximating Symmetric Norms

- ❖ Only care about number of coordinates in each range $[\xi^i, \xi^{i+1})$ for some $\xi > 1$ a function of the desired accuracy parameter α

$$\xi = 2$$

$$v = [0,0,1,2,2,3,4,5]$$

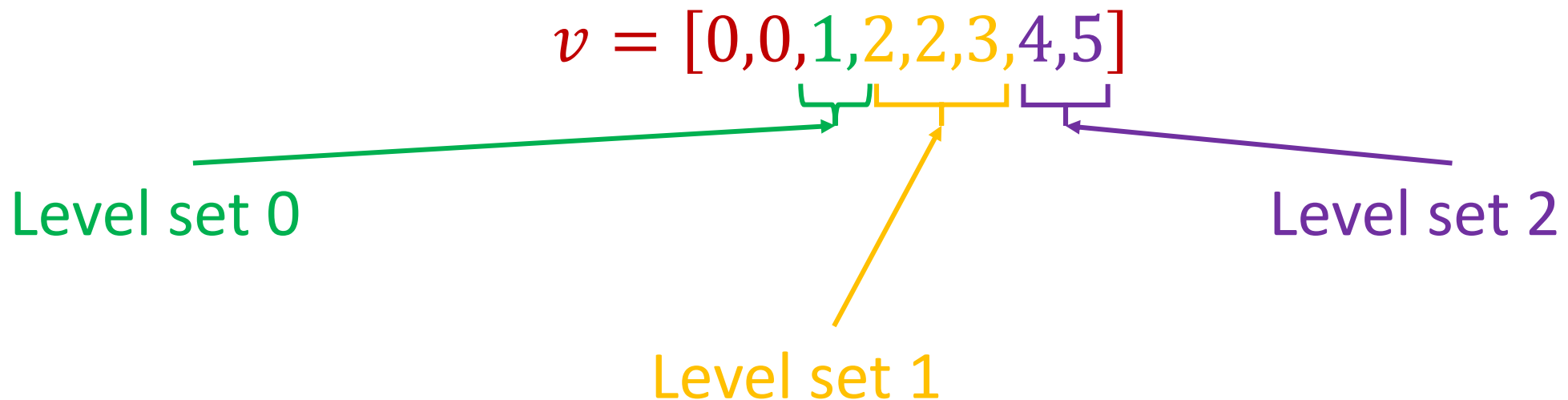
#coordinates in $[1,2)$: 1

#coordinates in $[2,4)$: 3

#coordinates in $[4,8)$: 2

Level Sets and Contribution

- ❖ Level set i is the set of coordinates with magnitude in range $[\xi^i, \xi^{i+1})$
- ❖ The *contribution* of the level set is the “amount” the level set contributes to the norm of the entire frequency vector



Important Level Sets

❖ A level set is *important* if its contribution is an $\frac{\alpha}{O(\log m)}$ fraction of the norm of the entire frequency vector

❖ It suffices to estimate the contribution of the important level sets within $\left(1 + \frac{\alpha}{O(\log m)}\right)$ -approximation
[BlasiokBravermanChestnutKrauthgamerYang17]

Important Level Sets

- ❖ **Intuition:** Important level sets must either have large magnitude coordinates or a large number of coordinates

$$v = [1, 1, 1, \dots, 1, 1, 10000]$$

$$[1, 1, 1, \dots, 1, 1, 0]$$

$$[0, 0, 0, \dots, 0, 0, 10000]$$

- ❖ How to privately release important level sets?

Important Level Sets

❖ **Definition:** Define thresholds T_1 and T_2 . A level set i is “high” if $\xi^i \geq T_1$. A level set i is “medium” if $\xi^{i+1} \leq T_1$ and $\xi^i \geq T_2$. A level set i is “low” if $\xi^{i+1} \leq T_2$

❖ **Intuition:** Important **high** level sets have **large coordinates**, important **low** level sets have a **large number of coordinates**, important **medium** level sets have a **combination of the two**

Heavy-Hitters

- ❖ Given a set S of m elements from $[n]$, let f_i be the frequency of element i . (How often it appears)
- ❖ Let L_2 be the norm of the frequency vector:

$$L_2 = \sqrt{f_1^2 + f_2^2 + \dots + f_n^2}$$

- ❖ **Goal:** Given a set S of m elements from $[n]$ and a threshold α , output the elements i such that $f_i > \alpha L_2$...and no elements j such that $f_j < \frac{\alpha}{16} L_2$

CountSketch

- ❖ Given a threshold/accuracy parameter α , there exists a one-pass streaming algorithm COUNTSKETCH that outputs an estimated frequency for each element, with additive error $\alpha \cdot L_2(f)$
- ❖ The algorithm uses $O\left(\frac{1}{\alpha^2} \log^2 m\right)$ space

CountSketch

- ❖ COUNTSKETCH with threshold/accuracy parameter $O\left(\frac{\text{poly}(\alpha, \varepsilon)}{M \text{ poly log } m}\right)$ will find the important high level sets because their magnitude is so large, but it will miss the others

$$v = [1, 1, 1, \dots, 1, 1, 10000]$$

$$[1, 1, 1, \dots, 1, 1, 0]$$

$$[0, 0, 0, \dots, 0, 0, 10000]$$

Subsampling the Universe

- ❖ Sample coordinates of the universe with probability $\frac{1}{2^j}$ for $j = 0, 1, \dots, O(\log n)$ [IndykWoodruff05]

$$v = [1, 1, 1, 1, 1, 1, 1, \dots, 1, 1, 1, 1, 1, 10000]$$
$$[1, 0, 1, 0, 0, 1, 0, \dots, 1, 0, 0, 1, 1, 10000]$$
$$[1, 0, 0, 0, 0, 1, 0, \dots, 0, 0, 0, 1, 0, 0]$$

- ❖ The important medium and low level sets will be heavy-hitters in the subsampled streams!

Subsampling the Universe

- ❖ Sample coordinates of the universe with probability $\frac{1}{2^j}$ for $j = 0, 1, \dots, O(\log n)$ [IndykWoodruff05]

$$v = \begin{bmatrix} 1, 1, 1, 1, 1, 1, 1, \dots, 1, 1, 1, 1, 1, 10000 \\ 1, 0, 1, 0, 0, 1, 0, \dots, 1, 0, 0, 1, 1, 10000 \\ 1, 0, 0, 0, 0, 1, 0, \dots, 0, 0, 0, 1, 0, 0 \end{bmatrix}$$

- ❖ Will find the important medium and low level sets

Towards Privacy

- ❖ PRIVCOUNTSKETCH, private release of heavy-hitters, by adding Laplacian noise to each coordinate
- ❖ Even though PRIVCOUNTSKETCH estimates n frequencies, only $O\left(\frac{1}{\alpha^2}\right)$ frequencies are released, so only need to add Laplacian noise with scale $O\left(\frac{1}{\alpha^2}\right)$

Towards Privacy

- ❖ Even Laplacian noise with scale $O\left(\frac{1}{\alpha^2}\right)$ is too much noise for important low level sets
- ❖ Instead add Laplacian noise to the *size* of each important low level set

Important High
Level Sets

PRIVCOUNTSKETCH

Private magnitudes
of coordinates

Important
Medium Level
Sets

Subsampling

PRIVCOUNTSKETCH
+ Rescaling level set
sizes

Private sizes of level
sets

Important Low
Level Sets

Subsampling

Adding noise to
level set sizes

Private sizes of level
sets

