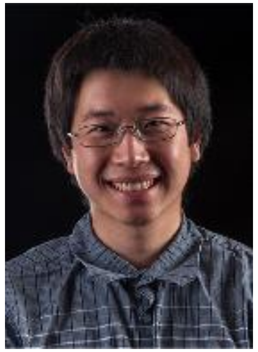
Separations for Estimating Large Frequency Moments on Data Streams

David P. Woodruff
Samson Zhou



Carnegie Mellon University

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

- Given a set S of m elements from $[n]$, let f_k be the frequency of element k (how often it appears)

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

Arbitrary-Order vs Random-Order Streams

- Arbitrary-order: Elements inducing f arrive sequentially in an arbitrary order (worst-case)

- Random-order: Elements inducing f arrive in a uniformly random order (average-case)

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

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

Frequency Moments

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

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

❖ Goal: Given a set S of m elements from $[n]$ and an accuracy parameter ε, output a $(1 + \varepsilon)$-approximation to F_p

❖ Motivation: Entropy estimation, network anomaly detection,…

Constant-Factor Approximation

- Space $O(\log n)$ algorithm for F_p with $p \in (0, 2]$ [BlasiokDingNelson17, BravermanViolaWoodruffYang18]
- Space $\widetilde{\Omega}\left(n^{1-2/p}\right)$ necessary for F_p with $p > 2$ [Ganguly12] on arbitrary-order streams, $\Omega\left(n^{1-2.5/p}\right)$ for random-order streams [ChakrabartiCormodeMcGregor16]

$(1+\varepsilon)$-Approximation for F_p with $p > 2$

- Space $\tilde{O}\left(\frac{1}{\varepsilon^2} n^{1-2/p}\right)$ algorithm [Ganguly11, GangulyWoodruff18]

- Space $\Omega\left(\frac{1}{\varepsilon^2} \frac{n^{1-2/p}}{\log n}\right)$ necessary for arbitrary-order streams [Ganguly12], $\Omega\left(n^{1-2.5/p} + \frac{1}{\varepsilon^2}\right)$ for random-order streams [ChakrabartiCormodeMcGregor16]

Our Results: F_p Moment Estimation, $p > 2$

- Space $\tilde{O}\left(\frac{1}{\varepsilon^{4/p}} n^{1-2/p}\right)$ algorithm for random-order insertion-only streams
- Space $\tilde{O}\left(\frac{1}{\varepsilon^{4/p}} n^{1-2/p}\right)$ algorithm for two-pass streams in arbitrary-order, even with turnstile updates
- Space $\Omega\left(\frac{1}{\varepsilon^{2}} n^{1-2/p}\right)$ necessary for one-pass arbitrary-order streams
- Results show separation between one-pass arbitrary-order and one-pass random-order, multi-pass arbitrary order

Level Sets

- Partition the coordinates $k \in [n]$ into *level sets* Λ_i based on the frequencies of each item, so that $k \in \Lambda_i$ if

$$f_k^p \in \left[\frac{F_p}{2^i}, \frac{2F_p}{2^i}\right]$$

- Define *contribution* C_i of level set Λ_i as the total contribution of all coordinates in Λ_i to F_p

$$C_i = \sum_{k \in \Lambda_i} f_k^p$$

Level Sets

❖ Intuition: Level sets Λ_i decompose F_p

$$F_p = \sum_{k \in [n]} f_k^p = \sum_i \sum_{k \in \Lambda_i} f_k^p = \sum_i C_i$$

❖ To obtain a $(1+\varepsilon)$-approximation to F_p, it suffices to obtain a $(1+\varepsilon)$-approximation $\widehat{C_i}$ to the contribution C_i of each level set Λ_i with $C_i \geq \varepsilon F_p$

Heavy-Hitters

- Let L_p be the norm of the frequency vector:

$$L_p = \left(f_1^p + f_2^p + \cdots + f_n^p\right)^{1/p}$$

- Goal: Given a set S of m elements from $[n]$ and a threshold ε, output the elements k such that $f_k > \varepsilon\, L_p$ and their approximate frequencies $\widehat{f_k}$
- Motivation: DDoS prevention, iceberg queries, moment estimation

Heavy-Hitters to Level Set Contributions

- If $f_k^p > \varepsilon^2 F_p$, then k is an L_2 heavy-hitter with threshold $\frac{\varepsilon^{2/p}}{n^{1/2-1/p}}$
  - $f_k^p > \varepsilon^2 F_p$ implies $f_k > \varepsilon^{2/p} L_p$ so $f_k^2 > \varepsilon^{4/p} L_p^2 > \frac{\varepsilon^{4/p}}{n^{1-2/p}} F_2$

- Use an L_2 heavy-hitter algorithm with threshold $\frac{\varepsilon^{2/p}}{n^{1/2-1/p}}$ to find k and obtain a $(1+\varepsilon)$ approximate frequency $\widehat{f_k}$

Level Sets with Large Frequencies

- Recall: To obtain a $(1+\varepsilon)$-approximation to F_p, it suffices to obtain a $(1+\varepsilon)$-approximation $\widehat{C_i}$ to the contribution C_i of each level set Λ_i with $C_i \geq \varepsilon F_p$
- If $f_k^p > \varepsilon^2 F_p$, then an L_2 heavy-hitter algorithm with threshold $\frac{\varepsilon^{2/p}}{n^{1/2-1/p}}$ can find k and obtain a $(1+\varepsilon)$ approximate frequency $\widehat{f_k}$
- In summary, we obtain a $(1+\varepsilon)$-approximation $\widehat{C_i}$ to the contribution C_i of each level set Λ_i with $i < \log\frac{1}{\varepsilon^2}$

Idealized Algorithm

1. Form $O(\log n)$ streams $S_0, S_1, S_2, S_3, \ldots$ by subsampling the universe $[n]$ at rate $\frac{1}{2^j}$ for $j = 0,1,2,3, \ldots$
2. Use L_2 heavy-hitter algorithms with threshold $\frac{\varepsilon^{2/p}}{n^{1/2-1/p}}$ on the substreams $S_0, S_1, S_2, S_3, \ldots$ and find $(1+\varepsilon)$-approximate frequencies $\widehat{f_k}$ to each reported heavy-hitter $k \in [n]$
3. Use the approximate frequencies $\widehat{f_k}$ to compute approximate contributions $\widehat{C_i}$ to C_i
4. Output $\sum_i \widehat{C_i}$

Space Complexity / Source of the Separation

- Space determined by L_2 heavy-hitter algorithms with threshold $\frac{\varepsilon^{2/p}}{n^{1/2-1/p}}$ on the substreams $S_0, S_1, S_2, S_3, \ldots$ to find $(1+\varepsilon)$-approximate frequencies $\widehat{f_k}$ to each reported heavy-hitter $k \in [n]$

- Can black-box heavy-hitter algorithms for one-pass random-order streams [BravermanGargWoodruff20]

- Similar results are NOT known for one-pass arbitrary-order streams

F_p Moment Estimation, $p > 2$

❖ Space $\tilde{O}\left(\frac{1}{\varepsilon^{4/p}}n^{1-2/p}\right)$ algorithm for random-order streams

❖ Space $\tilde{O}\left(\frac{1}{\varepsilon^{4/p}}n^{1-2/p}\right)$ algorithm for two-pass streams in arbitrary-order, even with turnstile updates

❖ Space $\Omega\left(\frac{1}{\varepsilon^{2}}n^{1-2/p}\right)$ necessary for one-pass arbitrary-order streams

❖ Results show separation between one-pass arbitrary-order and one-pass random-order, multi-pass arbitrary order

Level Sets with Small Frequencies

- Remains to approximate the contribution C_i of each level set Λ_i with $i \geq \log\frac{1}{\varepsilon^2}$ and $C_i \geq \varepsilon F_p$
- Suppose $C_i = F_p$ for some $i \geq T$, where $T = \log\frac{1}{\varepsilon^2}$
- Since $f_k^p \in \left[\frac{F_p}{2^i}, \frac{2F_p}{2^i}\right]$ for each $k \in \Lambda_i$, then $|\Lambda_i| \geq 2^{i-1}$

Level Sets with Small Frequencies

- If we sample the universe $[n]$ at a rate $\frac{1}{2^{i-T}}$, then $\frac{|\Lambda_i|}{2^{i-T}} \approx \frac{1}{\varepsilon^2}$ elements of Λ_i will be sampled
- Intuition: Use their approximate frequencies to estimate C_i

- Standard variance argument shows rescaling the sampled contribution by 2^{i-T} gives a $(1+\varepsilon)$-approximation $\widehat{C_i}$ to C_i, if we sample $\frac{1}{\varepsilon^2}$ elements of Λ_i

Level Sets with Small Frequencies

- How to compute approximate frequencies?
- If we sample the universe $[n]$ at a rate $\frac{1}{2^{i-T}}$, the frequency moment $U_p^{(i)}$ of the subsampled stream will be $\frac{F_p}{2^{i-T}}$ in expectation
- We have $f_k^p \in \left[\frac{F_p}{2^i}, \frac{2F_p}{2^i}\right]$, so $f_k^p \geq \frac{1}{2^T} U_p^{(i)}$ with $\frac{1}{2^T} = \varepsilon^2$

- In summary, we expect k to be a heavy-hitter with respect to $U_p^{(i)}$
- Use an L_2 heavy-hitter algorithm with threshold $\frac{\varepsilon^{2/p}}{n^{1/2-1/p}}$ w.r.t. $U_p^{(i)}$ to find k and obtain an approximate frequency $\widehat{f_k}$

Level Sets with Small Frequencies

- ❖ If $C_i = F_p$ for some $i \geq T$ and $f_k^p \in \left[\frac{F_p}{2^i}, \frac{2F_p}{2^i}\right]$ then an L_2 heavy-hitter algorithm on a substream that samples k with rate $\frac{1}{2^{i-T}}$ and threshold $\frac{\varepsilon^{2/p}}{n^{1/2-1/p}}$ can obtain an approximate frequency $\widehat{f_k}$

- ❖ If $\widehat{f_k}$ is a $(1+\varepsilon)$-approximation to f_k for all such k, then we can rescale and obtain a $(1+O(\varepsilon))$-approximation $\widehat{C_i}$ to C_i

Level Sets with Small Frequencies

- Recall: To obtain a $(1+\varepsilon)$-approximation to F_p, it suffices to obtain a $(1+\varepsilon)$-approximation $\widehat{C}_i$ to the contribution C_i of each level set Λ_i with $C_i \geq \varepsilon F_p$

- Previous argument shows $(1+\varepsilon)$-approximation $\widehat{C}_i$ to the contribution C_i if $C_i = F_p$
- Same argument will work if $C_i = \gamma_i F_p$ for some $\gamma_i \in [\varepsilon, 1]$, since we get $(1+\varepsilon/\gamma_i)$-approximation $\widehat{C}_i$ to the contribution C_i, which gives at most εF_p additive error to C_i

Lower Bound

- Space $\Omega\left(\frac{1}{\varepsilon^2} n^{1-2/p}\right)$ necessary for one-pass arbitrary-order streams

- We define the (t, ε, n)-player set disjointness estimation problem $(t, \varepsilon, n) - DisjInfty$ and show it has total communication cost $\Omega\left(\frac{n}{t}\right)$

- Set $t = \Theta\left(\frac{1}{\varepsilon} n^{1/p}\right)$ and show a reduction from $(1 + \varepsilon)$-approximation of F_p to $(t, \varepsilon, n) - DisjInfty$

Lower Bound

- Space $\Omega\left(\frac{1}{\varepsilon^2}n^{1-2/p}\right)$ necessary for one-pass arbitrary-order streams

- We define the (t,ε,n)-player set disjointness estimation problem $(t,\varepsilon,n)-DisjInfty$ and show it has total communication cost $\Omega\left(\frac{n}{t}\right)$

- Set $t=\Theta\left(\frac{1}{\varepsilon}n^{1/p}\right)$ and show a reduction from $(1+\varepsilon)$-approximation of F_p to $(t,\varepsilon,n)-DisjInfty$

(t,ε,n)-Player Set Disjointness Estimation

- There are $t+1$ players $P_1, P_2, \ldots, P_{t+1}$. For each $s \in [t]$, P_s receives a vector $v_s \in \{0,1\}^n$. Player P_{t+1} receives a "spike location" $j \in [n]$ and a bit $c \in \{0,1\}$.
- Let $u = \sum_s v_s$. The promise is that:
  - $u_i \leq 1$ for each $i \neq j$ (player sets are disjoint outside of coordinate j)
  - either $u_j \leq 1$ or $u_j = t$ (either player sets are disjoint at coordinate j or all players have j in their sets)
- Player P_{t+1} must differentiate between the three cases:
  - (1) $u_j + \frac{ct}{\varepsilon} \leq t$, (2) $u_j + \frac{ct}{\varepsilon} \in \left\{\frac{t}{\varepsilon}, \frac{t}{\varepsilon} + 1\right\}$, (3) $u_j + \frac{ct}{\varepsilon} = (1+\varepsilon)\frac{t}{\varepsilon}$

(t, ε, n)-Player Set Disjointness Estimation

- Intuition: Since $P_1, P_2, \dots, P_t$ do not know the spike location $j \in [n]$, they must solve the problem on all coordinates
- Solving multi-party set disjointness on a single coordinate is roughly solving the AND problem of t bits
- Hellinger distance argument shows the information complexity of AND is $\Omega\left(\frac{1}{t}\right)$ [Jayram09]
- Use direct sum embedding to show $(t, \varepsilon, n) - DisjInfty$ has total communication cost $\Omega\left(\frac{n}{t}\right)$

Reduction

- Recall that for each $s \in [t]$, P_s receives a vector $v_s \in \{0,1\}^n$. Player P_{t+1} receives a "spike location" $j \in [n]$ and a bit $c \in \{0,1\}$.

- For each $s \in [t]$, P_s inserts the coordinates of vector v_s into the stream
- P_{t+1} adds the vector $\frac{ct}{\varepsilon} e_j$, where e_j is the elementary vector corresponding to the spike location
- Intuition: Mass added to spike location provides F_p separation

Reduction

- Recall that for each $s \in [t]$, P_s receives a vector $v_s \in \{0,1\}^n$. Player P_{t+1} receives a "spike location" $j \in [n]$ and a bit $c \in \{0,1\}$.
- Let $u = \sum_s v_s$ and $t = \Theta\left(\frac{1}{\varepsilon} n^{1/p}\right)$

- If $c = 0$, then $\|x\|_p^p \leq n + \frac{C^p}{\varepsilon^p} n$ for some constant $C > 0$
- If $c = 1$ and $u_j = 1$, then $\frac{C^p}{\varepsilon^{2p}} n \leq \|x\|_p^p \leq n + p + \frac{pC^p}{\varepsilon^{2p}} n$
- If $c = 1$ and $u_j = t$, then $\|x\|_p^p \geq \left(1 + \frac{1}{\varepsilon}\right)^p \frac{pC^p}{\varepsilon^{2p}} n$

Lower Bound

- $(1+\varepsilon)$-approximation of F_p separates these three cases and thus solves $(t,\varepsilon,n) - DisjInfty$ for $t = \Theta\left(\frac{1}{\varepsilon} n^{1/p}\right)$
- $(1+\varepsilon)$-approximation of F_p requires $\Omega\left(\frac{1}{\varepsilon^2} n^{1-2/p}\right)$ space