Improved Sub-linear Time Moment Estimation Using Weighted Sampling

Anup Bhattacharya

Pinki Pradhan

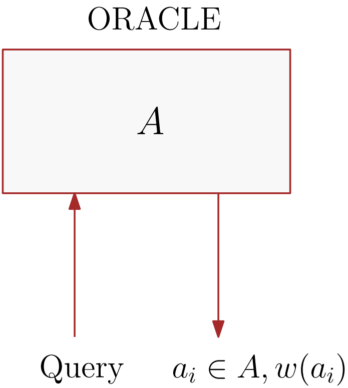
NISER, Bhubaneswar

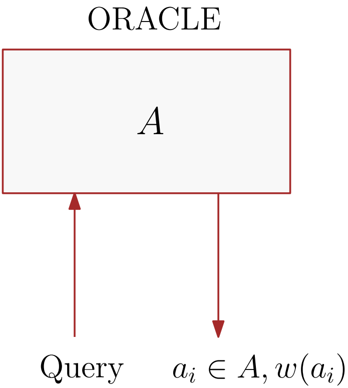
CIAC 2025, LUISS University, Rome, Italy

Introduction

- There is a set $A = \{a_1, a_2, \ldots, a_n\}$ of size n.
- Every $a_i \in A$ has a non-negative weight $w(a_i) \geq 0$.
- Given a parameter $t > 0$, the *t-th moment* of A, defined as

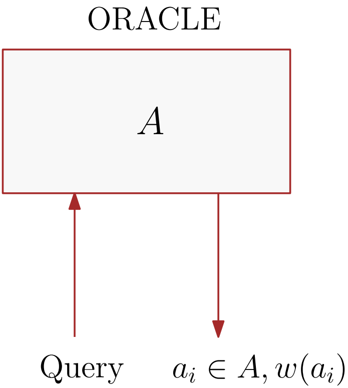
$$S_t = \sum_{i=1}^{n} w(a_i)^t.$$


ORACLE
A
Query
$a_i \in A, w(a_i)$



Question

Can we estimate S_t using a sublinear number of queries to the oracle?



Question

Can we estimate S_t using a sublinear number of queries to the oracle?

Observation

- $\Omega(n)$ queries are required to compute S_t exactly.

(ϵ, δ)-Estimator:

(ϵ, δ)-Estimator of S_t: For any $\epsilon, \delta \in (0,1)$, we say that $\widetilde{S}_t$ is an (ϵ, δ)-estimator of S_t if, with probability at least $1 - \delta$,

$$\widetilde{S}_t \in \left[(1-\epsilon)S_t,\ (1+\epsilon)S_t\right].$$

(ϵ, δ)-Estimator:

(ϵ, δ)-Estimator of S_t: For any $\epsilon, \delta \in (0,1)$, we say that $\widetilde{S}_t$ is an (ϵ, δ)-estimator of S_t if, with probability at least $1 - \delta$,

$$\widetilde{S}_t \in \left[(1-\epsilon)S_t,\ (1+\epsilon)S_t\right].$$

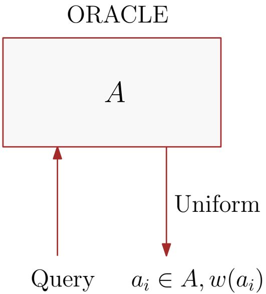
Question

How many queries are required to obtain an (ϵ, δ)-estimate of S_t?

Sampling Oracle Uniformly

Claim. $\Omega(n)$ queries are required to obtain an (ϵ, δ)-estimator of S_t using uniform sampling.

- A, where weights of the elements are $\{0, 0, \ldots, 0, n\}$.



Sampling Oracle Proportionally

- Probability of picking an element a_i proportionally:

$$p_i = \frac{w(a_i)}{W}, \quad \text{where } W = \sum_{i=1}^{n} w(a_i)$$

Sampling Oracle Proportionally

- Probability of picking an element a_i proportionally:

$$p_i = \frac{w(a_i)}{W}, \quad \text{where } W = \sum_{i=1}^{n} w(a_i)$$

Question

How many queries are required to obtain an (ϵ, δ)-estimator of S_t using proportional sampling?

Known Results

1. For $t = 1$, this is known as the **sum estimation problem**.

Known Results

1. For $t = 1$, this is known as the **sum estimation problem**.
2. Motwani, Panigrahy, and Xu [Motwani et al., 2007] first studied this problem using weighted sampling, providing:

$$\text{Upper bound: } \widetilde{O}\left(\frac{\sqrt{n}}{\epsilon^{7/2}}\right), \quad \text{Lower bound: } \Omega(\sqrt{n}).$$

Known Results

1. For $t = 1$, this is known as the **sum estimation problem**.
2. Motwani, Panigrahy, and Xu [Motwani et al., 2007] first studied this problem using weighted sampling, providing:

$$\text{Upper bound: } \widetilde{O}\left(\frac{\sqrt{n}}{\epsilon^{7/2}}\right), \quad \text{Lower bound: } \Omega(\sqrt{n}).$$

3. Beretta and Tětek [Beretta and Tětek, 2022] improved the result and established the **optimal bound**:

$$\Theta\left(\frac{\sqrt{n}}{\epsilon}\right).$$

Known Results

1. For $t = 1$, this is known as the **sum estimation problem**.
2. Motwani, Panigrahy, and Xu [Motwani et al., 2007] first studied this problem using weighted sampling, providing:

$$\text{Upper bound: } \widetilde{O}\left(\frac{\sqrt{n}}{\epsilon^{7/2}}\right), \quad \text{Lower bound: } \Omega(\sqrt{n}).$$

3. Beretta and Tětek [Beretta and Tětek, 2022] improved the result and established the **optimal bound**:

$$\Theta\left(\frac{\sqrt{n}}{\epsilon}\right).$$

4. Aliakbarpour et al. [Aliakbarpour et al., 2018] studied the estimation of t-stars in a graph assuming access to a **random edge sampling oracle**. They showed:

$$\text{Upper bound: } O\left(\frac{n^{1-1/t}\ln(1/\delta)}{\epsilon^2}\right), \quad \text{Lower bound: } \Omega\left(n^{1-1/t}\right).$$

Our Results

Results	Upper Bound	Lower Bound
For $t > 1$	$O\left(\frac{\sqrt{n}}{\epsilon} + \frac{n^{1-\frac{1}{t}}\ln(1/\delta)}{\epsilon^2}\right)$	$\Omega\left(\frac{n^{1-\frac{1}{t}}\ln(1/\delta)}{\epsilon^2}\right)$
For $0 < t < \frac{1}{2}$	–	$\Omega(n)$
For $\frac{1}{2} < t < 1$	$O\left(\frac{\sqrt{n}}{\epsilon} + \frac{n^{\frac{1}{t}-1}}{\epsilon^2}\right)$	–

Our Results

Results	Upper Bound	Lower Bound
For $t > 1$	$O\left(\frac{\sqrt{n}}{\epsilon} + \frac{n^{1-\frac{1}{t}}\ln(1/\delta)}{\epsilon^2}\right)$	$\Omega\left(\frac{n^{1-\frac{1}{t}}\ln(1/\delta)}{\epsilon^2}\right)$
For $0 < t < \frac{1}{2}$	–	$\Omega(n)$
For $\frac{1}{2} < t < 1$	$O\left(\frac{\sqrt{n}}{\epsilon} + \frac{n^{\frac{1}{t}-1}}{\epsilon^2}\right)$	–

Results	Upper Bound	Lower Bound
Hybrid sampling	–	$\Omega\left(\frac{n^{1-\frac{1}{t}}\ln(1/\delta)}{\epsilon^2}\right)$
Parameter ρ	$O\left(\left(\frac{\sqrt{n}}{\epsilon} + \frac{\rho}{\epsilon^2}\right)\ln\frac{1}{\delta}\right)$	$\Omega\left(\frac{\rho\ln(1/\delta)}{\epsilon}\right)$

Upper Bound for $t > 1$

Theorem 1. There exists an algorithm that given proportional sampling access to the weights of the elements in a set A and parameter $t > 1, \epsilon, \delta \in (0,1)$, provides an (ϵ, δ)-estimate of S_t using $O\left(\frac{\sqrt{n}}{\epsilon} + \frac{n^{1-1/t}\log 1/\delta}{\epsilon^2}\right)$ samples.

Proof sketch

- Let a_i be the first sampled element and $p_i = \frac{w(a_i)}{W}$.

Proof sketch

- Let a_i be the first sampled element and $p_i = \frac{w(a_i)}{W}$.
- $X_1 = w(a_i)^t$.

Proof sketch

- Let a_i be the first sampled element and $p_i = \frac{w(a_i)}{W}$.
- $X_1 = w(a_i)^t$.
- $\mathbb{E}[X_1] = \sum_{j=1}^{n} w(a_j)^t . \frac{w(a_j)}{W}$.

Proof Sketch

- $X_1 = \frac{w(a_i)^t}{p_i}$, where $p_i = \frac{w(a_i)}{W}$.

Proof Sketch

- $X_1 = \frac{w(a_i)^t}{p_i}$, where $p_i = \frac{w(a_i)}{W}$.
- $\mathbb{E}[X_1] = \sum_{j=1}^{n} \frac{w(a_j)^t}{p_j}.\frac{w(a_j)}{W} = S_t$.

Proof Sketch

- $X_1 = \frac{w(a_i)^t}{p_i}$, where $p_i = \frac{w(a_i)}{W}$.
- $\mathbb{E}[X_1] = \sum_{j=1}^{n} \frac{w(a_j)^t}{p_j} . \frac{w(a_j)}{W} = S_t$.
- Using result of [Beretta and Tětek, 2022], we can get estimator $\widetilde{W}$ of W, such that $(1-\epsilon_1)W \leq \widetilde{W} \leq (1+\epsilon_1)W$, where $\epsilon_1 = \frac{\epsilon}{2}$.

Proof Sketch

- $X_1 = \frac{w(a_i)^t}{p_i}$, where $p_i = \frac{w(a_i)}{W}$.
- $\mathbb{E}[X_1] = \sum_{j=1}^{n} \frac{w(a_j)^t}{p_j} . \frac{w(a_j)}{W} = S_t$.
- Using result of [Beretta and Tětek, 2022], we can get estimator $\widetilde{W}$ of W, such that $(1-\epsilon_1)W \leq \widetilde{W} \leq (1+\epsilon_1)W$, where $\epsilon_1 = \frac{\epsilon}{2}$.
- $(1-\epsilon_1)S_t \leq \mathbb{E}[X_1] \leq (1+\epsilon_1)S_t$.

Proof Sketch

- Let $X = \frac{\sum_{j=1}^{l} X_j}{l}$ denote the average over l independent samples.

Proof Sketch

- Let $X = \frac{\sum_{j=1}^{l} X_j}{l}$ denote the average over l independent samples.
- $(1-\epsilon_1)S_t \leq \mathbb{E}[X] \leq (1+\epsilon_1)S_t$.

Proof Sketch

- Let $X = \frac{\sum_{j=1}^{l} X_j}{l}$ denote the average over l independent samples.
- $(1-\epsilon_1)S_t \leq \mathbb{E}[X] \leq (1+\epsilon_1)S_t$.
- By Chebyshev's inequality, the failure probability is bounded by:

$$\frac{\mathrm{Var}[X]}{(\epsilon-\epsilon_1)^2 S_t^2} \leq \frac{(1+\epsilon_1)^2}{l(\epsilon-\epsilon_1)^2} \cdot n^{1-1/t}.$$

Proof Sketch

- Let $X = \frac{\sum_{j=1}^{l} X_j}{l}$ denote the average over l independent samples.
- $(1-\epsilon_1)S_t \leq \mathbb{E}[X] \leq (1+\epsilon_1)S_t$.
- By Chebyshev's inequality, the failure probability is bounded by:

$$\frac{\mathrm{Var}[X]}{(\epsilon-\epsilon_1)^2 S_t^2} \leq \frac{(1+\epsilon_1)^2}{l(\epsilon-\epsilon_1)^2} \cdot n^{1-1/t}.$$

- Therefore, choosing $l = O\left(\frac{(1+\epsilon_1)^2 \cdot n^{1-1/t}}{(\epsilon-\epsilon_1)^2}\right)$ ensures the failure probability is bounded by a constant.

Proof Sketch

- Let $X = \frac{\sum_{j=1}^{l} X_j}{l}$ denote the average over l independent samples.
- $(1-\epsilon_1)S_t \leq \mathbb{E}[X] \leq (1+\epsilon_1)S_t$.
- By Chebyshev's inequality, the failure probability is bounded by:

$$\frac{\mathrm{Var}[X]}{(\epsilon-\epsilon_1)^2 S_t^2} \leq \frac{(1+\epsilon_1)^2}{l(\epsilon-\epsilon_1)^2} \cdot n^{1-1/t}.$$

- Therefore, choosing $l = O\left(\frac{(1+\epsilon_1)^2 \cdot n^{1-1/t}}{(\epsilon-\epsilon_1)^2}\right)$ ensures the failure probability is bounded by a constant.
- Repeating this independently $O(\log(1/\delta))$ times and taking the median reduces the failure probability to at most δ.

Lower Bound

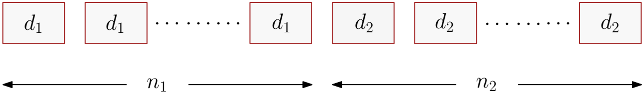
Theorem 2. For any $\epsilon, \delta \in (0, 1)$ and and $t > 1$, any randomized algorithm that computes an (ϵ, δ)-estimate of S_t requires $\Omega(\frac{n^{1-1/t} \ln 1/\delta}{\epsilon^2})$ proportional samples.

Lower Bound

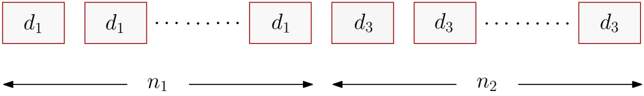
- To establish the lower bound, we apply **Yao's Minimax Principle**.

- To establish the lower bound, we apply **Yao's Minimax Principle**.

D_1



D_2



Lower Bound

D_1

$$n_1 = \frac{n^2}{n + \epsilon^{\frac{2t-1}{t-1}}}$$

$$d_1 = n^{1-1/t}\epsilon^{1/(t-1)}$$

$$n_2 = \frac{n\epsilon^{\frac{2t-1}{t-1}}}{n + \epsilon^{\frac{2t-1}{t-1}}}$$

$$d_2 = 0$$

D_2

$$n_1 = \frac{n^2}{n + \epsilon^{\frac{2t-1}{t-1}}}$$

$$d_1 = n^{1-1/t}\epsilon^{1/(t-1)}$$

$$n_2 = \frac{n\epsilon^{\frac{2t-1}{t-1}}}{n + \epsilon^{\frac{2t-1}{t-1}}}$$

$$d_3 = n$$

Lower Bound

1. $n_1 + n_2 = n$.

Lower Bound

1. $n_1 + n_2 = n$.
2. $n_1 \cdot d_1^t + n_2 \cdot d_3^t = (1 + \epsilon)(n_1 \cdot d_1^t)$.

Lower Bound

1. $n_1 + n_2 = n$.
2. $n_1 \cdot d_1^t + n_2 \cdot d_3^t = (1+\epsilon)(n_1 \cdot d_1^t)$.
3. The probability of seeing at least one d_3 using proportional sampling is: $\frac{n_2 d_3}{n_2 d_3 + n_1 d_1} = \frac{1}{1 + \frac{n^{1-1/t}}{\epsilon^2}}$.

Lower Bound

1. $n_1 + n_2 = n$.
2. $n_1 \cdot d_1^t + n_2 \cdot d_3^t = (1+\epsilon)(n_1 \cdot d_1^t)$.
3. The probability of seeing at least one d_3 using proportional sampling is: $\frac{n_2 d_3}{n_2 d_3 + n_1 d_1} = \frac{1}{1+\frac{n^{1-1/t}}{\epsilon^2}}$.
4. If $p = \frac{1}{1+\frac{n^{1-1/t}}{\epsilon^2}}$, then the number of samples require to be drawn from a Geom(p) to observe one success with probability at least $1-\delta$ is $\Omega\left(\frac{\ln(1/\delta)}{p}\right) = \Omega\left(\frac{n^{1-1/t} \cdot \ln(1/\delta)}{\epsilon^2}\right)$.

Upper Bound for $1/2 < t < 1$

Theorem 3. There exists an algorithm, that given proportional sampling access to the weights of the elements of a set A and a parameters $1/2 < t < 1, \epsilon \in (0, 1)$, provides an $(\epsilon, 1/3)$-estimate of S_t using $O\left(\frac{\sqrt{n}}{\epsilon} + \frac{n^{\frac{1}{t}-1}}{\epsilon^2}\right)$.

Lower Bound for $0 < t < 1/2$

Theorem 4. For any $\epsilon > 0$ and $t \leq 1/2$, any randomized algorithm that computes an $(\epsilon, 1/3)$-estimate of S_t requires $\Omega(n)$ proportional samples.

Characterization of Sample Complexity

Let $S_t = \sum_{a_i \in A} w(a_i)^t$ and $W = \sum_{a_i \in A} w(a_i)$.

Characterization of Sample Complexity

Let $S_t = \sum_{a_i \in A} w(a_i)^t$ and $W = \sum_{a_i \in A} w(a_i)$.

We define the **moment-density parameter** ρ as:

$$\rho = \max_{L \subseteq A} \frac{\dfrac{\sum\limits_{a_j \in L} w(a_j)^t}{S_t}}{\dfrac{\sum\limits_{a_j \in L} w(a_j)}{W}} = \max_{L \subseteq A} \left(\frac{\sum\limits_{a_j \in L} w(a_j)^t}{\sum\limits_{a_j \in L} w(a_j)} \cdot \frac{W}{S_t} \right)$$

Characterization of Sample Complexity

Theorem 5. There exists an algorithm that given weighted sampling access to the weights of elements of A having moment-density parameter ρ and parameters $t > 1, \epsilon, \delta \in (0, 1)$, provides an (ϵ, δ)-estimate of S_t using $O((\sqrt{n}/\epsilon + \rho/\epsilon^2) \ln 1/\delta)$ weighted samples.

Characterization of Sample Complexity

Theorem 5. There exists an algorithm that given weighted sampling access to the weights of elements of A having moment-density parameter ρ and parameters $t > 1, \epsilon, \delta \in (0,1)$, provides an (ϵ, δ)-estimate of S_t using $O((\sqrt{n}/\epsilon + \rho/\epsilon^2) \ln 1/\delta)$ weighted samples.

Theorem 6. For any ρ, $\epsilon, \delta > 0$ and $t > 1$, any randomized algorithm for an (ϵ, δ)-estimate of S_t on an instance with the moment-density parameter ρ requires $\Omega(\frac{\rho \ln 1/\delta}{\epsilon})$ weighted samples.

Summary of Results

- **Moment Estimation for $t > 1$:**
  - *Upper bound:*
    $$O\left(\frac{\sqrt{n}}{\epsilon} + \frac{n^{1-\frac{1}{t}\ln(1/\delta)}}{\epsilon^2}\right)$$
  - *Lower bound:*
    $$\Omega\left(\frac{n^{1-\frac{1}{t}\ln(1/\delta)}}{\epsilon^2}\right)$$
  - Tight bounds for $t \geq 2$.
- **Fractional Moments $\left(\frac{1}{2} < t < 1\right)$:**
  - *Upper bound:*
    $$O\left(\frac{\sqrt{n}}{\epsilon} + \frac{n^{\frac{1}{t}-1}}{\epsilon^2}\right)$$
- **Very Small Moments $\left(0 < t < \frac{1}{2}\right)$:** $\Omega(n)$ queries are required

Summary of Results

- **Hybrid Sampling:**

$$\Omega\left(\frac{n^{1-\frac{1}{t}\ln(1/\delta)}}{\epsilon^2}\right) \text{ queries are required}$$

- **Parameter ρ:**
  - *Upper bound:*

$$O\left(\left(\frac{\sqrt{n}}{\epsilon}+\frac{\rho}{\epsilon^2}\right)\ln\frac{1}{\delta}\right)$$

  - *Lower bound:*

$$\Omega\left(\frac{\rho\ln(1/\delta)}{\epsilon}\right)$$

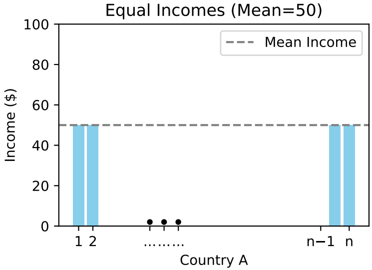
References I

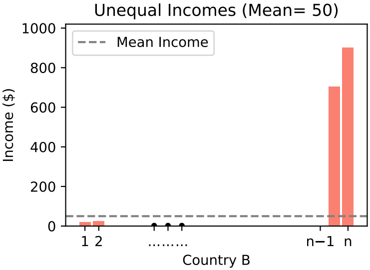
Aliakbarpour, M., Biswas, A. S., Gouleakis, T., Peebles, J., Rubinfeld, R., and Yodpinyanee, A. (2018).
Sublinear-time algorithms for counting star subgraphs via edge sampling.
Algorithmica, 80:668–697.

Beretta, L. and Tětek, J. (2022).
Better sum estimation via weighted sampling.
In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2303–2338. SIAM.

Motwani, R., Panigrahy, R., and Xu, Y. (2007).
Estimating sum by weighted sampling.
In *International Colloquium on Automata, Languages, and Programming (ICALP)*, pages 53–64. Springer.

Application: Wealth Inequality via Higher Moments





- Higher moments $S_2, S_3, \ldots$ do not grow rapidly.
- Wealth is evenly distributed across individuals.

- Higher moments $S_2, S_3, \ldots$ grow exponentially.
- Wealth is concentrated among a small fraction of individuals.