Communication and Memory Efficient Testing of Discrete Distributions

Themis Gouleakis
USC→MPI

July 21, 2019

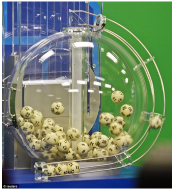
Joint work with: *Ilias Diakonikolas (USC), Daniel Kane (UCSD)* and *Sankeerth Rao (UCSD)*

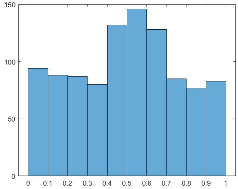
Motivation

- Datasets growing → too many samples needed!
- Can we do *property testing* distributedly?

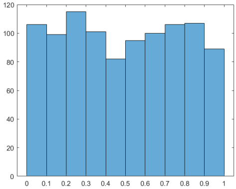
- Insufficient memory!
- Design *low memory* algorithms!

Is the lottery fair?



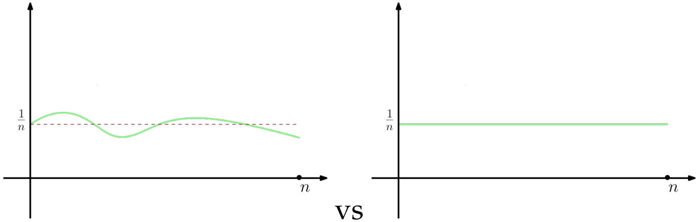


vs



- We can **learn** the distribution: $\Omega(n)$ samples.
- Centralized sampling/ unbounded memory: we can **test** (uniform vs ε-far) with $\Theta(\sqrt{n}/\varepsilon^2)$ samples.
- What if we have memory constraints/unavailable centralized sampling?

Definition and (centralized) prior work



Uniformity testing problem

Given samples from a probability distribution p, distinguish $p = U_n$ from $\|p - U_n\|_1 > \varepsilon$ with success probability at least $2/3$.

- Sample complexity: $\Theta\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ [Goldreich, Ron 00],[Batu, Fisher, Fortnow, Kumar, Rubinfeld, White 01],[Paninski 08], [Chan, Diakonikolas, Valiant, Valiant 14], [Diakonikolas, G, Peebles, Price 17]

Prior/related work

Distributed learning

- Parameter estimation [ZDJW13],[GMN14],[BGMNW16],[JLY16],[HOW18]
- Non-parametric [DGLNOS17],[HMOW18]

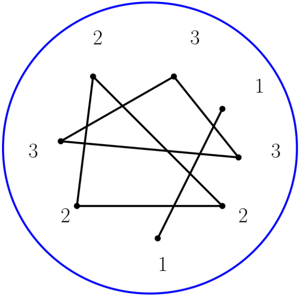
Distributed testing

- Single sample per machine with sublogarithmic size messages: [Acharya, Cannone, Tyagi 18]
- Two-party setting: [Andoni, Malkin, Nosatzki 18]
- LOCAL and CONGEST models: [Fisher, Meir, Oshman 18]

Centralized Collision-Based Algorithm

[Goldreich, Ron 00],[Batu, Fisher, Fortnow, Kumar, Rubinfeld, White 01]

Problem: Given distribution p over $[n]$, distinguish $p = U_n$ from $\|p - U_n\|_1 \geq \epsilon$.



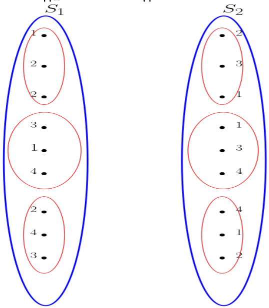
- m samples
- **Node labels:** i.i.d samples from p.
- **Edges:** $\{i, j\} \in E$ iff $L(i) = L(j)$

- Define statistic $Z = \sharp$edges $\Rightarrow \mathbb{E}[Z] = \binom{m}{2} \cdot \|p\|_2^2$
  - Minimized for $p = U_n$
- **Idea:** Draw *enough* samples and *compare* Z to some threshold.

Generic Bipartite Testing Algorithm

ℓ samples per machine

Problem: Given distribution p over $[n]$, distinguish $p = U_n$ from $\|p - U_n\|_1 \geq \epsilon$.

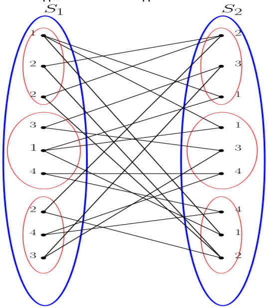


- ℓ samples **per machine**.
- **Node labels:** i.i.d samples from p.
- **Edges:** $\{i, j\} \in E$ iff $(i \in S_1) \wedge (j \in S_2) \wedge (L(i) = L(j))$

Generic Bipartite Testing Algorithm

ℓ samples per machine

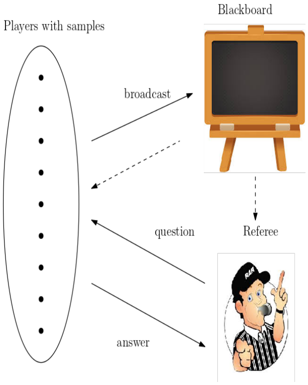
Problem: Given distribution p over $[n]$, distinguish $p = U_n$ from $\|p - U_n\|_1 \geq \epsilon$.



- ℓ samples **per machine**.
- **Node labels:** i.i.d samples from p.
- **Edges:** $\{i, j\} \in E$ iff $(i \in S_1) \wedge (j \in S_2) \wedge (L(i) = L(j))$

- Define statistic $Z = \sharp$edges $\Rightarrow \mathbb{E}[Z] = |S_1| \cdot |S_2| \cdot \|p\|_2^2$
  - Minimized for $p = U_n$
- **Remark:** *Suboptimal* sample complexity, but can lead to *optimal* communication complexity in certain cases.

Communication model

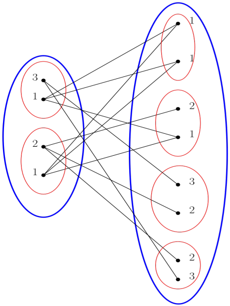


- *Unbounded* number of players
- Players can *broadcast* on the blackboard
- The referee asks questions to players and receives replies.

- **Goal:** Minimize total number of *bits* of communication.

A Communication efficient Algorithm

- **Idea:** Statistic $Z =$ sum of degrees on *one* side.
  - *Only* the opposite side needs to reveal samples exactly.

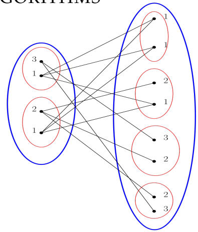


- **Broadcasted samples:** $\ell \cdot |S_1| = \frac{\sqrt{n/\ell}}{\epsilon^2 \sqrt{\log n}}$
  - **Not** enough for testing.
- And the samples on the right?
  - Only **degrees** d_k sent to the referee.
    - $O(1)$ bits/message w.l.o.g.

- **Communication complexity:** $O\left(\frac{\sqrt{n/\ell}\sqrt{\log n}}{\epsilon^2}\right)$ bits.
  - Matching lower bound of $\Omega\left(\frac{\sqrt{n/\ell}\sqrt{\log n}}{\epsilon^2}\right)$ bits for small ℓ.
- Better than naive $O\left(\frac{\sqrt{n}\log n}{\epsilon^2}\right)$ bits.

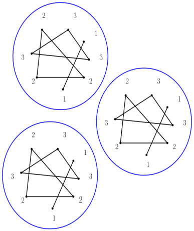
Communication efficient implementation

Two algorithms



Case I: $\ell = \tilde{O}(n^{1/3}/\varepsilon^{4/3})$ samples/ machine

- Use cross collisions - bipartite graph
- **Communication complexity:** $O\left(\frac{\sqrt{n/\ell}\sqrt{\log n}}{\epsilon^2}\right)$ bits.



Case II: $\ell = \tilde{\Omega}(n^{1/3}/\varepsilon^{4/3})$ samples/machine

- Each machine sends that number of **local** collisions and to the referee.
- The referee computes the total sum Z of the collisions.
  - $\mathbb{E}[Z] = \binom{\ell}{2}\|p\|_2^2$
  - Threshold: $(1+\varepsilon^2)\mathbb{E}[Z]$
- **Communication complexity:** $O\left(\frac{n \log n}{\ell^2 \epsilon^4}\right)$ bits.

Memory efficient implementation

In the one-pass streaming model

Model:

One-pass streaming algorithm: The samples arrive in a **stream** and the algorithm can access them **only once**.

Memory constraint: At most m bits for some $m \geq \log n/\varepsilon^6$

- Use $N_1 = m/2 \log n$ samples to get the multiset of labels S_1.
- Use collision information from $N_2 = \Theta\left(n \log n/(m\varepsilon^4)\right)$ other samples (i.e the multiset of labels S_2).

Remarks:

- We can store $\sum_{k=1}^{r} d_k, 1 \leq r \leq N_2$ in a single pass.
- For $m = \Omega(\sqrt{n} \log n/\varepsilon^2)$, we simply run the classical collision-based tester using the first $O(\sqrt{n}/\varepsilon^2)$ samples.

SUMMARY OF RESULTS

	Sample Complexity Bounds with Memory Constraints		
Property	**Upper Bound**	**Lower Bound 1**	**Lower Bound 2**
Uniformity	$O\left(\frac{n \log n}{m\varepsilon^4}\right)$	$\Omega\left(\frac{n \log n}{m\varepsilon^4}\right)$	$\Omega\left(\frac{n}{m\varepsilon^2}\right)$
Conditions	$n^{0.9} \gg m \gg \log(n)/\varepsilon^2$	$m = \tilde{\Omega}(\frac{n^{0.34}}{\varepsilon^{8/3}} + \frac{n^{0.1}}{\varepsilon^4})$	Unconditional
Closeness	$O(n\sqrt{\log(n)}/(\sqrt{m}\varepsilon^2))$	–	–
Conditions	$\tilde{\Theta}(\min(n, n^{2/3}/\varepsilon^{4/3})) \gg m \gg \log(n)$	–	–

	Communication Complexity Bounds				
Property	**UB 1**	**UB 2**	**LB 1**	**LB 2**	**LB 3**
Uniformity	$O\left(\frac{\sqrt{n\log(n)/\ell}}{\varepsilon^2}\right)$	$O\left(\frac{n\log(n)}{\ell^2\varepsilon^4}\right)$	$\Omega\left(\frac{\sqrt{n\log(n)/\ell}}{\varepsilon^2}\right)$	$\Omega(\frac{\sqrt{n/\ell}}{\varepsilon})$	$\Omega(\frac{n}{\ell^2\varepsilon^2\log n})$
Conditions	$\frac{\varepsilon^8 n}{\log n} \gg \ell \gg \frac{\varepsilon^{-4}}{n^{0.9}}$	$\ell \ll \frac{\sqrt{n}}{\varepsilon^2}$	$\varepsilon^{4/3}n^{0.3} \gg \ell$	$\ell = \tilde{O}\left(\frac{n^{1/3}}{\varepsilon^{4/3}}\right)$	$\ell = \tilde{\Omega}\left(\frac{n^{1/3}}{\varepsilon^{4/3}}\right)$
Closeness	$O\left(\frac{n^{2/3}\log^{1/3}(n)}{\ell^{2/3}\varepsilon^{4/3}}\right)$	-	-	-	-
Conditions	$n\varepsilon^4/\log(n) \gg \ell$	-	-	-	-

Lower Bounds (One Pass)

k SAMPLES, m BITS OF MEMORY, ℓ SAMPLES PER MACHINE

1. Memory:
   - $k \cdot m = \Omega(\frac{n}{\varepsilon^2})$
   - Under technical assumptions: $k \cdot m = \Omega(\frac{n \log n}{\varepsilon^4})$

Reduction (low communication $\Rightarrow$ low memory)

- samples/machine: ℓ
- bits of communication: t

Store samples of the **next player only** $\Rightarrow t + \ell \log n$-memory

2. Communication ($\ell = O\left(\frac{n^{1/3}}{\varepsilon^{4/3}(\log n)^{1/3}}\right)$)-one pass:
   - $\Omega\left(\frac{\sqrt{n/\ell}}{\varepsilon}\right)$ samples.
   - Under assumptions: $\Omega\left(\frac{\sqrt{n \log n/\ell}}{\varepsilon^2}\right)$
3. Communication ($\ell = \Omega\left(\frac{n^{1/3}}{\varepsilon^{4/3}(\log n)^{1/3}}\right)$)-one pass:
   - $\Omega\left(\frac{n}{\ell^2 \varepsilon^2 \log n}\right)$ samples.

Summary-Open problems

- We described a bipartite collision-based algorithm for uniformity.
  - Then applied it to memory constrained and distributed settings.
- Showed matching lower bounds for certain parameter regimes.
  - An asymptotically optimal algorithm becomes (provably) suboptimal as ℓ grows.

Open Problems:

- Do the lower bounds still hold if multiple passes are allowed?
- Is there an algorithm with a better communication-sample complexity trade-off?