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?

▶ We can learn the distribution: $\Omega(n)$ samples.
▶ Centralized sampling/ unbounded memory: we can test (uniform vs $\varepsilon$-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 \( \frac{2}{3} \).

- **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 \( \|p - U_n\|_1 \geq \epsilon \).

\[ m \text{ samples} \]
\[ \text{Node labels: i.i.d samples from } p. \]
\[ \text{Edges: } \{i, j\} \in E \text{ iff } L(i) = L(j) \]

Define statistic \( Z = \#\text{edges} \Rightarrow \mathbb{E}[Z] = \left(\frac{m}{2}\right) \cdot \|p\|_2^2 \)

Minimized for \( p = U_n \)

Idea: Draw enough samples and compare \( Z \) to some threshold.
**Problem:** Given distribution $p$ over $[n]$, distinguish $p = U_n$ from $\|p - U_n\|_1 \geq \epsilon$.

- $\ell$ samples per machine.
- **Node labels:** i.i.d samples from $p$.
- **Edges:** $\{i, j\} \in E$ iff $(i \in S_1) \land (j \in S_2) \land (L(i) = L(j))$
GENERAL Bipartite Testing Algorithm

\( \ell \) SAMPLES PER MACHINE

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

\( \ell \) samples per machine.

Node labels: i.i.d samples from \( p \).

Edges: \( \{i, j\} \in E \) iff 
\( (i \in S_1) \land (j \in S_2) \land (L(i) = L(j)) \)

Define statistic \( Z = \text{\#edges} \Rightarrow \mathbb{E}[Z] = |S_1| \cdot |S_2| \cdot \|p\|_2^2 \)

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 $Z = \text{sum of degrees on one side}$.  
  - *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 $\ell$.
  - 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:

$$O \left( \frac{\sqrt{n/\ell} \sqrt{\log n}}{\varepsilon^2} \right) \text{ bits.}$$

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.

▶ $E[Z] = \binom{\ell}{2} \|p\|_2^2$

▶ Threshold: $(1 + \varepsilon^2)E[Z]$

▶ Communication complexity:

$$O \left( \frac{n \log n}{\ell^2 \varepsilon^4} \right) \text{ bits.}$$
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 $N_2 = \Theta(n \log n/(m\varepsilon^4))$ 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

<table>
<thead>
<tr>
<th>Property</th>
<th>Upper Bound</th>
<th>Lower Bound 1</th>
<th>Lower Bound 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniformity</td>
<td>$O\left(\frac{n \log n}{m \varepsilon^4}\right)$</td>
<td>$\Omega\left(\frac{n \log n}{m \varepsilon^4}\right)$</td>
<td>$\Omega\left(\frac{n}{m \varepsilon^2}\right)$</td>
</tr>
<tr>
<td>Conditions</td>
<td>$n^{0.9} \gg m \gg \log(n)/\varepsilon^2$</td>
<td>$m = \tilde{\Omega}\left(\frac{n^{0.34}}{\varepsilon^{8/3}} + \frac{n^{0.1}}{\varepsilon^4}\right)$</td>
<td>Unconditional</td>
</tr>
<tr>
<td>Closeness</td>
<td>$O\left(n \sqrt{\log(n)}/(\sqrt{m} \varepsilon^2)\right)$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Conditions</td>
<td>$\tilde{\Theta}\left(\min(n, n^{2/3}/\varepsilon^{4/3})\right) \gg m \gg \log(n)$</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### Communication Complexity Bounds

<table>
<thead>
<tr>
<th>Property</th>
<th>UB 1</th>
<th>UB 2</th>
<th>LB 1</th>
<th>LB 2</th>
<th>LB 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniformity</td>
<td>$O\left(\frac{\sqrt{n \log(n)}/\ell}{\varepsilon^2}\right)$</td>
<td>$O\left(\frac{n \log(n)}{\ell^2 \varepsilon^4}\right)$</td>
<td>$\Omega\left(\frac{\sqrt{n \log(n)}/\ell}{\varepsilon^2}\right)$</td>
<td>$\Omega\left(\frac{\sqrt{n/\ell}}{\varepsilon}\right)$</td>
<td>$\Omega\left(\frac{n}{\ell^2 \varepsilon^2 \log n}\right)$</td>
</tr>
<tr>
<td>Conditions</td>
<td>$\varepsilon^{8n}/\log n \gg \ell \gg \varepsilon^{-4} n^{0.9}$</td>
<td>$\ell \ll \frac{\sqrt{n}}{\varepsilon^2}$</td>
<td>$\varepsilon^{4/3} n^{0.3} \gg \ell$</td>
<td>$\ell = \tilde{O}\left(\frac{n^{1/3}}{\varepsilon^{4/3}}\right)$</td>
<td>$\ell = \tilde{\Omega}\left(\frac{n^{1/3}}{\varepsilon^{4/3}}\right)$</td>
</tr>
<tr>
<td>Closeness</td>
<td>$O\left(\frac{n^{2/3} \log^{1/3}(n)}{\ell^{2/3} \varepsilon^{4/3}}\right)$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Conditions</td>
<td>$n \varepsilon^3/\log(n) \gg \ell$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
LOWER BOUNDS (ONE PASS)

$k$ samples, $m$ bits of memory, $\ell$ samples per machine

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

Reduction (low communication $\Rightarrow$ low memory)

   - samples/machine: $\ell$
   - 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 $\ell$ 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?