The coincidence based uniformity testing achieves $\ell_2$-tolerance

Sutanu Gayen (University of Nebraska-Lincoln, sutanugayen@gmail.com)
Joint work with A. Pavan (Iowa State University) and N. V. Vinodchandran (University of Nebraska-Lincoln)
Canonical distribution testing

1. Decide whether $D$ satisfies some properties of interest or not
2. Number of samples as small as possible
3. Success probability 51%
Distribution testing for a class $C$

- Let $D$ be an unknown discrete distribution over a known domain $\Omega$
- Let $C$ be a class of discrete distributions over $\Omega$
- Given independent samples from $D$, distinguish between (with success probability 51% in either case):
  - $D$ belongs to the class $C$
    versus
  - $D$ is far from every member of $C$, wrt a distance measure
- Example of $C$: singleton class containing an explicitly known distribution, unimodal, monotone …
- Distance measure (dist): $\ell_1, \ell_2, Hellinger, …$
Canonical tester for class $\mathcal{C}$

- Output ‘yes’ if $D \in \mathcal{C}$
- Output ‘no’ if $\text{dist}(D, \mathcal{C}) > \epsilon$
- $N(|\Omega|, \epsilon)$, the number of samples must be as small as possible
  - Eg. $O(|\Omega|^{1-\gamma}/\text{poly}(\epsilon))$
- Henceforth, $\Omega = [m] = \{1,2,\ldots,m\}$ wlog and $N = \text{the number of samples}$
Uniformity testing

- $C = \{U_m\}$, the singleton class having only the uniform distribution over $[m]$
- Distinguish between
  - $D = U_m$
  - $\text{dist}(D, U_m) > \epsilon$ for some distance function $\text{dist}$
Uniformity testing: variations

<table>
<thead>
<tr>
<th>Variation</th>
<th>‘yes’ class</th>
<th>‘no’ class</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\ell_1$ guarantee</td>
<td>$D = U_m$</td>
<td>$\ell_1(D, U_m) &gt; \epsilon$</td>
</tr>
<tr>
<td>$\ell_2$ guarantee</td>
<td>$D = U_m$</td>
<td>$\ell_2^2(D, U_m) &gt; \epsilon^2/m$</td>
</tr>
<tr>
<td>$\ell_2$ guarantee with $\ell_2$ ‘tolerance’</td>
<td>$\ell_2^2(D, U_m) &lt; \epsilon^2/10m$</td>
<td>$\ell_2^2(D, U_m) &gt; \epsilon^2/m$</td>
</tr>
</tbody>
</table>

- $\ell_2^2(D, U_m) / \ell_1^2(D, U_m) / m$, the testing becomes harder, as we move down.
- Third variation is known to have a $O(\sqrt{m}/\epsilon^2)$ sample tester.
- First variation is known to require $\Omega(\sqrt{m}/\epsilon^2)$ samples.
- $\ell_1$ guarantee with $\ell_1$ tolerance is known to require $\Omega(m / \log m)$ samples for constant $\epsilon$. 


## Uniformity testing: brief history

<table>
<thead>
<tr>
<th>tester, technique</th>
<th>guarantee, tolerance</th>
<th>Sample complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goldreich and Ron [ECCC, 2000]</td>
<td>$\ell_2$, nil</td>
<td>$O(\sqrt{m}/\epsilon^4)$</td>
</tr>
<tr>
<td>Paninski [IEEE Trans. Information Theory, 2008], $\epsilon = \Omega(1/m^{1/4})$</td>
<td>$\ell_1$, nil</td>
<td>$O(\sqrt{m}/\epsilon^2)$</td>
</tr>
<tr>
<td>Valiant and Valiant [FOCS 2014]</td>
<td>$\ell_2$, nil</td>
<td>$\Omega(\sqrt{m}/\epsilon^2)$</td>
</tr>
<tr>
<td>Diakonikolas, Kane and Nikishkin [SODA 2015]</td>
<td>$\ell_2$, nil</td>
<td>$O(\sqrt{m}/\epsilon^2)$</td>
</tr>
<tr>
<td>Acharya, Daskalakis and Kamath [NIPS 2015]</td>
<td>$\ell_1$, $\ell_2$</td>
<td>$O(\sqrt{m}/\epsilon^2)$</td>
</tr>
<tr>
<td>Diakonikolas and Kane [FOCS 2016]</td>
<td>$\ell_2$, $\ell_2$</td>
<td>$O(\sqrt{m}/\epsilon^2)$</td>
</tr>
<tr>
<td>Diakonikolas, Gouleakis, Peebles and Price [ECCC 2016]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diakonikolas, Gouleakis, Peebles and Price [ICALP 2018], success probability $1 - \delta$</td>
<td>$\ell_1$, nil</td>
<td>$\Theta(1/\epsilon^2(\sqrt{m \log 1/\delta} + \log 1/\delta))$</td>
</tr>
</tbody>
</table>
Our result

- For the regime $\epsilon = \Omega(1/m^{1/4})$, the tester of Diakonikolas, Gouleakis, Peebles and Price [ICALP 2018], is able to test in with $\ell_2$ guarantee and achieves $\ell_2$ tolerance.

- This is the optimal $\ell_2$ guarantee and $\ell_2$ tolerant tester with success probability $(1 - \delta)$ for this regime.

- Our analysis is elementary and simpler than the analysis of the known testers with the same guarantee.

- For simplicity, henceforth we focus on 51% success probability.
  - Improved guarantee from Paninski 2008.
Our tester

- Just counts the number of distinct items appearing in the sample set
- Intuitively, uniform distribution should produce more distinct items
Algorithm (for success probability 51%)

- **Input:** $\epsilon > \Omega(m^{-1/4})$, $S \leftarrow N = O(\sqrt{m}/\epsilon^2)$ iid samples from distribution $D$
- **Decide:** $||D - U||_2^2 \leq \epsilon^2/10m$ or $||D - U||_2^2 \geq \epsilon^2/m$

1. $Z(S) \leftarrow$ number of distinct items in $S$
2. $T = m(1 - (1 - 1/m)^N) - 0.22\left(\frac{N}{2}\right)\epsilon^2/m$
3. If $Z(S) > T$ Output YES
4. Else Output NO
Analysis
High Level Steps

1. A lower bound for the gap in the expected number of distinct items between the cases: 1) $D$ is close to uniform, and 2) $D$ is far from uniform

2. Upper bounds for the variances in both cases.

3. By Chebyshev’s inequality, with 51% probability, the two cases are separated by a threshold.
Lemma: Let $D = \{p_i\}_{i=1}^m$ be a probability distribution over $m$ items. Let $U$ be the uniform distribution over these items. Let $Z(S)$ denotes the number of distinct items in set $S$. Let $G = E_{S \sim U^N}[Z(S)] - E_{S \sim D^N}[Z(S)]$. Then for any $N \leq m$,

$$1.5 \binom{N}{2} \|D - U\|_2^2 \geq G \geq 0.36 \binom{N}{2} \|D - U\|_2^2$$
Analysis
Gap in expectation

$X_i = 1$ if item $i$ appears in these $N$ samples
$= 0$ otherwise

$E[X_i] = \Pr[X_i = 1] = 1 - \Pr[X_i = 0] = 1 - (1 - p_i)^N$

$E_{S \sim D^N}[Z(S)] = E[\sum_{i=1}^{m} X_i] = \sum_{i=1}^{m} E[X_i] = \sum_{i=1}^{m} [1 - (1 - p_i)^N]$

$E_{S \sim U^N}[Z(S)] = \sum_{i=1}^{m} [1 - (1 - 1/m)^N]$
Analysis
Gap in expectation

\[ G = \sum_{i=1}^{m} [(1 - p_i)^N - (1 - 1/m)^N] \]
\[ = \sum_{i=1}^{m} [(1 - 1/m + 1/m - p_i)^N - (1 - 1/m)^N] \]
\[ = \sum_{k=2}^{N} \binom{N}{k} (1 - 1/m)^{N-k} \left( \sum_{i=1}^{m} (1/m - p_i)^k \right) \]

\[ \geq 0 \text{ for any } k, \]  
\[ \text{from Jensen’s inequality} \]
\[ \geq \binom{N}{2} (1 - 1/m)^{N-2} \|D - U\|_2^2 \]
Analysis
Gap in expectation

\[ G = \sum_{i=1}^{m} [(1 - p_i)^N - (1 - 1/m)^N] \]
\[ = \sum_{i=1}^{N} [(1 - 1/m + 1/m - p_i)^N - (1 - 1/m)^N] \]
\[ = \sum_{k=2}^{N} \binom{N}{k} (1 - 1/m)^{N-k} \left( \sum_{i=1}^{m} (1/m - p_i)^k \right) \]

\[ \leq (1/m^{k-2}) \sum_{i=1}^{m} (1/m - p_i)^2 \text{ for any } k \]

\[ \leq \binom{N}{2} \sum_{k=2}^{N} (N/3m)^{k-2} \|D - U\|_2^2 \leq 1.5 \binom{N}{2} \|D - U\|_2^2 \text{ for } N \leq m \]
Analysis
Gap in expectation

- $1.5 \left( \frac{N}{2} \right) \|D - U\|_2^2 \geq G \geq 0.36 \left( \frac{N}{2} \right) \|D - U\|_2^2$

- For ‘yes’ class, $\|D - U\|_2^2 \leq \varepsilon^2 / 10m \Rightarrow G \leq 0.15 \left( \frac{N}{2} \right) \varepsilon^2 / m$
  - Hence, $E_{S-DN}[Z(S)] \geq E_{S-UN}[Z(S)] - 0.15 \left( \frac{N}{2} \right) \varepsilon^2 / m$

- For ‘no’ class, $\|D - U\|_2^2 \geq \varepsilon^2 / m \Rightarrow G \geq 0.36 \left( \frac{N}{2} \right) \varepsilon^2 / m$
  - Hence, $E_{S-DN}[Z(S)] \leq E_{S-UN}[Z(S)] - 0.36 \left( \frac{N}{2} \right) \varepsilon^2 / m$

- $\Omega(N^2 \varepsilon^2 / m)$ gap between the two cases
Analysis
Variance upper bound

- Lemma: \( \text{Var}_{S \sim \mathcal{D}^N} Z(S) \leq N(N - 1) \|D - U\|_2^2 + N^2 / m \), for any integer \( N \) and any distribution \( D \). \( U \) is the uniform distribution.

- For \( N = \Theta(\sqrt{m}/\epsilon^2) \) samples, noise is \( O(N^2 \epsilon^2 / m) \), with probability 51% by Chebyshev’s inequality.

- Since gap in the expectations is \( \Omega(N^2 \epsilon^2 / m) \), the following threshold works:
  - \( T = m(1 - (1 - 1/m)^N) - 0.22 \binom{N}{2} \epsilon^2 / m \)
Remarks

- High probability bounds using Diakonikolas, Gouleakis, Peebles and Price [ICALP 2018] instead of application of Chebyshev’s inequality
- Efficient algorithms for distinct elements are known across various models
- Future directions: $\Theta(N^2e^2/m)$ additive error for distinct elements is hard for small space streaming model. Algorithm/Lower bound? Crouch, McGregor, Valiant, Woodruff [ESA 2016]
THANK YOU

sutanugayen@gmail.com, CSA Lab 301
A poster with complete proofs is available at https://cse.unl.edu/~sgayen/