Optimal Distribution Testing via Reductions

Ilias Diakonikolas
USC

Joint work with
Daniel Kane (UCSD)
Distribution Testing

Given samples from one or more unknown probability distributions, decide whether they satisfy a certain property.

• Introduced by Karl Pearson (1899).

• Classical Problem in Statistics
  [Neyman-Pearson’33, Lehman-Romano’05]

• Last fifteen years (TCS): property testing
  [Goldreich-Ron’00, Batu et al. FOCS’00/JACM’13]
Notation

Basic object of study:
Probability distributions over finite domain.

\[ [n] \quad \text{or} \quad [n]^d \]

Notation:

\( p, q \): probability mass function
Example: Testing Closeness

- Let $\mathcal{D}$ be a family of probability distributions

**Testing Closeness Problem:**
- Distinguish between the cases $p=q$ and $\text{dist} (p, q) > \varepsilon$
- Minimize sample size, computation time

### Total Variation Distance

$$d_{TV}(p, q) = (1/2) \| p - q \|_1$$
This Work

Simple Framework for Distribution Testing:
Leads to *sample-optimal and computationally efficient*
estimators
for a *variety of properties*

Primarily based on:

**A New Approach for Testing Properties of Discrete Distributions**
(I. Diakonikolas and D. Kane, FOCS’16)
Outline

- Related and Prior Work
- Framework Overview and Statement of Results
- Case Study: Testing Identity, Closeness, and Independence
- Future Directions and Concluding Remarks
Outline

- Related and Prior Work
- Framework Overview and Statement of Results
- Case Study: Testing Identity, Closeness, and Independence
- Future Directions and Concluding Remarks
Prior Work: Identity Testing

Focus has been on arbitrary distributions over support of size $n$.

**Testing Identity to a *known* Distribution:**

- [Goldreich-Ron’00]: $O(\sqrt{n}/\epsilon^4)$ upper bound for *uniformity testing* (collision statistics)

- [Batu et al., FOCS’01]: $\tilde{O}(\sqrt{n}) \cdot \text{poly}(1/\epsilon)$ upper bound for testing identity to any *known* distribution.

- [Paninski ’03]: upper bound of $O(\sqrt{n}/\epsilon^2)$ for uniformity testing, assuming $\epsilon = \Omega(n^{-1/4})$. Lower bound of $\Omega(\sqrt{n}/\epsilon^2)$.

- [Valiant-Valiant, FOCS’14, D-Kane-Nikishkin, SODA’15]: upper bound of $O(\sqrt{n}/\epsilon^2)$ for identity testing to any known distribution.

- [D-Gouleakis-Peebles-Price’16]: [GR’00] tester is optimal!
Prior Work: Closeness Testing

Focus has been on arbitrary distributions over support of size $n$.

**Testing Closeness between two unknown distributions:**

- [Batu et al., FOCS’00]: $O(n^{2/3} \log n/\epsilon^{8/3})$ upper bound for testing closeness between two unknown discrete distributions.

- [P. Valiant, STOC’08]: lower bound of $\Omega(n^{2/3})$ for constant error.

- [Chan-D-Valiant-Valiant, SODA’14]: tight upper and lower bound of $O(\max\{n^{2/3}/\epsilon^{4/3}, n^{1/2}/\epsilon^2\})$

- [Bhatacharya-Valiant, NIPS’15]: tight bounds for different sample sizes (assuming $\epsilon > n^{-1/12}$).
Prior Work: Testing Independence

Focus has been on arbitrary distributions over support of size $n$.

**Testing Independence of a distribution on** $[n] \times [m]$.

- [Batu et al., FOCS’01]: $\tilde{O}(n^{2/3}m^{1/3} \cdot \text{poly}(1/\epsilon))$ upper bound.
- [Levi-Ron-Rubinfeld, ICS’11]: lower bounds for constant error $\Omega(m^{1/2}n^{1/2})$ and $\Omega(n^{2/3}m^{1/3})$, for $n = \Omega(m \log m)$
- [Acharya-Daskalakis-Kamath, NIPS’15]: upper bound of $O(n/\epsilon^2)$ for $n=m$. 
Outline

- Related and Prior Work
- Framework Overview and Statement of Results
- Case Study: Testing Identity, Closeness, and Independence
- Future Directions and Concluding Remarks
L2 Closeness Testing

**Lemma 1:** Let $p, q$ be unknown distributions on a domain of size $n$. There is an algorithm that uses $O(\min\{\|p\|_2, \|q\|_2\} n / \epsilon^2)$ samples from each of $p, q$, and with probability at least $2/3$ distinguishes between the cases that $p = q$ and $\|p - q\|_1 \geq \epsilon$.

**Basic Tester** [Chan-D-Valiant-Valiant’14]:

- Calculate $Z = \sum_i \{(X_i - Y_i)^2 - X_i - Y_i\}$
- If $Z > \epsilon^2 m^2$ then output “No” (different), otherwise, output “Yes” (same)

Collision-based estimator also works [D-Gouleakis-Peebles-Price’16]
Main New Idea

Solve *all* problems by reducing to this as a black-box.
Framework and Results

• **Approach**: Reduction of L1 Testing to L2 testing

1) Transform given distribution(s) to new distribution(s) (over potentially larger domain) with small L2 norm.

2) Use standard L2 tester as a black-box.

• Circumvents method of explicitly learning heavy elements [Batu et al., FOCS’00]
Algorithmic Applications

Sample Optimal Testers for:

- Identity to a Fixed Distribution
- Closeness between two Unknown Distributions
- (Nearly) Instance-optimal Identity Testing
- Closeness with unequal sample size
- Adaptive Closeness Testing
- Independence (in any dimension)
- Properties of Collections of Distributions (Sample & Query model)
- Testing Histograms
- Other Metrics (chi-squared, Hellinger)

All algorithms follow same pattern. Very simple analysis.
Outline

- Related and Prior Work
- Framework Overview and Statement of Results
  - Case Study: Testing Identity, Closeness, and Independence
- Future Directions and Concluding Remarks
Warm-up: Testing Identity to Fixed Distribution (I)

Let $p$ be unknown distribution and $q$ known distribution on $[n]$.

**Main Idea:** “Stretch” the domain size to make $L_2$ norm of $q$ small.

- For every bin $i \in [n]$ create set $S_i$ of $[nq_i]$ new bins.
- Subdivide the probability mass of bin $i$ equally within $S_i$.

Let $S$ be the new domain and $p', q'$ the resulting distributions over $S$.
Warm-up: Testing Identity to Fixed Distribution (II)

Let $p$ be unknown distribution and $q$ known distribution on $[n]$.

L1 Identity Tester

- Given $q$, construct new domain $S$.
- Use basic tester to distinguish between $p' = q'$ and $\|p' - q'\|_1 \geq \epsilon$.

We construct $q'$ explicitly. Can sample from $p'$ given sample from $p$.

Analysis:

Observation 1: $\|p' - q'\|_1 = \|p - q\|_1$

Observation 2: $|S| \leq 2n$ and $\|q'\|_2 = O(1/\sqrt{n})$

By Lemma 1, we can test identity between $p'$ and $q'$ with sample size

$$O(\|q'\|_2 |S| / \epsilon^2) = O(\sqrt{n}/\epsilon^2)$$
Identity Reduces to Uniformity

• Summary of Previous Slides:

Identity reduces to its special case when the explicit distribution has max probability $O(1/n)$.

• Recent Improvement:

[Oded Goldreich’16]:

Identity Reduces to Uniformity.
Testing Closeness (I)

Let $p, q$ be unknown distributions on $[n]$.

**Main Idea:** Use samples from $q$ to “stretch” the domain size.

- Draw a set $S$ of $\text{Poi}(k)$ samples from $q$.
- Let $a_i$ be the number of times we see $i \in [n]$ in $S$.
- Subdivide the mass of bin $i$ equally within $a_i + 1$ new bins.

Let $S'$ be the new domain and $p', q'$ the resulting distributions over $S'$.

We can sample from $p', q'$.

**Observation:** $\|p' - q'\|_1 = \|p - q\|_1$
Testing Closeness (II)

Let \( p, q \) be unknown distributions on \([n]\).

**L1 Closeness Tester**
- Draw a set \( S' \) of \( \text{Poi}(k) \) samples from \( q \), construct new domain \( S'' \).
- Use basic tester to distinguish between \( p' = q' \) and \( \|p' - q'\|_1 \geq \epsilon \).

**Claim:** Whp \( |S'| \leq n + O(k) \) and \( \|q'\|_2 = O(1/\sqrt{k}) \).

**Proof:**
\[
\|p'\|_2^2 = \sum_{i=1}^{n} p_i^2/(1 + a_i), \quad \mathbb{E}[1/(1 + a_i)] \leq 1/(kp_i).
\]

By Lemma 1, we can test identity between \( p' \) and \( q' \) with sample size
\[
O(\|q'\|_2 |S''|/\epsilon^2) = O(k^{-1/2} \cdot (n + k)/\epsilon^2).
\]

Total sample size
\[
O(k + k^{-1/2} \cdot (n + k)/\epsilon^2).
\]

Set \( k := \min\{n, n^{2/3} \epsilon^{-4/3}\} \).
Closeness with Unequal Samples

Let $p, q$ be unknown distributions on $[n]$.
Have $m_1 + m_2$ samples from $q$ and $m_2$ samples from $p$.

L1 Closeness Tester Unequal

- Set $k := \min\{n, m_1\}$.
- Draw $\text{Poi}(k)$ samples from $q$, construct new domain $S'$.
- Use basic tester to distinguish between $p' = q'$ and $\|p' - q'\|_1 \geq \epsilon$.

Claim: Whp $|S'| \leq n + O(k)$ and $\|q'\|_2 = O(1/\sqrt{k})$.

By Lemma 1, we can test identity between $p'$ and $q'$ with sample size

$$m_2 = O(\|q'\|_2 |S'|/\epsilon^2) = O(k^{-1/2} \cdot (n + k)/\epsilon^2).$$

By our choice of $k$, it follows

$$m_2 = O(\max\{nm_1^{-1/2} \epsilon^2, n^{1/2} / \epsilon^2\}).$$
Testing Independence in 2-d

Let $p$ be unknown distribution on $[n] \times [m]$. 
Let $q = p_1 \times p_2$.

**L1 Independence Tester**

- Set $k := \min\{n, n^{2/3}m^{1/3}\epsilon^{-4/3}\}$.
- Draw a set $S_1$ of $\text{Poi}(k)$ samples from $p_1$,
  and $S_2$ of $\text{Poi}(m)$ samples from $p_2$.
- Stretch domain in each dimension to obtain new support.
- Use basic tester to distinguish between $p' = q'$ and $\|p' - q'\|_1 \geq \epsilon$.

By Lemma 1, we can test identity between $p'$ and $q'$ with sample size

$$O(\|q'\|_2|S'|/\epsilon^2) = O(k^{-1/2}m^{-1/2} \cdot mn/\epsilon^2)$$

$$= O(\max\{n^{2/3}m^{1/3}\epsilon^{-4/3}, (mn)^{1/2}/\epsilon^2\})$$
Outline

- Introduction, Related and Prior Work
- Framework Overview and Statement of Results
- Case Study: Testing Identity, Closeness, and Independence
- Future Directions and Concluding Remarks
Future Directions (I)

**This Talk:** Unified Technique for Testing *Unstructured* Discrete Distributions.

Gives sample-optimal estimators for many properties in the literature.

**Game Over?**

- Recent line of work on Testing *Structured* Distributions
  [D-Kane-Nikishkin, SODA’15 / FOCS’15 / ICALP’16]

- Dependence on error probability? [D-Gouleakis-Peebles-Price’17]
  E.g., identity testing
  \[ O(\sqrt{n \log(1/\delta)/\epsilon^2 + \log(1/\delta)/\epsilon^2}) \]

- Optimal Constants? Practically relevant question; requires new insights.
  [Huang-Meyn IEEE ToIT’14]
Future Directions (II)

This Talk: Unified Technique for Testing *Unstructured* Discrete Distributions.

Future Directions:

- High-Dimensional *Structured* Distributions
  [Canonne-D-Kane-Stewart’16, Daskalakis-Pan’16, Daskalakis-Dikkala-Kamath’16, D-Kane-Stewart’17]

- Other criteria (privacy, communication, etc.)
  [Cai-Daskalakis-Kamath’17, Aliakbarpour-D-Rubinfeld’17, Acharya-Sun-Zhang’17, D-Grigorescu-Onak-Natarajan’16]

- Beyond Worst-Case Analysis

Thank you for your attention!