Testing Conditional Independence of Discrete Distributions

January 12, 2018

Clément Canonne (Stanford), Ilias Diakonikolas (USC), Daniel Kane (UCSD), and Alistair Stewart (USC)

An experiment

X,Y Boolean variables, **Z** in {1,2,…,195}

- **X**: "prefers pizza over deep dish pizza"
- **Y**: "prefers the Bulls over the Knicks"
- **Z**: City in the US

Clearly, X and Y are **not** independent.

But are they *after controlling for Z*?

Hypothesis Testing: Independence

Given realizations of (X,Y), are X and Y statistically independent?

- Fundamental question in **Statistics**: Pearson's chi-squared test, Fischer's exact test, G test, ...
- Studied in **Computer Science** (discrete data): [BFF+01, LRR11, ADK16, DK16]
- Falls **short** of what we want: conditional independence

Hypothesis Testing: Conditional Independence

Given realizations of (X,Y,Z), are X and Y statistically independent conditioned on Z?

- Strict generalization of the previous question
- Crucial in practice: Machine Learning (graphical models), natural sciences (controlling for a confounding factor), …
- Some tests used, e.g. Cochran–Mantel–Haenszel

Yet **no provable guarantee**

Hypothesis Testing: Conditional Independence

Given realizations of (X,Y,Z), are X and Y statistically independent conditioned on Z?

Can we get sample-efficient (fast), information-theoretically optimal algorithms to test conditional independence?

Formalization: distribution testing

Given realizations of (X,Y,Z), are X and Y statistically independent conditioned on Z?

Independent samples from (X,Y,Z) over domain $[\ell 1]\times[\ell 2]\times[n]$, distance parameter ε

- **Accept** if X and Y are independent conditioned on Z
- **Reject** if (X,Y,Z) is statistically far from every conditionally independent (X',Y',Z'):

$$TV(\ (X,Y,Z),\ (X',Y',Z')\ ) > \varepsilon$$

(where TV is the total variation distance between distributions)

Formalization: distribution testing

Our results (I)

Binary case: X,Y in {0,1} (Z can take n values)

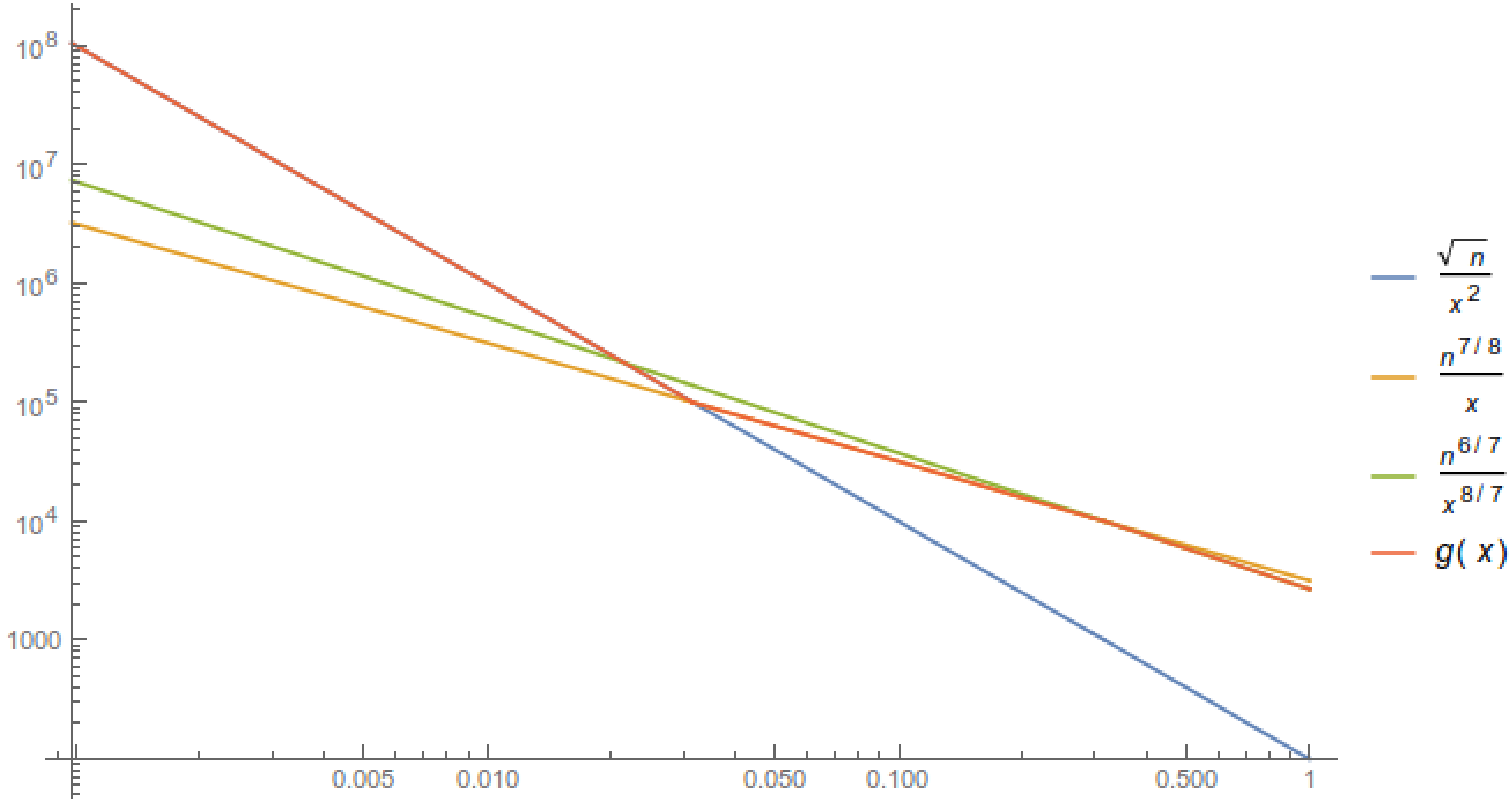
We give a **computationally efficient** tester with sample complexity

$$O\left(\max\left(n^{1/2}/\varepsilon^2, \min\left(n^{7/8}/\varepsilon, n^{6/7}/\varepsilon^{8/7}\right)\right)\right)$$

(which is **optimal**).

No o(n) sample tester previously known

The sample complexity: 3 regimes



The algorithm

Draw a multi-set S of samples

For all value z of Z with ≥ 4 samples

Use these samples to get an estimate $A(z)$ of the squared L2 distance from the conditional distribution of $(X,Y,Z=z)$ to the nearest independent distribution

If $\Sigma A(z) \leq \alpha$

Accept

Else

Reject

The lower bound

Information-theoretic (follows methodology of [DK16])

Formalizes the intuition: *unless at least 4 samples are observed in a given "bin" z, no information from these samples.*

Idea: construct two distributions (X,Y,Z) and (X',Y',Z') with matching first 3 moments *(yet resp. conditionally independent and far from it).*

Our results (II)

General case: X,Y in [ℓ1]×[ℓ2] (Z can take n values)

We give a **computationally efficient** tester with sample complexity

$$O\left(\max\left(\min\left(\frac{n^{7/8}\ell_1^{1/4}\ell_2^{1/4}}{\varepsilon}, \frac{n^{6/7}\ell_1^{2/7}\ell_2^{2/7}}{\varepsilon^{8/7}}\right), \frac{n^{3/4}\ell_1^{1/2}\ell_2^{1/2}}{\varepsilon}, \frac{n^{2/3}\ell_1^{2/3}\ell_2^{1/3}}{\varepsilon^{4/3}}, \frac{n^{1/2}\ell_1^{1/2}\ell_2^{1/2}}{\varepsilon^2}\right)\right)$$

(which we believe to be **optimal***).

* and we show is in several regimes of parameters.

The algorithm

The same original idea, but with crucial *(and somewhat painful)* additions to avoid paying polynomial dependencies on $\ell 1$, $\ell 2$.

In the process: establish new bounds on the variance of estimators of polynomials Q(p) in the probabilities p=(p1,..,pn)

Summary

- First sublinear algorithms for testing conditional independence of discrete (X,Y,Z)
- Information-theoretically optimal and computationally efficient
- Generalizes to other distance measures (conditional mutual information)

Future work: implement our algorithms (Python/Julia) and assess their performance in practice.


Thank
You.