ALICE AND BOB SHOW DISTRIBUTION TESTING LOWER BOUNDS

They don’t talk to each other anymore.

Clément Canonne (Columbia University)

December 8, 2016

Joint work with **Eric Blais** (UWaterloo) and **Tom Gur** (Weizmann Institute)

“DISTRIBUTION TESTING?”

Property testing of probability distributions:

WHY?

Property testing of probability distributions: sublinear, Property testing of probability distributions: sublinear, approximate,

WHY?

Property testing of probability distributions: sublinear, approximate, randomized

WHY?

Property testing of probability distributions: sublinear, approximate, randomized algorithms that take random samples

WHY?

Property testing of probability distributions: sublinear, approximate, randomized algorithms that take random samples

- Big Dataset: too big

WHY?

Property testing of probability distributions: sublinear, approximate, randomized algorithms that take random samples

- Big Dataset: too big
- Expensive access: pricey data

WHY?

Property testing of probability distributions: sublinear, approximate, randomized algorithms that take random samples

- Big Dataset: too big
- Expensive access: pricey data
- “Model selection”: many options

WHY?

Property testing of probability distributions: sublinear, approximate, randomized algorithms that take random samples

- Big Dataset: too big
- Expensive access: pricey data
- “Model selection”: many options

Need to infer information – one bit – from the data: fast, or with very few samples.


12
22
5
18
26
49
30

(Property) Distribution Testing:

(Property) Distribution Testing:

(Property) Distribution Testing:

in an (egg)shell.

HOW?

Known domain (here $[n] = \{1, \ldots, n\}$)

Property $\mathcal{P} \subseteq \Delta([n])$

Independent samples from unknown $p \in \Delta([n])$

Distance parameter $\varepsilon \in (0, 1]$

HOW?

Known domain (here $[n] = \{1, \dots, n\}$)

Property $\mathcal{P} \subseteq \Delta([n])$

Independent samples from unknown $p \in \Delta([n])$

Distance parameter $\varepsilon \in (0, 1]$

Must decide:

$$p \in \mathcal{P}$$

Known domain (here $[n] = \{1, \ldots, n\}$)

Property $\mathcal{P} \subseteq \Delta([n])$

Independent samples from unknown $p \in \Delta([n])$

Distance parameter $\varepsilon \in (0, 1]$

Must decide:

$$p \in \mathcal{P}, \text{ or } \ell_1(p, \mathcal{P}) > \varepsilon?$$

Known domain (here $[n] = \{1, \ldots, n\}$)

Property $\mathcal{P} \subseteq \Delta([n])$

Independent samples from unknown $p \in \Delta([n])$

Distance parameter $\varepsilon \in (0, 1]$

Must decide:

$$p \in \mathcal{P}, \text{ or } \ell_1(p, \mathcal{P}) > \varepsilon?$$

(and be correct on any p with probability at least 2/3)

AND?

Many results on many properties:

Many results on many properties:

- Uniformity [GR00, BFR$^+$00, Pan08]

AND?

Many results on many properties:

- Uniformity [GR00, BFR$^+$00, Pan08]
- Identity* [BFF$^+$01, VV14]

AND?

Many results on many properties:

- Uniformity [GR00, BFR$^+$00, Pan08]
- Identity* [BFF$^+$01, VV14]
- Equivalence [BFR$^+$00, Val11, CDVV14]

AND?

Many results on many properties:

- Uniformity [GR00, BFR+00, Pan08]
- Identity* [BFF+01, VV14]
- Equivalence [BFR+00, Val11, CDVV14]
- Independence [BFF+01, LRR13]

AND?

Many results on many properties:

- Uniformity [GR00, BFR$^+$00, Pan08]
- Identity* [BFF$^+$01, VV14]
- Equivalence [BFR$^+$00, Val11, CDVV14]
- Independence [BFF$^+$01, LRR13]
- Monotonicity [BKR04]

AND?

Many results on many properties:

- Uniformity [GR00, BFR$^+$00, Pan08]
- Identity* [BFF$^+$01, VV14]
- Equivalence [BFR$^+$00, Val11, CDVV14]
- Independence [BFF$^+$01, LRR13]
- Monotonicity [BKR04]
- Poisson Binomial Distributions [AD14]

AND?

Many results on many properties:

- Uniformity [GR00, BFR$^+$00, Pan08]
- Identity* [BFF$^+$01, VV14]
- Equivalence [BFR$^+$00, Val11, CDVV14]
- Independence [BFF$^+$01, LRR13]
- Monotonicity [BKR04]
- Poisson Binomial Distributions [AD14]
- Generic approaches for classes [CDGR15, ADK15]

AND?

Many results on many properties:

- Uniformity [GR00, BFR+00, Pan08]
- Identity* [BFF+01, VV14]
- Equivalence [BFR+00, Val11, CDVV14]
- Independence [BFF+01, LRR13]
- Monotonicity [BKR04]
- Poisson Binomial Distributions [AD14]
- Generic approachs for classes [CDGR15, ADK15]
- and more...

BUT?

Lower bounds...

... are quite tricky.

BUT?

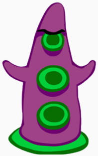
Lower bounds...

... are quite tricky. We want more methods. Generic if possible, applying to many problems at once.

BUT?

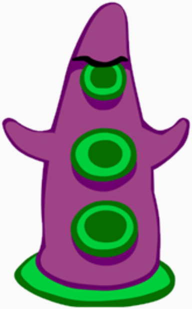
Lower bounds...

... are quite tricky. We want more methods. Generic if possible, applying to many problems at once.



Lower bounds...

... are quite tricky. We want more methods. Generic if possible, applying to many problems at once.



BUT?

Lower bounds...

... are quite tricky. We want more methods. Generic if possible, applying to many problems at once.

“COMMUNICATION COMPLEXITY?”

WHAT NOW?

$$f(x, y)$$

WHAT NOW?

$$f(x, y)$$

WHAT NOW?

$$f(x, y)$$

$$f(x, y)$$

$$f(x, y)$$

But communicating is hard.

- f known by all parties
- Alice gets x, Bob gets y
- Private randomness

Goal: minimize communication (worst case over x, y, randomness) to compute $f(x, y)$.

ALSO...

...in our setting, Alice and Bob do not get to communicate.

...in our setting, Alice and Bob do not get to communicate.

- f known by all parties
- Alice gets x, Bob gets y
- Both send one-way messages to a referee
- Private randomness

ALSO...

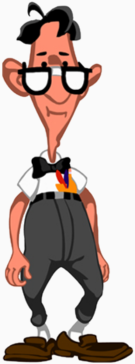
...in our setting, Alice and Bob do not get to communicate.

- f known by all parties
- Alice gets x, Bob gets y
- Both send one-way messages to a referee
- Private randomness

SMP

Simultaneous Message Passing model.

REFEREE MODEL (SMP).



Upshot

$$\mathrm{SMP}(\mathrm{EQ}_n) = \Omega(\sqrt{n})$$

(Only $O(\log n)$ with one-way communication!)

WELL, SURE, BUT WHY?

- Introduced by Blais, Brody, and Matulef [BBM12] for Boolean functions
- Elegant reductions, generic framework
- Carry over very strong communication complexity lower bounds

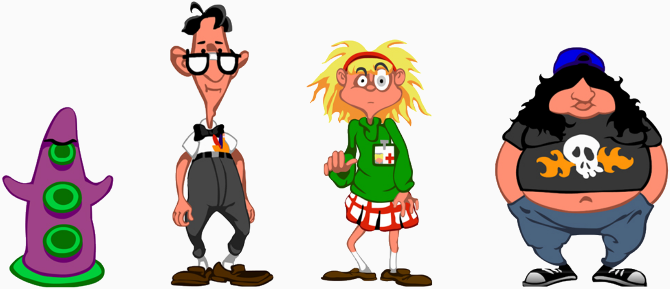
- Introduced by Blais, Brody, and Matulef [BBM12] for Boolean functions
- Elegant reductions, generic framework
- Carry over very strong communication complexity lower bounds

Can we...

... have the same for distribution testing?

DISTRIBUTION TESTING VIA COMM. COMPL.

THE TITLE SHOULD MAKE SENSE NOW.



1. The general methodology.
2. Application: testing uniformity, and the struggle for EQUALITY
3. Testing identity, an unexpected connection
   - The [VV14] result and the 2/3-pseudonorm
   - Our reduction, p-weighted codes, and the K-functional
   - Wait, what is this thing?
4. Conclusion

Theorem

Let $\varepsilon > 0$, and let Ω be a domain of size n. Fix a property $\Pi \subseteq \Delta(\Omega)$ and $f\colon \{0,1\}^k \times \{0,1\}^k \to \{0,1\}$. Suppose there exists a mapping $p\colon \{0,1\}^k \times \{0,1\}^k \to \Delta(\Omega)$ that satisfies the following conditions.

1. Decomposability: $\forall x, y \in \{0,1\}^k$, there exist $\alpha = \alpha(x), \beta = \beta(y) \in [0,1]$ and $p_A(x), p_B(y) \in \Delta(\Omega)$ such that
$$p(x,y) = \frac{\alpha}{\alpha+\beta} \cdot p_A(x) + \frac{\beta}{\alpha+\beta} \cdot p_B(y)$$
and α, β can each be encoded with $O(\log n)$ bits.
2. Completeness: For every $(x,y) = f^{-1}(1)$, it holds that $p(x,y) \in \Pi$.
3. Soundness: For every $(x,y) = f^{-1}(0)$, it holds that $p(x,y)$ is ε-far from Π in ℓ_1 distance.

Then, every ε-tester for Π needs $\Omega\left(\frac{\mathrm{SMP}(f)}{\log(n)}\right)$ samples.

Take the "equality" predicate $\textsc{Eq}_k$ as f:

Theorem (Newman and Szegedy [NS96])

For every $k \in \mathbb{N}$, $\mathrm{SMP}(\textsc{Eq}_k) = \Omega\left(\sqrt{k}\right)$.

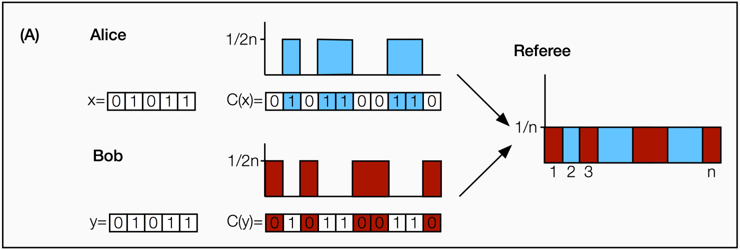
Take the “equality” predicate EQ_k as f:

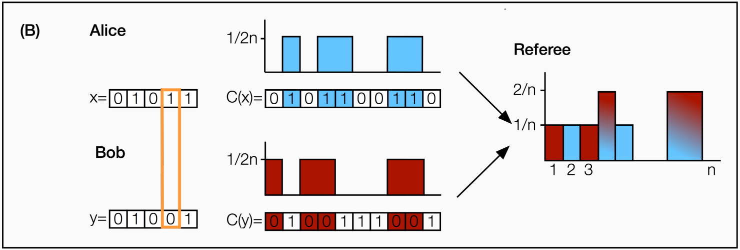
Theorem (Newman and Szegedy [NS96])

For every $k \in \mathbb{N}$, $\mathrm{SMP}(\mathrm{EQ}_k) = \Omega\left(\sqrt{k}\right)$.

Goal:

Will (re)prove an $\tilde{\Omega}(\sqrt{n})$ lower bound on testing uniformity.


(A)
Alice
1/2n
x= 0 1 0 1 1
C(x)= 0 1 0 1 1 0 0 1 1 0
Bob
1/2n
y= 0 1 0 1 1
C(y)= 0 1 0 1 1 0 0 1 1 0
Referee
1/n
1 2 3 n


(B)
Alice
1/2n
x= 0 1 0 1 1
C(x)= 0 1 0 1 1 0 0 1 1 0
Bob
1/2n
y= 0 1 0 0 1
C(y)= 0 1 0 0 1 1 1 0 0 1
Referee
2/n
1/n
1 2 3 n

Statement

Explicit description of $p \in \Delta([n])$, parameter $\varepsilon \in (0, 1]$. Given samples from (unknown) $q \in \Delta([n])$, decide

$$p = q \qquad \text{vs.} \qquad \ell_1(p, q) > \varepsilon$$

Statement

Explicit description of $p \in \Delta([n])$, parameter $\varepsilon \in (0,1]$. Given samples from (unknown) $q \in \Delta([n])$, decide

$$p = q \qquad \text{vs.} \qquad \ell_1(p,q) > \varepsilon$$

Theorem

Identity testing requires $\Omega\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ samples (and we just proved $\tilde{\Omega}(\sqrt{n})$).

Statement

Explicit description of $p \in \Delta([n])$, parameter $\varepsilon \in (0, 1]$. Given samples from (unknown) $q \in \Delta([n])$, decide

$$p = q \qquad \text{vs.} \qquad \ell_1(p, q) > \varepsilon$$

Theorem

Identity testing requires $\Omega\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ samples (and we just proved $\tilde{\Omega}(\sqrt{n})$).

Actually...

Theorem ([VV14])

Identity testing requires $\Omega\left(\frac{\|p_{-\varepsilon}^{-\max}\|_{2/3}}{\varepsilon^2}\right)$ samples (and this is "tight").

An issue: how to interpret this $\|p^{-\max}_{-\varepsilon}\|_{2/3}$?

An issue: how to interpret this $\|p^{-\max}_{-\varepsilon}\|_{2/3}$?

Goal:

Will prove an(other) $\tilde{\Omega}(\Phi(p, \varepsilon))$ lower bound on testing identity, via communication complexity.

An issue: how to interpret this $\|p^{-\max}_{-\varepsilon}\|_{2/3}$?

Goal:

Will prove an(other) $\tilde{\Omega}(\Phi(p,\varepsilon))$ lower bound on testing identity, via communication complexity.

(and it will be "tight" as well.)

- p-weighted codes

$$\mathrm{dist}_p(x,y) := \sum_{i=1}^{n} p(i)\cdot |x_i - y_i| \qquad\qquad (x,y\in\{0,1\}^n)$$

  A p-weighted code has distance guarantee w.r.t. this p-distance: $\mathrm{dist}_p(C(x),C(y)) > \gamma$ for all distinct $x,y\in\{0,1\}^k$.

- Volume of the p-ball:

$$\mathrm{Vol}_{\mathbb{F}_2^n,\mathrm{dist}_p}(\varepsilon) := \left|\{\, w\in\mathbb{F}_2^n : \ \mathrm{dist}_p(w,0^n)\le\varepsilon \,\}\right|.$$

Lemma (Balanced p-weighted exist)

Fix $p \in \Delta([n])$ and ε. There exists a p-weighted (nearly) balanced code $C\colon \{0,1\}^k \to \{0,1\}^n$ with relative distance ε such that $k = \Omega(n - \log \mathrm{Vol}_{\mathbb{F}_2^n, \mathrm{dist}_p}(\varepsilon))$.

Lemma (Balanced p-weighted exist)

Fix $p \in \Delta([n])$ and ε. There exists a p-weighted (nearly) balanced code $C\colon \{0,1\}^k \to \{0,1\}^n$ with relative distance ε such that $k = \Omega(n - \log \mathrm{Vol}_{\mathbb{F}_2^n, \mathrm{dist}_p}(\varepsilon))$.

(Sphere packing bound: must have $k \leq n - \log \mathrm{Vol}_{\mathbb{F}_2^n, \mathrm{dist}_p}(\varepsilon/2)$)

Lemma (Balanced p-weighted exist)

Fix $p \in \Delta([n])$ and ε. There exists a p-weighted (nearly) balanced code $C\colon \{0,1\}^k \to \{0,1\}^n$ with relative distance ε such that $k = \Omega(n - \log \mathrm{Vol}_{\mathbb{F}_2^n,\mathrm{dist}_p}(\varepsilon))$.

(Sphere packing bound: must have $k \leq n - \log \mathrm{Vol}_{\mathbb{F}_2^n,\mathrm{dist}_p}(\varepsilon/2)$)

Recall

Our reduction will give a lower bound of $\Omega\left(\frac{\sqrt{k}}{\log n}\right)$: so we need to analyze $\mathrm{Vol}_{\mathbb{F}_2^n,\mathrm{dist}_p}(\varepsilon)$.

ENTER CONCENTRATION INEQUALITIES

$$
\begin{aligned}
\mathrm{Vol}_{\mathbb{F}_2^n,\mathrm{dist}_p}(\gamma) &= \left| \left\{ w \in \mathbb{F}_2^n : \sum_{i=1}^{n} p_i w_i \leq \gamma \right\} \right| \\
&= 2^n \Pr_{Y \sim \{0,1\}^n} \left[\sum_{i=1}^{n} p_i Y_i \leq \gamma \right] \\
&= 2^n \Pr_{X \sim \{-1,1\}^n} \left[\sum_{i=1}^{n} p_i X_i \geq 1 - 2\gamma \right]
\end{aligned}
$$

$$\begin{aligned}
\mathrm{Vol}_{\mathbb{F}_2^n,\mathrm{dist}_p}(\gamma) &= \left|\left\{ w \in \mathbb{F}_2^n : \ \sum_{i=1}^n p_i w_i \le \gamma \right\}\right| \\
&= 2^n \Pr_{Y \sim \{0,1\}^n}\left[\sum_{i=1}^n p_i Y_i \le \gamma\right] \\
&= 2^n \Pr_{X \sim \{-1,1\}^n}\left[\sum_{i=1}^n p_i X_i \ge 1 - 2\gamma\right]
\end{aligned}$$

Concentration inequalities for weighted sums of Rademacher r.v.'s?

Definition (K-functional)

Fix any two Banach spaces $(X_0, \|\cdot\|_0), (X_1, \|\cdot\|_1)$. The K-functional between X_0 and X_1 is the function $K_{X_0,X_1} \colon (X_0 + X_1) \times (0, \infty) \to [0, \infty)$ defined by

$$K_{X_0,X_1}(x, t) := \inf_{\substack{(x_0,x_1)\in X_0\times X_1 \\ x_0+x_1=x}} \|x_0\|_0 + t\|x_1\|_1.$$

For $a \in \ell_1 + \ell_2$, we write κ_a for the function $t \mapsto K_{\ell_1,\ell_2}(a, t)$.

Theorem ([MS90])

Let $(X_i)_{i\geq 0}$ be a sequence of independent Rademacher random variables, i.e. uniform on $\{-1,1\}$. Then, for any $a \in \ell_2$ and $t > 0$,

$$\Pr\left[\sum_{i=1}^{\infty} a_i X_i \geq \kappa_a(t)\right] \leq e^{-\frac{t^2}{2}}. \tag{1}$$

and

$$\Pr\left[\sum_{i=1}^{\infty} a_i X_i \geq \frac{1}{2}\kappa_a(t)\right] \geq e^{-(2\ln 24)t^2}. \tag{2}$$

Theorem ([BCG16])

Identity testing to $p \in \Delta([n])$ requires $\Omega(t_\varepsilon / \log(n))$ samples, where $t_\varepsilon := \kappa_p^{-1}(1 - 2\varepsilon)$.

Theorem ([BCG16])

Identity testing to $p \in \Delta([n])$ requires $\Omega(t_\varepsilon / \log(n))$ samples, where $t_\varepsilon := \kappa_p^{-1}(1 - 2\varepsilon)$.

But...

...is it tight?

Theorem ([BCG16])

Identity testing to $p \in \Delta([n])$ can be done with $O\left(\frac{t_{\varepsilon/18}}{\varepsilon^2}\right)$ samples and requires $\Omega\left(\frac{t_\varepsilon}{\varepsilon}\right)$ of them, where $t_\varepsilon := \kappa_p^{-1}(1-2\varepsilon)$.

Theorem ([BCG16])

Identity testing to $p \in \Delta([n])$ can be done with $O\left(\frac{t_{\varepsilon/18}}{\varepsilon^2}\right)$ samples and requires $\Omega\left(\frac{t_\varepsilon}{\varepsilon}\right)$ of them, where $t_\varepsilon := \kappa_p^{-1}(1-2\varepsilon)$.

Upper bound established by a new connection between K-functional and "effective support size."

Theorem ([Ast10, MS90])

For arbitrary $a \in \ell_2$ and $t \in \mathbb{N}$, define the norm

$$\|a\|_{Q(t)} := \sup \left\{ \sum_{j=1}^{t} \left(\sum_{i \in A_j} a_i^2 \right)^{1/2} \ : \ (A_j)_{1 \leq j \leq t} \text{ partition of } \mathbb{N} \right\}.$$

Then, for any $a \in \ell_2$, and $t > 0$ such that $t^2 \in \mathbb{N}$, we have

$$\|a\|_{Q(t^2)} \leq \kappa_a(t) \leq \sqrt{2}\|a\|_{Q(t^2)}. \tag{3}$$

Lemma ([BCG16])

For any $a \in \ell_2$ and t such that $t^2 \in \mathbb{N}$, we have

$$\|a\|_{Q(t^2)} \leq \kappa_a(t) \leq \|a\|_{Q(2t^2)}. \tag{4}$$

Lemma (Sparsity Lemma)

If $\|p\|_{Q(T)} \geq 1 - 2\varepsilon$, then there is a subset S of T elements such that $p(S) \geq 1 - 6\varepsilon$.

Lemma (Sparsity Lemma)

If $\|p\|_{Q(T)} \geq 1 - 2\varepsilon$, then there is a subset S of T elements such that $p(S) \geq 1 - 6\varepsilon$.

Directly implies the upperbound, using $T := 2t^2_{O(\varepsilon)}$.

Lemma (Sparsity Lemma)

If $\|p\|_{Q(T)} \geq 1 - 2\varepsilon$, then there is a subset S of T elements such that $p(S) \geq 1 - 6\varepsilon$.

Directly implies the upperbound, using $T := 2t^2_{O(\varepsilon)}$.

Proof idea.

By monotonicity, $\sum_{j=1}^{T} \left(\sum_{i \in A_j} p_i^2\right)^{1/2} \leq \sum_{j=1}^{T} \sum_{i \in A_j} p_i = \|p\|_1 = 1$. So we have

$$1 - 2\varepsilon \leq \sum_{j=1}^{T} \left(\sum_{i \in A_j} p_i^2\right)^{1/2} \leq 1$$

which (morally) implies that p is "close to a singleton" on each A_j. $\square$

CONCLUSION

- New framework to prove distribution testing lower bounds: reduction from communication complexity

CONCLUSION

- New framework to prove distribution testing lower bounds: reduction from communication complexity
- Clean and simple

CONCLUSION

- New framework to prove distribution testing lower bounds: reduction from communication complexity
- Clean and simple
- Leads to new insights: "instance-optimal" identity testing, revisited

CONCLUSION

- New framework to prove distribution testing lower bounds: reduction from communication complexity
- Clean and simple
- Leads to new insights: “instance-optimal” identity testing, revisited
- unexpected connection to interpolation theory

- New framework to prove distribution testing lower bounds: reduction from communication complexity
- Clean and simple
- Leads to new insights: "instance-optimal" identity testing, revisited
- unexpected connection to interpolation theory
- **Codes are great!**

Thank you

Jayadev Acharya and Constantinos Daskalakis.
Testing Poisson Binomial Distributions.
In Proceedings of SODA, pages 1829–1840, 2014.

Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath.
Optimal testing for properties of distributions.
ArXiV, (abs/1507.05952), July 2015.

Sergey V. Astashkin.
Rademacher functions in symmetric spaces.
Journal of Mathematical Sciences, 169(6):725–886, sep 2010.

Eric Blais, Joshua Brody, and Kevin Matulef.
Property testing lower bounds via communication complexity.
Computational Complexity, 21(2):311–358, 2012.

Eric Blais, Clément L. Canonne, and Tom Gur.
Alice and Bob Show Distribution Testing Lower Bounds (They don't talk to each other anymore.).
Electronic Colloquium on Computational Complexity (ECCC), 23:168, 2016.

Tuğkan Batu, Eldar Fischer, Lance Fortnow, Ravi Kumar, Ronitt Rubinfeld, and Patrick White.
Testing random variables for independence and identity.
In Proceedings of FOCS, pages 442–451, 2001.

Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White.
Testing that distributions are close.
In Proceedings of FOCS, pages 189–197, 2000.

Tuğkan Batu, Ravi Kumar, and Ronitt Rubinfeld.
Sublinear algorithms for testing monotone and unimodal distributions.

In Proceedings of STOC, pages 381–390, New York, NY, USA, 2004. ACM.

Clément L. Canonne, Ilias Diakonikolas, Themis Gouleakis, and Ronitt Rubinfeld.
Testing Shape Restrictions of Discrete Distributions.
ArXiV, abs/1507.03558, July 2015.

Siu-On Chan, Ilias Diakonikolas, Gregory Valiant, and Paul Valiant.
Optimal algorithms for testing closeness of discrete distributions.
In Proceedings of SODA, pages 1193–1203. Society for Industrial and Applied Mathematics (SIAM), 2014.

Oded Goldreich and Dana Ron.
On testing expansion in bounded-degree graphs.
Electronic Colloquium on Computational Complexity (ECCC), 7:20, 2000.

Reut Levi, Dana Ron, and Ronitt Rubinfeld.
Testing properties of collections of distributions.
Theory of Computing, 9:295–347, 2013.

Stephen J. Montgomery-Smith.
The distribution of Rademacher sums.
Proceedings of the American Mathematical Society, 109(2):517–522, 1990.

Ilan Newman and Mario Szegedy.
Public vs. private coin flips in one round communication games.
In Proceedings of the twenty-eighth annual ACM symposium on Theory of computing, pages 561–570. ACM, 1996.

Liam Paninski.
A coincidence-based test for uniformity given very sparsely sampled discrete data.
IEEE Transactions on Information Theory, 54(10):4750–4755, 2008.

Jaak Peetre.
A theory of interpolation of normed spaces.
Notas de Matemática, No. 39. Instituto de Matemática Pura e Aplicada, Conselho Nacional de Pesquisas, Rio de Janeiro, 1968.

Paul Valiant.
Testing symmetric properties of distributions.
SIAM Journal on Computing, 40(6):1927–1968, 2011.

Gregory Valiant and Paul Valiant.
An automatic inequality prover and instance optimal identity testing.
In Proceedings of FOCS, 2014.