ALICE AND BOB SHOW DISTRIBUTION TESTING LOWER BOUNDS

They don’t talk to each other anymore.

Clément Canonne (Columbia University)
July 9, 2017

Joint work with Eric Blais (UWaterloo) and Tom Gur (Weizmann Institute, UC Berkeley)
“DISTRIBUTION TESTING?”
Property testing of probability distributions:
Property testing of probability distributions: sublinear,
Property testing of probability distributions: sublinear, approximate,
Property testing of probability distributions: sublinear, approximate, randomized
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
Property testing of probability distributions: sublinear, approximate, randomized algorithms that take random samples

- Big Dataset: too big
- Expensive access: pricey data
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
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.
(Property) Distribution Testing:
(Property) Distribution Testing:
(Property) Distribution Testing:

in an (egg)shell.
Known domain (here \([n] = \{1, \ldots, n\}\))

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

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

Distance parameter \(\varepsilon \in (0, 1]\)
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:

$p \in \mathcal{P}$
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)
Many results on many properties:
Many results on many properties:

- Uniformity [GR00, BFR\textsuperscript{+}00, Pan08]
Many results on many properties:

- Uniformity
- Identity*

[GR00, BFR⁺00, Pan08]
[BFF⁺01, VV14]
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 more…
Many results on many properties:

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

* Identity is marked with an asterisk.
Many results on many properties:

- Uniformity  \[GR00, BFR^+00, Pan08]\n- Identity*  \[BFF^+01, VV14]\n- Equivalence  \[BFR^+00, Val11, CDVV14]\n- Independence  \[BFF^+01, LRR13]\n- Monotonicity  \[BKR04]\n
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*
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]
Many results on many properties:

- Uniformity
- 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 more… [GR00, BFR+00, Pan08]
Lower bounds...

... are quite tricky.
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.
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.
“COMMUNICATION COMPLEXITY?”
f(x, y)
f(x, y)
WHAT NOW?

\[ f(x, y) \]
WHAT NOW?

\[ f(x, y) \]
$f(x, y)$
But communicating is hard.
WAS THAT A TOILET?

- 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).
...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
...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

\[ \text{SMP}(E_{Q_n}) = \Omega(\sqrt{n}) \]

(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 $\Omega$ be a domain of size $n$. Fix a property $\Pi \subseteq \Delta(\Omega)$ and $f: \{0, 1\}^k \times \{0, 1\}^k \rightarrow \{0, 1\}$. Suppose there exists a mapping $p: \{0, 1\}^k \times \{0, 1\}^k \rightarrow \Delta(\Omega)$ 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 $\alpha, \beta$ 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 $\varepsilon$-far from $\Pi$ in $\ell_1$ distance.

Then, every $\varepsilon$-tester for $\Pi$ needs $\Omega\left(\frac{\text{SMP}(f)}{\log(n)}\right)$ samples.
Take the “equality” predicate $\text{Eq}_k$ as $f$:

**Theorem (Newman and Szegedy [NS96])**

For every $k \in \mathbb{N}$, $\text{SMP}(\text{Eq}_k) = \Omega\left(\sqrt{k}\right)$. 
Take the “equality” predicate $\text{Eq}_k$ as $f$:

Theorem (Newman and Szegedy [NS96])

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

**Goal:**

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

(A) Alice
x=[0 1 0 1 1]
C(x)=[0 1 0 1 1 0 0 1 1 0]

Bob
y=[0 1 0 1 1]
C(y)=[0 1 0 1 1 0 0 1 1 0]

(B) Alice
x=[0 1 0 1 1]
C(x)=[0 1 0 1 1 0 0 1 1 0]

Bob
y=[0 1 0 0 1]
C(y)=[0 1 0 0 1 1 1 0 0 1]
Testing Identity

Statement

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

$$p = q \quad \text{vs.} \quad \ell_1(p, q) > \varepsilon$$
Statement
Explicit description of $p \in \Delta([n])$, parameter $\varepsilon \in (0, 1]$. Given samples from (unknown) $q \in \Delta([n])$, decide

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

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

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

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 $\Omega\left(\frac{\|p_{-\max}^\varepsilon\|_2/3}{\varepsilon^2}\right)$ samples (and this is “tight”).
An issue: how to interpret this $\|p^{\max}_{-\varepsilon}\|_{2/3}$?
An issue: how to interpret this $\|p_\varepsilon^\max\|_{2/3}$?

Goal:

Will prove an(other) $\tilde{\Omega}(\Phi(p, \varepsilon))$ lower bound on testing identity, via communication complexity.
An issue: how to interpret this $\| p_{\epsilon}^{\max} \|_{2/3}$?

Goal:

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

(and it will be “tight” as well.)
· **p-weighted** codes

\[
dist_p(x, y) := \sum_{i=1}^{n} p(i) \cdot |x_i - y_i| \quad (x, y \in \{0, 1\}^n)
\]

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

· **Volume of the p-ball**:

\[
\text{Vol}_{F_2^n, \text{dist}_p}(\varepsilon) := |\{ w \in F_2^n : \text{dist}_p(w, 0^n) \leq \varepsilon \}|.
\]
Lemma (Balanced $p$-weighted exist)

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

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

(Sphere packing bound: must have $k \leq n - \log \text{Vol}_{\mathbb{F}_2, \text{dist}_p}(\epsilon/2)$)
Lemma (Balanced \( p \)-weighted exist)

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

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

Recall

Our reduction will give a lower bound of \( \Omega\left(\frac{\sqrt{k}}{\log n}\right) \): so we need to analyze \( \text{Vol}_{\mathbb{F}_2^n, \text{dist}_p} (\varepsilon) \).
Concentration inequalities for weighted sums of Rademacher r.v.'s?

\[ \text{Vol}_{F_2^n, dist_p} (\gamma) = \left| \left\{ w \in F_2^n : \sum_{i=1}^{n} p_i w_i \leq \gamma \right\} \right| \]

\[ = 2^n \Pr_{Y \sim \{0,1\}^n} \left[ \sum_{i=1}^{n} p_i Y_i \leq \gamma \right] \]

\[ = 2^n \Pr_{X \sim \{-1,1\}^n} \left[ \sum_{i=1}^{n} p_i X_i \geq 1 - 2\gamma \right] \]
Concentration inequalities for weighted sums of Rademacher r.v.'s?

\[
\text{Vol}_{\mathbb{F}_2^n, \text{dist}_p} (\gamma) = \left| \left\{ w \in \mathbb{F}_2^n : \sum_{i=1}^{n} p_i w_i \leq \gamma \right\} \right|
\]

\[
= 2^n \Pr_{Y \sim \{0,1\}^n} \left[ \sum_{i=1}^{n} p_i Y_i \leq \gamma \right]
\]

\[
= 2^n \Pr_{X \sim \{-1,1\}^n} \left[ \sum_{i=1}^{n} p_i X_i \geq 1 - 2\gamma \right]
\]
**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 \(K_{X_0,X_1} : (X_0 + X_1) \times (0, \infty) \to [0, \infty)\) defined by

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

For \(a \in \ell_1 + \ell_2 = \ell_2\), we write \(\kappa_a\) 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

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

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

(3)

Lemma ([BCG16])

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

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

(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 $T := 2t^2_{O}(\varepsilon)$. 

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 $A_j$. □
· New framework to prove distribution testing lower bounds: reduction from *communication complexity*
• 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.

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.

Eric Blais, Joshua Brody, and Kevin Matulef.
Property testing lower bounds via communication complexity.

Eric Blais, Clément L. Canonne, and Tom Gur.
Alice and Bob Show Distribution Testing Lower Bounds (They don’t talk to each other anymore.).

Tuğkan Batu, Eldar Fischer, Lance Fortnow, Ravi Kumar, Ronitt Rubinfeld, and Patrick White.
Testing random variables for independence and identity.

Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White.
Testing that distributions are close.

Tuğkan Batu, Ravi Kumar, and Ronitt Rubinfeld.
Sublinear algorithms for testing monotone and unimodal distributions.
Jaak Peetre.
A theory of interpolation of normed spaces.

Paul Valiant.
Testing symmetric properties of distributions.

Gregory Valiant and Paul Valiant.
An automatic inequality prover and instance optimal identity testing.