PROPERTY TESTING AND PROBABILITY DISTRIBUTIONS: NEW TECHNIQUES, NEW MODELS, AND NEW GOALS

Ph.D. Defense

Clément Canonne

September 18, 2017

Columbia University, Computer Science Department

Background, Context, and Motivation

PROPERTY TESTING

Sublinear-time,

PROPERTY TESTING

Sublinear-time, approximate,

PROPERTY TESTING

Sublinear-time, approximate, randomized

PROPERTY TESTING

Sublinear-time, approximate, randomized decision algorithms that make local queries to their input.

PROPERTY TESTING

Sublinear-time, approximate, randomized decision algorithms that make local queries to their input.

- Big Dataset: too big

PROPERTY TESTING

Sublinear-time, approximate, randomized decision algorithms that make local queries to their input.

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

PROPERTY TESTING

Sublinear-time, approximate, randomized decision algorithms that make local queries to their input.

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

PROPERTY TESTING

Sublinear-time, approximate, randomized decision algorithms that make local queries to their input.

- Big Dataset: too big
- Expensive access: pricey data
- “Model selection”: many options
- Good Enough: a priori knowledge

PROPERTY TESTING

Sublinear-time, approximate, randomized decision algorithms that make local queries to their input.

- Big Dataset: too big
- Expensive access: pricey data
- “Model selection”: many options
- Good Enough: a priori knowledge

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

PROPERTY TESTING

Figure: Property Testing: Inside the yolk, or outside the egg.

Introduced by [RS96, GGR98] – has been a very active area since.

- Known space (e.g., $\{0,1\}^N$)
- Property $\mathcal{P} \subseteq \{0,1\}^N$
- Oracle access to unknown $x \in \{0,1\}^N$
- Proximity parameter $\varepsilon \in (0,1]$

Must decide

$$x \in \mathcal{P} \quad \text{vs.} \quad \operatorname{dist}(x, \mathcal{P}) > \varepsilon$$

(has the property, or is ε-far from it)

Many variants, subareas, with a plethora of results (see e.g. [Ron08, Ron10, Gol10, Gol17, BY17]).

DISTRIBUTION TESTING

Now, our “big object” is a probability distribution over a (discrete*) domain Ω (here, $|\Omega| = n$).

DISTRIBUTION TESTING

Now, our “big object” is a probability distribution over a (discrete*) domain Ω (here, $|\Omega| = n$).

- instead of queries: samples*

DISTRIBUTION TESTING

Now, our “big object” is a probability distribution over a (discrete*) domain Ω (here, $|\Omega| = n$).

- instead of queries: samples*
- instead of Hamming distance: total variation*

DISTRIBUTION TESTING

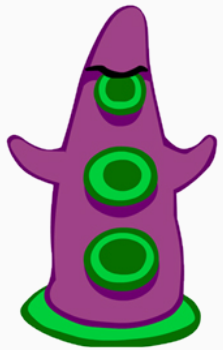
Now, our “big object” is a probability distribution over a (discrete*) domain Ω (here, $|\Omega| = n$).

- instead of queries: samples*
- instead of Hamming distance: total variation*
- instead of functions/graphs/strings: distributions

DISTRIBUTION TESTING

Now, our "big object" is a probability distribution over a (discrete*) domain Ω (here, $|\Omega| = n$).

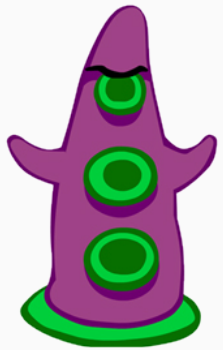
- instead of queries: samples*
- instead of Hamming distance: total variation*
- instead of functions/graphs/strings: distributions



DISTRIBUTION TESTING

Now, our "big object" is a probability distribution over a (discrete*) domain Ω (here, $|\Omega| = n$).

- instead of queries: samples*
- instead of Hamming distance: total variation*
- instead of functions/graphs/strings: distributions

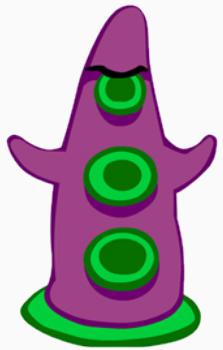


Focus on the sample complexity, with efficiency as ancillary goal.

DISTRIBUTION TESTING

Now, our "big object" is a probability distribution over a (discrete*) domain Ω (here, $|\Omega| = n$).

- instead of queries: samples*
- instead of Hamming distance: total variation*
- instead of functions/graphs/strings: distributions



Focus on the sample complexity, with efficiency as ancillary goal.

*usually.

Over the past 15 years, many results on many properties:

Over the past 15 years, many results on many properties:

- Uniformity [GR00, BFR+00, Pan08, DGPP16]

BACKGROUND

Over the past 15 years, many results on many properties:

- Uniformity [GR00, BFR+00, Pan08, DGPP16]
- Identity* [BFF+01, VV17]

BACKGROUND

Over the past 15 years, many results on many properties:

- Uniformity [GR00, BFR+00, Pan08, DGPP16]
- Identity* [BFF+01, VV17]
- Equivalence [BFR+00, Val11, CDVV14]

Over the past 15 years, many results on many properties:

- Uniformity [GR00, BFR+00, Pan08, DGPP16]
- Identity* [BFF+01, VV17]
- Equivalence [BFR+00, Val11, CDVV14]
- Independence [BFF+01, LRR13, DK16]

Over the past 15 years, many results on many properties:

- Uniformity [GR00, BFR+00, Pan08, DGPP16]
- Identity* [BFF+01, VV17]
- Equivalence [BFR+00, Val11, CDVV14]
- Independence [BFF+01, LRR13, DK16]
- Monotonicity [BKR04, BFRV11, ADK15]

Over the past 15 years, many results on many properties:

- Uniformity [GR00, BFR+00, Pan08, DGPP16]
- Identity* [BFF+01, VV17]
- Equivalence [BFR+00, Val11, CDVV14]
- Independence [BFF+01, LRR13, DK16]
- Monotonicity [BKR04, BFRV11, ADK15]
- Poisson Binomial Distributions [AD15]

Over the past 15 years, many results on many properties:

- Uniformity [GR00, BFR+00, Pan08, DGPP16]
- Identity* [BFF+01, VV17]
- Equivalence [BFR+00, Val11, CDVV14]
- Independence [BFF+01, LRR13, DK16]
- Monotonicity [BKR04, BFRV11, ADK15]
- Poisson Binomial Distributions [AD15]
- and more... [Rub12, Can15b]

Over the past 15 years, many results on many properties:

- Uniformity [GR00, BFR+00, Pan08, DGPP16]
- Identity* [BFF+01, VV17]
- Equivalence [BFR+00, Val11, CDVV14]
- Independence [BFF+01, LRR13, DK16]
- Monotonicity [BKR04, BFRV11, ADK15]
- Poisson Binomial Distributions [AD15]
- and more... [Rub12, Can15b]

Much has been done;

Over the past 15 years, many results on many properties:

- Uniformity [GR00, BFR+00, Pan08, DGPP16]
- Identity* [BFF+01, VV17]
- Equivalence [BFR+00, Val11, CDVV14]
- Independence [BFF+01, LRR13, DK16]
- Monotonicity [BKR04, BFRV11, ADK15]
- Poisson Binomial Distributions [AD15]
- and more... [Rub12, Can15b]

Much has been done; and yet...

MANY QUESTIONS REMAIN

Techniques

Most algorithms, results are somewhat ad hoc, and property-specific.

MANY QUESTIONS REMAIN

Techniques

Most algorithms, results are somewhat ad hoc, and property-specific.

Hardness

Most properties are depressingly hard to test: $\Omega(\sqrt{n})$ samples are required.

MANY QUESTIONS REMAIN

Techniques

Most algorithms, results are somewhat ad hoc, and property-specific.

Hardness

Most properties are depressingly hard to test: $\Omega(\sqrt{n})$ samples are required.

Beyond?

Testing is only a preliminary step! What if...

OUTLINE OF THE THESIS

OUTLINE OF THE THESIS

Three main axes:

OUTLINE OF THE THESIS

Three main axes:

1. Changing the way: general algorithms and approaches

OUTLINE OF THE THESIS

Three main axes:

1. Changing the way: general algorithms and approaches
2. Changing the rules: different, more powerful types of access

OUTLINE OF THE THESIS

Three main axes:

1. Changing the way: general algorithms and approaches
2. Changing the rules: different, more powerful types of access
3. Changing the goal: beyond testing – correcting the distributions

OUTLINE OF THE THESIS

Three main axes:

1. Changing the way: general algorithms and approaches
2. Changing the rules: different, more powerful types of access
3. Changing the goal: beyond testing – correcting the distributions

OUTLINE OF THE THESIS

Three main axes:

1. Changing the way: general algorithms and approaches
2. Changing the rules: different, more powerful types of access
3. Changing the goal: beyond testing – correcting the distributions

This defense: We will cover all three (focusing on the first).

SOME NOTATION

- Probability distributions over $[n] := \{1, \ldots, n\}$

$$\Delta([n]) = \left\{ p : [n] \to [0,1] : \sum_{i=1}^{n} p(i) = 1 \right\}$$

GLOSSARY

- Probability distributions over $[n] := \{1, \ldots, n\}$

$$\Delta([n]) = \left\{ p : [n] \to [0,1] : \sum_{i=1}^{n} p(i) = 1 \right\}$$

- Property (or class) of distributions over $[n]$:

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

GLOSSARY

- Probability distributions over $[n] := \{1, \ldots, n\}$

$$\Delta([n]) = \left\{ p : [n] \to [0,1] : \sum_{i=1}^{n} p(i) = 1 \right\}$$

- Property (or class) of distributions over $[n]$:

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

- Total variation distance (statistical distance, ℓ_1 distance):

$$d_{\mathrm{TV}}(p,q) = \sup_{S \subseteq \Omega}(p(S) - q(S)) = \frac{1}{2}\sum_{x \in \Omega} |p(x) - q(x)| \in [0,1]$$

GLOSSARY

- Probability distributions over $[n] := \{1, \ldots, n\}$

$$\Delta([n]) = \left\{ p : [n] \to [0,1] : \sum_{i=1}^{n} p(i) = 1 \right\}$$

- Property (or class) of distributions over $[n]$:

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

- Total variation distance (statistical distance, ℓ_1 distance):

$$d_{\mathrm{TV}}(p,q) = \sup_{S \subseteq \Omega}(p(S) - q(S)) = \frac{1}{2}\sum_{x \in \Omega} |p(x) - q(x)| \in [0,1]$$

Domain size $n \in \mathbb{N}$ is big ("goes to ∞").

GLOSSARY

- Probability distributions over $[n] := \{1, \ldots, n\}$

$$\Delta([n]) = \left\{ p : [n] \to [0,1] : \sum_{i=1}^{n} p(i) = 1 \right\}$$

- Property (or class) of distributions over $[n]$:

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

- Total variation distance (statistical distance, ℓ_1 distance):

$$d_{\mathrm{TV}}(p,q) = \sup_{S \subseteq \Omega}(p(S) - q(S)) = \frac{1}{2}\sum_{x \in \Omega} |p(x) - q(x)| \in [0,1]$$

Domain size $n \in \mathbb{N}$ is big ("goes to ∞"). Proximity parameter $\varepsilon \in (0,1]$ is small.

GLOSSARY

- Probability distributions over $[n] := \{1, \ldots, n\}$

$$\Delta([n]) = \left\{ p : [n] \to [0,1] : \sum_{i=1}^{n} p(i) = 1 \right\}$$

- Property (or class) of distributions over $[n]$:

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

- Total variation distance (statistical distance, ℓ_1 distance):

$$d_{\mathrm{TV}}(p,q) = \sup_{S \subseteq \Omega}(p(S) - q(S)) = \frac{1}{2}\sum_{x \in \Omega} |p(x) - q(x)| \in [0,1]$$

Domain size $n \in \mathbb{N}$ is big ("goes to ∞"). Proximity parameter $\varepsilon \in (0,1]$ is small. Lowercase Greek letters are in $(0,1]$.

GLOSSARY

- Probability distributions over $[n] := \{1, \ldots, n\}$

$$\Delta([n]) = \left\{ p : [n] \to [0,1] : \sum_{i=1}^{n} p(i) = 1 \right\}$$

- Property (or class) of distributions over $[n]$:

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

- Total variation distance (statistical distance, ℓ_1 distance):

$$d_{\mathrm{TV}}(p,q) = \sup_{S \subseteq \Omega}(p(S) - q(S)) = \frac{1}{2}\sum_{x \in \Omega} |p(x) - q(x)| \in [0,1]$$

Domain size $n \in \mathbb{N}$ is big ("goes to ∞"). Proximity parameter $\varepsilon \in (0,1]$ is small. Lowercase Greek letters are in $(0,1]$. Asymptotics $\tilde{O}$, $\tilde{\Omega}$, $\tilde{\Theta}$ hide logarithmic factors.*

CHANGING THE WAY: UNIFIED APPROACHES

What we want

General algorithms applying to all (or many) distribution testing problems.

What we want

General algorithms applying to all (or many) distribution testing problems.

Theorem (Wishful)

Let $\mathcal{P}$ be a class of distributions that all exhibit some "nice structure." If $\mathcal{P}$ can be tested with q queries, algorithm $\mathfrak{T}$ can too, with "roughly" q queries as well.

What we want

General algorithms applying to all (or many) distribution testing problems.

Theorem (Wishful)

Let $\mathcal{P}$ be a class of distributions that all exhibit some "nice structure." If $\mathcal{P}$ can be tested with q queries, algorithm $\mathfrak{T}$ can too, with "roughly" q queries as well.

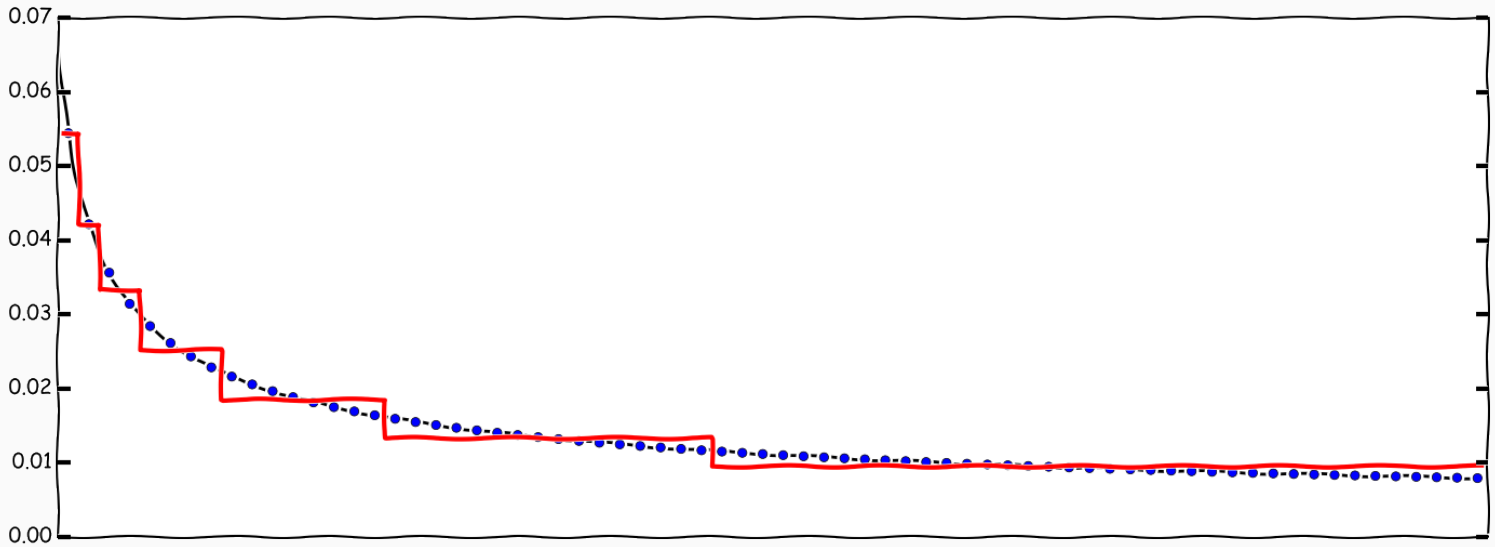
More formally, we want:

Goal

Design general-purpose testing algorithms that, when applied to a property $\mathcal{P}$, have (tight, or at least reasonable) sample complexity $q(\varepsilon, \tau)$ as long as $\mathcal{P}$ satisfies some structural assumption $\mathcal{S}_\tau$ parameterized by τ.

SWISS ARMY KNIVES: SHAPE RESTRICTIONS

Structural assumption $\mathcal{S}_\tau$: every distribution in $\mathcal{P}$ is well-approximated (in a specific ℓ_2-type sense) by a piecewise-constant distribution with $\mathrm{L}_{\mathcal{P}}(\tau)$ pieces.



SWISS ARMY KNIVES: SHAPE RESTRICTIONS

Structural assumption $\mathcal{S}_\tau$: every distribution in $\mathcal{P}$ is well-approximated (in a specific ℓ_2-type sense) by a piecewise-constant distribution with $L_{\mathcal{P}}(\tau)$ pieces.

Theorem ([CDGR16])

There exists an algorithm which, given sampling access to an unknown distribution p over [n] and parameter $\varepsilon \in (0, 1]$, can distinguish with probability 2/3 between **(a)** $p \in \mathcal{P}$ versus (b) $d_{\mathrm{TV}}(p, \mathcal{P}) > \varepsilon$, with $\tilde{O}(\sqrt{n L_{\mathcal{P}}(\varepsilon)}/\varepsilon^3 + L_{\mathcal{P}}(\varepsilon)/\varepsilon^2)$ samples.

SWISS ARMY KNIVES: SHAPE RESTRICTIONS

Outline: Abstracting ideas from [BKR04] (for monotonicity):

1. **decomposition step:** recursively build a partition Π of $[n]$ in $O(L_{\mathcal{P}}(\varepsilon))$ intervals s.t. p is roughly uniform on each piece. If successful, then p will be close to its "flattening" q on Π; if not, we have proof that $p \notin \mathcal{P}$ and we can reject.
2. **approximation step:** learn q. Can be done with few samples since Π has few intervals.
3. **projection step:** (computational) verify that $d_{\mathrm{TV}}(q, \mathcal{P}) < O(\varepsilon)$.

SWISS ARMY KNIVES: SHAPE RESTRICTIONS

Outline: Abstracting ideas from [BKR04] (for monotonicity):

1. **decomposition step:** recursively build a partition Π of $[n]$ in $O(L_{\mathcal{P}}(\varepsilon))$ intervals s.t. p is roughly uniform on each piece. If successful, then p will be close to its "flattening" q on Π; if not, we have proof that $p \notin \mathcal{P}$ and we can reject.
2. **approximation step:** learn q. Can be done with few samples since Π has few intervals.
3. **projection step:** (computational) verify that $d_{\mathrm{TV}}(q, \mathcal{P}) < O(\varepsilon)$.

Applications

- monotonicity
- unimodality
- k-modality
- k-histograms
- log-concavity
- Poisson Binomial
- Monotone Hazard Rate

...

Structural assumption $\mathcal{S}_\tau$: every distribution in $\mathcal{P}$ has sparse Fourier and effective support: $\exists M_{\mathcal{P}}(\tau), S_{\mathcal{P}}(\tau)$ s.t. $\forall p \in \mathcal{P}$, $\exists I_p \subseteq [n]$ with $|I_p| \leq M_{\mathcal{P}}(\tau)$

$$\left\|\hat{p}\mathbf{1}_{\overline{S_{\mathcal{P}}(\varepsilon)}}\right\|_2 \leq O(\varepsilon), \quad \left\|p\mathbf{1}_{\overline{I_p}}\right\|_1 \leq O(\varepsilon)$$

Theorem ([CDS17])

There exists an algorithm which, given sampling access to an unknown distribution p over $[n]$ and parameter $\varepsilon \in (0, 1]$, can distinguish with probability $2/3$ between **(a)** $p \in \mathcal{P}$ versus (b) $d_{\mathrm{TV}}(p, \mathcal{P}) > \varepsilon$, with $\tilde{O}(\sqrt{|S_{\mathcal{P}}(\varepsilon)|\, M_{\mathcal{P}}(\varepsilon)}/\varepsilon^2 + |S_{\mathcal{P}}(\varepsilon)|/\varepsilon^2)$ samples.

SWISS ARMY KNIVES: FOURIER SPARSITY

Outline:

1. **effective support test:** take samples to identify a candidate I_p, and check $|I_p| \leq M(\varepsilon)$
2. **Fourier effective support test:** invoke a Fourier sparsity subroutine to check that $\|\hat{p}\mathbf{1}_{\overline{S_{\mathcal{P}}(\varepsilon)}}\|_2 \leq O(\varepsilon)$ (if so learn q, inverse Fourier transform of $\hat{p}\mathbf{1}_{\overline{S_{\mathcal{P}}(\varepsilon)}}$)
3. **projection step:** (computational) verify that $d_{\mathrm{TV}}(q, \mathcal{P}) < O(\varepsilon)$.

SWISS ARMY KNIVES: FOURIER SPARSITY

Outline:

1. **effective support test:** take samples to identify a candidate I_p, and check $|I_p| \leq M(\varepsilon)$
2. **Fourier effective support test:** invoke a Fourier sparsity subroutine to check that $\|\hat{p}\mathbf{1}_{\overline{S_{\mathcal{P}}(\varepsilon)}}\|_2 \leq O(\varepsilon)$ (if so learn q, inverse Fourier transform of $\hat{p}\mathbf{1}_{\overline{S_{\mathcal{P}}(\varepsilon)}}$)
3. **projection step:** (computational) verify that $d_{\mathrm{TV}}(q, \mathcal{P}) < O(\varepsilon)$.

Applications

- k-SIIRVS
- Poisson Binomial
- Poisson Multinomial
- log-concavity

Reductions: if you can learn, you can't test

Theorem (Wishful)

Let $\mathcal{P}$ be a class of distributions. Then, testing $\mathcal{P}$ is at least as hard as testing any $\mathcal{P}_{\mathrm{Hard}} \subseteq \mathcal{P}$.

Very intuitive...

Reductions: if you can learn, you can't test

Theorem (Wishful)

Let $\mathcal{P}$ be a class of distributions. Then, testing $\mathcal{P}$ is at least as hard as testing any $\mathcal{P}_{\mathrm{Hard}} \subseteq \mathcal{P}$.

Very intuitive... and very false.

Reductions: If you can learn, you can't test

Theorem (Wishful)

Let $\mathcal{P}$ be a class of distributions. Then, testing $\mathcal{P}$ is at least as hard as testing any $\mathcal{P}_{\mathrm{Hard}} \subseteq \mathcal{P}$.

Very intuitive... and very false. But we can make it "true-er":

Theorem ([CDGR16])

Let $\mathcal{P}$ be a class of distributions that has a sample-efficient agnostic* learner. Then, testing $\mathcal{P}$ is at least as hard as testing any $\mathcal{P}_{\mathrm{Hard}} \subseteq \mathcal{P}$.

Reductions: if you can learn, you can't test

Outline: "testing-by-narrowing"

1. Use tester for $\mathcal{P}$ on p: if p $\in \mathcal{P}_{\mathrm{Hard}}$, it accepts (and if does not reject, p close to $\mathcal{P}$)
2. Use agnostic learner to learn q (will be close to p by the previous step)
3. Check offline if q is close to $\mathcal{P}_{\mathrm{Hard}}$, accept only if it is the case

Reductions: if you can learn, you can't test

Outline: "testing-by-narrowing"

1. Use tester for $\mathcal{P}$ on p: if p $\in \mathcal{P}_{\mathrm{Hard}}$, it accepts (and if does not reject, p close to $\mathcal{P}$)
2. Use agnostic learner to learn q (will be close to p by the previous step)
3. Check offline if q is close to $\mathcal{P}_{\mathrm{Hard}}$, accept only if it is the case

Applications

- monotonicity
- unimodality
- k-modality
- k-histograms
- log-concavity
- Poisson Binomial
- MHR
- k-SIIRVS
- PMD

REDUCTIONS: IF YOU CAN LEARN, YOU CAN'T TEST

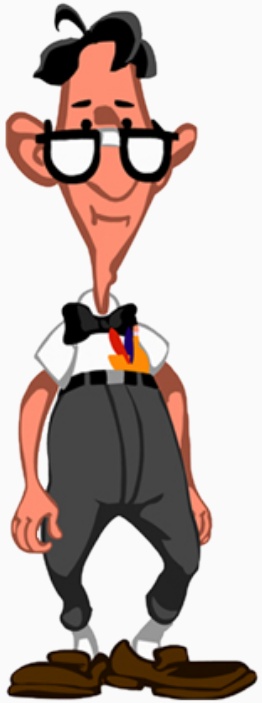
Outline: "testing-by-narrowing"

1. Use tester for $\mathcal{P}$ on p: if p $\in \mathcal{P}_{\mathrm{Hard}}$, it accepts (and if does not reject, p close to $\mathcal{P}$)
2. Use agnostic learner to learn q (will be close to p by the previous step)
3. Check offline if q is close to $\mathcal{P}_{\mathrm{Hard}}$, accept only if it is the case

Applications

- monotonicity
- unimodality
- k-modality
- k-histograms
- log-concavity
- Poisson Binomial
- MHR
- k-SIIRVS
- PMD
- +tolerant testing

REDUCTIONS: ALICE, BOB, AND THE REFEREE



REDUCTIONS: ALICE, BOB, AND THE REFEREE

- Communication complexity: huge literature, culminating in very strong lower bounds

REDUCTIONS: ALICE, BOB, AND THE REFEREE

- Communication complexity: huge literature, culminating in very strong lower bounds
- Can be leveraged to prove hardness results in streaming, data structures, decision tree complexity, ...

REDUCTIONS: ALICE, BOB, AND THE REFEREE

- Communication complexity: huge literature, culminating in very strong lower bounds
- Can be leveraged to prove hardness results in streaming, data structures, decision tree complexity, ...
- Elegant framework of Blais, Brody, and Matulef [BBM12]: actually, in property testing too!

REDUCTIONS: ALICE, BOB, AND THE REFEREE

- Communication complexity: huge literature, culminating in very strong lower bounds
- Can be leveraged to prove hardness results in streaming, data structures, decision tree complexity, ...
- Elegant framework of Blais, Brody, and Matulef [BBM12]: actually, in property testing too!

Can we do the same in distribution testing? ("Tap into" the hard-earned lower bounds results from communication complexity)

Reductions: Alice, Bob, and the Referee

Theorem ([BCG17])

Yes. (Even better, from the Simultaneous Message Passing (SMP) model.)

REDUCTIONS: ALICE, BOB, AND THE REFEREE

Theorem ([BCG17])

Yes. (Even better, from the Simultaneous Message Passing (SMP) model.)

Applications

- monotonicity
- unimodality
- k-modality
- log-concavity
- Poisson Binomial
- symmetric sparse support
- MHR
- junta

REDUCTIONS: ALICE, BOB, AND THE REFEREE

Theorem ([BCG17])

Yes. (Even better, from the Simultaneous Message Passing (SMP) model.)

Applications

- monotonicity
- unimodality
- k-modality
- log-concavity
- Poisson Binomial
- symmetric sparse support
- MHR
- junta
- identity

Changing the rules: what if...

Fact

In the standard sampling model, most (natural) properties are “hard” to test; i.e., require a strong dependence on n (at least $\Omega(\sqrt{n})$).

Fact

In the standard sampling model, most (natural) properties are "hard" to test; i.e., require a strong dependence on n (at least $\Omega(\sqrt{n})$).

Example

Testing uniformity has $\Theta(\sqrt{n}/\varepsilon^2)$ sample complexity [GR00, BFR+13, Pan08], identity $\Theta(\sqrt{n}/\varepsilon^2)$ [BFF+01, Pan08]; equivalence $\Theta(n^{2/3})$ [BFR+13, Val11, CDVV14]...

Fact

In the standard sampling model, most (natural) properties are "hard" to test; i.e., require a strong dependence on n (at least $\Omega(\sqrt{n})$).

Example

Testing uniformity has $\Theta(\sqrt{n}/\varepsilon^2)$ sample complexity [GR00, BFR+13, Pan08], identity $\Theta(\sqrt{n}/\varepsilon^2)$ [BFF+01, Pan08]; equivalence $\Theta(n^{2/3})$ [BFR+13, Val11, CDVV14]...

and more depressing for tolerant testing: $\Omega\left(n^{1-o(1)}\right)$ for entropy, support size...even for uniformity! [VV11, VV10a, JVYW15, WY16]

THE LIMITATIONS OF THE STANDARD MODEL

- Not practical or even affordable in many situations

THE LIMITATIONS OF THE STANDARD MODEL

- Not practical or even affordable in many situations
- Ways to “get around” somewhat unsatisfactory (add strong structural assumptions on p) [BKR04, DDS+13, DKN15b, DKN15a]

THE LIMITATIONS OF THE STANDARD MODEL

- Not practical or even affordable in many situations
- Ways to “get around” somewhat unsatisfactory (add strong structural assumptions on p) [BKR04, DDS+13, DKN15b, DKN15a]
- Standard sampling model too conservative?

THE LIMITATIONS OF THE STANDARD MODEL

- Not practical or even affordable in many situations
- Ways to "get around" somewhat unsatisfactory (add strong structural assumptions on p) [BKR04, DDS+13, DKN15b, DKN15a]
- Standard sampling model too conservative?

"In practice"

The algorithm may have stronger type of access to the data. Does that help?

THE LIMITATIONS OF THE STANDARD MODEL

- Not practical or even affordable in many situations
- Ways to "get around" somewhat unsatisfactory (add strong structural assumptions on p) [BKR04, DDS+13, DKN15b, DKN15a]
- Standard sampling model too conservative?

"In practice"

The algorithm may have stronger type of access to the data. Does that help? How to model it?

THE LIMITATIONS OF THE STANDARD MODEL

- Not practical or even affordable in many situations
- Ways to "get around" somewhat unsatisfactory (add strong structural assumptions on p) [BKR04, DDS+13, DKN15b, DKN15a]
- Standard sampling model too conservative?

"In practice"

The algorithm may have stronger type of access to the data. Does that help? How to model it?

(Also, can this help "in theory" as well?)

Conditional sampling models [CRS14, CRS15, ACK15, Can15a]

Definition (Conditional oracle)

Fix a distribution p over $[n]$. A conditional oracle for p, denoted COND_p, is defined as follows: given as input a query set $S \subseteq [n]$ that has $p(S) > 0$, it returns an element $i \in S$, where $i \in S$ is returned with probability $p_S(i) = p(i)/p(S)$ (independently of all previous calls).

Conditional sampling models [CRS14, CRS15, ACK15, Can15a]

Definition (Conditional oracle)

Fix a distribution p over $[n]$. A conditional oracle for p, denoted COND_p, is defined as follows: given as input a query set $S \subseteq [n]$ that has $p(S) > 0$, it returns an element $i \in S$, where $i \in S$ is returned with probability $p_S(i) = p(i)/p(S)$ (independently of all previous calls).

Remarks

Conditional sampling models [CRS14, CRS15, ACK15, Can15a]

Definition (Conditional oracle)

Fix a distribution p over $[n]$. A conditional oracle for p, denoted COND_p, is defined as follows: given as input a query set $S \subseteq [n]$ that has $p(S) > 0$, it returns an element $i \in S$, where $i \in S$ is returned with probability $p_S(i) = p(i)/p(S)$ (independently of all previous calls).

Remarks

- generalizes the SAMP oracle ($S = [n]$), but allows adaptivity;

Conditional sampling models [CRS14, CRS15, ACK15, Can15a]

Definition (Conditional oracle)

Fix a distribution p over $[n]$. A conditional oracle for p, denoted COND_p, is defined as follows: given as input a query set $S \subseteq [n]$ that has $p(S) > 0$, it returns an element $i \in S$, where $i \in S$ is returned with probability $p_S(i) = p(i)/p(S)$ (independently of all previous calls).

Remarks

- generalizes the SAMP oracle ($S = [n]$), but allows adaptivity;
- variants which only allow some specific types of subsets to be queried: PCOND (either $[n]$ or pairs $\{i, j\}$) and ICOND (intervals);

Conditional sampling models [CRS14, CRS15, ACK15, Can15a]

Definition (Conditional oracle)

Fix a distribution p over $[n]$. A conditional oracle for p, denoted COND_p, is defined as follows: given as input a query set $S \subseteq [n]$ that has $p(S) > 0$, it returns an element $i \in S$, where $i \in S$ is returned with probability $p_S(i) = p(i)/p(S)$ (independently of all previous calls).

Remarks

- generalizes the SAMP oracle ($S = [n]$), but allows adaptivity;
- variants which only allow some specific types of subsets to be queried: PCOND (either $[n]$ or pairs $\{i, j\}$) and ICOND (intervals);
- not defined for sets S with zero probability under p.

(Similar model introduced in [CFGM13].)

Conditional sampling models [CRS14, CRS15, ACK15, Can15a]

Practical situations: more control

- Experiment: can influence the outcome by tuning the conditions
- Polling: surveys and stratified sampling

Conditional sampling models [CRS14, CRS15, ACK15, Can15a]

Practical situations: more control

- Experiment: can influence the outcome by tuning the conditions
- Polling: surveys and stratified sampling

New insights

- Understanding where the core difficulties of problems lie: e.g., identity is easy [CRS14], [FJO+15], equivalence stays "hard" [ACK15]

Conditional sampling models [CRS14, CRS15, ACK15, Can15a]

Practical situations: more control

- Experiment: can influence the outcome by tuning the conditions
- Polling: surveys and stratified sampling

New insights

- Understanding where the core difficulties of problems lie: e.g., identity is easy [CRS14], [FJO+15], equivalence stays "hard" [ACK15]

... and more

- Uses in other areas (Algorithms) [GTZ17]

CONDITIONAL SAMPLING MODELS [CRS14, CRS15, ACK15, CAN15A]

Problem	Conditional model		Standard model
uniformity	COND_p	$\Omega\left(\frac{1}{\varepsilon^2}\right)$	$\Theta\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ [GR00, BFR+13, Pan08]
	PCOND_p	$\tilde{O}\left(\frac{1}{\varepsilon^2}\right)$	
	ICOND_p	$\tilde{O}\left(\frac{\log^3 n}{\varepsilon^3}\right), \Omega\left(\frac{\log n}{\log\log n}\right)$	
identity	COND_p	$\tilde{O}\left(\frac{1}{\varepsilon^2}\right)$ [FJO+15]	
	PCOND_p	$\tilde{O}\left(\frac{\log^4 n}{\varepsilon^4}\right), \Omega\left(\sqrt{\frac{\log n}{\log\log n}}\right)$	$\Theta\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ [BFF+01, Pan08, VV17]
equivalence	COND_{p_1,p_2}	$\tilde{O}\left(\frac{\log\log n}{\varepsilon^5}\right)$ [FJO+15], $\tilde{O}\left(\frac{\log^5 n}{\varepsilon^4}\right)$ $\Omega\left(\sqrt{\log\log n}\right)$	$\Theta\left(\max\left(\frac{n^{2/3}}{\varepsilon^{4/3}}, \frac{\sqrt{n}}{\varepsilon^2}\right)\right)$ [BFR+13, Val11, CDVV14]
	PCOND_{p_1,p_2}	$\tilde{O}\left(\frac{\log^6 n}{\varepsilon^{21}}\right)$	
monotonicity	COND_p	$\tilde{O}\left(\frac{1}{\varepsilon^{22}}\right)$	$\Theta\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ [BKR04, CDGR16, ADK15, Pan08]
	PCOND_p	$\tilde{O}\left(\frac{\log^2 n}{\varepsilon^4}\right)$	
	ICOND_p	$\tilde{O}\left(\frac{\log^5 n}{\varepsilon^4}\right)$	
tolerant uniformity	PCOND_p	$\tilde{O}\left(\frac{1}{\varepsilon^{20}}\right)$	$\Theta\left(\frac{1}{\varepsilon^2}\frac{n}{\log(n/\varepsilon)}\right)$ [VV11, VV10b, VV10a, JYW17]

Table: Comparison between COND and SAMP on a sample of problems.

Dual access models [CR14, Can15a]

Definition (Dual oracle)

Fix a distribution p over [n]. A dual oracle for p is a pair of oracles $(\mathrm{SAMP}_p, \mathrm{EVAL}_p)$:

- sampling oracle: SAMP_p returns $i \in [n]$ drawn from p;
- evaluation oracle: EVAL_p takes $j \in [n]$, and returns $p(j)$.

Definition (Cumulative Dual oracle)

A cumulative dual oracle for p is a pair of oracles $(\mathrm{SAMP}_p, \mathrm{CEVAL}_p)$:

- sampling oracle: SAMP_p as above;
- cumulative evaluation oracle: CEVAL_p takes $j \in [n]$, and returns $p([j]) = \sum_{i=1}^{j} p(i)$.

(Considered, although not systematically, in [BDKR05, GMV06] and [BKR04].)

Dual access models [CR14, Can15a]

Broke Arthur & Greedy Merlin

- Free but huge dataset out there
- Long and expensive analysis of it held by Merlin
- Computationally limited Arthur working on the data

Dual access models [CR14, Can15a]

Broke Arthur & Greedy Merlin

- Free but huge dataset out there
- Long and expensive analysis of it held by Merlin
- Computationally limited Arthur working on the data

Someone did the work!

- Google n-gram data: pdf for sequences of n words + samples of sequences
- Sorted files: samples in $O(1)$ time, cdf and pdf queries $O(\log n)$

Dual access models [CR14, Can15a]

Broke Arthur & Greedy Merlin

- Free but huge dataset out there
- Long and expensive analysis of it held by Merlin
- Computationally limited Arthur working on the data

Someone did the work!

- Google n-gram data: pdf for sequences of n words + samples of sequences
- Sorted files: samples in $O(1)$ time, cdf and pdf queries $O(\log n)$

...and more.

- Connection to streaming algorithms [GMV06]
- Uses in other areas of Theory [CGG+17]

DUAL ACCESS MODELS [CR14, CAN15A]

Problem	EVAL	Dual access	Cumulative Dual
uniformity	$O\left(\frac{1}{\varepsilon}\right)$ [RS09], $\Omega\left(\frac{1}{\varepsilon}\right)$	$\Theta\left(\frac{1}{\varepsilon}\right)$	$\Theta\left(\frac{1}{\varepsilon}\right)$
Testing identity			
equivalence	$\Omega(n)$		
tolerant uniformity	$\Omega\left(\frac{1}{(\varepsilon_2-\varepsilon_1)^2}\right)$	$\Theta\left(\frac{1}{(\varepsilon_2-\varepsilon_1)^2}\right)$	$O\left(\frac{1}{(\varepsilon_2-\varepsilon_1)^2}\right)$
Tolerant identity			
tolerant equivalence			
monotonicity	$O\left(\frac{\log n}{\varepsilon}+\frac{1}{\varepsilon^2}\right), \Omega\left(\frac{\log n}{\log\log n}\right)$	(upper bound from EVAL)	$\tilde{O}\left(\frac{1}{\varepsilon^4}\right)$
tolerant monotonicity		$O\left(\frac{\log n}{\varepsilon_2^3}\right)$ for $\varepsilon_2 > 3.1\varepsilon_1$	$O\left(\frac{1}{\varepsilon_2^2}+\frac{\log n}{\varepsilon_2}\right)$ for $\varepsilon_2 > 3.1\varepsilon_1$

Table: Comparison between EVAL, Dual access and Cumulative Dual.

Changing the goal: Be the change that…

MOTIVATION

Often, the distribution on the data has particular, useful structure that algorithms can exploit (monotone pmf, uniform distribution, independent components...)

MOTIVATION

Often, the distribution on the data has particular, useful structure that algorithms can exploit (monotone pmf, uniform distribution, independent components...)

But in many situations, sample data comes from noisy or imperfect sources, tampering with these properties.

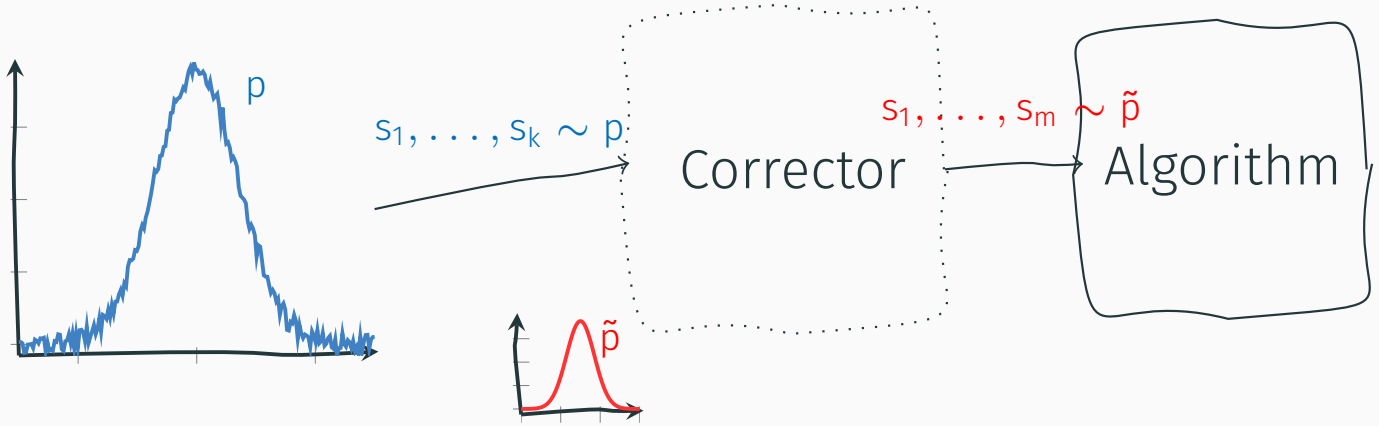
MOTIVATION

Often, the distribution on the data has particular, useful structure that algorithms can exploit (monotone pmf, uniform distribution, independent components...)

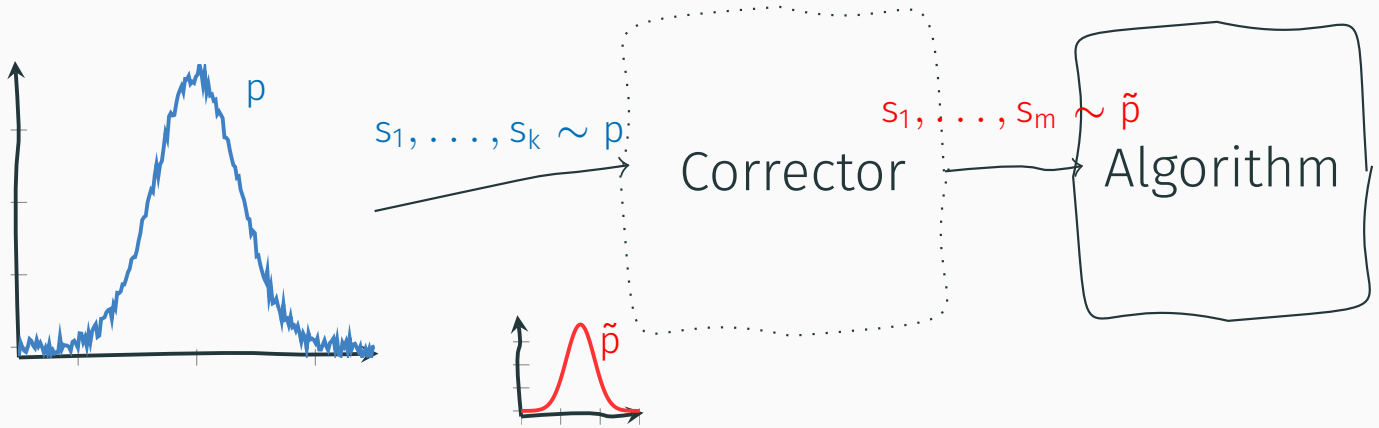
But in many situations, sample data comes from noisy or imperfect sources, tampering with these properties.

Can we still exploit the structure that the distribution should have had?

CORRECTING DISTRIBUTIONS



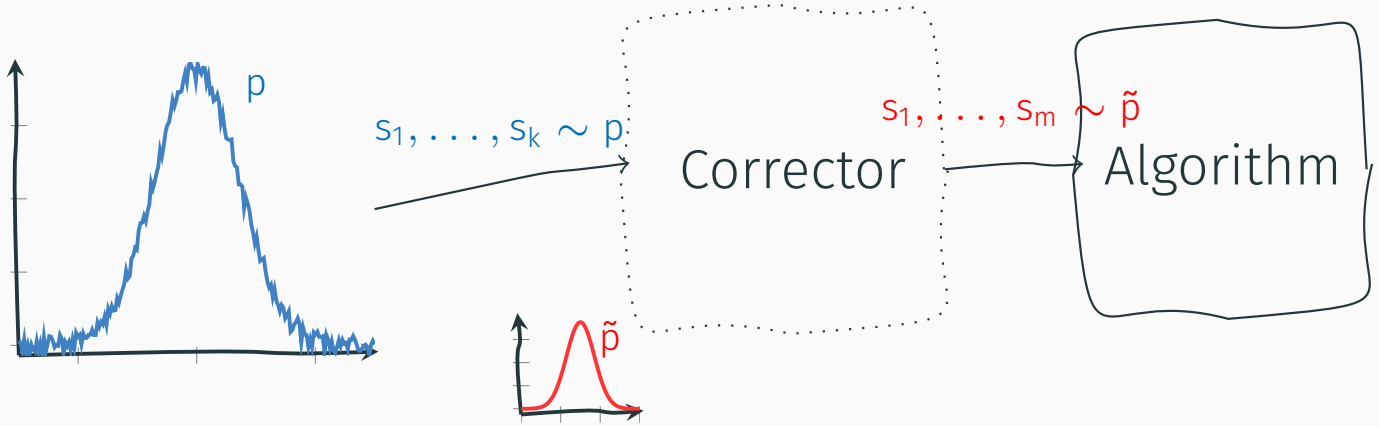
CORRECTING DISTRIBUTIONS



Fix a specific property $\mathcal{P}$ of distributions. (application-dependent)

- independent samples from a p promised to be ε-close to $\mathcal{P}$
- want independent samples from some $\tilde{p}$ which:
  - has the property: $\tilde{p} \in \mathcal{P}$;
  - remains faithful to the data: $d_{\mathrm{TV}}(\tilde{p}, p) = O(\varepsilon)$.

CORRECTING DISTRIBUTIONS



Fix a specific property $\mathcal{P}$ of distributions. (application-dependent)

- independent samples from a p promised to be ε-close to $\mathcal{P}$
- want independent samples from some $\tilde{p}$ which:
  - has the property: $\tilde{p} \in \mathcal{P}$;
  - remains faithful to the data: $d_{\mathrm{TV}}(\tilde{p}, p) = O(\varepsilon)$.

Similar in spirit to the "local filters" for functions [ACCL08, SS10, JR11, BGJ+12].

CORRECTING DISTRIBUTIONS

- neither learning nor testing

CORRECTING DISTRIBUTIONS

- neither learning nor testing
- connections to both (and others: tolerant testing, agnostic learning...)

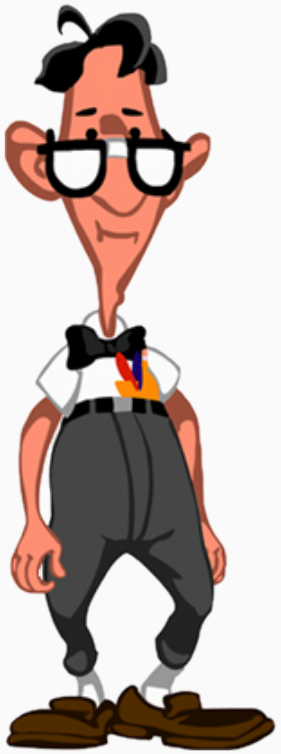
CORRECTING DISTRIBUTIONS

- neither learning nor testing
- connections to both (and others: tolerant testing, agnostic learning...)
- systematic investigation of this model

CORRECTING DISTRIBUTIONS

- neither learning nor testing
- connections to both (and others: tolerant testing, agnostic learning...)
- systematic investigation of this model
- (much more left to explore)

FUTURE DIRECTIONS: WHAT NOW?



Thank you.

And thank you, Akash, Alaa, Alex, Alice, Alistair, Amit, Anindya, Dana, Daniel, Dimitris, Elena, Eric, Erik, Fay, Fernando, Gautam, Igor, Ilias, Ilya, Jayadev, Jerry, Jessica, Joschi, Juba, Karl, Laurent, Li-Yang, Madhu, Mihalis, Narges, Omri, Raghu, Rémi, Reut, Rocco, Ronitt, Tal, Talya, Themis, Tom, (other) Tom, Tuğkan, Venkat, Vicky, Xi, Yumi. Thank you, Nandini, Maman, Papa, Thomas, Cécile, Marion, Nicolas, Quentin, Margot, Sacha, and tiny Lola.

Nir Ailon, Bernard Chazelle, Seshadhri Comandur, and Ding Liu.
Property-preserving data reconstruction.
Algorithmica, 51(2):160–182, 2008.

Jayadev Acharya, Clément L. Canonne, and Gautam Kamath.
A chasm between identity and equivalence testing with conditional queries.
In APPROX-RANDOM, volume 40 of LIPIcs, pages 449–466, 2015.

Jayadev Acharya and Constantinos Daskalakis.
Testing Poisson Binomial Distributions.
In Proceedings of SODA, pages 1829–1840, 2015.

Jayadev Acharya, Constantinos Daskalakis, and Gautam C. Kamath.
Optimal Testing for Properties of Distributions.
In C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, R. Garnett, and R. Garnett, editors, Advances in Neural Information Processing Systems 28, pages 3577–3598. Curran Associates, Inc., 2015.

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.

Distribution testing lower bounds via reductions from communication complexity.

In Computational Complexity Conference, volume 79 of LIPIcs, pages 28:1–28:40. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2017.

Tuğkan Batu, Sanjoy Dasgupta, Ravi Kumar, and Ronitt Rubinfeld.

The complexity of approximating the entropy.

SIAM Journal on Computing, 35(1):132–150, 2005.

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, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White.

Testing closeness of discrete distributions.

Journal of the ACM, 60(1):4:1–4:25, 2013.

This is a long version of [BFR+00].


Arnab Bhattacharyya, Eldar Fischer, Ronitt Rubinfeld, and Paul Valiant.

Testing monotonicity of distributions over general partial orders.

In Proceedings of ITCS, pages 239–252, 2011.


Arnab Bhattacharyya, Elena Grigorescu, Madhav Jha, Kyomin Jung, Sofya Raskhodnikova, and David P. Woodruff.

Lower bounds for local monotonicity reconstruction from transitive-closure spanners.

SIAM Journal on Discrete Mathematics, 26(2):618–646, 2012.


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.


Arnab Bhattacharyya and Yuichi Yoshida.

Property Testing.

Forthcoming, 2017.


Clément L. Canonne.

Big Data on the Rise? Testing Monotonicity of Distributions.

In Proceedings of ICALP, pages 294–305. Springer, 2015.


Clément L. Canonne.

A Survey on Distribution Testing: your data is Big. But is it Blue?

Electronic Colloquium on Computational Complexity (ECCC), 22:63, April 2015.


Clément L. Canonne, Ilias Diakonikolas, Themis Gouleakis, and Ronitt Rubinfeld.
Testing Shape Restrictions of Discrete Distributions.
In Proceedings of STACS, 2016.
See also [CDGR17] (full version).


Clément L. Canonne, Ilias Diakonikolas, Themis Gouleakis, and Ronitt Rubinfeld.
Testing shape restrictions of discrete distributions.
Theory of Computing Systems, pages 1–59, 2017.


Yu Cheng, Ilias Diakonikolas, and Alistair Stewart.
Playing anonymous games using simple strategies.
In Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, Proceedings of SODA, pages 616–631, Philadelphia, PA, USA, 2017. Society for Industrial and Applied Mathematics.


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, 2014.


Sourav Chakraborty, Eldar Fischer, Yonatan Goldhirsh, and Arie Matsliah.
On the power of conditional samples in distribution testing.
In Proceedings of ITCS, pages 561–580, New York, NY, USA, 2013. ACM.


Clément L. Canonne, Elena Grigorescu, Siyao Guo, Akash Kumar, and Karl Wimmer.

Testing k-monotonicity.

In 8th Innovations in Theoretical Computer Science (ITCS). ACM, 2017.


Clément L. Canonne and Ronitt Rubinfeld.

Testing probability distributions underlying aggregated data.

In Proceedings of ICALP, pages 283–295, 2014.


Clément L. Canonne, Dana Ron, and Rocco A. Servedio.

Testing equivalence between distributions using conditional samples.

In Proceedings of SODA, pages 1174–1192. Society for Industrial and Applied Mathematics (SIAM), 2014.


Clément L. Canonne, Dana Ron, and Rocco A. Servedio.

Testing probability distributions using conditional samples.

SIAM Journal on Computing, 44(3):540–616, 2015.


Constantinos Daskalakis, Ilias Diakonikolas, Rocco A. Servedio, Gregory Valiant, and Paul Valiant.

Testing k-modal distributions: Optimal algorithms via reductions.

In Proceedings of SODA, pages 1833–1852. Society for Industrial and Applied Mathematics (SIAM), 2013.


Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price.
Collision-based testers are optimal for uniformity and closeness.
Electronic Colloquium on Computational Complexity (ECCC), 23:178, 2016.


Ilias Diakonikolas and Daniel M. Kane.
A new approach for testing properties of discrete distributions.
In Proceedings of FOCS. IEEE Computer Society, 2016.


Ilias Diakonikolas, Daniel M. Kane, and Vladimir Nikishkin.
Optimal algorithms and lower bounds for testing closeness of structured distributions.
In Proceedings of FOCS, 2015.


Ilias Diakonikolas, Daniel M. Kane, and Vladimir Nikishkin.
Testing Identity of Structured Distributions.
In Proceedings of SODA, 2015.


Moein Falahatgar, Ashkan Jafarpour, Alon Orlitsky, Venkatadheeraj Pichapathi, and Ananda Theertha Suresh.
Faster algorithms for testing under conditional sampling.
In Proceedings of COLT, JMLR Proceedings, pages 607–636, 2015.


Oded Goldreich, Shafi Goldwasser, and Dana Ron.
Property testing and its connection to learning and approximation.

Journal of the ACM, 45(4):653–750, July 1998.

Sudipto Guha, Andrew McGregor, and Suresh Venkatasubramanian.

Streaming and sublinear approximation of entropy and information distances.

In Proceedings of SODA, pages 733–742, Philadelphia, PA, USA, 2006. Society for Industrial and Applied Mathematics (SIAM).

Oded Goldreich, editor.

Property Testing: Current Research and Surveys.

Springer, 2010.

LNCS 6390.

Oded Goldreich.

Introduction to Property Testing.

Forthcoming, 2017.

Oded Goldreich and Dana Ron.

On testing expansion in bounded-degree graphs.

Technical Report TR00-020, Electronic Colloquium on Computational Complexity (ECCC), 2000.

Themistoklis Gouleakis, Christos Tzamos, and Manolis Zampetakis.

Faster sublinear algorithms using conditional sampling.

In Proceedings of SODA, pages 1743–1757. SIAM, 2017.

Madhav Jha and Sofya Raskhodnikova.
Testing and reconstruction of Lipschitz functions with applications to data privacy.
In Proceedings of FOCS, pages 433–442, Oct 2011.

Jiantao Jiao, Kartik Venkat, Han Yanjun, and Tsachy Weissman.
Minimax estimation of functionals of discrete distributions.
IEEE Transactions on Information Theory, 61(5):2835–2885, May 2015.

Jiantao Jiao, Han Yanjun, and Tsachy Weissman.
Minimax Estimation of the L_1 Distance.
ArXiv e-prints, May 2017.

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

Ilan Newman and Mario Szegedy.
Public vs. private coin flips in one round communication games.
In Proceedings of STOC, 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.

Dana Ron.

Property Testing: A Learning Theory Perspective.

Foundations and Trends in Machine Learning, 1(3):307–402, 2008.

Dana Ron.

Algorithmic and analysis techniques in property testing.

Foundations and Trends in Theoretical Computer Science, 5:73–205, 2010.

Ronitt Rubinfeld and Madhu Sudan.

Robust characterization of polynomials with applications to program testing.

SIAM Journal on Computing, 25(2):252–271, 1996.

Ronitt Rubinfeld and Rocco A. Servedio.

Testing monotone high-dimensional distributions.

Random Structures and Algorithms, 34(1):24–44, January 2009.

Ronitt Rubinfeld.

Taming big probability distributions.

XRDS: Crossroads, The ACM Magazine for Students, 19(1):24, sep 2012.

Michael Saks and Comandur Seshadhri.

Local monotonicity reconstruction.

SIAM Journal on Computing, 39(7):2897–2926, 2010.


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


Gregory Valiant and Paul Valiant.
A CLT and tight lower bounds for estimating entropy.
Electronic Colloquium on Computational Complexity (ECCC), 17:179, 2010.


Gregory Valiant and Paul Valiant.
Estimating the unseen: A sublinear-sample canonical estimator of distributions.
Electronic Colloquium on Computational Complexity (ECCC), 17:180, 2010.


Gregory Valiant and Paul Valiant.
The power of linear estimators.
In Proceedings of FOCS, pages 403–412, October 2011.
See also [VV10a] and [VV10b].


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


Gregory Valiant and Paul Valiant.
An automatic inequality prover and instance optimal identity testing.
SIAM Journal on Computing, 46(1):429–455, 2017.

IEEE Transactions on Information Theory, 62(6):3702–3720, June 2016.

BACKUP SLIDES

Theorem ([BCG17])

Let $\mathcal{P} \subseteq \Delta([n])$, predicate $f\colon \{0,1\}^k \times \{0,1\}^k \to \{0,1\}$, and $\varepsilon > 0$. Suppose there exists $p\colon \{0,1\}^k \times \{0,1\}^k \to \Delta([n])$ satisfying:

1. Decomposability: For every $x, y \in \{0,1\}^k$, there exist constants* $\alpha = \alpha(x), \beta = \beta(y) \in [0,1]$ and distributions $p_A(x), p_B(y)$ s.t.
$$p(x,y) = \frac{\alpha}{\alpha+\beta} \cdot p_A(x) + \frac{\beta}{\alpha+\beta} \cdot p_B(y)$$
2. Completeness: For all $(x,y) \in f^{-1}(1)$, $p(x,y) \in \mathcal{P}$.
3. Soundness: For all $(x,y) \in f^{-1}(0)$, $p(x,y)$ is ε-far from $\mathcal{P}$.

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

REDUCTIONS: ALICE, BOB, AND THE REFEREE

Take the "equality" predicate EQ_k as f:

Theorem (Newman and Szegedy [NS96])

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

REDUCTIONS: ALICE, BOB, AND THE REFEREE

Take the "equality" predicate EQ_k as f:

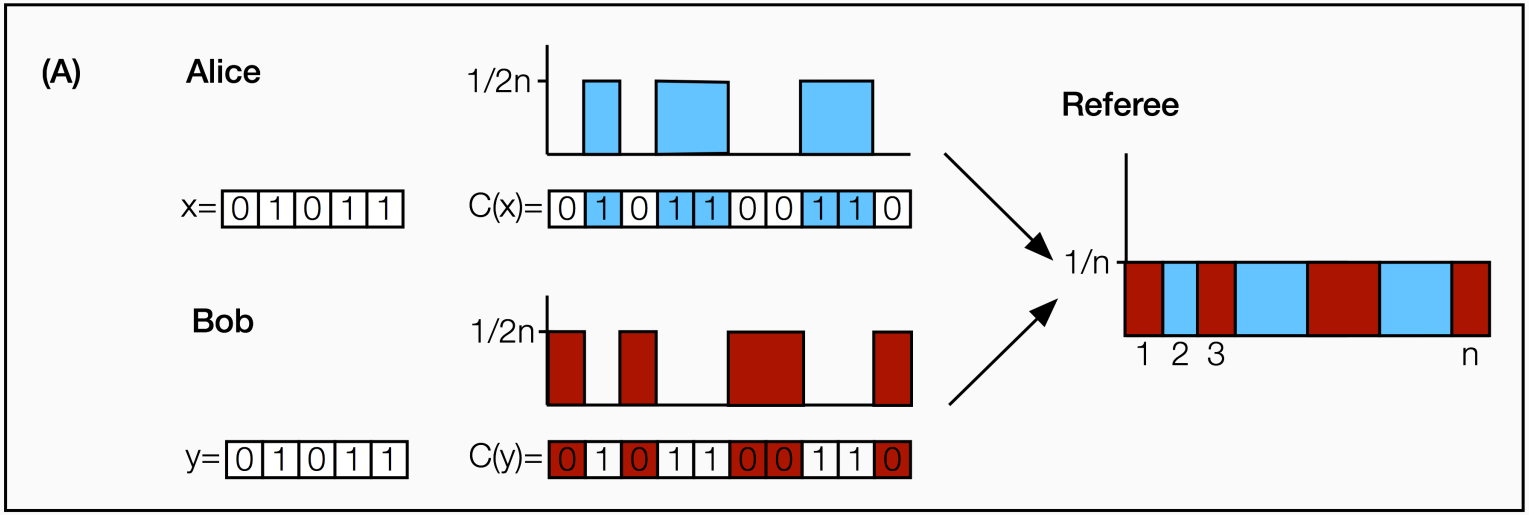
Theorem (Newman and Szegedy [NS96])

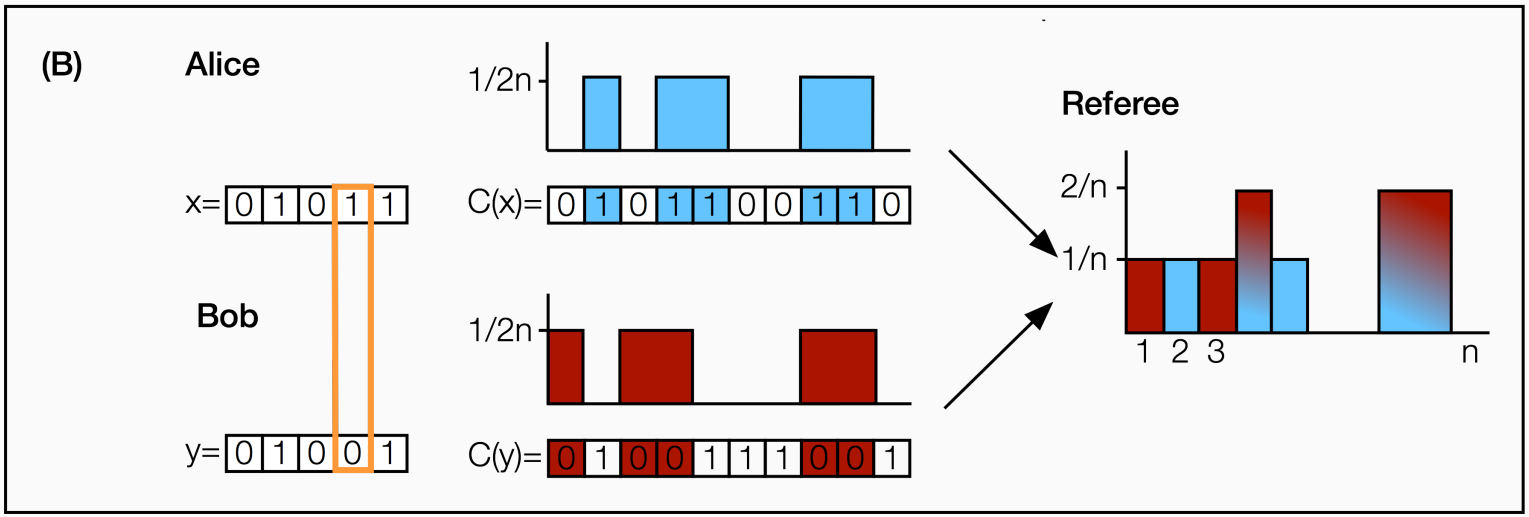
For every $k \in \mathbb{N}$, $\mathrm{SMP}(\mathrm{EQ}_k) = \Omega\left(\sqrt{k}\right)$.

Example:

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

REDUCTIONS: ALICE, BOB, AND THE REFEREE





Applications

Testing monotonicity, unimodality, k-modality, log-concavity, Poisson Binomial Distributions, symmetric sparse support,

Applications

Testing monotonicity, unimodality, k-modality, log-concavity, Poisson Binomial Distributions, symmetric sparse support, identity*...

Applications

Testing monotonicity, unimodality, k-modality, log-concavity, Poisson Binomial Distributions, symmetric sparse support, identity*...

The * above

- unexpected connection to interpolation theory
- led to new insights: "instance-optimal" identity testing [VV17], revisited:

$$\|q^{-\max}_{-\Theta(\varepsilon)}\|_{2/3} \text{ [VV17]} \quad \rightsquigarrow \quad \kappa_q^{-1}(1-\Theta(\varepsilon)) \text{ [BCG17]}$$