Testing Distributions

Candidacy Talk

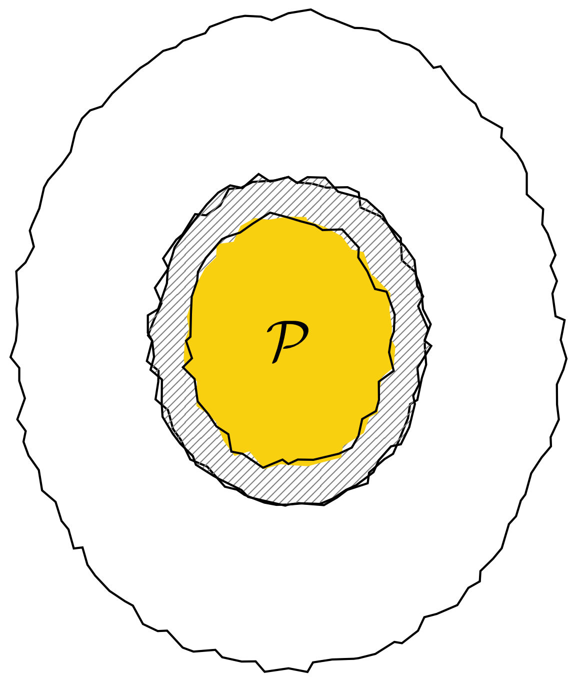
Clément Canonne

Columbia University – 2015

Introduction

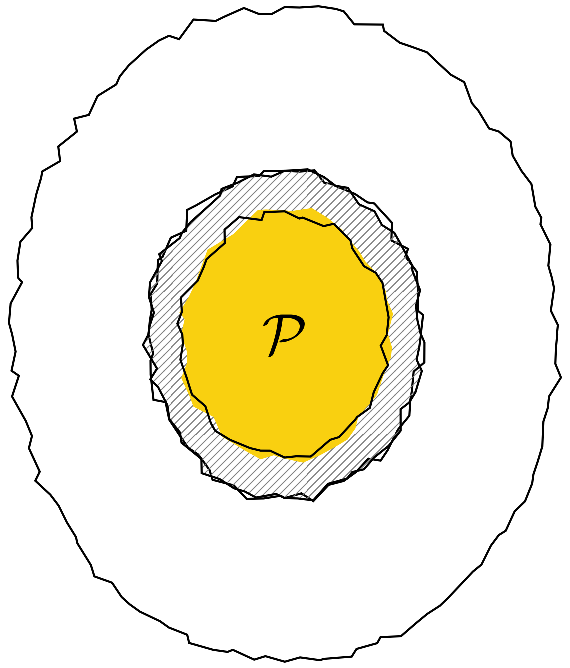
Introduction

Property testing: what can we say about an object while barely looking at it?



Introduction

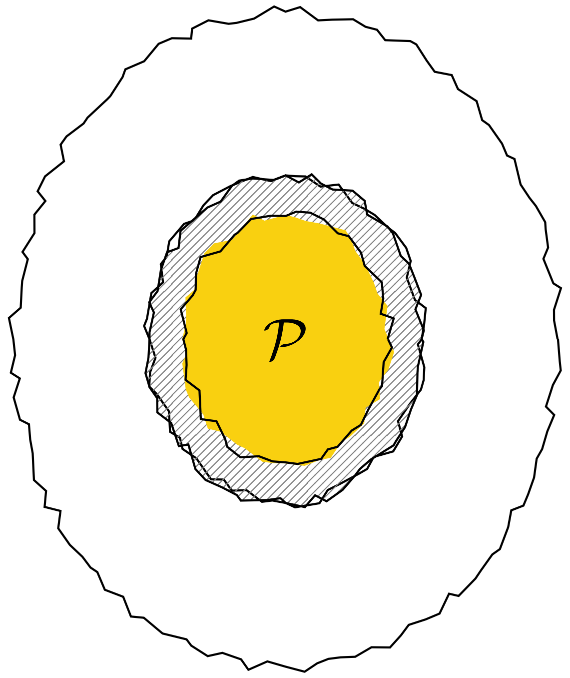
Property testing: what can we say about an object while barely looking at it?



"Is it in the yolk?"

Introduction

Property testing: what can we say about an object while barely looking at it?

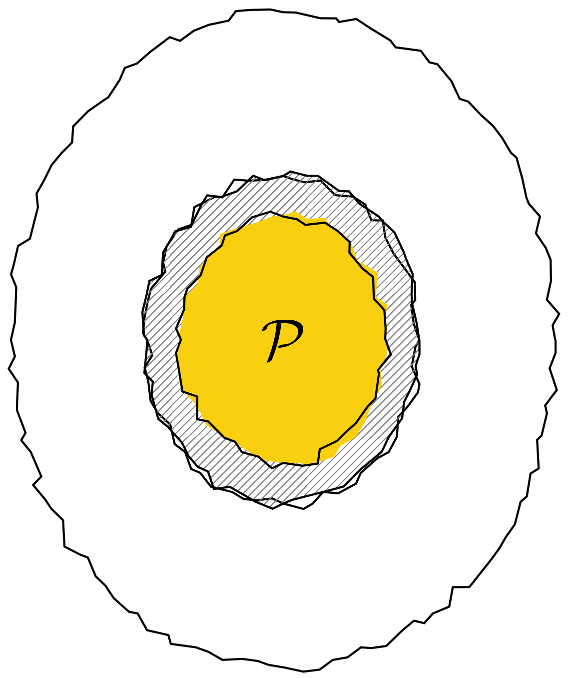


"Is it in the yolk?"

This talk: distribution testing, for various types of properties and settings.

Introduction

Property testing: what can we say about an object while barely looking at it?



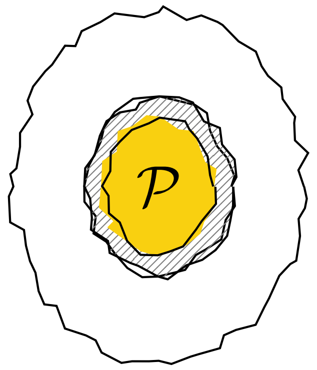
“Is it in the yolk?”

This talk: distribution testing, for various types of properties and settings.
(what is known, what is impossible, and under which assumptions can it still be done)

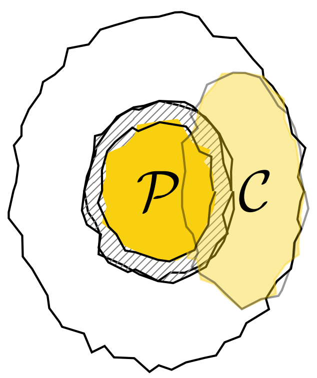
Outline of the talk

Plan in more detail

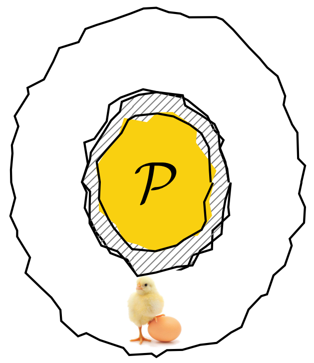
- Testing From Samples: the standard model, upper and lower bounds



- Testing Under Assumptions: "testing for $\mathcal{P}$ while knowing $\mathcal{C}$"



- Testing Differently: some other access (stronger or incomparable), or some other goal



Testing From Samples

The setting

$\Delta(\Omega)$: all distributions over (finite) domain Ω of size n. **Property:** subset $\mathcal{P} \subseteq \Delta(\Omega)$. **Tester:** randomized algorithm (knows n, $\mathcal{P}$).

Given independent samples from a distribution $D \in \Delta(\Omega)$, and parameter $\varepsilon \in (0,1)$, output accept or reject:

- If $D \in \mathcal{P}$, accept with probability at least $2/3$; *(in the yolk)*
- If $\ell_1(D, \mathcal{P}) > \varepsilon$, reject with probability at least $2/3$; *(definitely white)*
- otherwise, whatever (make an omelet).

Goal: take $o(n)$ samples, ideally $O_\varepsilon(1)$.

The setting

$\Delta(\Omega)$: all distributions over (finite) domain Ω of size n. **Property:** subset $\mathcal{P} \subseteq \Delta(\Omega)$. **Tester:** randomized algorithm (knows n, $\mathcal{P}$).

Given independent samples from a distribution $D \in \Delta(\Omega)$, and parameter $\varepsilon \in (0, 1)$, output accept or reject:

- If $D \in \mathcal{P}$, accept with probability at least $2/3$; *(in the yolk)*
- If $\ell_1(D, \mathcal{P}) > \varepsilon$, reject with probability at least $2/3$; *(definitely white)*
- otherwise, whatever (make an omelet).

Goal: take $o(n)$ samples, ideally $O_\varepsilon(1)$.
(time efficiency is secondary, yet not frowned upon.)

The setting

$\Delta(\Omega)$: all distributions over (finite) domain Ω of size n. **Property:** subset $\mathcal{P} \subseteq \Delta(\Omega)$. **Tester:** randomized algorithm (knows n, $\mathcal{P}$).

Given independent samples from a distribution $D \in \Delta(\Omega)$, and parameter $\varepsilon \in (0,1)$, output accept or reject:

- If $D \in \mathcal{P}$, accept with probability at least $2/3$; *(in the yolk)*
- If $\ell_1(D, \mathcal{P}) > \varepsilon$, reject with probability at least $2/3$; *(definitely white)*
- otherwise, whatever (make an omelet).

Goal: take $o(n)$ samples, ideally $O_\varepsilon(1)$.
(time efficiency is secondary, yet not frowned upon.)

[BFF+01, BKR04, BFR+10, GGR98]

Results: a representative sample

Is it possible?

Testing uniformity has sample complexity $O(\sqrt{n}/\varepsilon^2)$ [GR00, Pan08].

Results: a representative sample

Is it possible?

Testing uniformity has sample complexity $O(\sqrt{n}/\varepsilon^2)$ [GR00, Pan08]. Testing identity has sample complexity $O(\sqrt{n}/\varepsilon^2)$ [BFF$^+$01, VV14, DKN15].

Results: a representative sample

Is it possible?

Testing uniformity has sample complexity $O(\sqrt{n}/\varepsilon^2)$ [GR00, Pan08]. Testing identity has sample complexity $O(\sqrt{n}/\varepsilon^2)$ [BFF+01, VV14, DKN15]. Testing closeness has sample complexity $O(n^{2/3}/\varepsilon^{4/3})$ [BFR+10, CDVV14].

Results: a representative sample

Is it possible?

Testing uniformity has sample complexity $O(\sqrt{n}/\varepsilon^2)$ [GR00, Pan08]. Testing identity has sample complexity $O(\sqrt{n}/\varepsilon^2)$ [BFF+01, VV14, DKN15]. Testing closeness has sample complexity $O(n^{2/3}/\varepsilon^{4/3})$ [BFR+10, CDVV14].

Can we do better?

Testing uniformity has sample complexity $\Omega(\sqrt{n}/\varepsilon^2)$ [GR00, Pan08].

Results: a representative sample

Is it possible?

Testing uniformity has sample complexity $O(\sqrt{n}/\varepsilon^2)$ [GR00, Pan08]. Testing identity has sample complexity $O(\sqrt{n}/\varepsilon^2)$ [BFF$^+$01, VV14, DKN15]. Testing closeness has sample complexity $O(n^{2/3}/\varepsilon^{4/3})$ [BFR$^+$10, CDVV14].

Can we do better?

Testing uniformity has sample complexity $\Omega(\sqrt{n}/\varepsilon^2)$ [GR00, Pan08]. Testing identity has sample complexity $\Omega(\sqrt{n}/\varepsilon^2)$ [ibid].

Results: a representative sample

Is it possible?

Testing uniformity has sample complexity $O(\sqrt{n}/\varepsilon^2)$ [GR00, Pan08]. Testing identity has sample complexity $O(\sqrt{n}/\varepsilon^2)$ [BFF+01, VV14, DKN15]. Testing closeness has sample complexity $O(n^{2/3}/\varepsilon^{4/3})$ [BFR+10, CDVV14].

Can we do better?

Testing uniformity has sample complexity $\Omega(\sqrt{n}/\varepsilon^2)$ [GR00, Pan08]. Testing identity has sample complexity $\Omega(\sqrt{n}/\varepsilon^2)$ [ibid]. Testing closeness has sample complexity $\Omega(n^{2/3}/\varepsilon^{4/3})$ [BFR+10, Val11, CDVV14].

Any more ~~good~~ news?

Tolerant testing uniformity (and a range of other interesting properties) has sample complexity $\Theta(n/\log n)$ [Pan04, RRSS09, Val11, VV10a, VV10b, VV11].

Results: a representative sample

Is it possible?

Testing uniformity has sample complexity $O(\sqrt{n}/\varepsilon^2)$ [GR00, Pan08]. Testing identity has sample complexity $O(\sqrt{n}/\varepsilon^2)$ [BFF+01, VV14, DKN15]. Testing closeness has sample complexity $O(n^{2/3}/\varepsilon^{4/3})$ [BFR+10, CDVV14].

Can we do better?

Testing uniformity has sample complexity $\Omega(\sqrt{n}/\varepsilon^2)$ [GR00, Pan08]. Testing identity has sample complexity $\Omega(\sqrt{n}/\varepsilon^2)$ [ibid]. Testing closeness has sample complexity $\Omega(n^{2/3}/\varepsilon^{4/3})$ [BFR+10, Val11, CDVV14].

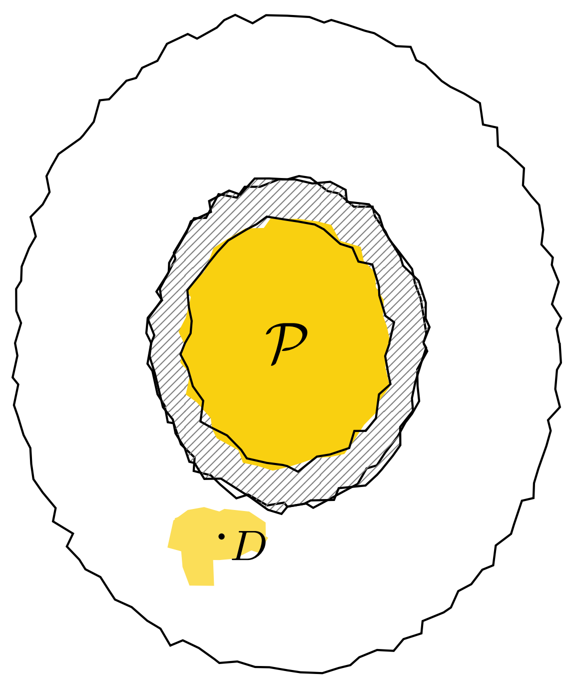
Any more ~~good~~ news?

Tolerant testing uniformity (and a range of other interesting properties) has sample complexity $\Theta(n/\log n)$ [Pan04, RRSS09, Val11, VV10a, VV10b, VV11]. (And that's what we would like to do [PRR06].)

Testing Under Assumptions: Changing The Goal

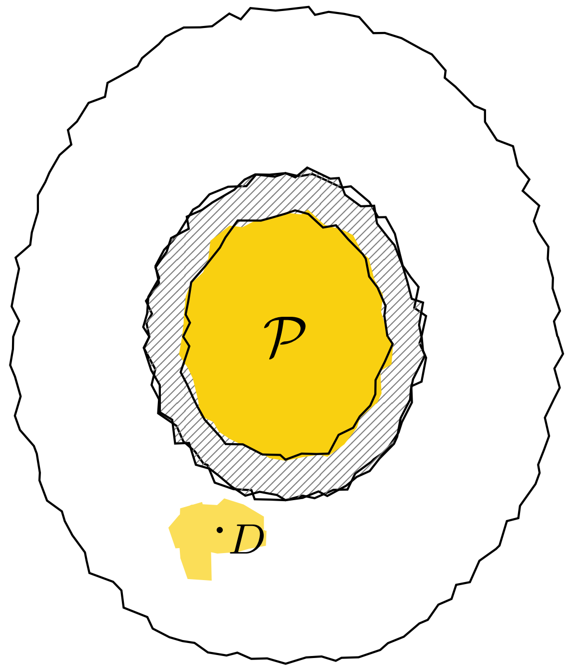
The twist

Intuitively, one key difficulty here is that in the negative case, *the distribution could be absolutely anything.* No structure to exploit!



The twist

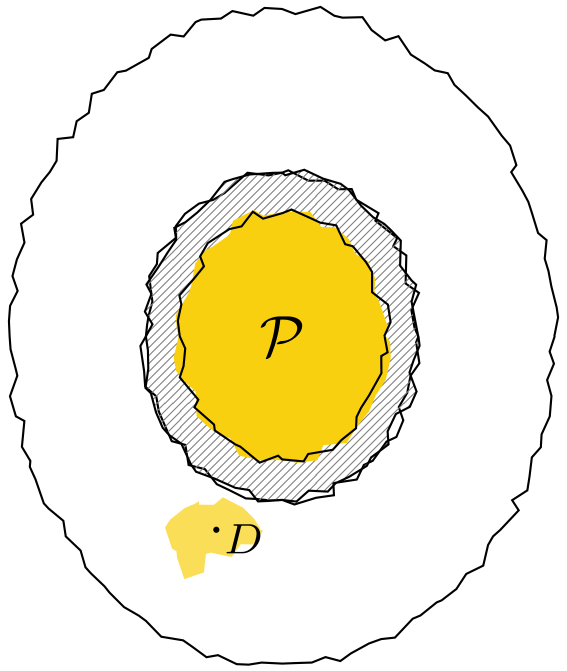
Intuitively, one key difficulty here is that in the negative case, *the distribution could be absolutely anything.* No structure to exploit!



"It surely looks like yolk, but..."

The twist

Intuitively, one key difficulty here is that in the negative case, *the distribution could be absolutely anything.* No structure to exploit!



"It surely looks like yolk, but…"

But what if D was not arbitrary? E.g., the distribution is *known* to have some structure $\mathcal{C}$ – does it make it easier to test if it *also* has the property $\mathcal{P}$?

Under shape assumptions: it's exponentially sublinear!

For monotone distributions

Testing uniformity has sample complexity $\Theta(1/\varepsilon^2)$.

Under shape assumptions: it's exponentially sublinear!

For monotone distributions

Testing uniformity has sample complexity $\Theta(1/\varepsilon^2)$. Testing identity has sample complexity $O\left(\sqrt{\log n}/\varepsilon^{5/2}\right)$ [BKR04, Bir87, DDS+13].

Under shape assumptions: it's exponentially sublinear!

For monotone distributions

Testing uniformity has sample complexity $\Theta(1/\varepsilon^2)$. Testing identity has sample complexity $O\left(\sqrt{\log n}/\varepsilon^{5/2}\right)$ [BKR04, Bir87, DDS+13].
Testing closeness has sample complexity $O(\log^{2/3} n/\varepsilon^2)$ [ibid].

Under shape assumptions: it's exponentially sublinear!

For monotone distributions

Testing uniformity has sample complexity $\Theta(1/\varepsilon^2)$. Testing identity has sample complexity $O\left(\sqrt{\log n}/\varepsilon^{5/2}\right)$ [BKR04, Bir87, DDS+13]. Testing closeness has sample complexity $O\left(\log^{2/3} n/\varepsilon^2\right)$ [ibid]. Tolerant testing closeness has sample complexity $O\left(\frac{\log n}{\varepsilon^3 \log\log n}\right)$ [DDS+13].

Under shape assumptions: it's exponentially sublinear!

For monotone distributions

Testing uniformity has sample complexity $\Theta(1/\varepsilon^2)$. Testing identity has sample complexity $O(\sqrt{\log n}/\varepsilon^{5/2})$ [BKR04, Bir87, DDS+13]. Testing closeness has sample complexity $O(\log^{2/3} n/\varepsilon^2)$ [ibid]. Tolerant testing closeness has sample complexity $O(\frac{\log n}{\varepsilon^3 \log\log n})$ [DDS+13].

For k-modal distributions [DDS+13]

Testing identity has sample complexity $O(\frac{k^2}{\varepsilon^4} + \frac{\sqrt{k \log n}}{\varepsilon^3})$ (†).

Under shape assumptions: it's exponentially sublinear!

For monotone distributions

Testing uniformity has sample complexity $\Theta(1/\varepsilon^2)$. Testing identity has sample complexity $O(\sqrt{\log n}/\varepsilon^{5/2})$ [BKR04, Bir87, DDS+13]. Testing closeness has sample complexity $O(\log^{2/3} n/\varepsilon^2)$ [ibid]. Tolerant testing closeness has sample complexity $O(\frac{\log n}{\varepsilon^3 \log\log n})$ [DDS+13].

For k-modal distributions [DDS+13]

Testing identity has sample complexity $O(\frac{k^2}{\varepsilon^4} + \frac{\sqrt{k \log n}}{\varepsilon^3})$ (†). Testing closeness has sample complexity $O(\frac{k^2}{\varepsilon^4} + \frac{(k\log n)^{2/3}}{\varepsilon^{8/3}})$.

Under shape assumptions: it's exponentially sublinear!

For monotone distributions

Testing uniformity has sample complexity $\Theta(1/\varepsilon^2)$. Testing identity has sample complexity $O\big(\sqrt{\log n}/\varepsilon^{5/2}\big)$ [BKR04, Bir87, DDS+13]. Testing closeness has sample complexity $O\big(\log^{2/3} n/\varepsilon^2\big)$ [ibid]. Tolerant testing closeness has sample complexity $O\big(\frac{\log n}{\varepsilon^3 \log\log n}\big)$ [DDS+13].

For k-modal distributions [DDS+13]

Testing identity has sample complexity $O\big(\frac{k^2}{\varepsilon^4} + \frac{\sqrt{k\log n}}{\varepsilon^3}\big)$ (†). Testing closeness has sample complexity $O\big(\frac{k^2}{\varepsilon^4} + \frac{(k\log n)^{2/3}}{\varepsilon^{8/3}}\big)$. Tolerant testing closeness has sample complexity $O\big(\frac{k^2}{\varepsilon^4} + \frac{k\log n}{\varepsilon^4 \log(k\log n)}\big)$.

Under shape assumptions: it's exponentially sublinear!

For monotone distributions

Testing uniformity has sample complexity $\Theta(1/\varepsilon^2)$. Testing identity has sample complexity $O\big(\sqrt{\log n}/\varepsilon^{5/2}\big)$ [BKR04, Bir87, DDS+13]. Testing closeness has sample complexity $O\big(\log^{2/3} n/\varepsilon^2\big)$ [ibid]. Tolerant testing closeness has sample complexity $O\big(\frac{\log n}{\varepsilon^3 \log\log n}\big)$ [DDS+13].

For k-modal distributions [DDS+13]

Testing identity has sample complexity $O\big(\frac{k^2}{\varepsilon^4} + \frac{\sqrt{k\log n}}{\varepsilon^3}\big)$ (†). Testing closeness has sample complexity $O\big(\frac{k^2}{\varepsilon^4} + \frac{(k\log n)^{2/3}}{\varepsilon^{8/3}}\big)$. Tolerant testing closeness has sample complexity $O\big(\frac{k^2}{\varepsilon^4} + \frac{k\log n}{\varepsilon^4 \log(k\log n)}\big)$.

Can we do better?

All of the above is essentially tight [DDS+13], up to logarithmic factors and dependence on ε.

Under shape assumptions: it's exponentially sublinear!

For monotone distributions

Testing uniformity has sample complexity $\Theta(1/\varepsilon^2)$. Testing identity has sample complexity $O(\sqrt{\log n}/\varepsilon^{5/2})$ [BKR04, Bir87, DDS+13]. Testing closeness has sample complexity $O(\log^{2/3} n/\varepsilon^2)$ [ibid]. Tolerant testing closeness has sample complexity $O(\frac{\log n}{\varepsilon^3 \log\log n})$ [DDS+13].

For k-modal distributions [DDS+13]

Testing identity has sample complexity $O(\frac{k^2}{\varepsilon^4} + \frac{\sqrt{k\log n}}{\varepsilon^3})$ (†). Testing closeness has sample complexity $O(\frac{k^2}{\varepsilon^4} + \frac{(k\log n)^{2/3}}{\varepsilon^{8/3}})$. Tolerant testing closeness has sample complexity $O(\frac{k^2}{\varepsilon^4} + \frac{k\log n}{\varepsilon^4 \log(k\log n)})$.

Can we do better?

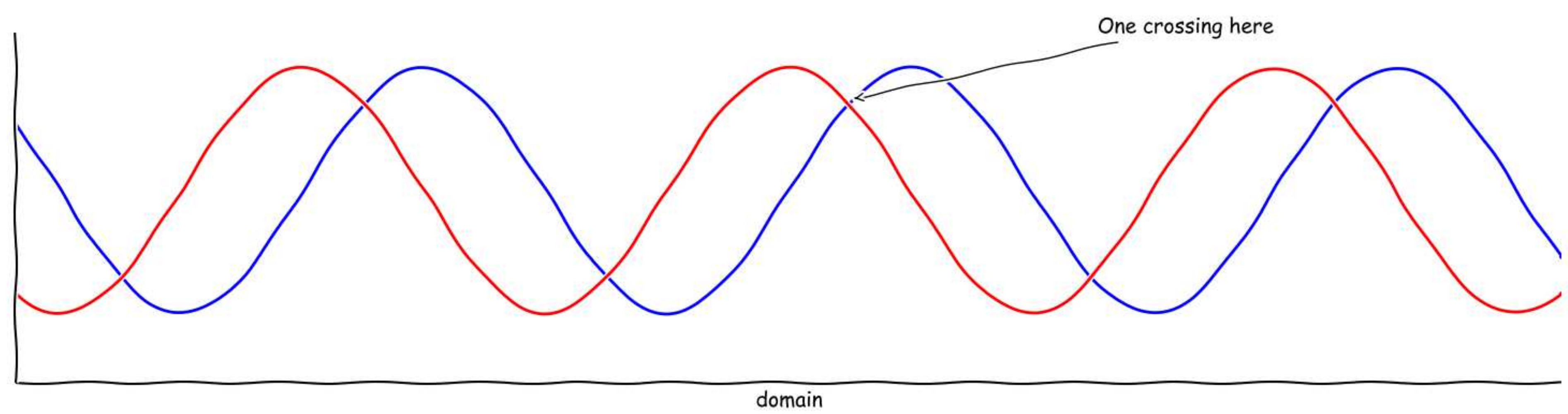
All of the above is essentially tight [DDS+13], up to logarithmic factors and dependence on ε.

Main idea: reductions to (and from) the general case *via* structural results.

Under structural assumptions: m is the new n.

A different flavor of results is obtained in [DKN15]: no assumption on the "shape," but rather on the "structure" of the unknown distribution.

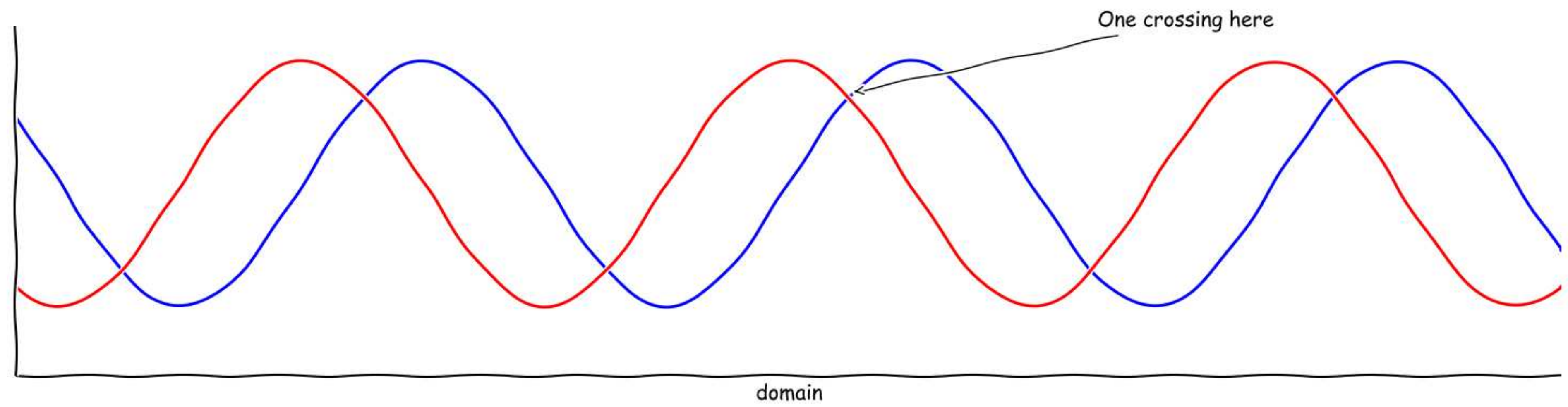
Theorem: Let $\mathcal{C} \subseteq \Delta([n])$ be a distribution class such that the probability mass functions (pmf) of any two $D, D' \in \mathcal{C}$ cross "essentially" at most m times. Then, given sampling access to an unknown $D \in \mathcal{C}$, one can test identity to an explicit D^* with $O(\sqrt{m}/\varepsilon^2)$ samples.



Under structural assumptions: m is the new n.

A different flavor of results is obtained in [DKN15]: no assumption on the "shape," but rather on the "structure" of the unknown distribution.

Theorem: Let $\mathcal{C} \subseteq \Delta([n])$ be a distribution class such that the probability mass functions (pmf) of any two $D, D' \in \mathcal{C}$ cross "essentially" at most m times. Then, given sampling access to an unknown $D \in \mathcal{C}$, one can test identity to an explicit D^* with $O(\sqrt{m}/\varepsilon^2)$ samples.

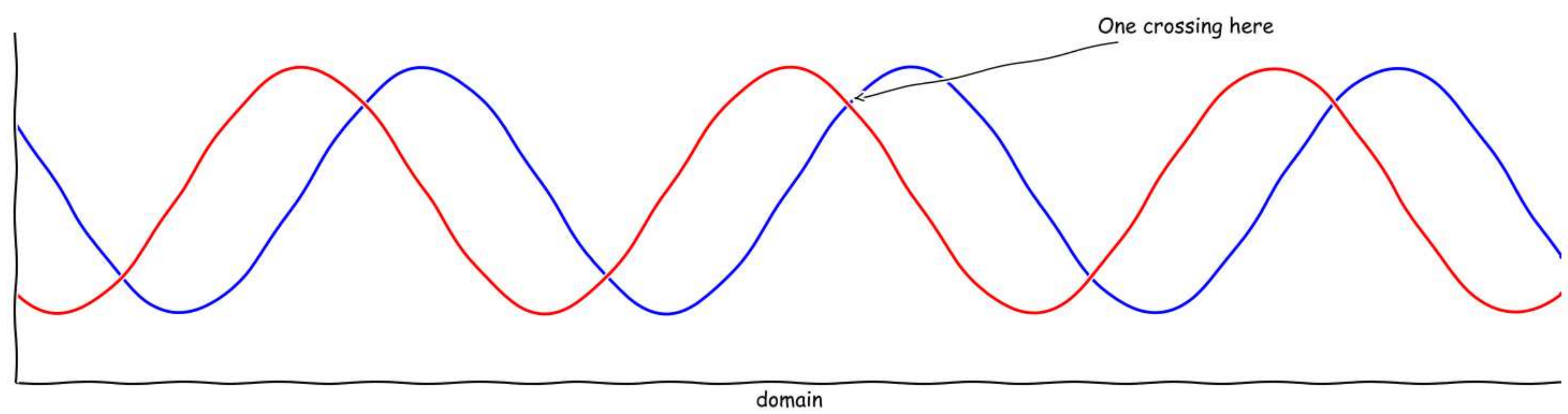


Applies to testing identity for k-modal $O(\sqrt{k \log n}/\varepsilon^{5/2})$, log-concave $\tilde{O}(1/\varepsilon^{9/4})$, monotone hazard risks $O(\sqrt{\log(n/\varepsilon)}/\varepsilon^{5/2})$, k-histograms $O(\sqrt{k}/\varepsilon^2)$, and mixtures thereof.

Under structural assumptions: m is the new n.

A different flavor of results is obtained in [DKN15]: no assumption on the "shape," but rather on the "structure" of the unknown distribution.

Theorem: Let $\mathcal{C} \subseteq \Delta([n])$ be a distribution class such that the probability mass functions (pmf) of any two $D, D' \in \mathcal{C}$ cross "essentially" at most m times. Then, given sampling access to an unknown $D \in \mathcal{C}$, one can test identity to an explicit D^* with $O(\sqrt{m}/\varepsilon^2)$ samples.



Applies to testing identity for k-modal $O(\sqrt{k \log n}/\varepsilon^{5/2})$, log-concave $\tilde{O}(1/\varepsilon^{9/4})$, monotone hazard risks $O(\sqrt{\log(n/\varepsilon)}/\varepsilon^{5/2})$, k-histograms $O(\sqrt{k}/\varepsilon^2)$, and mixtures thereof.

Main idea: "testing in $\mathcal{A}_m$-norm," and reduction to testing uniformity in this distance (over a much bigger domain).

Testing Differently: Changing the Rules

Let's twist again!

Instead of changing the assumptions on $D \in \Delta(\Omega)$, changing the model of access to D:

Let's twist again!

Instead of changing the assumptions on $D \in \Delta(\Omega)$, changing the model of access to D:

- with evaluation queries to the pmf: [RS09] ("property-testing"-style)

$$x \in \Omega \rightsquigarrow D(x)$$

Let's twist again!

Instead of changing the assumptions on $D \in \Delta(\Omega)$, changing the model of access to D:

- with evaluation queries to the pmf: [RS09] ("property-testing"-style)

$$x \in \Omega \rightsquigarrow D(x)$$

- with sampling and evaluation queries to the pmf: [BDKR05, GMV06, CR14]

$$? \rightsquigarrow x \sim D \quad \text{and} \quad x \in \Omega \rightsquigarrow D(x)$$

Let's twist again!

Instead of changing the assumptions on $D \in \Delta(\Omega)$, changing the model of access to D:

- with evaluation queries to the pmf: [RS09] ("property-testing"-style)

$$x \in \Omega \rightsquigarrow D(x)$$

- with sampling and evaluation queries to the pmf: [BDKR05, GMV06, CR14]

$$? \rightsquigarrow x \sim D \quad \text{and} \quad x \in \Omega \rightsquigarrow D(x)$$

- with sampling and evaluation queries to the cdf: [BKR04, CR14]

$$? \rightsquigarrow j \sim D \quad \text{and} \quad j \in [n] \rightsquigarrow \sum_{i=1}^{j} D(i)$$

Let's twist again!

Instead of changing the assumptions on $D \in \Delta(\Omega)$, changing the model of access to D:

- with evaluation queries to the pmf: [RS09] ("property-testing"-style)

$$x \in \Omega \rightsquigarrow D(x)$$

- with sampling and evaluation queries to the pmf: [BDKR05, GMV06, CR14]

$$? \rightsquigarrow x \sim D \quad \text{and} \quad x \in \Omega \rightsquigarrow D(x)$$

- with sampling and evaluation queries to the cdf: [BKR04, CR14]

$$? \rightsquigarrow j \sim D \quad \text{and} \quad j \in [n] \rightsquigarrow \sum_{i=1}^{j} D(i)$$

- with conditional sampling: [CFGM13, CRS15]

$$S \subseteq \Omega \rightsquigarrow x \sim D_S$$

Results: the Sunny Side (Up)

Informally: across the models and flavors, exponential sample complexity improvements – sometimes even from $n^{\Omega(1)}$ to constant. Some hardness remains, still – and most importantly, *all rules of thumbs are down.*

Results: the Sunny Side (Up)

Informally: across the models and flavors, exponential sample complexity improvements – sometimes even from $n^{\Omega(1)}$ to constant. Some hardness remains, still – and most importantly, *all rules of thumbs are down.*

Conditional sampling: identity and closeness testing are no longer related ($O_\varepsilon(1)$ vs. $(\log\log n)^{\Omega(1)}$).

Results: the Sunny Side (Up)

Informally: across the models and flavors, exponential sample complexity improvements – sometimes even from $n^{\Omega(1)}$ to constant. Some hardness remains, still – and most importantly, *all rules of thumbs are down.*

Conditional sampling: identity and closeness testing are no longer related ($O_\varepsilon(1)$ vs. $(\log\log n)^{\Omega(1)}$). Tolerant uniformity testing and entropy estimation are, similarly, worlds apart.

Results: the Sunny Side (Up)

Informally: across the models and flavors, exponential sample complexity improvements – sometimes even from $n^{\Omega(1)}$ to constant. Some hardness remains, still – and most importantly, *all rules of thumbs are down.*

Conditional sampling: identity and closeness testing are no longer related ($O_\varepsilon(1)$ vs. $(\log\log n)^{\Omega(1)}$). Tolerant uniformity testing and entropy estimation are, similarly, worlds apart.

Testing with queries: Testing uniformity, identity and closeness becomes easy: the challenge now seems to lie in tolerant testing, or in testing against classes.

Results: the Sunny Side (Up)

Informally: across the models and flavors, exponential sample complexity improvements – sometimes even from $n^{\Omega(1)}$ to constant. Some hardness remains, still – and most importantly, *all rules of thumbs are down.*

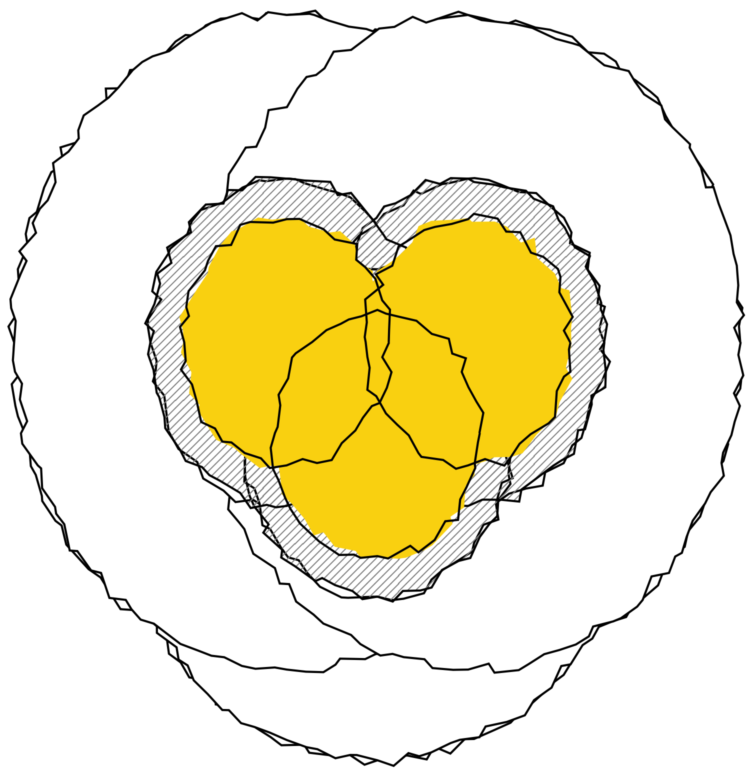
Conditional sampling: identity and closeness testing are no longer related ($O_\varepsilon(1)$ vs. $(\log\log n)^{\Omega(1)}$). Tolerant uniformity testing and entropy estimation are, similarly, worlds apart.

Testing with queries: Testing uniformity, identity and closeness becomes easy: the challenge now seems to lie in tolerant testing, or in testing against classes.

Challenges: Understanding the intrinsic power and limitations of the models, how they relate, and whether there exist generic tools to analyze them.

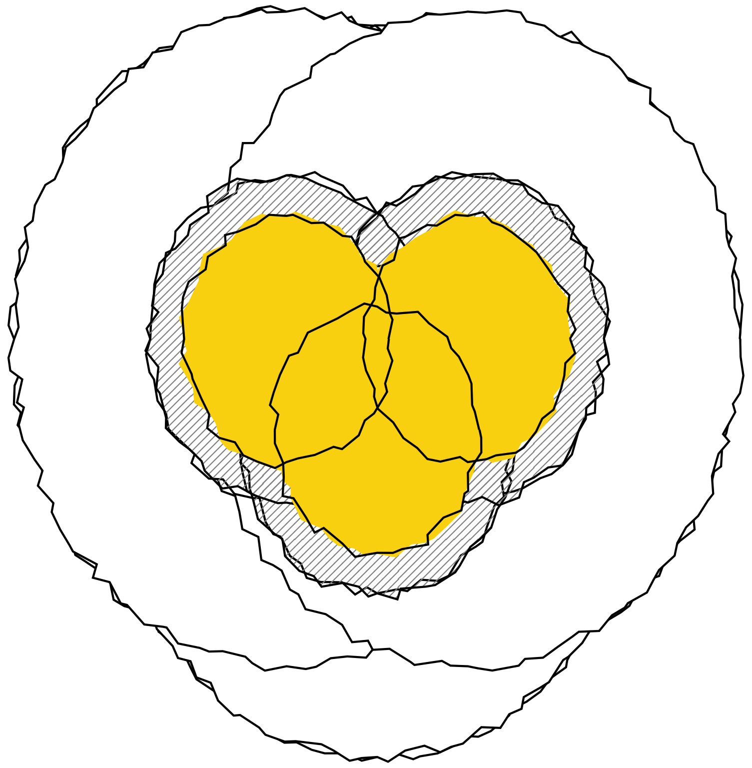
A Collections of Eggs

Instead of changing the assumptions on $D \in \Delta(\Omega)$, changing the number of D's:



A Collections of Eggs

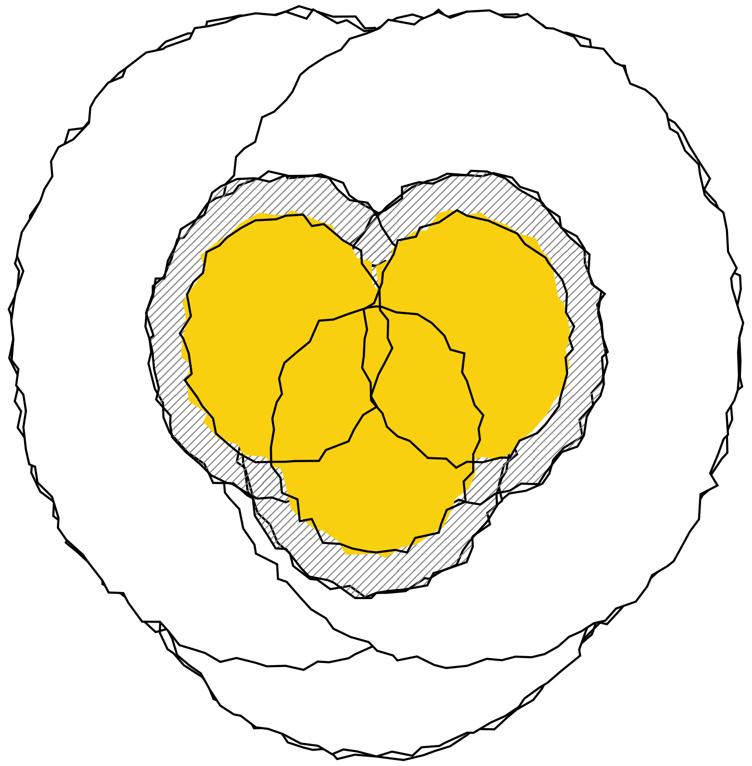
Instead of changing the assumptions on $D \in \Delta(\Omega)$, changing the number of D's:



Testing collections: given "sampling access" to a family $\mathcal{D} = (D_1, \dots, D_m)$ of m distributions over Ω, test whether they satisfy a joint property or are far from it (in average ℓ_1 distance). E.g., equivalence: $D_1 = \cdots = D_m$.

A Collections of Eggs

Instead of changing the assumptions on $D \in \Delta(\Omega)$, changing the number of D's:



Testing collections: given "sampling access" to a family $\mathcal{D} = (D_1, \dots, D_m)$ of m distributions over Ω, test whether they satisfy a joint property or are far from it (in average ℓ_1 distance). E.g., equivalence: $D_1 = \cdots = D_m$.

[LRR13] (equivalence and clustering), [LRR14] (similarity of means)

Some collected results

Testing equivalence has sample complexity $\tilde{O}(\min(m^{1/3}n^{2/3}, m^{1/2}n^{1/2}))$. *(Beats the naive approach based on the sampling model.)*

Some collected results

Testing equivalence has sample complexity $\tilde{O}\big(\min(m^{1/3}n^{2/3}, m^{1/2}n^{1/2})\big)$. *(Beats the naive approach based on the sampling model.)* It has sample complexity $\Omega\big(m^{1/2}n^{1/2}\big)$; and $\Omega\big(m^{1/3}n^{2/3}\big)$ as long as $n = \tilde{\Omega}\left(m\right)$.

Some collected results

Testing equivalence has sample complexity $\tilde{O}(\min(m^{1/3}n^{2/3}, m^{1/2}n^{1/2}))$. *(Beats the naive approach based on the sampling model.)* It has sample complexity $\Omega(m^{1/2}n^{1/2})$; and $\Omega(m^{1/3}n^{2/3})$ as long as $n = \tilde{\Omega}(m)$.

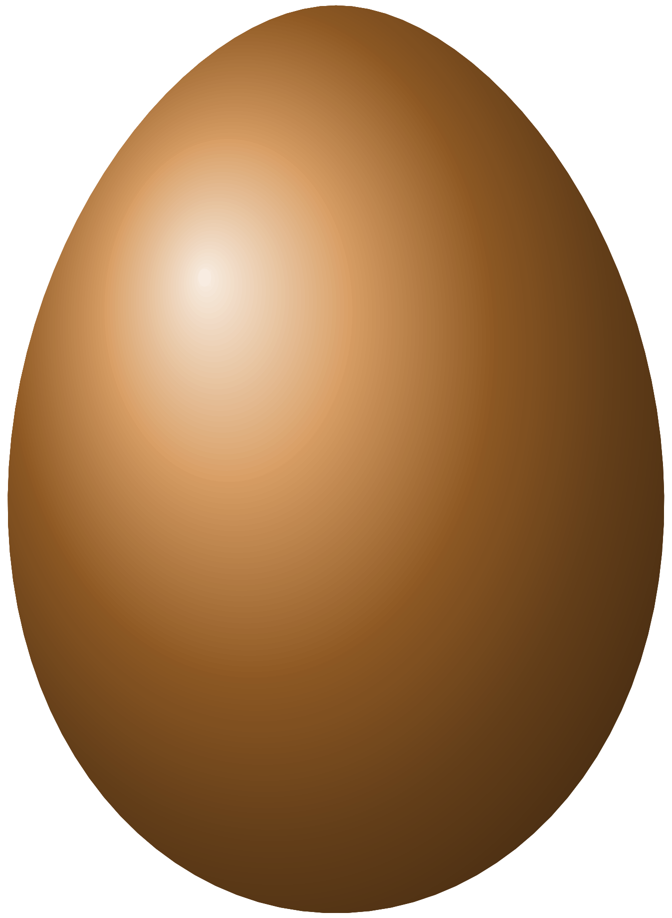
Bonus: Strong connections between equivalence for collections and independence for distributions.

Some collected results

Testing equivalence has sample complexity $\tilde{O}(\min(m^{1/3}n^{2/3}, m^{1/2}n^{1/2}))$. *(Beats the naive approach based on the sampling model.)* It has sample complexity $\Omega(m^{1/2}n^{1/2})$; and $\Omega(m^{1/3}n^{2/3})$ as long as $n = \tilde{\Omega}(m)$.

Bonus: Strong connections between equivalence for collections and independence for distributions. Testing collections related to conditional sampling.

That's All, (Y)olks!



Thank you.

Bibliography (1)

[AD14] J. Acharya and C. Daskalakis. Testing Poisson Binomial Distributions. In *SODA*, 2014.

[BDKR05] T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld. The complexity of approximating the entropy. *SICOMP*, 35(1):132–150, 2005.

[BFF+01] T. Batu, E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White. Testing random variables for independence and identity. In *FOCS*, 2001.

[BFR+10] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing closeness of discrete distributions. (abs/1009.5397), 2010.

[Bir87] L. Birgé. On the risk of histograms for estimating decreasing densities. *The Annals of Statistics*, 15(3), 1987.

[BKR04] T. Batu, R. Kumar, and R. Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *STOC*, 2004.

[CDGR15] C. L. Canonne, I. Diakonikolas, T. Gouleakis, and R. Rubinfeld. Testing Shape Restrictions, 2015. Manuscript.

[CDVV14] S-O. Chan, I. Diakonikolas, G. Valiant, and P. Valiant. Optimal algorithms for testing closeness of discrete distributions. In *SODA*, 2014.

[CFGM13] S. Chakraborty, E. Fischer, Y. Goldhirsh, and A. Matsliah. On the power of conditional samples in distribution testing. In *ITCS*, 2013.

[CR14] C. L. Canonne and R. Rubinfeld. Testing probability distributions underlying aggregated data. In *ICALP*, 2014.

[CRS15] C. L. Canonne, D. Ron, and R. A. Servedio. Testing probability distributions using conditional samples. *SICOMP*, 2015. To appear. Also available on arXiv at abs/1211.2664.

[DDS12] C. Daskalakis, I. Diakonikolas, and R. A. Servedio. Learning k-modal distributions via testing. In *SODA*, 2012.

[DDS+13] C. Daskalakis, I. Diakonikolas, R. A. Servedio, G. Valiant, and P. Valiant. Testing k-modal distributions: Optimal algorithms via reductions. In *SODA*, 2013.

Bibliography (2)

[DKN15] I. Diakonikolas, D. M. Kane, and V. Nikishkin. Testing Identity of Structured Distributions. In *SODA*, 2015.

[GGR98] O. Goldreich, S. Goldwasser, and D. Ron. Property testing and its connection to learning and approximation. *JACM*, 45(4):653–750, July 1998.

[GMV06] S. Guha, A. McGregor, and S. Venkatasubramanian. Streaming and sublinear approximation of entropy and information distances. In *SODA*, 2006.

[GR00] O. Goldreich and D. Ron. On testing expansion in bounded-degree graphs. Technical Report TR00-020, ECCC, 2000.

[LRR13] R. Levi, D. Ron, and R. Rubinfeld. Testing properties of collections of distributions. *Theory Comput.*, 9:295–347, 2013.

[LRR14] R. Levi, D. Ron, and R. Rubinfeld. Testing similar means. *SIDMA*, 28(4):1699–1724, 2014.

[Pan04] L. Paninski. Estimating entropy on m bins given fewer than m samples. *IEEE-IT*, 50(9), 2004.

[Pan08] L. Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE-IT*, 54(10), 2008.

[PRR06] M. Parnas, D. Ron, and R. Rubinfeld. Tolerant property testing and distance approximation. *JCSS*, 72(6):1012–1042, 2006.

[RRSS09] S. Raskhodnikova, D. Ron, A. Shpilka, and A. Smith. Strong lower bounds for approximating distributions support size and the distinct elements problem. *SICOMP*, 39(3):813–842, 2009.

[RS09] R. Rubinfeld and R. A. Servedio. Testing monotone high-dimensional distributions. *RSA*, 34(1):24–44, January 2009.

[Val11] P. Valiant. Testing symmetric properties of distributions. *SICOMP*, 40(6):1927–1968, 2011.

[VV10a] G. Valiant and P. Valiant. A CLT and tight lower bounds for estimating entropy. *ECCC*, 17:179, 2010.

[VV10b] G. Valiant and P. Valiant. Estimating the unseen: A sublinear-sample canonical estimator of distributions. *ECCC*, 17:180, 2010.

[VV11] G. Valiant and P. Valiant. The power of linear estimators. In *FOCS*, 2011. See also [VV10a] and [VV10b].

[VV14] G. Valiant and P. Valiant. An automatic inequality prover and instance optimal identity testing. In *FOCS*, 2014.