Big Data on the Rise?
Testing Monotonicity of Distributions

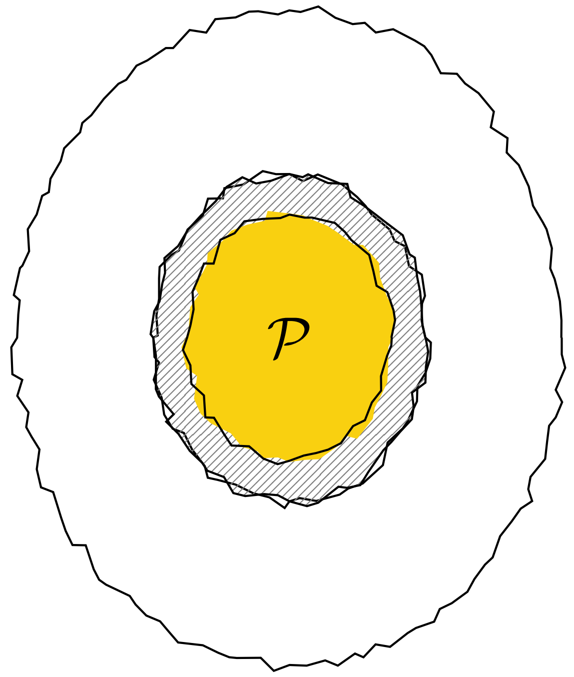
Clément Canonne

ICALP – 2015, July 8$^{\text{th}}$

Introduction

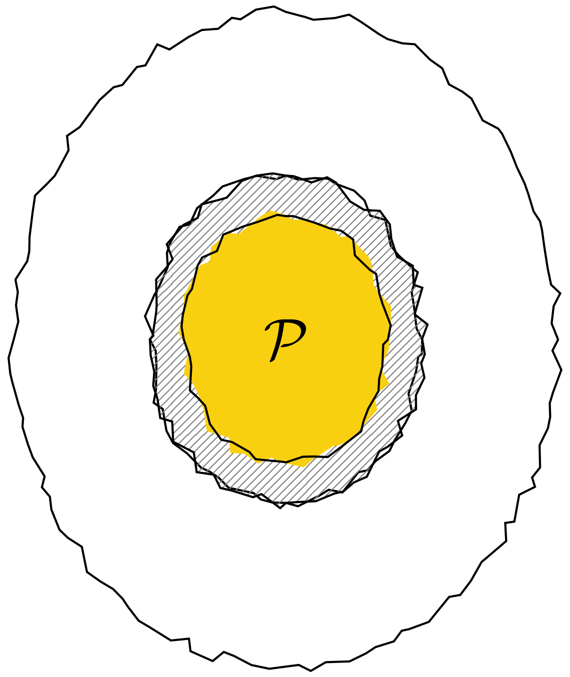
Introduction

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



Introduction

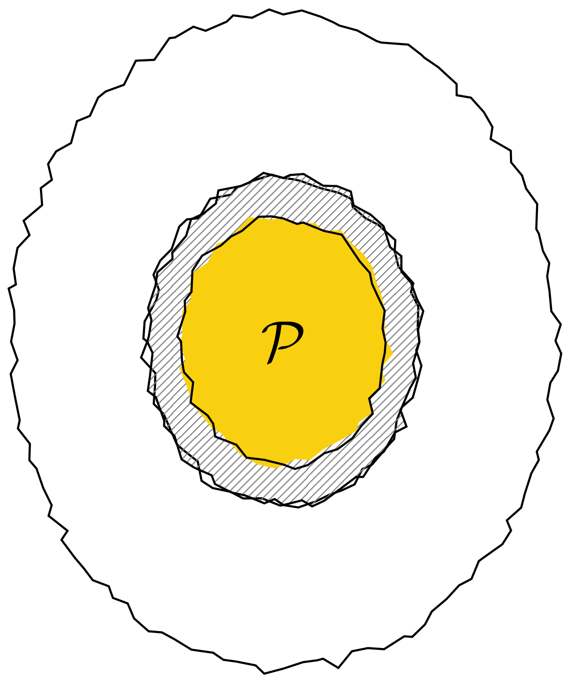
Property testing: what can we say about an object while barely looking at it?



"Is it in the yolk?"

Introduction

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

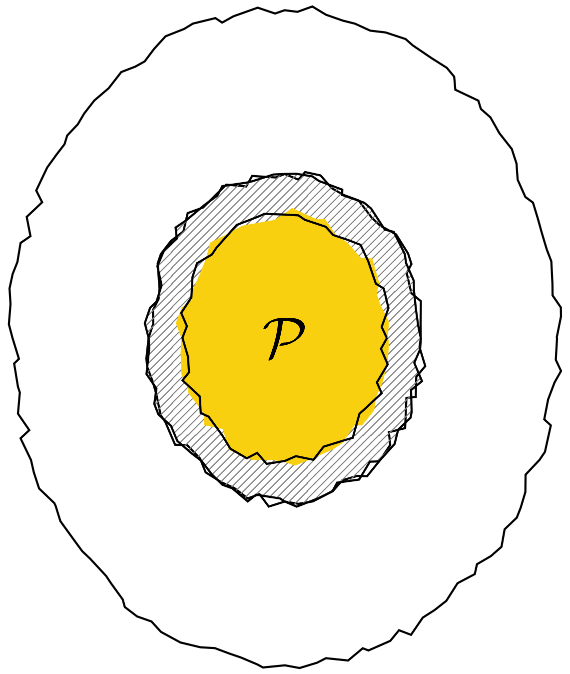


"Is it in the yolk?"

This talk: distribution testing, for *one* property ("class") and various 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 *one* property ("class") and various settings.
(some of the puns will be made on purpose)

Outline of the talk

Introduction

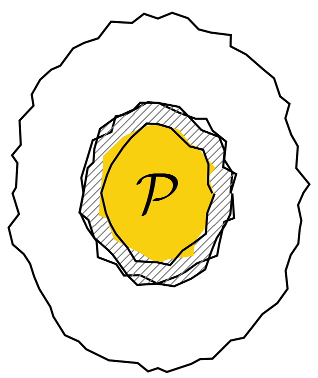
Testing For Monotonicity

Testing From Samples

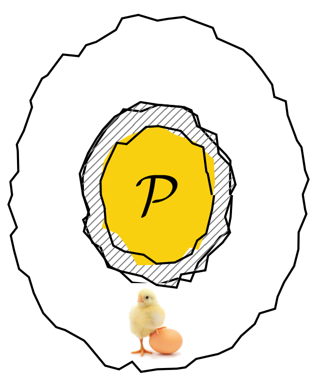
Testing Differently: Changing the Rules

Plan in more detail

- What We Are Doing Here: "testing for monotonicity"
- Testing From Samples: the standard model, upper and lower bounds



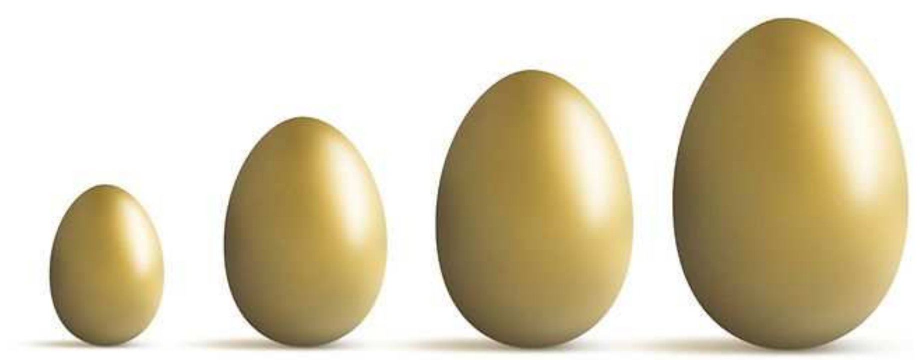
- Testing Differently: some other access (stronger or incomparable), or some other goal



Testing For Monotonicity

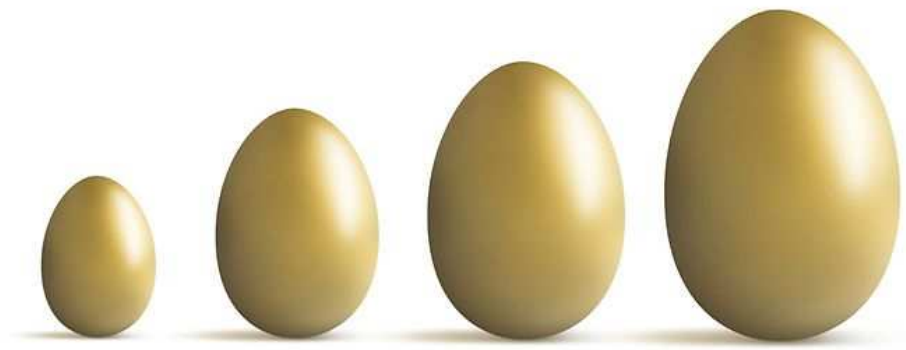
Monotone distributions

A probability distribution D on $[n] = \{1, \ldots, n\}$ is *monotone* (non-increasing) if its pmf is: $D(1) \geq D(2) \geq \cdots \geq D(n)$.



Monotone distributions

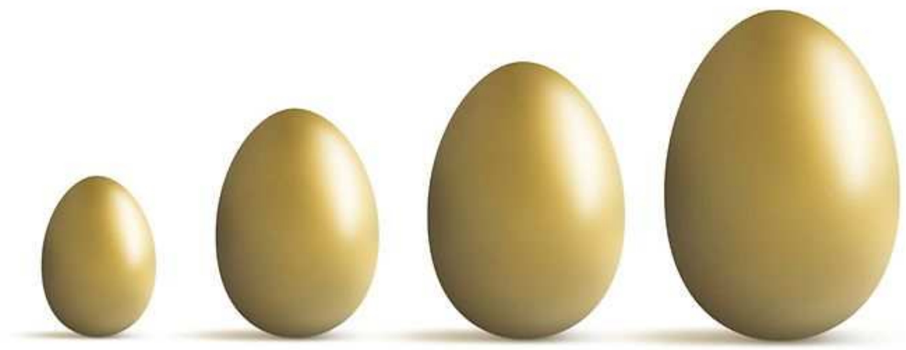
A probability distribution D on $[n] = \{1, \ldots, n\}$ is *monotone* (non-increasing) if its pmf is: $D(1) \geq D(2) \geq \cdots \geq D(n)$.



Why: pervasive,

Monotone distributions

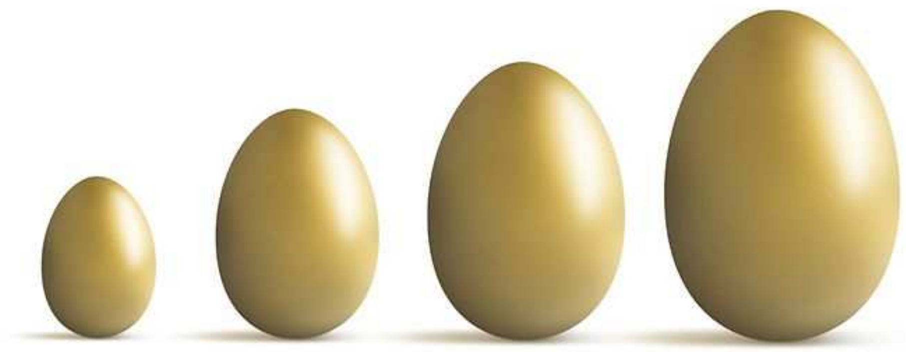
A probability distribution D on $[n] = \{1, \dots, n\}$ is *monotone* (non-increasing) if its pmf is: $D(1) \geq D(2) \geq \cdots \geq D(n)$.



Why: pervasive, simple enough,

Monotone distributions

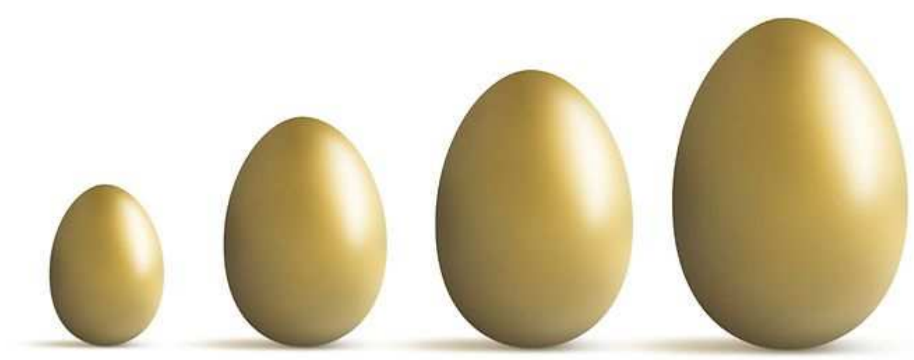
A probability distribution D on $[n] = \{1, \ldots, n\}$ is *monotone* (non-increasing) if its pmf is: $D(1) \geq D(2) \geq \cdots \geq D(n)$.



Why: pervasive, simple enough, complex enough,

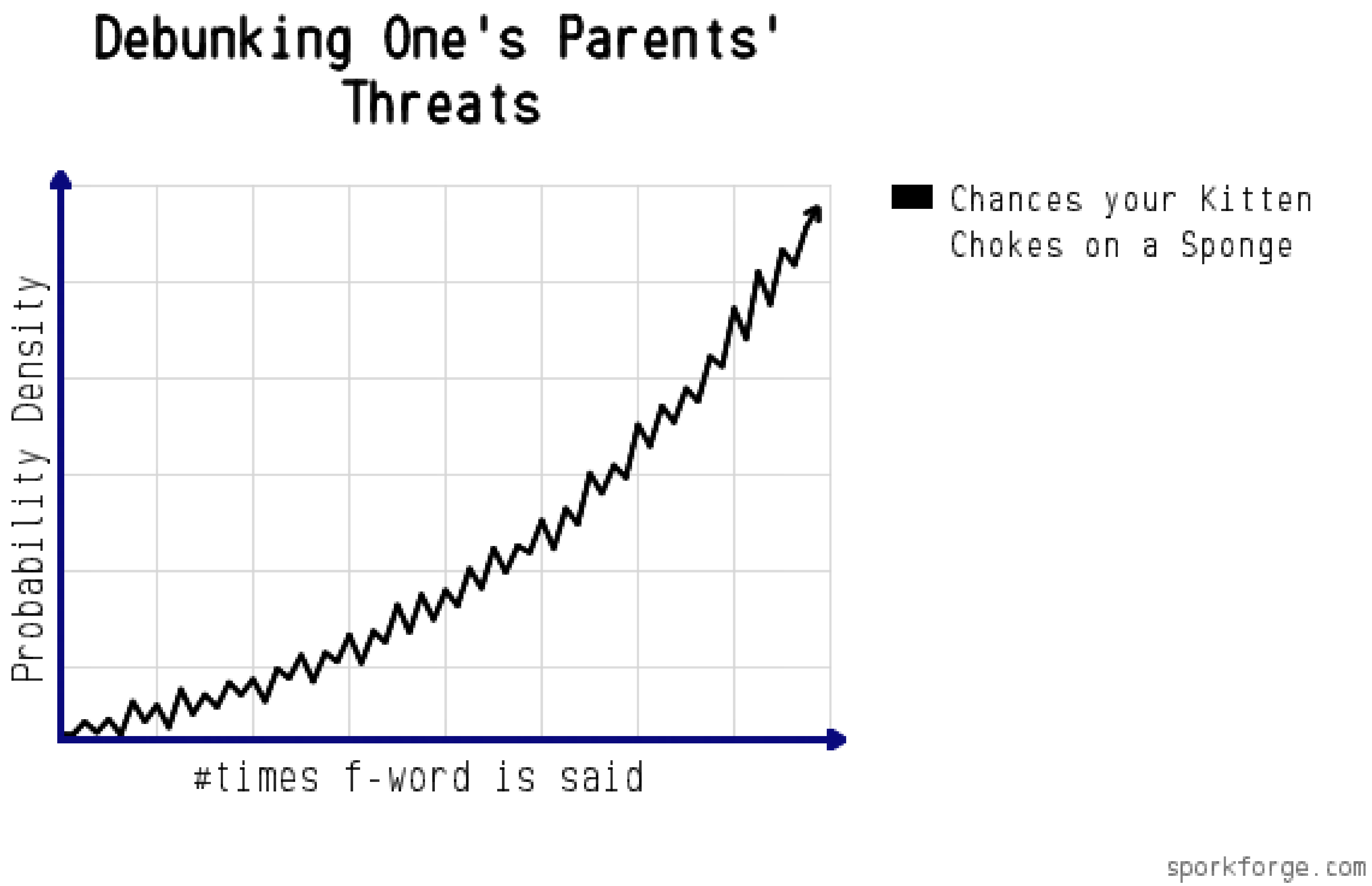
Monotone distributions

A probability distribution D on $[n] = \{1, \ldots, n\}$ is *monotone* (non-increasing) if its pmf is: $D(1) \geq D(2) \geq \cdots \geq D(n)$.



Why: pervasive, simple enough, complex enough, fundamental.

Monotone distributions



Testing From Samples

The setting

$\Delta(\Omega)$: all distributions over (finite) domain Ω of size n, $[n]$ (ordered) in this talk. **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, $[n]$ (ordered) in this talk. **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, $[n]$ (ordered) in this talk. **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]

Previous Results: a representative sample

Is it possible?

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

Previous Results: a representative sample

Is it possible?

Testing uniformity has sample complexity $O(\sqrt{n}/\varepsilon^2)$ [GR00, Pan08]. Testing monotonicity has sample complexity $\tilde{O}(\sqrt{n}/\varepsilon^6)$ [BKR04].

Previous Results: a representative sample

Is it possible?

Testing uniformity has sample complexity $O(\sqrt{n}/\varepsilon^2)$ [GR00, Pan08]. Testing monotonicity has sample complexity $\tilde{O}(\sqrt{n}/\varepsilon^6)$ [BKR04]. $\tilde{O}(\sqrt{n}/\varepsilon^{7/2})$, actually [CDGR15].

Previous Results: a representative sample

Is it possible?

Testing uniformity has sample complexity $O(\sqrt{n}/\varepsilon^2)$ [GR00, Pan08]. Testing monotonicity has sample complexity $\tilde{O}(\sqrt{n}/\varepsilon^6)$ [BKR04]. $\tilde{O}(\sqrt{n}/\varepsilon^{7/2})$, actually [CDGR15]. Let's even say $O(\sqrt{n}/\varepsilon^2)$ [ADK15].

Previous Results: a representative sample

Is it possible?

Testing uniformity has sample complexity $O(\sqrt{n}/\varepsilon^2)$ [GR00, Pan08]. Testing monotonicity has sample complexity $\tilde{O}(\sqrt{n}/\varepsilon^6)$ [BKR04]. $\tilde{O}(\sqrt{n}/\varepsilon^{7/2})$, actually [CDGR15]. Let's even say $O(\sqrt{n}/\varepsilon^2)$ [ADK15].

Can we do better?

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

Is it possible?

Testing uniformity has sample complexity $O(\sqrt{n}/\varepsilon^2)$ [GR00, Pan08]. Testing monotonicity has sample complexity $\tilde{O}(\sqrt{n}/\varepsilon^6)$ [BKR04]. $\tilde{O}(\sqrt{n}/\varepsilon^{7/2})$, actually [CDGR15]. Let's even say $O(\sqrt{n}/\varepsilon^2)$ [ADK15].

Can we do better?

Testing uniformity has sample complexity $\Omega(\sqrt{n}/\varepsilon^2)$ [GR00, Pan08]. Testing monotonicity has sample complexity $\Omega(\sqrt{n}/\varepsilon^2)$ [BKR04].

Any more ~~good~~ news?

Well, it's tight. And everything else (closeness testing, etc.) has sample complexity at least $n^{\Omega(1)}$.

Previous Results: a representative sample

Is it possible?

Testing uniformity has sample complexity $O(\sqrt{n}/\varepsilon^2)$ [GR00, Pan08]. Testing monotonicity has sample complexity $\tilde{O}(\sqrt{n}/\varepsilon^6)$ [BKR04]. $\tilde{O}(\sqrt{n}/\varepsilon^{7/2})$, actually [CDGR15]. Let's even say $O(\sqrt{n}/\varepsilon^2)$ [ADK15].

Can we do better?

Testing uniformity has sample complexity $\Omega(\sqrt{n}/\varepsilon^2)$ [GR00, Pan08]. Testing monotonicity has sample complexity $\Omega(\sqrt{n}/\varepsilon^2)$ [BKR04].

Any more ~~good~~ news?

Well, it's tight. And everything else (closeness testing, etc.) has sample complexity at least $n^{\Omega(1)}$. Worse – tolerant testing uniformity (let alone monotonicity) has sample complexity $\Theta(n/\log n)$ [Pan04, RRSS09, Val11, VV10a, VV10b, VV11]

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]

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

These models: Everything is Scrambled

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.*

These models: Everything is Scrambled

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. $(\log\log n)^{\Omega(1)}$).

These models: Everything is Scrambled

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.

These models: Everything is Scrambled

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.

These models: Everything is Scrambled

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.

And monotonicity? Subject of this work.

New Results: the Sunny Side (Up)

Model	Upper bound	Lower bound
SAMP	$O\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$	$\Omega\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$
COND	$\tilde{O}\left(\frac{1}{\varepsilon^{22}}\right)$, $\tilde{O}\left(\frac{\log^2 n}{\varepsilon^3} + \frac{\log^4 n}{\varepsilon^2}\right)$	$\Omega\left(\frac{1}{\varepsilon^2}\right)$
INTCOND	$\tilde{O}\left(\frac{\log^5 n}{\varepsilon^4}\right)$	$\Omega\left(\sqrt{\frac{\log n}{\log\log n}}\right)$
EVAL	$O\left(\max\left(\frac{\log n}{\varepsilon}, \frac{1}{\varepsilon^2}\right)\right)^*$	$\Omega\left(\frac{\log n}{\varepsilon}\right)^*$, $\Omega\left(\frac{\log n}{\log\log n}\right)$
Cumulative Dual	$\tilde{O}\left(\frac{1}{\varepsilon^4}\right)$	$\Omega\left(\frac{1}{\varepsilon}\right)$

Table 1: Highlighted results are new; bounds with * hold for non-adaptive testers.

New Results: the Sunny Side (Up)

Model	Upper bound	Lower bound
SAMP	$O(\frac{\sqrt{n}}{\varepsilon^2})$	$\Omega(\frac{\sqrt{n}}{\varepsilon^2})$
COND	$\tilde{O}(\frac{1}{\varepsilon^{22}})$, $\tilde{O}(\frac{\log^2 n}{\varepsilon^3} + \frac{\log^4 n}{\varepsilon^2})$	$\Omega(\frac{1}{\varepsilon^2})$
INTCOND	$\tilde{O}(\frac{\log^5 n}{\varepsilon^4})$	$\Omega(\sqrt{\frac{\log n}{\log\log n}})$
EVAL	$O(\max\left(\frac{\log n}{\varepsilon}, \frac{1}{\varepsilon^2}\right))^*$	$\Omega(\frac{\log n}{\varepsilon})^*$, $\Omega(\frac{\log n}{\log\log n})$
Cumulative Dual	$\tilde{O}(\frac{1}{\varepsilon^4})$	$\Omega(\frac{1}{\varepsilon})$

Table 1: Highlighted results are new; bounds with * hold for non-adaptive testers.

Upshot: depending on the model, monotonicity testing can become over easy, or still medium hard.

A glimpse at the techniques

Upper bounds

INTCOND: Transposing [BKR04], with a twist – testing *uniformity* is much cheaper now (but careful! We need a bit more.)

A glimpse at the techniques

Upper bounds

INTCOND: Transposing [BKR04], with a twist – testing *uniformity* is much cheaper now (but careful! We need a bit more.)
COND: Using the "Birgé decomposition" [Bir87] to shrink the size of the domain – *but* this changes the property as well.

A glimpse at the techniques

Upper bounds

INTCOND: Transposing [BKR04], with a twist – testing *uniformity* is much cheaper now (but careful! We need a bit more.)
COND: Using the "Birgé decomposition" [Bir87] to shrink the size of the domain – *but* this changes the property as well.
EVAL: Showing we can learn monotone distributions with this many queries – and (conveniently!) do tolerant testing afterwards.

A glimpse at the techniques

Upper bounds

INTCOND: Transposing [BKR04], with a twist – testing *uniformity* is much cheaper now (but careful! We need a bit more.)
COND: Using the "Birgé decomposition" [Bir87] to shrink the size of the domain – *but* this changes the property as well.
EVAL: Showing we can learn monotone distributions with this many queries – and (conveniently!) do tolerant testing afterwards.
Cumulative Dual: Same approach as for COND, but getting simpler – one of the primitives is now much easier to get.

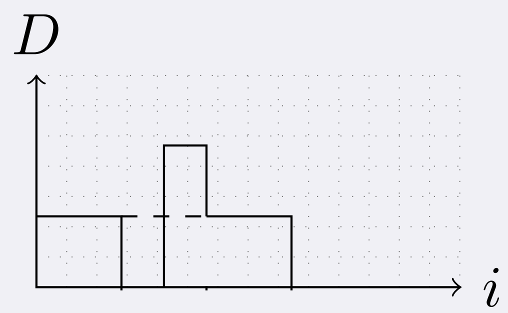
A glimpse at the techniques

Upper bounds

INTCOND: Transposing [BKR04], with a twist – testing *uniformity* is much cheaper now (but careful! We need a bit more.)
COND: Using the "Birgé decomposition" [Bir87] to shrink the size of the domain – *but* this changes the property as well.
EVAL: Showing we can learn monotone distributions with this many queries – and (conveniently!) do tolerant testing afterwards.
Cumulative Dual: Same approach as for COND, but getting simpler – one of the primitives is now much easier to get.

Lower bounds

EVAL, **non-adaptive:** "Needle-in-a-haystack" approach: random support size, with random "chunk" of violations hidden inside.



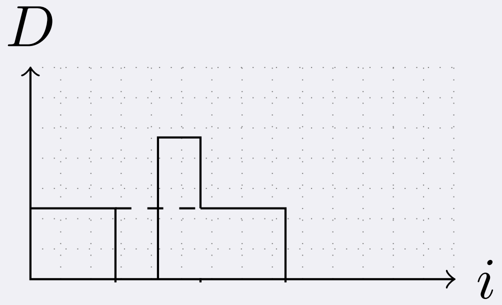
A glimpse at the techniques

Upper bounds

INTCOND: Transposing [BKR04], with a twist – testing *uniformity* is much cheaper now (but careful! We need a bit more.)
COND: Using the "Birgé decomposition" [Bir87] to shrink the size of the domain – *but* this changes the property as well.
EVAL: Showing we can learn monotone distributions with this many queries – and (conveniently!) do tolerant testing afterwards.
Cumulative Dual: Same approach as for COND, but getting simpler – one of the primitives is now much easier to get.

Lower bounds

EVAL, non-adaptive: "Needle-in-a-haystack" approach: random support size, with random "chunk" of violations hidden inside.



EVAL, adaptive: Reduction from another question: "what is the sum of a monotone sequence?"

Open Questions

- Closing the gaps? (COND, INTCOND, Cumulative Dual)

Open Questions

- Closing the gaps? (COND, INTCOND, Cumulative Dual)
- EVAL: "it should be $\Omega(\frac{\log n}{\varepsilon})$ for adaptive as well"

Open Questions

- Closing the gaps? (COND, INTCOND, Cumulative Dual)
- EVAL: "it should be $\Omega(\frac{\log n}{\varepsilon})$ for adaptive as well"
- Other models?

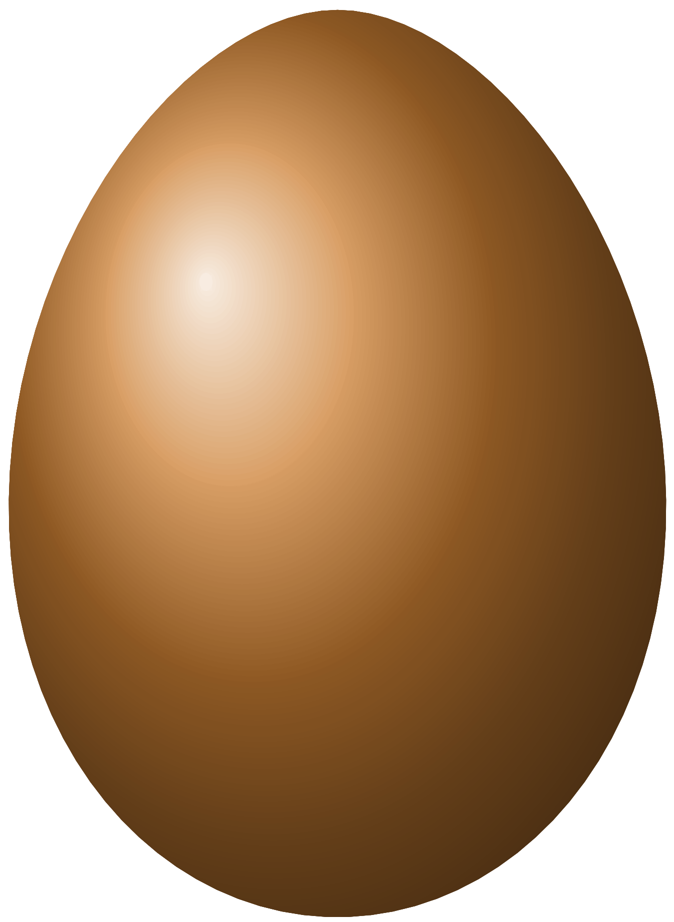
Open Questions

- Closing the gaps? (COND, $\mathsf{INTCOND}$, Cumulative Dual)
- EVAL: "it should be $\Omega(\frac{\log n}{\varepsilon})$ for adaptive as well"
- Other models? Other classes?

Open Questions

- Closing the gaps? (COND, INTCOND, Cumulative Dual)
- EVAL: "it should be $\Omega(\frac{\log n}{\varepsilon})$ for adaptive as well"
- Other models? Other classes?
- *What about Dual access*? (in between EVAL and Cumulative Dual– but *where*?)

That's All, (Y)olks!



Thank you.

Bibliography

[ADK15] J. Acharya, C. Daskalakis, and G. Kamath. Optimal Testing for Families of Distributions. Manuscript, 2015.

[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 of Discrete Distributions. Manuscript, 2015.

[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.

[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.

[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.

[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].