Testing probability distributions with more oracles

"Please, sir, I want some more."

Who? Clément Canonne* Ronitt Rubinfeld†

From? *Columbia University

†MIT

When? February 7, 2014

Plan of the talk

Introduction: distribution testing

Two new models: Dual and Cumulative Dual access

Spoiler: the results

Main techniques: an application to testing entropy

Background and motivation

Linear is the new exponential.

"Recently there has been a lot of glorious hullabaloo about Big Data and how it is going to revolutionize the way we work, play, eat and sleep." (R. Servedio)

Background and motivation

What is distribution testing?

Property testing

Given a big, hidden "object" X one can only access by local, expensive inspections (e.g., oracle queries), and a property $\mathcal{P}$, the goal is to check in sublinear number of inspections if (a) X has the property or (b) X is "far" from all objects having the property.[1]

[1] wrt to some specified metric, and parameter $\varepsilon > 0$ given to the tester.

Background and motivation

What is distribution testing?

Property testing

Given a big, hidden "object" X one can only access by local, expensive inspections (e.g., oracle queries), and a property $\mathcal{P}$, the goal is to check in sublinear number of inspections if (a) X has the property or (b) X is "far" from all objects having the property.[1]

Testing distributions (standard model)

X is an unknown probability distribution D over some n-element set; the testing algorithm has blackbox sample access to D.

[1] wrt to some specified metric, and parameter $\varepsilon > 0$ given to the tester.

Distribution testing (1)

In more details.

Distance criterion: total variation distance ($\propto \ell_1$)

$$d_{\mathrm{TV}}(D_1, D_2) \stackrel{\text{def}}{=} \frac{1}{2}\|D_1 - D_2\|_1 = \frac{1}{2}\sum_{i\in[n]} |D_1(i) - D_2(i)|.$$

Definition (Testing algorithm) Let $\mathcal{P}$ be a property of distributions over $[n]$, and ORACLE_D be some type of oracle which provides access to D. A $q(\varepsilon, n)$-query ORACLE testing algorithm for $\mathcal{P}$ is an algorithm T which, given ε, n as input parameters and oracle access to ORACLE_D, for any distribution D over $[n]$ makes at most $q(\varepsilon, n)$ calls to ORACLE_D, and:

- if $D \in \mathcal{P}$ then, w.p. at least $2/3$, T outputs ACCEPT;
- if $d_{\mathrm{TV}}(D, \mathcal{P}) \geq \varepsilon$ then, w.p. $2/3$, T outputs REJECT.

Distribution testing (2)

Comments

A few remarks

- tester is randomized;

Distribution testing (2)

Comments

A few remarks

- tester is randomized;
- "gray" area for $d_{\mathrm{TV}}(D, \mathcal{P}) \in (0, \varepsilon)$;

Distribution testing (2)

Comments

A few remarks

- tester is randomized;
- “gray” area for $d_{\mathrm{TV}}(D, \mathcal{P}) \in (0, \varepsilon)$;
- $2/3$ is completely arbitrary;

Distribution testing (2)

Comments

A few remarks

- tester is randomized;
- "gray" area for $d_{\mathrm{TV}}(D, \mathcal{P}) \in (0, \varepsilon)$;
- $2/3$ is completely arbitrary;
- extends to several oracles and distributions;

Distribution testing (2)

Comments

A few remarks

- tester is randomized;
- "gray" area for $d_{\mathrm{TV}}(D, \mathcal{P}) \in (0, \varepsilon)$;
- $2/3$ is completely arbitrary;
- extends to several oracles and distributions;
- focuses on the sample complexity (*not* the runtime).

Distribution testing (3)

Now robust – tolerant testing.

Definition (Tolerant testing algorithm)

Let $\mathcal{P}$ and ORACLE_D be as before. A $q(\varepsilon_1, \varepsilon_2, n)$-query *tolerant* testing algorithm for $\mathcal{P}$ is an algorithm T which is given $\varepsilon_1 < \varepsilon_2, n$ and oracle access to ORACLE_D, such that:

- if $d_{\mathrm{TV}}(D, \mathcal{P}) \leq \varepsilon_1$ then, w.p. $2/3$, T outputs ACCEPT;
- if $d_{\mathrm{TV}}(D, \mathcal{P}) \geq \varepsilon_2$ then, w.p. $2/3$, T outputs REJECT.

Distribution testing (3)

Now robust – tolerant testing.

Definition (Tolerant testing algorithm)

Let $\mathcal{P}$ and ORACLE_D be as before. A $q(\varepsilon_1, \varepsilon_2, n)$-query *tolerant* testing algorithm for $\mathcal{P}$ is an algorithm T which is given $\varepsilon_1 < \varepsilon_2, n$ and oracle access to ORACLE_D, such that:

- if $d_{\mathrm{TV}}(D, \mathcal{P}) \leq \varepsilon_1$ then, w.p. $2/3$, T outputs ACCEPT;
- if $d_{\mathrm{TV}}(D, \mathcal{P}) \geq \varepsilon_2$ then, w.p. $2/3$, T outputs REJECT.

Comments

- still some "gray" in $(\varepsilon_1, \varepsilon_2)$;

Distribution testing (3)

Now robust – tolerant testing.

Definition (Tolerant testing algorithm)

Let $\mathcal{P}$ and ORACLE$_D$ be as before. A $q(\varepsilon_1, \varepsilon_2, n)$-query *tolerant* testing algorithm for $\mathcal{P}$ is an algorithm T which is given $\varepsilon_1 < \varepsilon_2, n$ and oracle access to ORACLE$_D$, such that:

- if $d_{\mathrm{TV}}(D, \mathcal{P}) \leq \varepsilon_1$ then, w.p. $2/3$, T outputs ACCEPT;
- if $d_{\mathrm{TV}}(D, \mathcal{P}) \geq \varepsilon_2$ then, w.p. $2/3$, T outputs REJECT.

Comments

- still some "gray" in $(\varepsilon_1, \varepsilon_2)$;
- essentially equivalent to *distance estimation*;

Distribution testing (3)

Now robust – tolerant testing.

Definition (Tolerant testing algorithm)

Let $\mathcal{P}$ and ORACLE_D be as before. A $q(\varepsilon_1, \varepsilon_2, n)$-query *tolerant* testing algorithm for $\mathcal{P}$ is an algorithm T which is given $\varepsilon_1 < \varepsilon_2, n$ and oracle access to ORACLE_D, such that:

- if $d_{\mathrm{TV}}(D, \mathcal{P}) \leq \varepsilon_1$ then, w.p. $2/3$, T outputs ACCEPT;
- if $d_{\mathrm{TV}}(D, \mathcal{P}) \geq \varepsilon_2$ then, w.p. $2/3$, T outputs REJECT.

Comments

- still some "gray" in $(\varepsilon_1, \varepsilon_2)$;
- essentially equivalent to *distance estimation*;
- usually much harder than testing.

Distribution testing (4)

Concrete example: testing uniformity

Can identity any property $\mathcal{P}$ with the set $\mathcal{S}_\mathcal{P}$ of distributions with this property.

$$\mathcal{P} = \text{Uniformity} \Leftrightarrow \mathcal{S}_\mathcal{P} = \{\mathcal{U}\}$$

Distance to $\mathcal{P}$:

$$\mathrm{d}_{\mathrm{TV}}(D, \mathcal{S}_\mathcal{P}) = \min_{D' \in \mathcal{S}_\mathcal{P}} \mathrm{d}_{\mathrm{TV}}\big(D, D'\big) \underset{\text{here}}{=} \mathrm{d}_{\mathrm{TV}}(D, \mathcal{U})$$

General outline

1. Draw a bunch of samples from D;
2. "Process" them (e.g. counting the number of points drawn more than once *(collisions)*);
3. Compare the result to what one would expect from the uniform distribution $\mathcal{U}$;
4. Reject if it differs too much; accept otherwise.

Background and motivation

So what is the problem with that?

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

Background and motivation

So what is the problem with that?

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

Example Testing *uniformity* has $\Theta(\sqrt{n}/\varepsilon^2)$ sample complexity [GR00, BFR+10, Pan08], *equivalence to a known distribution* $\tilde{\Theta}(\sqrt{n}/\varepsilon^2)$ [BFF+01, Pan08]; *equivalence of two unknown distributions* $\Omega(n^{2/3})$ [BFR+10, Val11, CDVV14] (essentially tight). . .

Background and motivation

So what is the problem with that?

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

Example Testing *uniformity* has $\Theta(\sqrt{n}/\varepsilon^2)$ sample complexity [GR00, BFR+10, Pan08], *equivalence to a known distribution* $\tilde{\Theta}(\sqrt{n}/\varepsilon^2)$ [BFF+01, Pan08]; *equivalence of two unknown distributions* $\Omega(n^{2/3})$ [BFR+10, Val11, CDVV14] (essentially tight)...

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

Background and motivation

Bypassing the lower bounds: changing the adversary

First idea Focusing on subclasses of distributions: structure may help!

Shape: monotone distributions, k-modal, log-concave...

Mixtures: Gaussian mixtures, Poisson Binomial Distributions, SIIRVs...

([BKR04, DDS+13], [DDS12, DDO+13] (learning)...)

Background and motivation

Bypassing the lower bounds: changing the rules

Second idea

What if the oracle itself was too weak?

Background and motivation

Bypassing the lower bounds: changing the rules

Second idea — What if the oracle itself was too weak?

COND — can ask for samples *conditioned on a subset* $S \subseteq [n]$
[CFGM13, CRS12, CRS14]

Background and motivation

Bypassing the lower bounds: changing the rules

Second idea	What if the oracle itself was too weak?
COND	can ask for samples *conditioned on a subset* $S \subseteq [n]$ [CFGM13, CRS12, CRS14]
This work	can sample from D and *query* it: have either PMF or CDF access.

Definition (Dual oracle)

Fix a distribution D over $[n]$. A dual oracle for D is a pair of oracles $(\mathrm{SAMP}_D, \mathrm{EVAL}_D)$ defined as follows:

- when queried, the *sampling* oracle SAMP_D returns an element $i \in [n]$ drawn from D independently of all previous calls to the oracles;
- the *evaluation* oracle EVAL_D takes as input a query element $j \in [n]$, and returns its probability weight $D(j)$.

Definition (Cumulative Dual oracle)

A cumulative dual oracle for D is a pair of oracles $(\mathrm{SAMP}_D, \mathrm{CEVAL}_D)$ defined as follows:

- the *sampling* oracle SAMP_D behaves as above;
- the *evaluation* oracle CEVAL_D takes as input a query element $j \in [n]$, and returns its cumulative weight $D([j]) = \sum_{i=1}^{j} D(i)$.

Dual and Cumulative Dual access models

A couple remarks

- EVAL-only model considered in [RS09]; CEVAL-only in [BKR04]; $(\mathrm{SAMP}, \mathrm{EVAL})$ in part of [BDKR05, GMV05]

Dual and Cumulative Dual access models

A couple remarks

- EVAL-only model considered in [RS09]; CEVAL-only in [BKR04]; $(\mathsf{SAMP}, \mathsf{EVAL})$ in part of [BDKR05, GMV05]
- $\mathsf{SAMP} \preceq (\mathsf{SAMP}, \mathsf{EVAL}) \preceq (\mathsf{SAMP}, \mathsf{CEVAL})$

Dual and Cumulative Dual access models

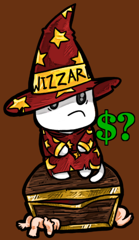
A couple remarks

- EVAL-only model considered in [RS09]; CEVAL-only in [BKR04]; (SAMP, EVAL) in part of [BDKR05, GMV05]
- $\mathsf{SAMP} \preceq (\mathsf{SAMP}, \mathsf{EVAL}) \preceq (\mathsf{SAMP}, \mathsf{CEVAL})$
- *How to motivate such a model?*

Dual and Cumulative Dual access models

Is that even a thing?

Broke Arthur & Greedy Merlin



- Huge dataset out there one can freely sample
  (collection of all words in vampire-related chick-lit)
- Painstakingly long and expensive analysis of this dataset held by a private party – selling access to it
  (say, by a multinational corporation famous for their search engine)
- Computationally limited Arthur working on this dataset
  (but no time nor will to analyze all of it)

Dual and Cumulative Dual access models

Is that even a thing?

Responsible Curator

- Database D of highly sensitive records *(healthcare information, financial records...)*
- (Untrusted) people in need of statistics about those *(scientific community, policy makers...)*
- Curator: release sanitized yet good-enough version $\tilde{D}$ of the records? [BLR13]

$\rightsquigarrow$ *tolerant testing of* $\tilde{D} \approx$ *tolerant testing of* D

Dual and Cumulative Dual access models

Is that even a thing?

. . . and more.

- Connection between dual model and datastream algorithms [GMV05]
- Further understanding of distribution testing (*what* is hard in it, and *why*?)

Our results

(and comparison with the original sampling model)

Problem	SAMP	Dual	Cumulative Dual
Testing uniformity	$\Theta\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$	$\Theta\left(\frac{1}{\varepsilon}\right)$	$\Theta\left(\frac{1}{\varepsilon}\right)$
Testing $\equiv D^*$	$\tilde{\Theta}\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$		
Testing $D_1 \equiv D_2$	$\Theta\left((\max\left(\frac{N^{2/3}}{\varepsilon^{4/3}}, \frac{\sqrt{N}}{\varepsilon^2}\right)\right)$		
Tolerant uniformity	$O\left(\frac{1}{(\varepsilon_2-\varepsilon_1)^2}\frac{n}{\log n}\right)$ $\Omega\left(\frac{n}{\log n}\right)$	$\Theta\left(\frac{1}{(\varepsilon_2-\varepsilon_1)^2}\right)$	$O\left(\frac{1}{(\varepsilon_2-\varepsilon_1)^2}\right)$
Tolerant D^*	$\Omega\left(\frac{n}{\log n}\right)$		
Tolerant D_1, D_2			
Estimating entropy to $\pm\Delta$	$\Theta\left(\frac{n}{\log n}\right)$	$O\left(\frac{\log^2\frac{n}{\Delta}}{\Delta^2}\right)$ $\Omega(\log n)$	$O\left(\frac{\log^2\frac{n}{\Delta}}{\Delta^2}\right)$
Estimating support size to $\pm\varepsilon n$	$\Theta\left(\frac{n}{\log n}\right)$	$\Theta\left(\frac{1}{\varepsilon^2}\right)$	$O\left(\frac{1}{\varepsilon^2}\right)$

Techniques (1)

Lower bounds: if I had a hammer...

Fact *To distinguish between D^+ and D^- with constant probability, any* SAMP *algorithm needs*

$$\Omega\left(\frac{1}{h^2(D^+, D^-)}\right) = \Omega\left(\frac{1}{\mathrm{d_{TV}}(D^+, D^-)}\right)$$

samples, where h is the Hellinger distance between two distributions ($\propto \|\sqrt{D^+} - \sqrt{D^-}\|_2$).

Techniques (1)

Lower bounds: if I had a hammer...

Fact *To distinguish between D^+ and D^- with constant probability, any* SAMP *algorithm needs*

$$\Omega\left(\frac{1}{h^2(D^+, D^-)}\right) = \Omega\left(\frac{1}{\mathrm{d_{TV}}(D^+, D^-)}\right)$$

samples, where h is the Hellinger distance between two distributions ($\propto \|\sqrt{D^+} - \sqrt{D^-}\|_2$).

Sad fact ...no longer true in our extended models, and no similar all-powerful tool. Must make do with Yao's lemma, *ad hoc* indistiguishability arguments

Techniques (1)

Lower bounds: if I had a hammer...

Fact

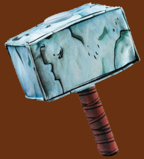
To distinguish between D^+ and D^- with constant probability, any SAMP *algorithm needs*

$$\Omega\left(\frac{1}{h^2(D^+, D^-)}\right) = \Omega\left(\frac{1}{\mathrm{d_{TV}}(D^+, D^-)}\right)$$

samples, where h is the Hellinger distance between two distributions ($\propto \|\sqrt{D^+} - \sqrt{D^-}\|_2$).

Sad fact

... no longer true in our extended models, and no similar all-powerful tool. Must make do with Yao's lemma, *ad hoc* indistiguishability arguments and biased coins.



Techniques (2)

Upper bounds: Well, I've got a hammer!

Main technique

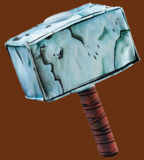
With Dual access: rewrite the quantity to estimate as

$$\mathbb{E}_{i\sim D}\left[\Phi(i, D(i))\right]$$

for *bounded* Φ.

Techniques (2)

Upper bounds: Well, I've got a hammer!



Main technique With Dual access: rewrite the quantity to estimate as

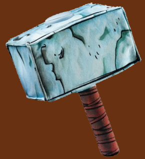
$$\mathbb{E}_{i\sim D}\left[\Phi(i, D(i))\right]$$

for *bounded* Φ.

Examples Entropy, support size, distance to D^* or D_2...

Techniques (2)

Upper bounds: Well, I've got a hammer!



Main technique

With Dual access: rewrite the quantity to estimate as

$$\mathbb{E}_{i\sim D}\left[\Phi(i, D(i))\right]$$

for *bounded* Φ.

Examples

Entropy, support size, distance to D^* or D_2...

(there is a catch)

Separation

Is Cumulative Dual any better?

Question

Do we have $(\mathsf{SAMP}, \mathsf{EVAL}) \not\preceq (\mathsf{SAMP}, \mathsf{CEVAL})$?

Separation

Is Cumulative Dual any better?

Question Do we have $(\mathrm{SAMP}, \mathrm{EVAL}) \not\preceq (\mathrm{SAMP}, \mathrm{CEVAL})$?

Intuition Can only be the case with properties using the order structure of $[n]$.

Separation

Is Cumulative Dual any better?

Question: Do we have $(\mathsf{SAMP}, \mathsf{EVAL}) \not\preceq (\mathsf{SAMP}, \mathsf{CEVAL})$?

Intuition: Can only be the case with properties using the order structure of $[n]$.

Answer: Yes: for entropy of monotone distributions.

Separation

Is Cumulative Dual any better?

Question Do we have $(\mathrm{SAMP}, \mathrm{EVAL}) \not\preceq (\mathrm{SAMP}, \mathrm{CEVAL})$?

Intuition Can only be the case with properties using the order structure of $[n]$.

Answer Yes: for entropy of close to monotone distributions.

Conclusion

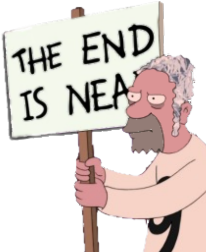
- Two new models for studying distributions
- Significant savings for property testing
- A general technique to get upper bounds with dual access

Conclusion

- Two new models for studying distributions
- Significant savings for property testing
- A general technique to get upper bounds with dual access

- Stronger separation between dual and cumulative dual oracles?
- More lower bounds for cumulative dual?
- What about other properties? (monotonicity (†), log-concavity. . .)
- What about learning? What about a "Lower Bound Hammer"?

Thank you.



References I

Tugkan Batu, Sanjoy Dasgupta, Ravi Kumar, and Ronitt Rubinfeld, *The complexity of approximating the entropy*, SIAM Journal on Computing **35** (2005), no. 1, 132–150.

T. Batu, E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White, *Testing random variables for independence and identity*, Proceedings of FOCS, 2001, pp. 442–451.

T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White, *Testing that distributions are close*, Proceedings of FOCS, 2000, pp. 189–197.

______, *Testing closeness of discrete distributions*, Tech. Report abs/1009.5397, ArXiv, 2010, This is a long version of [BFR$^+$00].

T. Batu, R. Kumar, and R. Rubinfeld, *Sublinear algorithms for testing monotone and unimodal distributions*, Proceedings of STOC, 2004, pp. 381–390.

Avrim Blum, Katrina Ligett, and Aaron Roth, *A learning theory approach to noninteractive database privacy*, J. ACM **60** (2013), no. 2, 12.

S.-O. Chan, I. Diakonikolas, G. Valiant, and P. Valiant, *Optimal Algorithms for Testing Closeness of Discrete Distributions*, Proceedings of SODA, 2014.

References II

- Sourav Chakraborty, Eldar Fischer, Yonatan Goldhirsh, and Arie Matsliah, *On the power of conditional samples in distribution testing*, Proceedings of the 4th conference on Innovations in Theoretical Computer Science (New York, NY, USA), ITCS '13, ACM, 2013, pp. 561–580.
- Clément Canonne, Dana Ron, and Rocco A. Servedio, *Testing probability distributions using conditional samples*, Tech. Report abs/1211.2664, ArXiV, November 2012.
- ______, *Testing equivalence between distributions using conditional samples*, Proceedings of SODA, 2014.
- Constantinos Daskalakis, Ilias Diakonikolas, Ryan O'Donnell, Rocco A. Servedio, and Li-Yang Tan, *Learning sums of independent integer random variables*, FOCS, 2013, pp. 217–226.
- Constantinos Daskalakis, Ilias Diakonikolas, and Rocco A. Servedio, *Learning poisson binomial distributions*, Proceedings of the 44th Symposium on Theory of Computing (New York, NY, USA), STOC '12, ACM, 2012, pp. 709–728.
- C. Daskalakis, I. Diakonikolas, R. Servedio, G. Valiant, and P. Valiant, *Testing k-modal distributions: Optimal algorithms via reductions*, Proceedings of SODA, 2013.

References III

- Sudipto Guha, Andrew McGregor, and Suresh Venkatasubramanian, *Streaming and sublinear approximation of entropy and information distances*, CoRR **abs/cs/0508122** (2005).
- O. Goldreich and D. Ron, *On testing expansion in bounded-degree graphs*, Tech. Report TR00-020, ECCC, 2000.
- L. Paninski, *A coincidence-based test for uniformity given very sparsely sampled discrete data*, IEEE-IT **54** (2008), no. 10, 4750–4755.
- R. Rubinfeld and R. A. Servedio, *Testing monotone high-dimensional distributions*, RSA **34** (2009), no. 1, 24–44.
- P. Valiant, *Testing symmetric properties of distributions*, SICOMP **40** (2011), no. 6, 1927–1968.
- G. Valiant and P. Valiant, *A CLT and tight lower bounds for estimating entropy*, Tech. Report TR10-179, ECCC, 2010.
- ______, *Estimating the unseen: A sublinear-sample canonical estimator of distributions*, Tech. Report TR10-180, ECCC, 2010.
- ______, *Estimating the unseen: an $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs*, Proceedings of STOC, 2011, See also [VV10a] and [VV10b], pp. 685–694.