Testing probability distributions underlying aggregated data

"Please, sir, I want some more."

Who? Clément Canonne* Ronitt Rubinfeld†

From? *Columbia University

†MIT and Tel Aviv University

When? July 11th, 2014

Plan of the talk

Introduction: distribution testing

Two new models: Dual and Cumulative Dual access

Spoiler: the results

Main techniques

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

Big, hidden "object" X only accessible by local, expensive inspections (queries), and property $\mathcal{P}$: check in sublinear number of queries if (a) X has the property or (b) X is "far" from all objects having it.

Background and motivation

What is distribution testing?

Property testing

Big, hidden "object" X only accessible by local, expensive inspections (queries), and property $\mathcal{P}$: check in sublinear number of queries if (a) X has the property or (b) X is "far" from all objects having it.

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.

Distribution testing (1)

In more details.

Distance: total variation distance ($\propto \ell_1$). ORACLE_D: type of access to D (e.g. sampling).

Definition (Tester)

Tester for property $\mathcal{P}$: algorithm T which is given ε, n, makes $q(\varepsilon, n)$ calls to ORACLE_D, and:

- if $D \in \mathcal{P}$ then w.h.p. T outputs Yes;
- if $d_{\mathrm{TV}}(D, \mathcal{P}) \geq \varepsilon$ then w.h.p. T outputs No.

Distribution testing (1)

In more details.

Distance: total variation distance ($\propto \ell_1$). ORACLE_D: type of access to D (e.g. sampling).

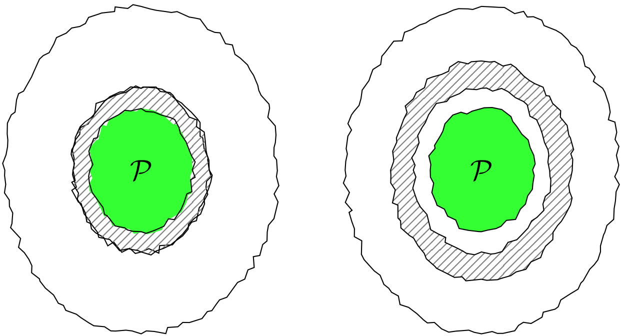
Definition (*Tolerant* tester)

Tolerant tester for property $\mathcal{P}$: algorithm T which is given $\varepsilon_1, \varepsilon_2, n$, makes $q(\varepsilon_1, \varepsilon_2, , n)$ calls to ORACLE_D, and:

- if $d_{\mathrm{TV}}(D, \mathcal{P}) \leq \varepsilon_1$ then w.h.p. T outputs Yes;
- if $d_{\mathrm{TV}}(D, \mathcal{P}) \geq \varepsilon_2$ then w.h.p. T outputs No.

Distribution testing (2)

Testing vs. Tolerant testing, in an egg-shell.



Distribution testing (3)

Comments

A few remarks

- “gray” area for $d_{\mathrm{TV}}(D, \mathcal{P}) \in (0, \varepsilon)$

Distribution testing (3)

Comments

A few remarks

- "gray" area for $d_{\mathrm{TV}}(D, \mathcal{P}) \in (0, \varepsilon)$
- tolerant testing usually much harder than testing.

Distribution testing (3)

Comments

A few remarks

- "gray" area for $d_{\mathrm{TV}}(D, \mathcal{P}) \in (0, \varepsilon)$
- tolerant testing usually much harder than testing.
- focuses on the sample complexity (*not* the runtime).

Distribution testing (4)

Concrete example: testing uniformity

General outline

1. Draw a bunch of samples from D;
2. "Process" them (e.g. count the number of points seen more than once *(collisions)*);
3. Compare to what the uniform distribution would give;
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* $\Theta(\sqrt{n}/\varepsilon^2)$ [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* $\Theta(\sqrt{n}/\varepsilon^2)$ [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, Sums of Independent Integer R.V.s. . .

([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 (probability mass function) or CDF (cumulative distribution function) access.

Definition (Dual oracle)

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

- *sampling* oracle SAMP_D returns $i \in [n]$ drawn from D;
- *evaluation* oracle EVAL_D takes $j \in [n]$, and returns $D(j)$.

Definition (Cumulative Dual oracle)

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

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

Dual and Cumulative Dual access models

A couple remarks

- $\mathsf{SAMP} \preceq (\mathsf{SAMP}, \mathsf{EVAL}) \preceq (\mathsf{SAMP}, \mathsf{CEVAL})$

Dual and Cumulative Dual access models

A couple remarks

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

Dual and Cumulative Dual access models

A couple remarks

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

  SAMP: can get $D(i) \pm \varepsilon$
  COND: can get $(1 \pm \varepsilon)D(i)$
  here: can get $D(i)$

Dual and Cumulative Dual access models

A couple remarks

- $\mathrm{SAMP} \preceq (\mathrm{SAMP}, \mathrm{EVAL}) \preceq (\mathrm{SAMP}, \mathrm{CEVAL})$
- EVAL-only model considered in [RS09]; CEVAL-only in [BKR04]; $(\mathrm{SAMP}, \mathrm{EVAL})$ in part of [BDKR05, GMV05]
- Key point
  SAMP: can get $D(i) \pm \varepsilon$
  COND: can get $(1 \pm \varepsilon) D(i)$
  here: can get $D(i)$
- *How to motivate such a model?*

Dual and Cumulative Dual access models

Is that even a thing?

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 and Cumulative Dual access models

Is that even a thing?

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 in $O(\log n)$

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)

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

Techniques (1)

Upper bounds: Hey, we'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 (1)

Upper bounds: Hey, we'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 (1)

Upper bounds: Hey, we'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...

$$H(D) = -\sum_{i\in[n]} D(i)\log D(i) = -\mathbb{E}_{i\sim D}\left[\log D(i)\right]$$

Techniques (2)

Lower bounds: if I had a hammer…

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

$$\Omega\left(\frac{1}{\mathrm{d}_{\mathrm{TV}}(D^+, D^-)}\right)$$

samples.

$\rightsquigarrow$ *nice way to show lower bounds in the* SAMP *model!*

Techniques (2)

Lower bounds: if I had a hammer...

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

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

samples.

$\rightsquigarrow$ *nice way to show lower bounds in the* SAMP *model!*

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

Techniques (2)

Lower bounds: if I had a hammer...

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

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

samples.

$\rightsquigarrow$ *nice way to show lower bounds in the* SAMP *model!*

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

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.

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.

References IV

———, *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.