The Price of Tolerance in Distribution Testing

Clément Canonne (U Sydney)

WALD(O) 2021

(Joint work with Ayush Jain (UCSD), Gautam Kamath (U Waterloo), and Jerry Li (MSR))

arXiv:2106.13414

Distribution Testing?

Given m i.i.d. samples from some unknown p over $[n]$, a reference q, and distance parameter ε, distinguish between

- Completeness: $p = q$
- Soundness: $TV(p,q) \geq \varepsilon$

with probability at least 1/10, where

$$TV(p,q) := \sup(p(S) - q(S)) = \frac{1}{2} \ell_1(p,q)$$

Distribution Testing?

Given m i.i.d. samples from some unknown p over $[n]$, a reference q, and distance parameter ε, distinguish between

- Completeness: $p = q$
- Soundness: $\mathrm{TV}(p,q) \geq \varepsilon$

with probability at least 1/10, where

$$\mathrm{TV}(p,q) := \sup(p(S) - q(S)) = \tfrac{1}{2}\,\ell_1(p,q)$$

Identity testing/one-sample goodness-of-fit

[GR00, BFFKRW01, Pan08, VV14, DKN15, ADK15, DGPP18...]

Distribution Testing?

Given m i.i.d. samples from some unknown p over $[n]$, a reference q, and distance parameter ε, distinguish between

- Completeness: $p = q$
- Soundness: $\mathrm{TV}(p,q) \geq \varepsilon$

with probability at least 1/10, where

$$\mathrm{TV}(p,q) := \sup(p(S) - q(S)) = \tfrac{1}{2}\,\ell_1(p,q)$$

$$m = \Theta(\sqrt{n}/\varepsilon^2)$$

Identity testing/one-sample goodness-of-fit

[GR00, BFFKRW01, Pan08, VV14, DKN15, ADK15, DGPP18...]

Distribution Testing?

Given m i.i.d. samples from some unknown p over $[n]$, a reference q, and distance parameter ε, distinguish between

- Completeness: $TV(p,q) = 0$
- Soundness: $TV(p,q) \geq \varepsilon$

with probability at least 1/10, where

$$TV(p,q) := \sup(p(S) - q(S)) = \frac{1}{2} \ell_1(p,q)$$

Distribution Testing?

Given m i.i.d. samples from some unknown p over $[n]$, a reference q, and distance parameters ε_1, ε_2, distinguish between

- Completeness: $TV(p,q) \leq \varepsilon_1$
- Soundness: $TV(p,q) \geq \varepsilon_2$

with probability at least 1/10, where

$$TV(p,q) := \sup(p(S) - q(S)) = \tfrac{1}{2}\,\ell_1(p,q)$$

Distribution Testing?

Given m i.i.d. samples from some unknown p over $[n]$, a reference q, and distance parameters ε_1, ε_2, distinguish between

- Completeness: $TV(p,q) \leq \varepsilon_1$
- Soundness: $TV(p,q) \geq \varepsilon_2$

with probability at least 1/10, where

$$TV(p,q) := \sup(p(S) - q(S)) = \tfrac{1}{2}\, \ell_1(p,q)$$

Tolerant identity testing

[PRR06,VV10,VV11,HJW16,JVHW17,JHW18…]

Distribution Testing?

Given m i.i.d. samples from some unknown p over $[n]$, a reference q, and distance parameters ε_1, ε_2, distinguish between

- Completeness: $TV(p,q) \leq \varepsilon_1$
- Soundness: $TV(p,q) \geq \varepsilon_2$

with probability at least 1/10, where

$$TV(p,q) := \sup(p(S) - q(S)) = \tfrac{1}{2}\,\ell_1(p,q)$$

$$m = \Theta(n/\log n)$$

Tolerant identity testing

[PRR06,VV10,VV11,HJW16,JVHW17,JHW18...]

Distribution Testing?

Given m i.i.d. samples from some unknown p over $[n]$, a reference q, and distance parameters ε_1, ε_2, distinguish between

- Completeness: $TV(p,q) \leq \varepsilon_1$
- Soundness: $TV(p,q) \geq \varepsilon_2$

with probability at least 1/10, where

$$TV(p,q) := \sup(p(S) - q(S)) = \tfrac{1}{2}\, \ell_1(p,q)$$

$m = \Theta(n/\log n)$

Tolerant identity testing

[PRR06,VV10,VV11,HJW16,JVHW17,JHW18...]

So, erm, we're done?

So, erm, we're done?

Not quite.

So, erm, we're done?

$O(n/((\varepsilon_2-\varepsilon_1)^2 \log n))$

$\Theta(n/\log n)$ for constant $\varepsilon_1 < \varepsilon_2$

$\Omega(n/((\varepsilon_2-\varepsilon_1)^2 \log n))$ for distance estimation

$\Omega(\sqrt{n}/\varepsilon_2^2)$ always

$O(\sqrt{n}/\varepsilon_2^2)$ for $\varepsilon_1 \leq \varepsilon_2/\sqrt{n}$

So, erm, we're done?

$O(n/((\varepsilon_2-\varepsilon_1)^2 \log n))$

$\Theta(n/\log n)$ for
constant $\varepsilon_1 < \varepsilon_2$

🤔

$\Omega(n/((\varepsilon_2-\varepsilon_1)^2 \log n))$
for distance
estimation

$\Omega(\sqrt{n}/\varepsilon_2^2)$
always

$O(\sqrt{n}/\varepsilon_2^2)$
for $\varepsilon_1 \leq \varepsilon_2/\sqrt{n}$

This work:

$O(n/((\varepsilon_2-\varepsilon_1)^2 \log n))$

$\Theta(n/\log n)$ for constant $\varepsilon_1 < \varepsilon_2$

$\Omega(n/((\varepsilon_2-\varepsilon_1)^2 \log n))$ for distance estimation

$\Omega(\sqrt{n}/\varepsilon_2^2)$ always

$O(\sqrt{n}/\varepsilon_2^2)$ for $\varepsilon_1 \leq \varepsilon_2/\sqrt{n}$

This work: 

Tight sample complexity* as a function of n, ε_2, ε_1, in all parameter regimes.

This work: 

Tight sample complexity* as a function of n, ε_2, ε_1, in all parameter regimes.

$$\tilde{\Theta}\left(\frac{\sqrt{n}}{\varepsilon_2^2} + \frac{n}{\log n} \cdot \max\left\{\frac{\varepsilon_1}{\varepsilon_2^2}, \left(\frac{\varepsilon_1}{\varepsilon_2^2}\right)^2\right\}\right)$$

This work:

Tight sample complexity* as a function of n, ε_2, ε_1, in all parameter regimes.

$$\tilde{\Theta}\left(\frac{\sqrt{n}}{\varepsilon_2^2}+\frac{n}{\log n}\cdot\max\left\{\frac{\varepsilon_1}{\varepsilon_2^2},\left(\frac{\varepsilon_1}{\varepsilon_2^2}\right)^2\right\}\right)$$

*Up to a log n in the upper bound.

This work: 

Tight sample complexity* as a function of n, ε_2, ε_1, in all parameter regimes.

$$\tilde{\Theta}\left(\frac{\sqrt{n}}{\varepsilon_2^2} + \frac{n}{\log n} \cdot \max\left\{\frac{\varepsilon_1}{\varepsilon_2^2}, \left(\frac{\varepsilon_1}{\varepsilon_2^2}\right)^2\right\}\right)$$

+ Analogous statement for tolerant **closeness** testing

Some remarks

New result even in the "low tolerance" regime: Cauchy–Schwarz wasn't tight!

$$\Theta(\sqrt{n}/\varepsilon_2^2) \text{ for } \varepsilon_1 \leq \varepsilon_2/\sqrt{n} \quad \rightsquigarrow \quad \Theta(\sqrt{n}/\varepsilon_2^2) \text{ for } \varepsilon_1 \leq \mathbf{1/\sqrt{n}}$$

Some remarks

New result even in the "low tolerance" regime: Cauchy–Schwarz wasn't tight!

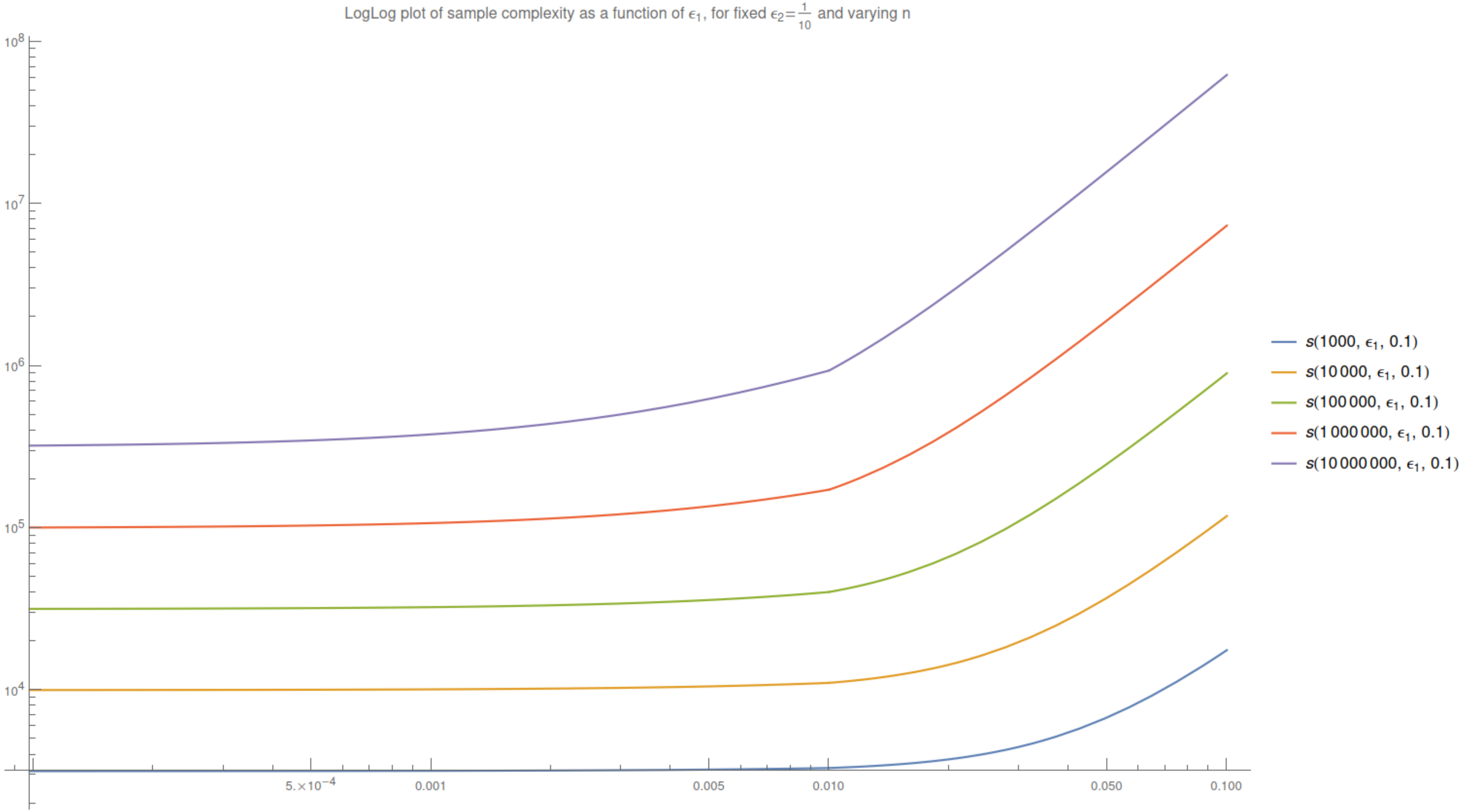
$\Theta(\sqrt{n}/\varepsilon_2^2)$ for $\varepsilon_1 \leq \varepsilon_2/\sqrt{n}$ $\leadsto$ $\Theta(\sqrt{n}/\varepsilon_2^2)$ for $\varepsilon_1 \leq \mathbf{1/\sqrt{n}}$

Some remarks

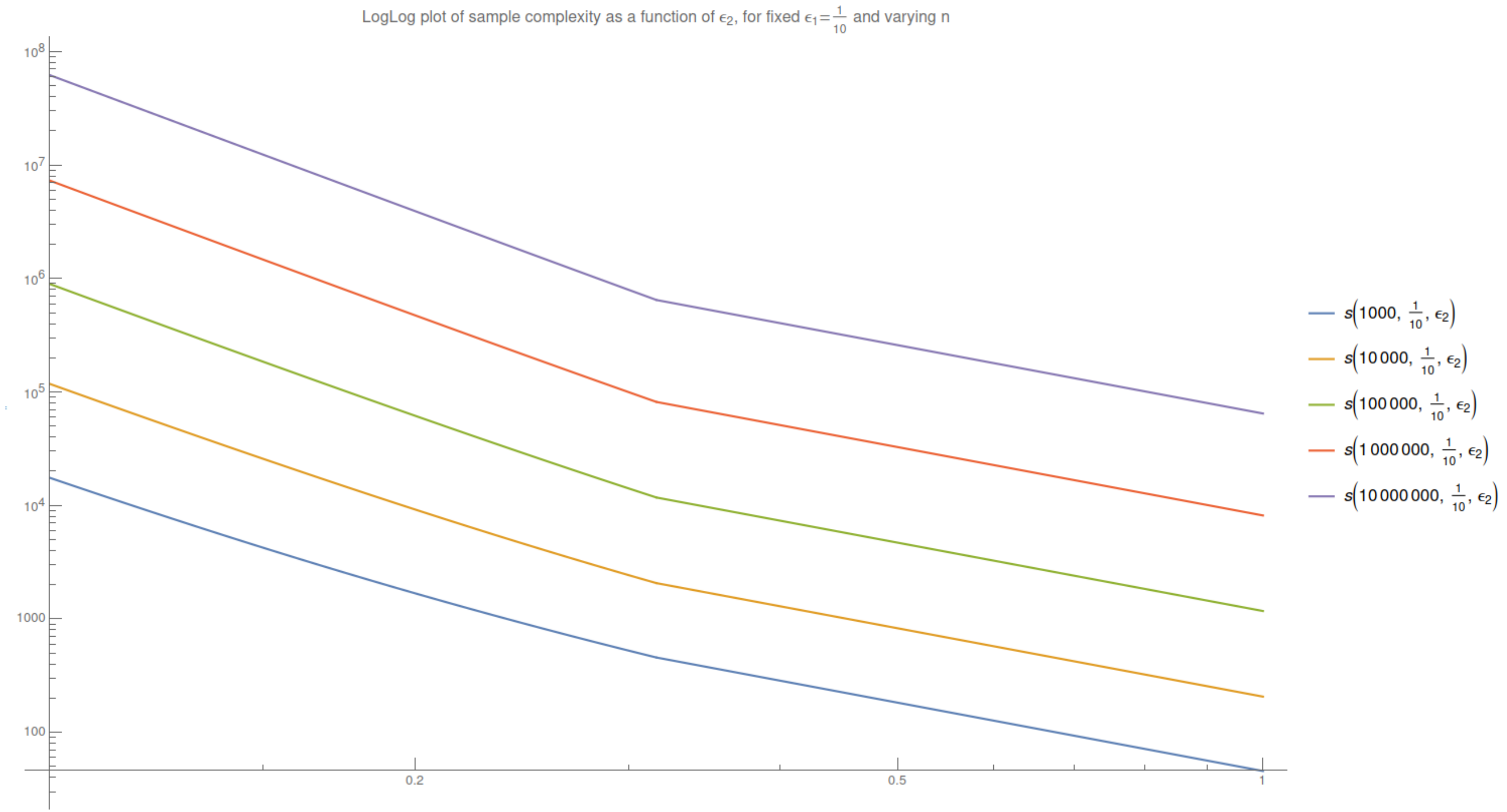
New result even in the "low tolerance" regime: Cauchy–Schwarz wasn't tight! 😮

$$\Theta(\sqrt{n}/\varepsilon_2^2) \text{ for } \varepsilon_1 \leq \varepsilon_2/\sqrt{n} \quad \rightsquigarrow \quad \Theta(\sqrt{n}/\varepsilon_2^2) \text{ for } \varepsilon_1 \leq \mathbf{1/\sqrt{n}}$$

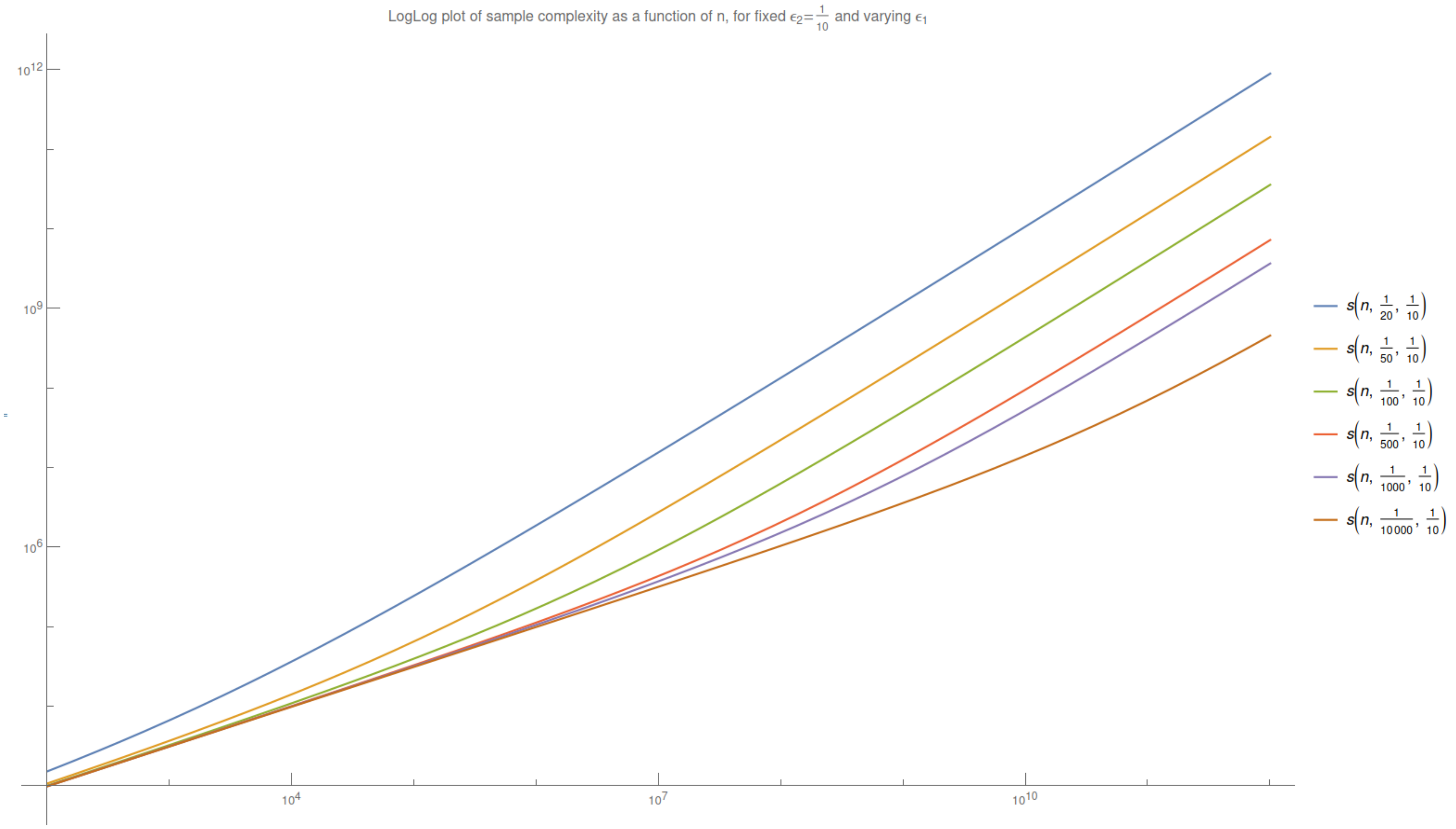
Some remarks plots



Some remarks plots



Some remarks plots



More remarks

Can use our results on tolerant **identity** and **closeness** to obtain tolerant testing for/under **structure**:

- Monotonicity, unimodality, log-concavity, independence
- Under structural assumptions ("A_k distances")
- Instance-optimal tolerant testing
- (insert favourite property here)

A glimpse of our techniques 👀

Upper bounds via carefully **rescaled** χ^2/ℓ_2-type tester

A glimpse of our techniques

Upper bounds via carefully **rescaled** χ^2/ℓ_2-type tester

Algorithm 1 Tolerant testing algorithm.

Require: $0 \leq \varepsilon_1 < \varepsilon_2 \leq 1$, m, n, two sets of $\mathrm{Poi}(m)$ samples from both p and q

Set the threshold

$$\tau \leftarrow c \cdot \min\left(\frac{m^{3/2}\varepsilon_2}{n^{1/2}}, \frac{m^2\varepsilon_2^2}{n}\right)$$

Compute Z from the sets of samples, as per (1).

if $Z \geq \tau$ **then return** $\|p-q\|_1 \geq \varepsilon_2$

else return $\|p-q\|_1 \leq \varepsilon_1$

end if

A glimpse of our techniques

Upper bounds via carefully **rescaled** χ^2/ℓ_2-type tester

Algorithm 1 Tolerant testing algorithm.

Require: $0 \leq \varepsilon_1 < \varepsilon_2 \leq 1$, m, n, two sets of $\mathrm{Poi}(m)$ samples from both p and q

Set the threshold

$$Z_i := (X_i - Y_i)^2 - X_i - Y_i$$

$$Z := \sum_{i=1}^{n} \frac{Z_i}{\widehat{f}_i}.$$

$$\tau \leftarrow c \cdot \min\left(\frac{m^{3/2}\varepsilon_2}{n^{1/2}}, \frac{m^2\varepsilon_2^2}{n}\right)$$

Compute Z from the sets of samples, as per (1).

if $Z \geq \tau$ **then return** $\|p-q\|_1 \geq \varepsilon_2$

else return $\|p-q\|_1 \leq \varepsilon_1$

end if

$$\widehat{f}_i := \begin{cases} \max\left\{\frac{|\tilde{X}_i - \tilde{Y}_i|}{\sqrt{m/n}}, \frac{\tilde{X}_i + \tilde{Y}_i}{m/n}, 1\right\} & \text{if } m \geq n \\ \max\{\tilde{X}_i + \tilde{Y}_i, 1\} & \text{if } m < n. \end{cases}$$

A glimpse of our techniques

Upper bounds via carefully **rescaled** χ^2/ℓ_2-type tester

Algorithm 1 Tolerant testing algorithm.

Require: $0 \le \varepsilon_1 < \varepsilon_2 \le 1$, m, n, two sets of $\mathrm{Poi}(m)$ samples from both p and q

Set the threshold

$$Z_i := (X_i - Y_i)^2 - X_i - Y_i$$

$$Z := \sum_{i=1}^{n} \frac{Z_i}{\widehat{f}_i}.$$

$$\tau \leftarrow c \cdot \min\left(\frac{m^{3/2}\varepsilon_2}{n^{1/2}}, \frac{m^2\varepsilon_2^2}{n}\right)$$

Compute Z from the sets of samples, as per (1).

if $Z \ge \tau$ **then return** $\|p-q\|_1 \ge \varepsilon_2$

else return $\|p-q\|_1 \le \varepsilon_1$

end if

$$\widehat{f}_i := \begin{cases} \max\left\{\frac{|\tilde{X}_i - \tilde{Y}_i|}{\sqrt{m/n}}, \frac{\tilde{X}_i + \tilde{Y}_i}{m/n}, 1\right\} & \text{if } m \ge n \\ \max\{\tilde{X}_i + \tilde{Y}_i, 1\} & \text{if } m < n. \end{cases}$$

Does not require knowledge of ε_1!

A glimpse of our techniques 👀

Lower bounds via moment-matching technique *à la* Wu—Yang

A glimpse of our techniques

Lower bounds via moment-matching technique *à la* Wu—Yang

- Reduce the problem to designing 2 r.v.'s U,U' with matching moments 🐟
- Write this as an LP ✍: need to lower bound its optimal value
- Massage this LP, take the dual 🪄
- Try to lower bound the dual's optimal value
  - Weep 😭, as the asymmetry *(testing, not estimation!)* rules out using the same polynomial uniform approximation results as in [WY]
  - Obtain the lower bound without this machinery 🏋

Lower bounds via moment-matching technique *à la* Wu—Yang

- Reduce the problem to designing 2 r.v.'s U,U' with matching moments 🧩
- Write this as an LP ✍️: need to lower bound its optimal value
- Massage this LP, take the dual 🪄
- Try to lower bound the dual's optimal value
  - Weep 😭, as the asymmetry *(testing, not estimation!)* rules out using the same polynomial uniform approximation results as in [WY]
  - Obtain the lower bound without this machinery 🏋️

Recap

Tight sample complexity* for tolerant identity** testing as a function of n, ε_2, ε_1, in all parameter regimes.

$$\tilde{\Theta}\left(\frac{\sqrt{n}}{\varepsilon_2^2} + \frac{n}{\log n} \cdot \max\left\{\frac{\varepsilon_1}{\varepsilon_2^2}, \left(\frac{\varepsilon_1}{\varepsilon_2^2}\right)^2\right\}\right)$$

arXiv:2106.13414

* Up to a log n in the upper bound.

** And closeness.

Recap

Tight sample complexity* for tolerant identity** testing as a function of n, ε_2, ε_1, in all parameter regimes.

$$\tilde{\Theta}\left(\frac{\sqrt{n}}{\varepsilon_2^2} + \frac{n}{\log n} \cdot \max\left\{\frac{\varepsilon_1}{\varepsilon_2^2}, \left(\frac{\varepsilon_1}{\varepsilon_2^2}\right)^2\right\}\right)$$

Thank You!

arXiv:2106.13414

* Up to a log n in the upper bound.

** And closeness.