

The Price of Tolerance in Distribution Testing

Clément Canonne

WALD(O) 2021

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



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: $TV(p, q) \geq \epsilon$

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: $TV(p, q) \geq \epsilon$

with probability at least $1/10$, where

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

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: $TV(p, q) \geq \epsilon$

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

with probability at least $1/10$, where

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

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

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: $TV(p, q) \leq \epsilon_1$
- Soundness: $TV(p, q) \geq \epsilon_2$

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: $TV(p, q) \leq \epsilon_1$
- Soundness: $TV(p, q) \geq \epsilon_2$

with probability at least $1/10$, where

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

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: $TV(p, q) \leq \epsilon_1$
- Soundness: $TV(p, q) \geq \epsilon_2$

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

with probability at least $1/10$, where

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

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: $TV(p, q) \leq \epsilon_1$
- Soundness: $TV(p, q) \geq \epsilon_2$

with probability at least $1/10$, where

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

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



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?

$$O(n/((\epsilon_2 - \epsilon_1)^2 \log n))$$

$\Theta(n/\log n)$ for
constant $\epsilon_1 < \epsilon_2$

$\Omega(n/((\epsilon_2 - \epsilon_1)^2 \log n))$
for distance
estimation

$\Omega(\sqrt{n}/\epsilon_2^2)$
always

$O(\sqrt{n}/\epsilon_2^2)$
for $\epsilon_1 \leq \epsilon_2/\sqrt{n}$

So, erm, we're done?

$$O(n/((\epsilon_2 - \epsilon_1)^2 \log n))$$

$\Theta(n/\log n)$ for
constant $\epsilon_1 < \epsilon_2$



$\Omega(n/((\epsilon_2 - \epsilon_1)^2 \log n))$
for distance
estimation

$\Omega(\sqrt{n}/\epsilon_2^2)$
always

$O(\sqrt{n}/\epsilon_2^2)$
for $\epsilon_1 \leq \epsilon_2/\sqrt{n}$

This work:

$$O(n/((\epsilon_2 - \epsilon_1)^2 \log n))$$

$\Theta(n/\log n)$ for
constant $\epsilon_1 < \epsilon_2$



$\Omega(n/((\epsilon_2 - \epsilon_1)^2 \log n))$
for distance
estimation

$\Omega(\sqrt{n}/\epsilon_2^2)$
always

$O(\sqrt{n}/\epsilon_2^2)$
for $\epsilon_1 \leq \epsilon_2/\sqrt{n}$

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!

$$\begin{array}{ccc} \Theta(\sqrt{n}/\varepsilon_2^2) & \rightsquigarrow & \Theta(\sqrt{n}/\varepsilon_2^2) \\ \text{for } \varepsilon_1 \leq \varepsilon_2/\sqrt{n} & & \text{for } \varepsilon_1 \leq \mathbf{1}/\sqrt{n} \end{array}$$

Some remarks

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

$$\begin{array}{ccc} \Theta(\sqrt{n}/\varepsilon_2^2) & \rightsquigarrow & \Theta(\sqrt{n}/\varepsilon_2^2) \\ \text{for } \varepsilon_1 \leq \varepsilon_2/\sqrt{n} & & \text{for } \varepsilon_1 \leq \mathbf{1}/\sqrt{n} \end{array}$$

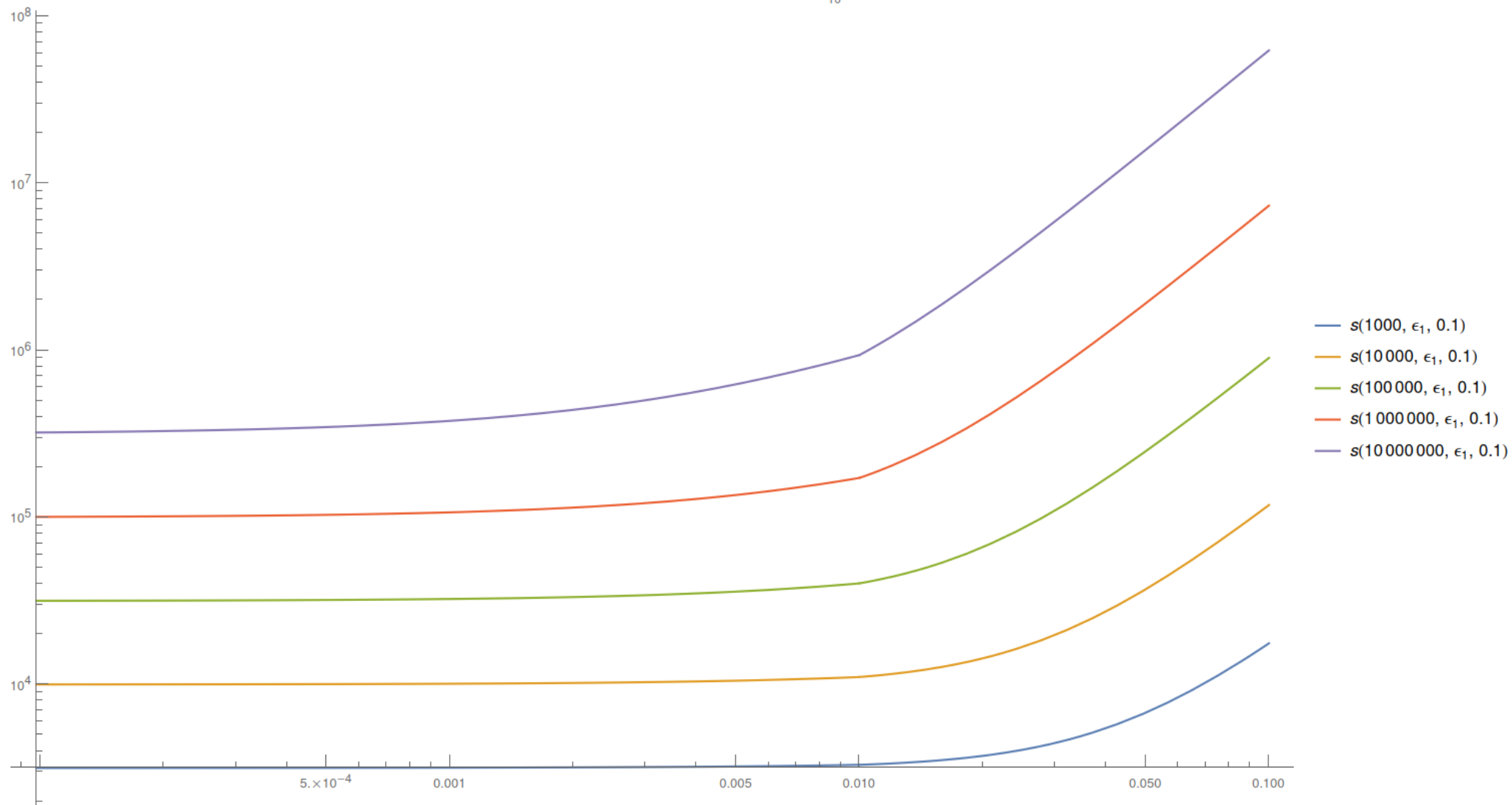
Some remarks

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

$$\begin{array}{ccc} \Theta(\sqrt{n}/\varepsilon_2^2) & \rightsquigarrow & \Theta(\sqrt{n}/\varepsilon_2^2) \\ \text{for } \varepsilon_1 \leq \varepsilon_2/\sqrt{n} & & \text{for } \varepsilon_1 \leq \mathbf{1}/\sqrt{n} \end{array}$$

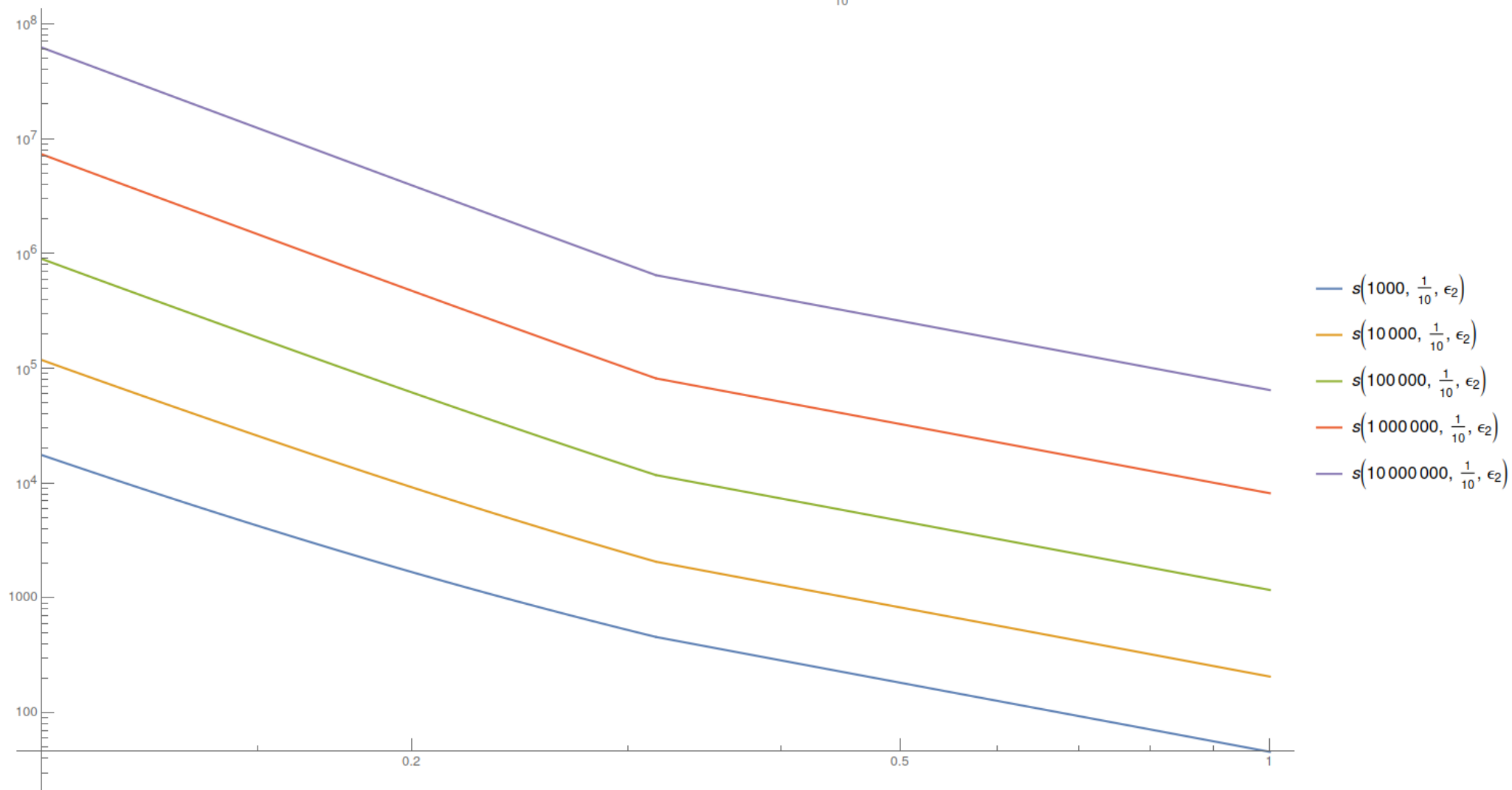
Some remarks plots

LogLog plot of sample complexity as a function of ϵ_1 , for fixed $\epsilon_2 = \frac{1}{10}$ and varying n



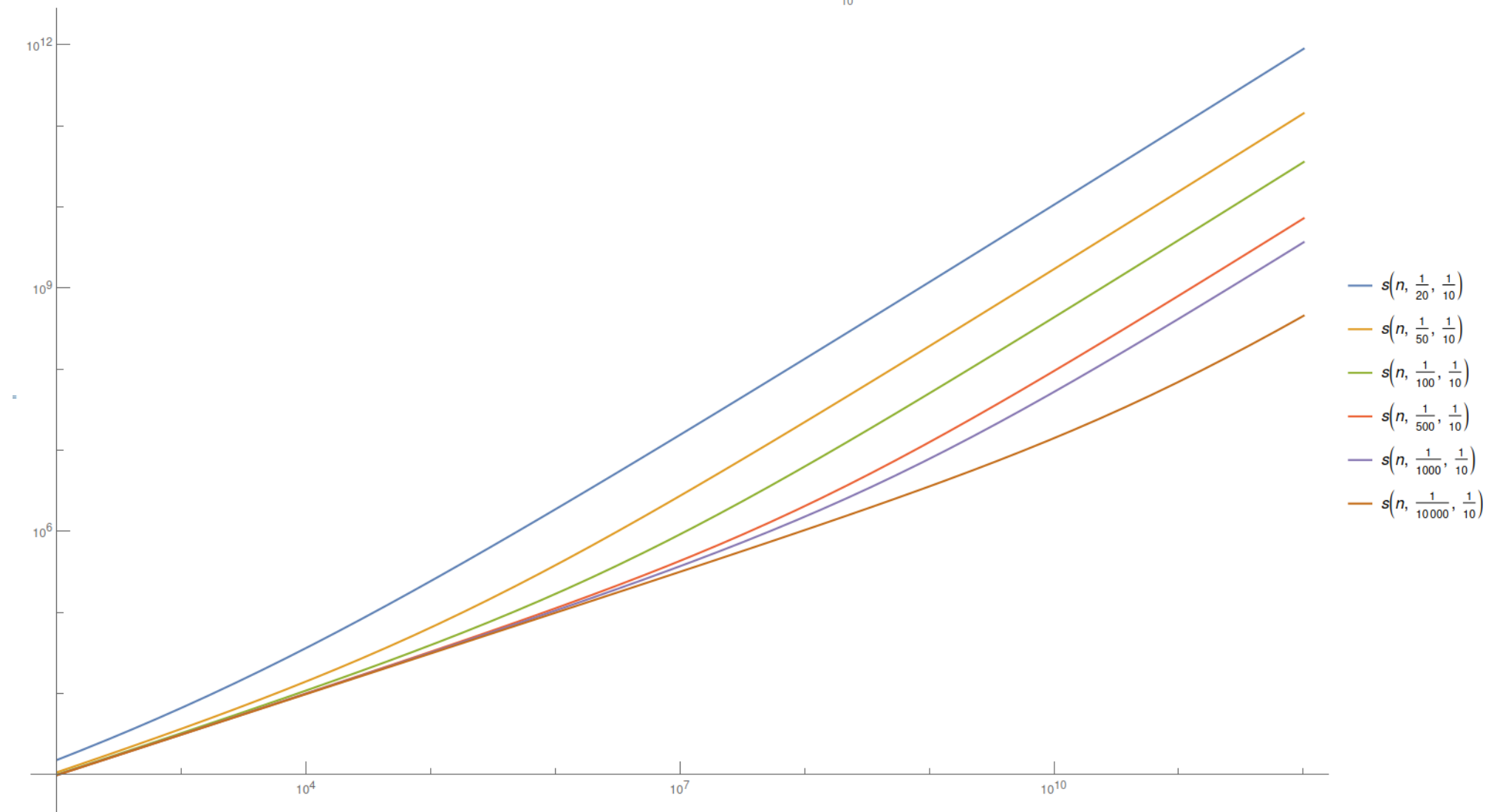
Some remarks plots

LogLog plot of sample complexity as a function of ϵ_2 , for fixed $\epsilon_1 = \frac{1}{10}$ and varying n



Some remarks plots


LogLog plot of sample complexity as a function of n , for fixed $\epsilon_2 = \frac{1}{10}$ and varying ϵ_1



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 $\text{Poi}(m)$ 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 $\text{Poi}(m)$ samples from both p and q

Set the threshold

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

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

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

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

Set the threshold

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

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

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

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

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

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.

$$\tilde{\Theta} \left(\frac{\sqrt{n}}{\epsilon_2^2} + \frac{n}{\log n} \cdot \max \left\{ \frac{\epsilon_1}{\epsilon_2^2}, \left(\frac{\epsilon_1}{\epsilon_2^2} \right)^2 \right\} \right)$$

Recap

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

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

Thank You!