Sharp Constants in Uniformity Testing via the Huber Statistic

Shivam Gupta (UT Austin), Eric Price (UT Austin)

July 18, 2023

Uniformity Testing

Definition

Given n samples from a discrete distribution q on $[m]$, determine whether q is the uniform distribution u, or ε-far from u in TV distance, with probability $1 - \delta$

Uniformity Testing

Definition

Given n samples from a discrete distribution q on $[m]$, determine whether q is the uniform distribution u, or ε-far from u in TV distance, with probability $1 - \delta$

- First introduced by Goldreich and Ron in the context of testing whether a bounded-degree regular graph is an expander

Uniformity Testing

Definition

Given n samples from a discrete distribution q on $[m]$, determine whether q is the uniform distribution u, or ε-far from u in TV distance, with probability $1 - \delta$

- First introduced by Goldreich and Ron in the context of testing whether a bounded-degree regular graph is an expander
- Is used as a basic building block for identity testing

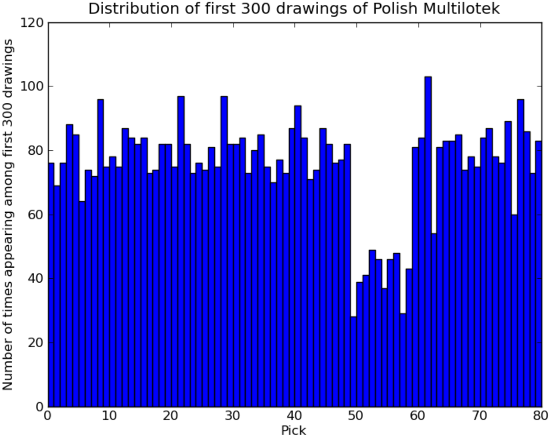
Uniformity Testing

Definition

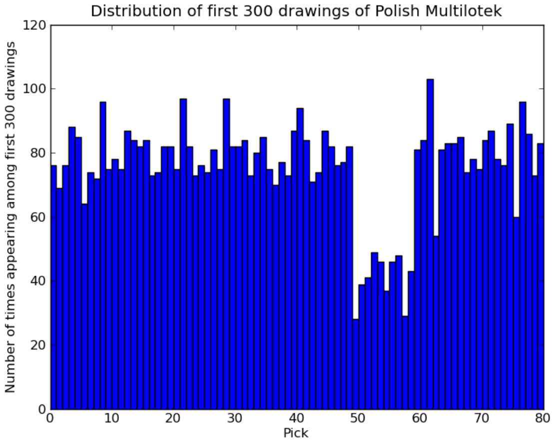
Given n samples from a discrete distribution q on $[m]$, determine whether q is the uniform distribution u, or ε-far from u in TV distance, with probability $1 - \delta$

- First introduced by Goldreich and Ron in the context of testing whether a bounded-degree regular graph is an expander
- Is used as a basic building block for identity testing
- Very well-studied [Goldreich and Ron, 2011; Batu, Fischer, Fortnow, Kumar, Rubinfeld, White, 2000; Paninski, 2008; Diakonikolas, Gouleakis, Peebles, Price, 2018; Diakonikolas, Gouleakis, Peebles, Price, 2019] various testers (collisions, TV, singleton) considered in the literature, matching upper and lower bounds known [DGPP18].

Motivation


Distribution of first 300 drawings of Polish Multilotek
Number of times appearing among first 300 drawings
Pick
0
10
20
30
40
50
60
70
80
0
20
40
60
80
100
120

Motivation



Question

How fast could the Polish lottery error be detected?

Notation

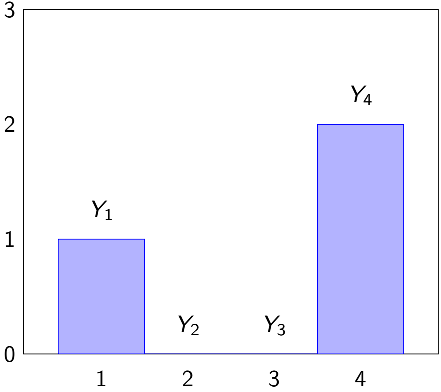
- n is the number of samples

Notation

- n is the number of samples
- m is the domain of the distribution

Notation

- n is the number of samples
- m is the domain of the distribution
- Let Y_j be the number of samples drawn that are equal to j.



Strategy

- Compute some test statistic $S = \sum_{j=1}^{m} f(Y_j)$. All existing statistics are of this form.

Strategy

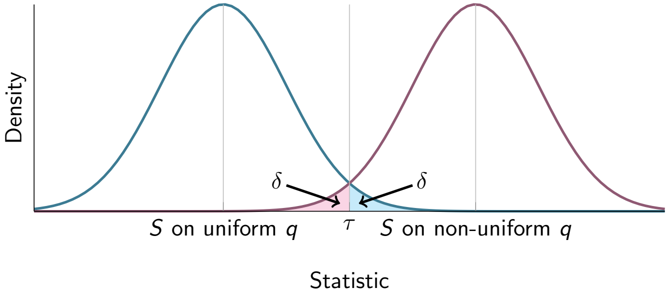
- Compute some test statistic $S = \sum_{j=1}^{m} f(Y_j)$. All existing statistics are of this form.
- Accept if $S \leq \tau$, and reject otherwise, for some threshold τ.

Strategy

- Compute some test statistic $S = \sum_{j=1}^{m} f(Y_j)$. All existing statistics are of this form.
- Accept if $S \leq \tau$, and reject otherwise, for some threshold τ.

Strategy

- Compute some test statistic $S = \sum_{j=1}^{m} f(Y_j)$. All existing statistics are of this form.
- Accept if $S \leq \tau$, and reject otherwise, for some threshold τ.



Sample Complexity Intuition

- Need $\Omega\left(\frac{m}{\varepsilon^2}\right)$ samples to *learn* the distribution

Sample Complexity Intuition

- Need $\Omega\left(\frac{m}{\varepsilon^2}\right)$ samples to *learn* the distribution
- Birthday paradox tells us that under the uniform distribution, we will start to see collisions after $O(\sqrt{m})$ samples

Sample Complexity Intuition

- Need $\Omega\left(\frac{m}{\varepsilon^2}\right)$ samples to *learn* the distribution
- Birthday paradox tells us that under the uniform distribution, we will start to see collisions after $O(\sqrt{m})$ samples
- Should see more collisions under any ε-far distribution

Overview of Existing Results

Tester	Test Statistic	Sample Complexity	Notes
Collisions	$\sum_{j=1}^{m} \binom{Y_j}{2}$	$\Theta\left(\frac{\sqrt{m}}{\varepsilon^2} \log \frac{1}{\delta}\right)$	[BFR+00, GR11, DGPP19]
Singletons	$\sum_{j=1}^{m} \mathbb{1}_{Y_j=1}$	$\Theta\left(\frac{\sqrt{m \log \frac{1}{\delta}}}{\varepsilon^2}\right)$	[Pan08], when $n = o(m)$
TV	$\left\lVert \frac{Y}{m} - u \right\rVert_{TV}$	$\Theta\left(\frac{\sqrt{m \log \frac{1}{\delta}}}{\varepsilon^2} + \frac{\log \frac{1}{\delta}}{\varepsilon^2}\right)$	[DGPP18]

Table: Existing Testers

Overview of Existing Results

Tester	Test Statistic	Sample Complexity	Notes
Collisions	$\sum_{j=1}^{m} \binom{Y_j}{2}$	$\Theta\left(\frac{\sqrt{m}}{\varepsilon^2} \log \frac{1}{\delta}\right)$	[BFR+00, GR11, DGPP19]
Singletons	$\sum_{j=1}^{m} \mathbb{1}_{Y_j=1}$	$\Theta\left(\frac{\sqrt{m \log \frac{1}{\delta}}}{\varepsilon^2}\right)$	[Pan08], when $n = o(m)$
TV	$\left\lVert \frac{Y}{m} - u \right\rVert_{TV}$	$\Theta\left(\frac{\sqrt{m \log \frac{1}{\delta}}}{\varepsilon^2} + \frac{\log \frac{1}{\delta}}{\varepsilon^2}\right)$	[DGPP18]

Table: Existing Testers

- There is a lower bound that matches the sample complexity of the TV tester (up to constants) [DGPP18].

Overview of Existing Results

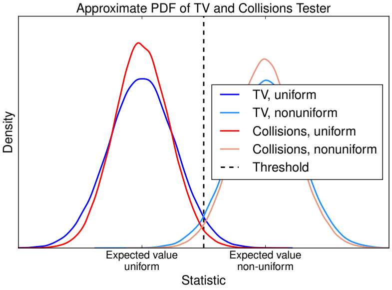
Tester	Test Statistic	Sample Complexity	Notes
Collisions	$\sum_{j=1}^{m} \binom{Y_j}{2}$	$\Theta\left(\frac{\sqrt{m}}{\varepsilon^2} \log \frac{1}{\delta}\right)$	[BFR+00, GR11, DGPP19]
Singletons	$\sum_{j=1}^{m} \mathbb{1}_{Y_j=1}$	$\Theta\left(\frac{\sqrt{m \log \frac{1}{\delta}}}{\varepsilon^2}\right)$	[Pan08], when $n = o(m)$
TV	$\left\lVert \frac{Y}{m} - u \right\rVert_{TV}$	$\Theta\left(\frac{\sqrt{m \log \frac{1}{\delta}}}{\varepsilon^2} + \frac{\log \frac{1}{\delta}}{\varepsilon^2}\right)$	[DGPP18]

Table: Existing Testers

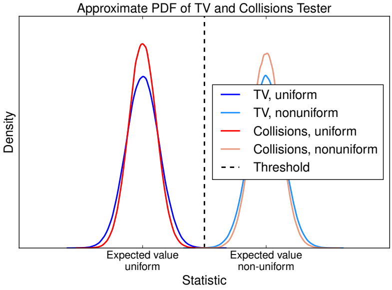
- There is a lower bound that matches the sample complexity of the TV tester (up to constants) [DGPP18].

These results suggest that one should use the TV tester in practice

Empirical Study



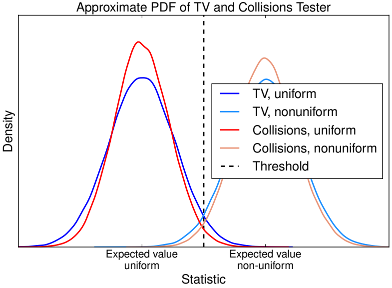
(a) The collisions tester has 1.7% error rate and the TV tester has 3.3%. ($m = n = 10000$ and $\varepsilon = 0.125$)



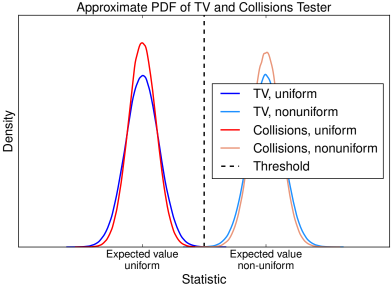
(b) The collisions tester has 10^{-5} error rate and the TV tester has 10^{-4}. ($m = n = 10^5$ and $\varepsilon = 0.1$)

Figure: Observed performance of TV vs collisions

Empirical Study



(a) The collisions tester has 1.7% error rate and the TV tester has 3.3%. ($m = n = 10000$ and $\varepsilon = 0.125$)

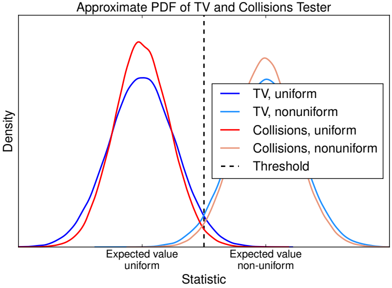


(b) The collisions tester has 10^{-5} error rate and the TV tester has 10^{-4}. ($m = n = 10^5$ and $\varepsilon = 0.1$)

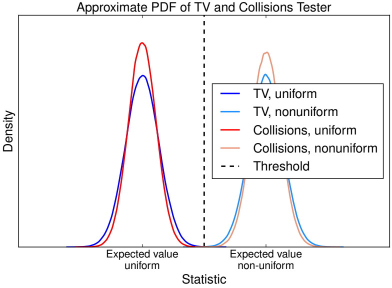
Figure: Observed performance of TV vs collisions

- Theory: TV Tester optimal, Collisions tester asymptotically bad in δ

Empirical Study



(a) The collisions tester has 1.7% error rate and the TV tester has 3.3%. ($m = n = 10000$ and $\varepsilon = 0.125$)



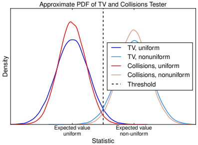
(b) The collisions tester has 10^{-5} error rate and the TV tester has 10^{-4}. ($m = n = 10^5$ and $\varepsilon = 0.1$)

Figure: Observed performance of TV vs collisions

- Theory: TV Tester optimal, Collisions tester asymptotically bad in δ
- Practice: Collisions tester is better, even for tiny δ

What's going on?

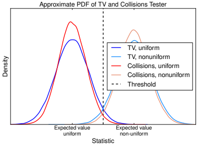
$$S = \sum_{j=1}^{m} f(Y_j)$$



- The distribution *looks* Gaussian, since Y_j are mostly independent

What's going on?

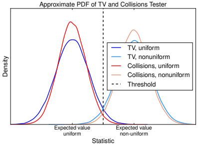
$$S = \sum_{j=1}^{m} f(Y_j)$$



- The distribution *looks* Gaussian, since Y_j are mostly independent
- If it *were* Gaussian, all that matters is the variance (after normalizing the expectation gap)

What's going on?

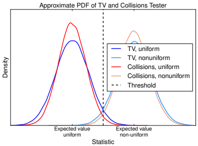
$$S = \sum_{j=1}^{m} f(Y_j)$$



- The distribution *looks* Gaussian, since Y_j are mostly independent
- If it *were* Gaussian, all that matters is the variance (after normalizing the expectation gap)
- We show that the collisions statistic optimizes this, while TV has 44% larger variance when $n = m \to \infty$.

What's going on?

$$S = \sum_{j=1}^{m} f(Y_j)$$



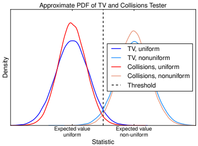
- The distribution *looks* Gaussian, since Y_j are mostly independent
- If it *were* Gaussian, all that matters is the variance (after normalizing the expectation gap)
- We show that the collisions statistic optimizes this, while TV has 44% larger variance when $n = m \to \infty$.

Theorem (informal)

The collisions statistic has minimum variance over all separable statistics.

What's going on?

$$S = \sum_{j=1}^{m} f(Y_j)$$



- The distribution *looks* Gaussian, since Y_j are mostly independent
- If it *were* Gaussian, all that matters is the variance (after normalizing the expectation gap)
- We show that the collisions statistic optimizes this, while TV has 44% larger variance when $n = m \to \infty$.

Theorem (informal)

The collisions statistic has minimum variance over all separable statistics.

- The collisions statistic actually has *exponential* tails when you go far enough away from the mean

Overview of Results

- The sample complexity of a tester can be expressed as

$$n = (C + o(1))\frac{\sqrt{m \log \frac{1}{\delta}}}{\varepsilon^2} + O\left(\frac{\log \frac{1}{\delta}}{\varepsilon^2}\right)$$

Overview of Results

- The sample complexity of a tester can be expressed as

$$n = (C + o(1))\frac{\sqrt{m \log \frac{1}{\delta}}}{\varepsilon^2} + O\left(\frac{\log \frac{1}{\delta}}{\varepsilon^2}\right)$$

Regime	**Dominant Term**
Sublinear	$\frac{\sqrt{m \log \frac{1}{\delta}}}{\varepsilon^2}$
Superlinear	$\frac{\log \frac{1}{\delta}}{\varepsilon^2}$

Table: Regimes

Overview of Results

- The sample complexity of a tester can be expressed as

$$n = (C + o(1))\frac{\sqrt{m \log \frac{1}{\delta}}}{\varepsilon^2} + O\left(\frac{\log \frac{1}{\delta}}{\varepsilon^2}\right)$$

Regime	**Dominant Term**
Sublinear	$\frac{\sqrt{m \log \frac{1}{\delta}}}{\varepsilon^2}$
Superlinear	$\frac{\log \frac{1}{\delta}}{\varepsilon^2}$

Table: Regimes

- We will focus on the sublinear regime in this talk

Overview of Results

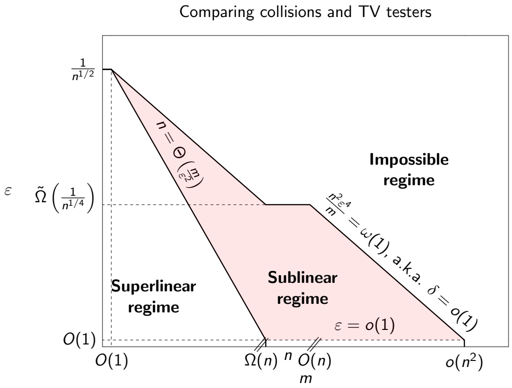


Figure: The constant C in different (n, ε, δ) parameter regimes.

Overview of Results

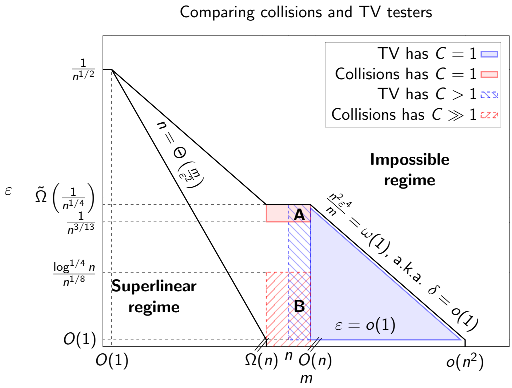


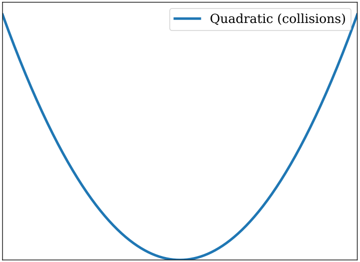
Figure: The constant C in different (n, ε, δ) parameter regimes.

In region **A**, the collisions tester performs better than the TV tester. In region **B**, both the collisions and TV tester have $C > 1$. When $n = m$, $C \approx 1.2$ for the TV tester.

Huber statistic

$$S = \sum_j f(Y_j)$$

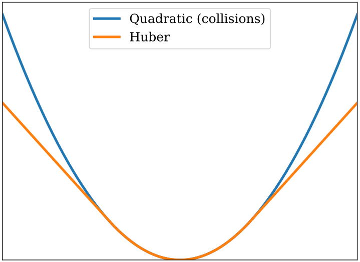
Statistic	Variance	Tails
Collisions	Optimal	Heavy
TV	Suboptimal	Gaussian



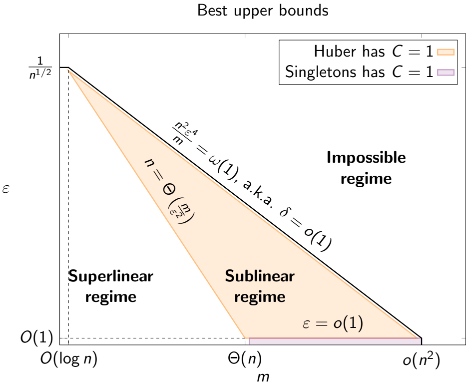
Huber statistic

$$S = \sum_j f(Y_j)$$

Statistic	Variance	Tails
Collisions	Optimal	Heavy
TV	Suboptimal	Gaussian
Huber	Optimal	Gaussian



Overview of Results



The Huber statistic achieves the best constant over the Sublinear regime when $\varepsilon = o(1)$. It matches the Gaussian approximation to the test statistic with optimal variance.

Analysis

- **Recap:** For a statistic $S = \sum_{j=1}^{m} f(Y_j)$, we need to understand the false negative probability under uniform distribution u

$$\delta_- := \Pr_u[S \geq \tau]$$

Analysis

- **Recap:** For a statistic $S = \sum_{j=1}^{m} f(Y_j)$, we need to understand the false negative probability under uniform distribution u

$$\delta_- := \Pr_u[S \geq \tau]$$

MGF $\longrightarrow$ Poissonization $\longrightarrow$ Depoissonization $\longrightarrow$ Chernoff

Analysis - Computing the MGF

MGF → Poissonization → Depoissonization → Chernoff

Analysis - Computing the MGF

MGF → Poissonization → Depoissonization → Chernoff

- We want to analyze the MGF given by

$$M_S(t) = \mathbb{E}[\exp(tS)] = \mathbb{E}\left[\prod_{j=1}^{m} \exp(t \cdot f(Y_j))\right]$$

Analysis - Computing the MGF

MGF → Poissonization → Depoissonization → Chernoff

- We want to analyze the MGF given by

$$M_S(t) = \mathbb{E}[\exp(tS)] = \mathbb{E}\left[\prod_{j=1}^{m} \exp(t \cdot f(Y_j))\right]$$

- Since the Y_j's are not independent, the expectation and product cannot be interchanged, and this is difficult to compute

Analysis - Poissonization

MGF → **Poissonization** → Depoissonization → Chernoff

Analysis - Poissonization

MGF $\longrightarrow$ Poissonization $\longrightarrow$ Depoissonization $\longrightarrow$ Chernoff

- To overcome this problem, let $S_{Poi(\lambda)}$ be the Poissonized statistic, i.e., statistic S with the number of samples sampled from $Poi(\lambda)$. Let $Z \sim Poi(\lambda)$ be the number of samples sampled

Analysis - Poissonization

MGF → **Poissonization** → Depoissonization → Chernoff

- To overcome this problem, let $S_{Poi(\lambda)}$ be the Poissonized statistic, i.e., statistic S with the number of samples sampled from $Poi(\lambda)$. Let $Z \sim Poi(\lambda)$ be the number of samples sampled
- Its MGF is given by

$$A_\lambda(t) := \mathbb{E}[\exp(t \cdot S_{Poi(\lambda)})]$$

This turns out to be easy to compute

Analysis - Depoissonization

MGF → Poissonization → **Depoissonization** → Chernoff

- Unfortunately, the Poissonized statistic does not concentrate as well

Analysis - Depoissonization

MGF $\longrightarrow$ Poissonization $\rightarrow$ **Depoissonization** $\longrightarrow$ Chernoff

- Unfortunately, the Poissonized statistic does not concentrate as well

$$A_\lambda(t) = \mathbb{E}[\exp(t \cdot S_{Poi(\lambda)})] = \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} \mathbb{E}[\exp(t \cdot S_{Poi(\lambda)}) | Z = k]$$

Analysis - Depoissonization

$$\boxed{\text{MGF}} \longrightarrow \boxed{\text{Poissonization}} \longrightarrow \boxed{\text{Depoissonization}} \longrightarrow \boxed{\text{Chernoff}}$$

- Unfortunately, the Poissonized statistic does not concentrate as well

$$A_\lambda(t) = \mathbb{E}[\exp(t \cdot S_{Poi(\lambda)})] = \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} \mathbb{E}[\exp(t \cdot S_{Poi(\lambda)}) | Z = k]$$

- Here, since $(S_{Poi(\lambda)} | Z = n)$ is precisely our original statistic S, the coefficient of λ^n in $e^\lambda A_\lambda(t)$ is

$$\frac{1}{n!} M_S(t)$$

Analysis - Depoissonization

MGF → Poissonization → Depoissonization → Chernoff

- Unfortunately, the Poissonized statistic does not concentrate as well

$$A_\lambda(t) = \mathbb{E}[\exp(t \cdot S_{Poi(\lambda)})] = \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} \mathbb{E}[\exp(t \cdot S_{Poi(\lambda)}) | Z = k]$$

- Here, since $(S_{Poi(\lambda)} | Z = n)$ is precisely our original statistic S, the coefficient of λ^n in $e^\lambda A_\lambda(t)$ is

$$\frac{1}{n!} M_S(t)$$

- For the statistics we analyze, $A_\lambda(t)$ is holomorphic in λ, and so, we can compute M_S using Cauchy's integral formula:

$$M_S(t) = \frac{n!}{2\pi i} \oint e^\lambda A_\lambda(t) \frac{d\lambda}{\lambda^{n+1}}$$

Analysis - Chernoff Bound

MGF → Poissonization → Depoissonization → Chernoff

- Once we have the MGF, Chernoff-type arguments imply

$$\delta_- < \inf_{t \geq 0} \frac{M_S(t)}{e^{t\tau}}$$

Analysis - Chernoff Bound

MGF → Poissonization → Depoissonization → **Chernoff**

- Once we have the MGF, Chernoff-type arguments imply

$$\delta_- < \inf_{t \geq 0} \frac{M_S(t)}{e^{t\tau}}$$

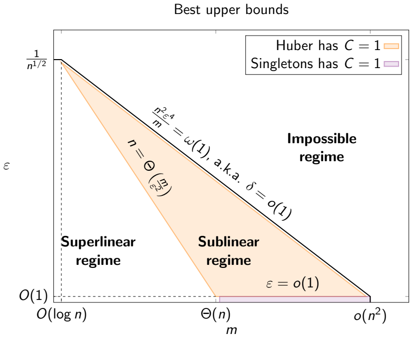
- We have focused on the false negative case in this talk. The false positive case is similar, but makes use of tools from [DGPP18] to restrict the class of alternative distributions

Formal Theorem

Huber Theorem

The Huber statistic for $n/m \ll 1/\varepsilon^2$, $\varepsilon, \delta \ll 1$, and $m \geq C \log n$ for sufficiently large constant C achieves sample complexity

$$n = (1 + o(1)) \frac{1}{\varepsilon^2} \sqrt{m \log \frac{1}{\delta}}$$



Experimental Results

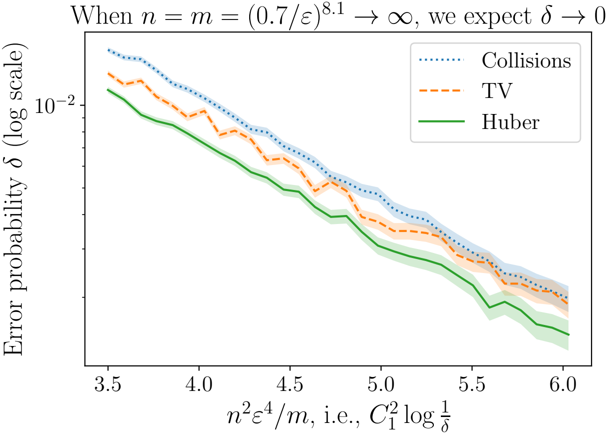


Figure: The Huber tester performs better than existing testers in practice

Summary

- Analyzing constant factors gives better understanding of actual performance

Summary

- Analyzing constant factors gives better understanding of actual performance
- Collisions tester has optimal variance of any statistic

Summary

- Analyzing constant factors gives better understanding of actual performance
- Collisions tester has optimal variance of any statistic
- Huber tester combines optimal variance with good tails

References I

Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D Smith, and Patrick White.
Testing that distributions are close.
In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 259–269. IEEE, 2000.

Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price.
Sample-optimal identity testing with high probability.
In *45th International Colloquium on Automata, Languages, and Programming (ICALP 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price.
Collision-based testers are optimal for uniformity and closeness.
Chicago Journal of Theoretical Computer Science, 2019:1, 2019.

References II

Oded Goldreich and Dana Ron.
On testing expansion in bounded-degree graphs.
In *Studies in Complexity and Cryptography. Miscellanea on the Interplay between Randomness and Computation*, pages 68–75. Springer, 2011.

Liam Paninski.
A coincidence-based test for uniformity given very sparsely sampled discrete data.
IEEE Transactions on Information Theory, 54(10):4750–4755, 2008.