# Testing Monotone Continuous Distributions on High-dimensional Real Cubes

Artur Czumaj

Department of Computer Science &

**DIMAP** (Centre for Discrete Maths and it Applications)

University of Warwick

Joint work with Michal Adamaszek & Christian Sohler

# Testing properties of distributions

- General question:
  – Test if a given probability distribution has a given property

> Distribution is available by accessing only samples drawn from the distribution

Examples:

- is given distribution uniform?
- are two distributions identical?
- are two distributions independent?

# Testing properties of distributions

Lots of research in statistics

Some recent research in algorithms

- Typical result:
  - Given a probability distribution on n points, we can test if it's uniform after seeing $\sim\sqrt{n}$ random samples

[Batu et al '01]

Testing = distinguish between uniform distribution and distributions which are $\epsilon$-far from uniform

$\epsilon$-far from uniform:

error probab. $\leq$ 1/3

$$\sum_{x\in} |\Pr[x] - \frac{1}{n}| \geq \epsilon$$

# Testing properties of distributions

- Typical result:
  - Given a probability distribution on n points, we can test if it's uniform after seeing $\sim\sqrt{n}$ random samples

[Batu et al '01]

- Similar bounds for testing
  - if a distribution is monotone
  - if two distributions are independent
  - …

# Testing properties of distributions

- Typical result:
  - Given a probability distribution on n points, we can test if it's uniform after seeing $\sim\sqrt{n}$ random samples

[Batu et al '01]

Many properties of distributions can be tested in time sublinear in the domain/support size (typically, with $n^{O(1)}$ samples)

# Testing properties of distributions

- Typical result:
  - Given a probability distribution on n points, we can test if it's uniform after seeing $\sim\sqrt{n}$ random samples

[Batu et al '01]

- What if distribution has infinite support?
- Continuous probability distributions?

# Testing properties of continuous distributions

- Typical result:
  - Given a probability distribution on n points, we can test if it's uniform after seeing $\Theta(\sqrt{n})$ random samples
  - $\Theta(\sqrt{n})$ random samples are necessary

- Given a continuous probability distribution on [0,1], can we test if it's uniform?

- Impossible
  - Follows from lower bound for discrete case with n$\rightarrow\infty$

# Testing properties of continuous distributions

- What can be tested?

- First question:

  test if the distribution is indeed continuous

# Testing properties of continuous distributions

- Dual question:

  Test if a probability distribution is discrete

- Prob. distribution D on $\Omega$ is discrete on N points if there is a set $X \subseteq \Omega$, $|X| \leq N$, st. $Pr_D[X]=1$

- D is $\epsilon$-far from discrete on N points if

  $\forall X \subseteq \Omega, |X| \leq N$

  $Pr_D[X] \leq 1-\epsilon$

# Testing if distribution is discrete on N points

- We repeatedly draw random points from D
- All what can we see:
  - Count frequency of each point
  - Count number of points drawn

For some D (eg, uniform or close):

- we need - $(\sqrt{N})$ to see first multiple occurrence

Gives a hope that can be solved in sublinear-time

Shows that we cannot be better than - $(\sqrt{N})$

# Testing if distribution is discrete on N points

Raskhodnikova et al '07 (Valiant'08):

## Distinct Elements Problem:

- D discrete with each element with prob. $\geq 1/N$
- Estimate the support size

**$\Omega(N^{1-o(1)})$** queries are needed to distinguish instances with $\leq N/100$ and $\geq N/11$ support size

Key property:

- two distributions that have identical first $\log^{\Theta(1)} N$ moments
- their expected frequencies up to $\log^{\Theta(1)} N$ are identical

# Testing if distribution is discrete on N points

Raskhodnikova et al '07 (Valiant'08):

**Distinct Elements Problem:**

- D discrete with each element with prob. $\geq 1/N$
- Estimate the support size

$\Omega(N^{1-o(1)})$ queries are needed to distinguish instances with $\leq N/100$ and $\geq N/11$ support size

Corollary: Testing if a distribution is discrete on N points requires $\Omega(N^{1-o(1)})$ samples

# Testing if distribution is discrete on N points

- We repeatedly draw random points from D
- All what can we see:
  - Count frequency of each point
  - Count number of points drawn

- Can we get O(N) time?

# Testing if distribution is discrete on N points

- Testing if a distribution is discrete on N points:

•Draw a sample $S = (s_1, \ldots, s_t)$ with $t = 2N/\epsilon$
•If S has more than N distinct elements
then REJECT
else ACCEPT

- If D is discrete on N points then we will accept D
- We only have to prove that
  - if D is $\epsilon$-far from discrete on N points, then we will reject with probability >2/3

# Testing if distribution is discrete on N points



- Testing if a distribution is discrete on N points:

•Draw a sample S = ($s_1$, …, $s_t$) with t = 2N/$\epsilon$
•If S has more than N distinct elements
then REJECT
else ACCEPT

D is $\epsilon$-far from discrete on N points, then reject with prob >2/3

$D$ is $\epsilon$-far from discrete on $N$ points $\Rightarrow$

$D$ is $\epsilon$-far from discrete on $N$ points iff
$\forall X \subseteq \Omega$, if $|X| \le N$ then $Pr_D[\Omega \setminus X] \geq \epsilon$

- Assuming that we haven't chosen $n$ points yet, we choose a new point with probability at least $\epsilon$

After $(1 + o(1))N/\epsilon$ samples, we choose $N + 1$ points with prob. $\geq 0.99$

# Testing if distribution is discrete on N points

- Testing if a distribution is discrete on N points:

•Draw a sample $S = (s_1, \ldots, s_t)$ with $t = 2N/\epsilon$
•If S has more than N distinct elements
then REJECT
else ACCEPT

Can we do better (if we only count distinct elements)?

D: has 1 point with prob. $1-4\epsilon$ and 2N points with prob. $2\epsilon/N$

D is $\epsilon$-far from discrete on N points

We need $\Omega(N/\epsilon)$ samples to see at least N points

# Open problem

What is the complexity of testing if **distribution is discrete on N points**?

Upper bound: **$O(N/\epsilon)$**
Lower bound: **$\Omega(N^{1-o(1)})$**

**Open problem: close the gap**

# Testing continuous probability distributions

- What can we test efficiently?
  - Complexity for discrete distributions should be “independent” on the support size
- Uniform distribution … under some conditions
- Rubinfeld & Servedio’05:
  - testing monotone distributions for uniformity

# Testing uniform distributions (discrete)

Rubinfeld & Servedio'05:

- Testing monotone distributions for uniformity

D: distribution on n-dimensional cube; $D:\{0,1\}^n \to \mathbf{R}$

$x,y \in \{0,1\}^n$, $x \preceq y$ iff $\forall i$: $x_i \le y_i$

D is monotone if $x \preceq y \Rightarrow \Pr[x] \le \Pr[y]$

Goal: test if a monotone distribution is uniform

Rubinfeld & Servedio'05:
Testing if a monotone distribution on n-dimensional binary cube is uniform:

- Can be done with $O(n \log(1/\epsilon)/\epsilon^2)$ samples
- Requires $\Omega(n/\log^2 n)$ samples

THE UNIVERSITY OF WARWICK

DIMAP

# Testing continuous distributions

Rubinfeld & Servedio'05:

- Testing monotone distributions for uniformity

D: distribution on n-dimensional cube; D:$\{0,1\}^n \to \mathbf{R}$

x,y $\in \{0,1\}^n$, x $\preceq$ y iff $\forall$i: $x_i \leq y_i$

D is monotone if x $\preceq$ y ➔ Pr[x] $\leq$ Pr[y]

Goal: test if a monotone distribution is uniform

D: distribution on n-dimensional cube;

density function $f:[0,1]^n \to \mathbf{R}$

x,y $\in [0,1]^n$, x $\preceq$ y iff $\forall$i: $x_i \leq y_i$

D is monotone if x $\preceq$ y ➔ f(x) $\leq$ f(y)

THE UNIVERSITY OF WARWICK

DIMAP

# Testing continuous distributions

Lower bounds holds for n-dimensional real cubes
Upper bound: ???

Rubinfeld & Servedio'05:
Testing if a monotone distribution on n-dimensional binary cube is uniform:

- Can be done with $O(n \log(1/\epsilon)/\epsilon^2)$ samples
- Requires $\Omega(n/\log^2 n)$ samples

# Testing monotone distributions for uniformity

D is $\epsilon$-far from uniform if $\frac{1}{2}\int_{x\in} |f(x) - 1| dx \geq \epsilon$

$L_1$ distance between f and uniform distribution

To test uniformity, we need to characterize monotone distributions that are $\epsilon$-far from uniform

On the high level:

- we follow approach of Rubinfeld & Servedio'05;
- details are different

# Testing monotone distributions for uniformity

D is $\epsilon$-far from uniform if $\frac{1}{2}\int_{x\in} |f(x)-1|dx \geq \epsilon$

**Key Technical Lemma:**

Let g:$[0,1]^n \rightarrow \mathbf{R}$ be a monotone function with $\int_x$ g(x) dx = 0 then

$$\int_x \|x\|_1 \cdot g(x)dx \geq \frac{1}{4}\int_x |g(x)|dx$$

Key Lemma follows from Key Technical Lemma with g(x) = f(x)-1

**Key Lemma:**

If D is a monotone distribution on $[0,1]^n$ with density function f and which is $\epsilon$-far from uniform then

$$E_f[\|x\|_1] = \int_x \|x\|_1 \cdot f(x)dx \geq \frac{n}{2} + \frac{\epsilon}{2}$$

# Testing monotone distributions for uniformity

**Key Lemma:**

If D is a monotone distribution on $[0,1]^n$ with density function f and which is $\epsilon$-far from uniform then

$$E_f[\|x\|_1] = \int_x \|x\|_1 \cdot f(x)dx \geq \frac{n}{2} + \frac{\epsilon}{2}$$

**Uniform distribution:**

If D is uniform on $[0,1]^n$ with density function f then

$$E_f[\|x\|_1] = \int_x \|x\|_1 \cdot f(x)dx = \frac{n}{2}$$

# Testing monotone distributions for uniformity

**Key Lemma:**
If D is a monotone distribution on $[0,1]^n$ with density function f and which is $\epsilon$-far from uniform then

$$E_f[\|x\|_1] = \int_x \|x\|_1 \cdot f(x)dx \geq \frac{n}{2} + \frac{\epsilon}{2}$$

s = $cn/\epsilon^2$
Repeat 20 times
  Draw a sample $S=(x_1,\ldots,x_s)$ from $[0,1]^n$
  If $\sum_i ||x_i||_1 \geq s\ (n/2+\epsilon/4)$ then REJECT and exit
ACCEPT

# Testing monotone distributions for uniformity

**Theorem:**
The algorithm below tests if D is uniform.
Its complexity is $O(n/\epsilon^2)$.

Slightly better bound than the one by RS'05

$s = cn/\epsilon^2$
Repeat 20 times
Draw a sample $S=(x_1,\ldots,x_s)$ from $[0,1]^n$
If $\sum_i ||x_i||_1 \geq s\,(n/2+\epsilon/4)$ then REJECT and exit
ACCEPT

# Testing monotone distributions for uniformity

$s = cn/\epsilon^2$

Repeat 20 times

Draw a sample $S=(x_1,\ldots,x_s)$ from $[0,1]^n$

If $\sum_i ||x_i||_1 \geq s\,(n/2+\epsilon/4)$ then REJECT and exit

ACCEPT

Lemma 1: If D is uniform then

$$\Pr[\sum_i ||x_i||_1 \geq s(n/2+\epsilon/4)] \leq 0.01$$

Easy application of Chernoff bound

Lemma 2: If D is $\epsilon$-far from uniform then

$$\Pr[\sum_i ||x_i||_1 < s(n/2+\epsilon/4)] \leq 12/13$$

By Key Lemma + Feige lemma

THE UNIVERSITY OF WARWICK

DIMAP

# Testing monotone distributions for uniformity

**Key Technical Lemma:**

Let $g:[0,1]^n \to \mathbf{R}$ be a monotone function with $\int_x g(x)\,dx = 0$ then

$$\int_x \|x\|_1 \cdot g(x) dx \geq \frac{1}{4} \int_x |g(x)| dx$$

Why such a bound:

Tight for $g(x) = sgn(x_1 - ½)$

$$\int_{x:x_1>\frac{1}{2}} \|x\|_1 \cdot g(x)\ dx \quad = \quad \frac{1}{2}\int_{x:x_1>\frac{1}{2}} (x_1 + \ldots + x_n)\ dx = \frac{1}{2}\left(\frac{3}{4} + \frac{1}{2} + \ldots + \frac{1}{2}\right)\ dx = \frac{n}{4} + \frac{1}{8}\ .$$

Similarly,

$$\int_{x:x_1<\frac{1}{2}} \|x\|_1 \cdot g(x)\ dx \quad = \quad \frac{1}{2}\left(\frac{1}{4} + \frac{1}{2} + \ldots + \frac{1}{2}\right) = \frac{n}{4} - \frac{1}{8}\ ,$$

and hence,

$$\int_x \|x\|_1 \cdot g(x)\ dx \quad = \quad \int_{x:x_1>\frac{1}{2}} \|x\|_1 \cdot g(x)\ dx - \int_{x:x_1<\frac{1}{2}} \|x\|_1 \cdot g(x)\ dx = \frac{1}{4} = \frac{1}{4} \cdot \int_x |g(x)|\ dx\ .$$

# Testing monotone distributions for uniformity

**Key Technical Lemma:**

Let $g:[0,1]^n \to \mathbf{R}$ be a monotone function with $\int_x g(x)\,dx = 0$ then

$$\int_x \|x\|_1 \cdot g(x)dx \geq \frac{1}{4}\int_x |g(x)|dx$$

# Testing monotone distributions for uniformity

Let $P = \{\mathbf{x} : g(\mathbf{x}) \geq 0\}$ and $N = \{\mathbf{x} : g(\mathbf{x}) < 0\}$. Consider:

$$\int_{\mathbf{x}\in N} \int_{\mathbf{y}\in P} |g(\mathbf{x}) - g(\mathbf{y})| \; dy \; dx \; .$$

For $g(\mathbf{x}) < 0 \cdot \; g(\mathbf{y})$, we have $|g(\mathbf{x}) - g(\mathbf{y})| = |g(\mathbf{x})| + |g(\mathbf{y})|$.
Moreover $\int_{\mathbf{x}\in N} |g(\mathbf{x})|dx = \int_{\mathbf{y}\in P} |g(\mathbf{y})|dy = \frac{1}{2}\int_{\mathbf{x}} |g(\mathbf{x})|dx$.
Hence:

$$= \int_{\mathbf{x}\in N} \int_{\mathbf{y}\in P} (|g(\mathbf{x})| + |g(\mathbf{y})|) \;=\; \int_{\mathbf{y}\in P} \int_{\mathbf{x}\in N} |g(\mathbf{x})| + \int_{\mathbf{x}\in N} \int_{\mathbf{y}\in P} |g(\mathbf{y})|$$

$$= \frac{1}{2}\int_{\mathbf{y}\in P} \int_{\mathbf{x}} |g(\mathbf{x})| + \frac{1}{2}\int_{\mathbf{x}\in N} \int_{\mathbf{y}} |g(\mathbf{y})| \;=\; \frac{1}{2}\int_{\mathbf{y}} \int_{\mathbf{x}} |g(\mathbf{x})| \;=\; \frac{1}{2}\int_{\mathbf{x}} |g(\mathbf{x})| \; .$$

Since every pair $(\mathbf{x}, \mathbf{y})$ can satisfy at most one of the conditions $(\mathbf{x}, \mathbf{y}) \in P \times N$ and $(\mathbf{x}, \mathbf{y}) \in N \times P$, we obtain:

$$\int_{\mathbf{x}\in N} \int_{\mathbf{y}\in P} |g(\mathbf{x}) - g(\mathbf{y})| \; dy \; dx \cdot \; \frac{1}{2}\int\int_{\mathbf{x},\mathbf{y}} |g(\mathbf{x}) - g(\mathbf{y})| \; dy \; dx \; .$$

Hence:

$$\frac{1}{2}\int_{\mathbf{x}} |g(\mathbf{x})| \; dx = \int_{\mathbf{x}\in N} \int_{\mathbf{y}\in P} |g(\mathbf{x}) - g(\mathbf{y})| \; dx \; dy \cdot \; \frac{1}{2}\int\int_{\mathbf{x},\mathbf{y}} |g(\mathbf{x}) - g(\mathbf{y})| \; dx \; dy \; .$$

# Testing monotone distributions for uniformity

By considering all the possible relative placements of $\mathbf{x}$ and $\mathbf{y}$ within $[0,1]^n$ and splitting the domain accordingly, one can prove that

$$\int\int_{\mathbf{x},\mathbf{y}} |g(\mathbf{x}) - g(\mathbf{y})| \; dy \; dx \cdot \int\int_{\mathbf{x}\prec\mathbf{y}} \left( \sum_{(\mathbf{u},\mathbf{v})\in D(\mathbf{x},\mathbf{y})} |g(\mathbf{u}) - g(\mathbf{v})| \right) \; dy \; dx \;\; ,$$

where $D(\{0,1\}^n)$ is the set of all **main diagonals** of **discrete** cube $\{0,1\}^n$:

$$D(\{0,1\}^n) = \{(\mathbf{x},\mathbf{y}) \in \{0,1\}^n \times \{0,1\}^n : x_i = 1 - y_i \text{ for every } i\}$$

# Testing monotone distributions for uniformity

## Key Technical Lemma:

Let g:[0,1]$^n$→**R** be a monotone function with $\int_x$ g(x) dx = 0 then

$$\int_x \|x\|_1 \cdot g(x)dx \geq \frac{1}{4}\int_x |g(x)|dx$$

Key inequalities in the proof:

$$\frac{1}{4}\int_{\mathbf{x}} |g(\mathbf{x})|\ dx \quad \cdot \quad \frac{1}{4}\int\int_{\mathbf{x},\mathbf{y}} |g(\mathbf{x}) - g(\mathbf{y})|dxdy$$

$$\cdot \quad \frac{1}{4}\int\int_{\mathbf{x}\prec\mathbf{y}} \left( \sum_{(\mathbf{u},\mathbf{v})\in D(\mathbf{x},\mathbf{y})} |g(\mathbf{u}) - g(\mathbf{v})| \right)\ dxdy$$

$$\cdot \quad \frac{1}{2}\sum_{i=1}^{n}\int\int_{\mathbf{x}\prec\mathbf{y}} \left( \sum_{(\mathbf{u},\mathbf{v})\in E_i(\mathbf{x},\mathbf{y})} |g(\mathbf{u}) - g(\mathbf{v})| \right)\ dxdy$$

$$\cdot \quad \frac{1}{2}\sum_{i=1}^{n}\int_{\mathbf{x}} (2x_i - 1)g(\mathbf{x})dx$$

$$\cdot \quad \int_{\mathbf{x}} \|\mathbf{x}\|_1 g(\mathbf{x})dx$$

# Testing monotone continuous distributions

Rubinfeld & Servedio'05:
Testing if a monotone distribution on n-dimensional binary cube is uniform:

- Can be done with $O(n \log(1/\epsilon)/\epsilon^2)$ samples
- Requires $\Omega(n/\log^2 n)$ samples

Here:
Testing if a monotone distribution on n-dimensional continuous cube is uniform :

- Can be done with $O(n/\epsilon^2)$ samples
- (Requires $\Omega(n/\log^2 n)$ samples)

THE UNIVERSITY OF WARWICK

DIMAP

# Testing monotone continuous distributions

## Further extension/application:

Testing if a monotone distribution on n-dimensional discrete cube $\{0,1,2,...,k\}^n$ is uniform:

- Can be done with $O(n/\epsilon^2)$ samples

Here:

Testing if a monotone distribution on n-dimensional continuous cube is uniform :

- Can be done with $O(n/\epsilon^2)$ samples
- (Requires $\Omega(n/\log^2 n)$ samples)

# Conclusions

- Testing distributions on infinite/uncountable support is different from testing discrete distributions
  - Continuous distributions are harder
- Challenge: understand when it's possible to test
  - Usually some additional conditions are to be imposed
- Tight(er) bounds?

THE UNIVERSITY OF WARWICK

DIMAP

# Conclusions

- Continuous distributions are harder
- Is the $L_1$-norm the right one?
  - It doesn't work well for continuous distributions
- Earth mover norm?
  - How much mass has to be moved and how far to obtain a given distribution
  - Ba, Nguyen, Nguyen, Rubinfeld 2009:
  - Testing uniformity on [0,1] can be done in time $f(1/\epsilon)$
  - Framework (holds for a variety of properties):
    reduction to the problem on the support of size $f(1/\epsilon)$