Testing Continuous Distributions

Artur Czumaj

DIMAP (Centre for Discrete Maths and its Applications) & Department of Computer Science
University of Warwick

Joint work with A. Adamaszek & C. Sohler
Testing probability distributions

• General question:
  - Test a given property of a given probability distribution
    • distribution is available by accessing only samples drawn from the distribution

Examples:
- is given probability uniform?
- are two prob. distributions independent?
Testing probability distributions

For more details/introduction:
see R. Rubinfeld’s talk on Wednesday

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

\[ \text{Testing} = \text{distinguish between uniform distribution and distributions which are } \epsilon \text{-far from uniform} \]

\[ \epsilon \text{-far from uniform: } \sum_{x \in \Omega} |\Pr[x] - \frac{1}{n}| \geq \epsilon \]
Testing probability distributions

For more details/introduction:
see R. Rubinfeld’s talk on Wednesday

• 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 continuous probability distributions

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

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

- Impossible
  - Follows from the lower bound for discrete case with \( n \to \infty \)
Testing continuous probability distributions

- More direct proof:
  - Suppose tester A distinguishes in at most $t$ steps between uniform distribution and $\epsilon$-far from uniform

- $D_1$ - uniform distribution

- $D_2$ is $\frac{1}{2}$-far from uniform and is defined as follows:
  - Partition $[0,1]$ into $t^3$ interval of identical length
  - Split each interval into two halves
  - Randomly choose one half:
    - the chosen half gets uniform distribution
    - the other half has zero probability
  - In $t$ steps, no interval will be chosen more than once in $D_2$

A cannot distinguish between $D_1$ and $D_2$
• What can be tested?

• First question:
  test if the distribution is indeed continuous
Testing continuous probability distributions

- 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
  \[ \forall X \subseteq \Omega, |X| \leq N \]
  \[ Pr_D[X] < 1 - \epsilon \]
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 $\Omega(\sqrt{N})$ to see first multiple occurrence

Gives a hope that can be solved in sublinear-time
Testing if distribution is discrete on N points

Raskhodnikova et al '07 (Valiant'08):

Distinct Elements Problem:
• D discrete with each element with prob. $\geq \frac{1}{N}$
• 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 step: two distributions that have identical first $\log^{\Theta(1)}N$ moments
• their expected frequencies up to $\log^{\Theta(1)}N$ are identical
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 $S = (s_1, ..., s_t)$ with $t = cN/\epsilon$
  - 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 \( t = \frac{cN}{\epsilon} \)
  - If \( S \) has more than \( N \) distinct elements, then \text{REJECT}
  - Else \text{ACCEPT}

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

\( D \): has 1 point with prob. \( 1-4\epsilon \)
- 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
Testing if distribution is discrete on \( N \) points

Assume \( D \) is \( \epsilon \)-far from discrete on \( N \) points

Order points in \( \Omega \) so that \( \Pr[X_i] = p_i \) and \( p_i \geq p_{i+1} \)

\( A = \{X_1, \ldots, X_N\} \), \( B = \) other points from the support

\[
p_1 + p_2 + \ldots + p_N < 1 - \epsilon
\]

\( \alpha = \# \) points from \( A \) drawn by the algorithm

\( \beta = \# \) points from \( B \) drawn by the algorithm

We consider 3 cases (all bounds are with prob. > 0.99):

1) \( p_N < \frac{\epsilon}{2N} \Rightarrow \beta > N \)
   - all points in \( B \) have small prob. \( \Rightarrow \) not too many repetitions

2) \( p_N \geq c \frac{N}{\epsilon} \Rightarrow \beta \geq \epsilon/2p_N \)
   - points in \( B \) have small prob. \( \Rightarrow \) bound for \#distinct points

3) \( p_N \geq \frac{\epsilon}{2N} \Rightarrow \alpha \geq N - \epsilon/2p_N \)
   - either many distinct points from \( A \) or \( p_N \) is very small (then \( \beta \) will be large)
Testing if distribution is discrete on N points

Assume D is $\epsilon$-far from discrete on N points
Order points in $\Omega$ so that $\Pr[X_i] = p_i$ and $p_i \geq p_{i+1}$
$A = \{X_1, \ldots, X_N\}$, $B = \text{other points from the support}$
$\alpha = \# \text{ points from A drawn by the algorithm}$
$\beta = \# \text{ points from B drawn by the algorithm}$

Main ideas:
Case 2) $p_N \geq cN/\epsilon \Rightarrow \beta \geq \epsilon/2p_N$
  
  - Worst case: all points in B have uniform and maximum distrib. = $p_N$
  - $Z_i = \text{random variable: number of steps to get ith new point from B}$
  - We have to prove that with prob. $> 0.99$: $\sum_{i=1}^{\epsilon/2p_N} Z_i < t$
  - $Z_1, Z_2, \ldots$ - geometric distribution: $E[Z_i] = \frac{1}{(r-i)p_N}$, $r = \text{number of points in B}$
  
  $\sum_{i=1}^{\epsilon/2p_N} E[Z_i] \leq \frac{2}{p_N}$

  $\rightarrow$ Markov gives with prob. $\geq 0.99$: $\sum_{i=1}^{\epsilon/2p_N} Z_i < t$
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

By sampling $O(N/\epsilon)$ points one can distinguish between
• distributions discrete on N points and
• those $\epsilon$-far from discrete on N points

The algorithm may fail with prob. < 1/3
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; $D: \{0,1\}^n \rightarrow \mathbb{R}$

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

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

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
Rubinfeld & Servedio'05:
• Testing monotone distributions for uniformity

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

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

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

Goal: test if a monotone distribution is uniform

D: distribution on n-dimensional cube; density function $f : [0,1]^n \rightarrow \mathbb{R}$

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

D is monotone if $x \preceq y \Rightarrow f(x) \leq f(y)$
Testing continuous probability distributions

Lower bounds holds for continuous cubes
Upper bound: ???
• is it a function of the dimension or the support?

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

$$\frac{1}{2} \int_{x \in \Omega} |f(x) - 1| dx \geq \epsilon$$

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

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

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

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

$$E_f[\|x\|_1] = \int_x \|x\|_1 f(x) \, dx \geq \frac{n}{2} + \frac{\varepsilon}{2}$$
Key Lemma:
If $D$ is a monotone distribution on $[0,1]^n$ with density function $f$ and which is $\epsilon$-far from uniform then

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

$s = cn/\epsilon^2$
Repeat 20 times
Draw a sample $S=(x_1,\ldots,x_s)$ from $[0,1]^n$
If $\sum_i \|x_i\|_1 \geq s \left( \frac{n}{2} + \frac{\epsilon}{4} \right)$ then REJECT and exit
ACCEPT
Theorem:
The algorithm below tests if $D$ is uniform. It’s 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

\[ s = \frac{cn}{\epsilon^2} \]

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

\[ \text{ACCEPT} \]

Lemma 1: If \( D \) is uniform then
\[ \text{Pr}[\sum_i \|x_i\|_1 \geq s(n/2 + \epsilon/4)] \leq 0.01 \]

Easy application of Chernoff bound

Lemma 2: If \( D \) is \( \epsilon \)-far from uniform then
\[ \text{Pr}[\sum_i \|x_i\|_1 < s(n/2 + \epsilon/4)] \leq 12/13 \]

By Key Lemma + Feige lemma
Testing monotone distributions for uniformity

$s = \frac{c n}{\epsilon^2}$
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 2: If $D$ is $\epsilon$-far from uniform then $\Pr[\sum_i ||x_i||_1 < s(n/2 + \epsilon/4)] \leq 12/13$

Proof:
D is $\epsilon$-far from uniform $\Rightarrow E[\sum_i ||x_i||_1] \geq s(n+\epsilon)/2$
Feige’s lemma: $Y_1, \ldots, Y_s$ independent r.v., $Y_i \geq 0$, $E[Y_i \leq 1] \Rightarrow$
$\Pr[\sum_i Y_i < s + 1/12] \geq 1/13$
Choose $Y_i = 2 - 2||x_i||_1/(n+\epsilon)$
Then, Feige’s lemma yields the desired claim
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

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

$s = cn/\epsilon^2$

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

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

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

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

$$\int_x \|x\|_1 g(x) dx \geq \frac{1}{4} \int_x |g(x)| dx$$
Testing monotone distributions for uniformity

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

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

Why such a bound:
Tight for $g(x) = \text{sgn}(x_1 - \frac{1}{2})$

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

Similarly,

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

and hence,

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

Can be easily extended to $\{0,1,\ldots,k\}^n$ cubes
Conclusions

• Testing continuous distributions is different from testing discrete distributions

• Continuous distributions are harder

• More examples when it’s possible to test
  – Usually some additional conditions are to be imposed

• Tight(er) bounds?