Introduction to Property Testing

Sourav Chakraborty

Indian Statistical Institute
Kolkata, India

Outline

Outline

Various kinds of Algorithms

Various kinds of Algorithms

- Determinstic Algorithm: Solves the problem exactly.

Various kinds of Algorithms

- Determinstic Algorithm: Solves the problem exactly.
- Randomized Algorithm: Solves the problem correctly with high probabilty. Saves running time.

Various kinds of Algorithms

- Determinstic Algorithm: Solves the problem exactly.
- Randomized Algorithm: Solves the problem correctly with high probabilty. Saves running time.
- Approximation Algorithm: Gives an approximate sollution to the problem. Saves running time.

Various kinds of Algorithms

- Determinstic Algorithm: Solves the problem exactly.
- Randomized Algorithm: Solves the problem correctly with high probabilty. Saves running time.
- Approximation Algorithm: Gives an approximate sollution to the problem. Saves running time.
- Parametrized Algorithms: Solves the problem exactly and quickly if the input has certain parameter "small".

Various kinds of Algorithms

- Determinstic Algorithm: Solves the problem exactly.
- Randomized Algorithm: Solves the problem correctly with high probabilty. Saves running time.
- Approximation Algorithm: Gives an approximate sollution to the problem. Saves running time.
- Parametrized Algorithms: Solves the problem exactly and quickly if the input has certain parameter "small".

One Main Goal: Have running time polynomial.

What about sub-linear?

What about sub-linear?

Cannot even read the whole input!

What about sub-linear?

Cannot even read the whole input!

But sometimes it is very important for various reasons:

- Want the answer in very small time (possibly constant time).
- Accessing the input can be costly affair or even impossible.

What about sub-linear?

Cannot even read the whole input!

But sometimes it is very important for various reasons:

- Want the answer in very small time (possibly constant time).
- Accessing the input can be costly affair or even impossible.

Property Testing

In Property testing we are usually interested in sub-linear query complexity, that is, we want to read a small fraction of the input.

But how is it possible?

We have to give up on something. We will assume some promise on the input.

Example: Checking Equality

Example: Checking Equality

Equality of strings

Given two strings $x, y \in \{0,1\}^n$ check if $x = y$, that is, for every $i \in \{1, \ldots, n\}$ is $x_i = y_i$.

Example: Checking Equality

Equality of strings

Given two strings $x, y \in \{0,1\}^n$ check if $x = y$, that is, for every $i \in \{1, \ldots, n\}$ is $x_i = y_i$.

The goal is to answer it in CONSTANT time and hence can't even read the whole input.

Example: Checking Equality

Equality of strings

Given two strings $x, y \in \{0,1\}^n$ check if $x = y$, that is, for every $i \in \{1, \ldots, n\}$ is $x_i = y_i$.

The goal is to answer it in CONSTANT time and hence can't even read the whole input. $--$ Not Possible

Example: Checking Equality

Equality of strings

Given two strings $x, y \in \{0,1\}^n$ check if $x = y$, that is, for every $i \in \{1, \ldots, n\}$ is $x_i = y_i$.

The goal is to answer it in CONSTANT time and hence can't even read the whole input. $--$ Not Possible

But, say, there is a promise that either $x = y$ OR x and y differ at more than $1/4$ fraction of the indices. Then ...

Simple sampling algorithm for testing of equality

Algorithm

Randomly pick 4 indices $\{i_1, i_2, i_3, i_4\}$ uniformly and independently at random. If

$$x_{i_1} = y_{i_1}, x_{i_2} = y_{i_2}, x_{i_3} = y_{i_3}, x_{i_4} = y_{i_4},$$

then ACCEPT otherwise REJECT.

Simple sampling algorithm for testing of equality

Algorithm

Randomly pick 4 indices $\{i_1, i_2, i_3, i_4\}$ uniformly and independently at random. If

$$x_{i_1} = y_{i_1}, x_{i_2} = y_{i_2}, x_{i_3} = y_{i_3}, x_{i_4} = y_{i_4},$$

then ACCEPT otherwise REJECT.

- If $x = y$ then the algorithm always ACCEPTS.

Simple sampling algorithm for testing of equality

Algorithm

Randomly pick 4 indices $\{i_1, i_2, i_3, i_4\}$ uniformly and independently at random. If

$$x_{i_1} = y_{i_1}, x_{i_2} = y_{i_2}, x_{i_3} = y_{i_3}, x_{i_4} = y_{i_4},$$

then ACCEPT otherwise REJECT.

- If $x = y$ then the algorithm always ACCEPTS.
- If x and y differ at $1/4$ fraction of the indices then the algorithm ACCEPTS with probability at most $1/3$.

Example: Exit poll

Example: Exit poll

Election

Given a set of n voter (voting for Party A or Party B) check if Party A has more votes than Party B.

Example: Exit poll

Election

Given a set of n voter (voting for Party A or Party B) check if Party A has more votes than Party B.

If the goal is to sample a small part of the voters then its not possible always to give the right answer (even with high probability).

Example: Exit poll

Election

Given a set of n voter (voting for Party A or Party B) check if Party A has more votes than Party B.

If the goal is to sample a small part of the voters then its not possible always to give the right answer (even with high probability).

But if we want to distinguish between whether Party A wins by a big margin or Party B wins by a big margin: then statistical sample works.

Example: Checking Bipartiteness.

2-colorability

Given an undirected graph G can we color the vertices of G with 2 colors such that no adjacent vertices are of the same color? Or in other words is it bipartite.

Example: Checking Bipartiteness.

2-colorability

Given an undirected graph G can we color the vertices of G with 2 colors such that no adjacent vertices are of the same color? Or in other words is it bipartite.

In general it may require us to look at the whole graph to answer but can we look at a very small fraction of the graph and distinguish

- The graph is bipartite
- A "lot" of edges have to be removed to make it bipartite.

Formal Definitions: Property and distance

- Let $x \in \{0,1\}^n$ be an input.

Formal Definitions: Property and distance

- Let $x \in \{0,1\}^n$ be an input.
- A property $\mathcal{P}$ is a subset of $\{0,1\}^n$.

Formal Definitions: Property and distance

- Let $x \in \{0,1\}^n$ be an input.
- A property $\mathcal{P}$ is a subset of $\{0,1\}^n$.
- For two strings $x, y \in \{0,1\}^n$, $dist(x,y)$ is the fraction of indices where they differ.

$$dist(x,y) = |\{i | x_i \neq y_i\}|/n.$$

Formal Definitions: Property and distance

- Let $x \in \{0,1\}^n$ be an input.
- A property $\mathcal{P}$ is a subset of $\{0,1\}^n$.
- For two strings $x, y \in \{0,1\}^n$, $dist(x,y)$ is the fraction of indices where they differ.

$$dist(x,y) = |\{i | x_i \neq y_i\}|/n.$$

- For a input x and a property $\mathcal{P}$, $dist(x,\mathcal{P}) = \min_{y \in \mathcal{P}} dist(x,y)$.

Formal Definitions: Property and distance

- Let $x \in \{0,1\}^n$ be an input.
- A property $\mathcal{P}$ is a subset of $\{0,1\}^n$.
- For two strings $x, y \in \{0,1\}^n$, $dist(x,y)$ is the fraction of indices where they differ.

$$dist(x,y) = |\{i | x_i \neq y_i\}|/n.$$

- For a input x and a property $\mathcal{P}$, $dist(x,\mathcal{P}) = \min_{y \in \mathcal{P}} dist(x,y)$.
- x is ϵ-far from being a property if $dist(x,\mathcal{P}) > \epsilon$.

Formal Definitions: Property and distance

- Let $x \in \{0,1\}^n$ be an input.
- A property $\mathcal{P}$ is a subset of $\{0,1\}^n$.
- For two strings $x, y \in \{0,1\}^n$, $dist(x,y)$ is the fraction of indices where they differ.

$$dist(x,y) = |\{i | x_i \neq y_i\}|/n.$$

- For a input x and a property $\mathcal{P}$, $dist(x,\mathcal{P}) = \min_{y \in \mathcal{P}} dist(x,y)$.
- x is ϵ-far from being a property if $dist(x,\mathcal{P}) > \epsilon$.

Promise Problem

For a property $\mathcal{P}$ and a distance parameter ϵ, given an input x distinguish between the two cases:
(a) Is $x \in \mathcal{P}$, OR (b) Is x ϵ-far from $\mathcal{P}$.

Informal Definition

Under some ASSUMPTION on the input can we make some intelligent deductions in SUB-LINEAR query/sample (or even time and space) complexity.

Use of Property Testing techniques

Many areas of research has been using techniques from property testing, including:

- Machine Learning
- Program Checking
- Communication Complexity
- Coding theory and cryptography

Use of Property Testing techniques

Many areas of research has been using techniques from property testing, including:

- Machine Learning
- Program Checking
- Communication Complexity
- Coding theory and cryptography

I usually classify problems in problems in property testing (and related areas) into 4 categoried: Function properties, Graph properties, Geometric properties and Distribution properties

Examples of Function properties

- **Linearity Testing:** Given a truth-table of a function f test is the function f linear OR the function has to be changed at at-least ϵ fraction of the domain to make it linear.
- **Branching Program Testing:** Given a truth-table of a function f test is the function is accepted by a constant depth read-once branching program OR is far from being accepted by a constant depth read-once branching program.
- **Isomorphism Testing:** Given two functions $\mathcal{F}_1$ and $\mathcal{F}_2$ test are the two functions isomorphic OR far-from being isomorphic.
- **Monotonicity Testing** Given a function from $\mathbb{R}^n$ to $\mathbb{R}$ is it monotonic.
- **Learning of a Function** Given access to the truth-table of a functions can one learn it quickly.

Examples of Graph properties

- **k-colorability** Given a graph is it k-colorable OR far-from being connected.
- **Connectivity** Given a graph test is it connected OR far-from being connected.
- **Isomorphism Testing:** Given two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ test are the two graphs isomorphic OR far-from being isomorphic.
- **Structure of Big Graphs:** Understanding the structures of massive graphs (like the internet graph).

Examples of Geometric Properties

- **Clustering:** Given a set of point are they clusterable into k clusters.
- **Classification:** Can one learn the classifier easily.
- **Dimension Reduction:** Can we reduce the dimension of the data in hand.

Examples of Properties of Distribution

- **Uniformity Testing:** Is a given distribution uniform OR is the ℓ_1 distance from uniform more than ϵ?
- **Equivalence Testing:** Is a given distribution identical to a known distribtuion OR are their ℓ_1 distance from uniform more than ϵ?
- **Independence Testing:** Given a joint distribution uniform are the two individual distributions independent?
- **Learning:** Given a distribution can we learn the distribution.

Our Goal ...

- We want to design a randomized algorithm that answers the promise problem correctly with high probability.

Our Goal ...

- We want to design a randomized algorithm that answers the promise problem correctly with high probability.
- We want to look at a very small portion of the input.

Our Goal ...

- We want to design a randomized algorithm that answers the promise problem correctly with high probability.
- We want to look at a very small portion of the input.

In the rest of the talk we would not consider the running time of an algorithm but rather the number of bits of the input that is read. Accessing each bit of the input is called a QUERY.

Property tester

Definition

Let $\mathcal{P}$ be a property. A tester for $\mathcal{P}$ is a *randomized* algorithm $\mathcal{A}$ with black box access to an input x and satisfies:

- If $x \in \mathcal{P} \Rightarrow \Pr[\mathcal{A} \text{ accepts}] \geq 2/3$.
- If x is ϵ-far from $\mathcal{P} \Rightarrow \Pr[\mathcal{A} \text{ rejects}] \geq 2/3$.

We allow the algorithm to be *adaptive* (queries may depend on the outcome of previous queries).

Query Complexity

Query complexity for the tester $\mathcal{A}$ is the maximum number of queries queried by the tester on any input.

Query Complexity

Query complexity for the tester $\mathcal{A}$ is the maximum number of queries queried by the tester on any input.

Query complexity of a property $\mathcal{P}$ is the query complexity of the tester that has the minimum query complexity.

Query Complexity

Query complexity for the tester $\mathcal{A}$ is the maximum number of queries queried by the tester on any input.

Query complexity of a property $\mathcal{P}$ is the query complexity of the tester that has the minimum query complexity.

Trivial example: let $\mathcal{P}$ be the property "$x \equiv 0$". Then taking $O(1/\epsilon)$ independent samples works w.h.p.

Other connected areas...

- Statistical estimation - In property testing we consider more combinatorial objects like properties of Boolean functions and graphs. .
- Evasiveness and Certificate Complexity.
- Probabilistically Checkable Proofs (PCP).
- Locally Decodable Codes.

Different Models

Different Models

There are different models depending on:

- Restricted error. [One-sided error or two-sided error]

Different Models

There are different models depending on:

- Restricted error. [One-sided error or two-sided error]
- How the input is represented? For example, is the graph given as adjacency matrix or adjacency list or some other way. [Dense graph model, sparse graph model, orientation model in graph testing]

Different Models

There are different models depending on:

- Restricted error. [One-sided error or two-sided error]
- How the input is represented? For example, is the graph given as adjacency matrix or adjacency list or some other way. [Dense graph model, sparse graph model, orientation model in graph testing]
- How the queries are made? [Classical, quantum]

Different Models

There are different models depending on:

- Restricted error. [One-sided error or two-sided error]
- How the input is represented? For example, is the graph given as adjacency matrix or adjacency list or some other way. [Dense graph model, sparse graph model, orientation model in graph testing]
- How the queries are made? [Classical, quantum]
- Do we also want to accept inputs that are "close" to the property? [Tolerant model and Intolerant Model]

What kind of questions to ask?

- Given a property $\mathcal{P}$ what is the query complexity for testing $\mathcal{P}$.
  - Design a property tester that tests $\mathcal{P}$ using $O(q)$ number of queries.
  - Prove that no property tester can test using less than $\Omega(q)$ number of queries.

What kind of questions to ask?

- Given a property $\mathcal{P}$ what is the query complexity for testing $\mathcal{P}$.
  - Design a property tester that tests $\mathcal{P}$ using $O(q)$ number of queries.
  - Prove that no property tester can test using less than $\Omega(q)$ number of queries.
- Classify the set of properties that can be tested using constant number of queries.

What kind of questions to ask?

- Given a property $\mathcal{P}$ what is the query complexity for testing $\mathcal{P}$.
  - Design a property tester that tests $\mathcal{P}$ using $O(q)$ number of queries.
  - Prove that no property tester can test using less than $\Omega(q)$ number of queries.
- Classify the set of properties that can be tested using constant number of queries.
- Come up with the right model for testing.

Outline

1-sided error testers

1-sided-error property tester

Let $\mathcal{P}$ be a property. A 1-sided-error property tester for $\mathcal{P}$ is a *randomized* algorithm $\mathcal{A}$ with black box access to an input x and satisfies:

- (Completeness) If $x \in \mathcal{P} \Rightarrow \Pr[\mathcal{A} \text{ accepts}] = 1$.
- (Soundness) If x is ϵ-far from $\mathcal{P} \Rightarrow \Pr[\mathcal{A} \text{ rejects}] \geq 2/3$.

We allow the algorithm to be *adaptive* (queries may depend on the outcome of previous queries).

1-sided-error tester has its hands tied

- The tester has to ACCEPT if the input satisfies the property.

1-sided-error tester has its hands tied

- The tester has to ACCEPT if the input satisfies the property.
- Hence, the only way the tester can reject is if it find a PROOF that the input does not satisfy the property.

1-sided-error tester has its hands tied

- The tester has to ACCEPT if the input satisfies the property.
- Hence, the only way the tester can reject is if it find a PROOF that the input does not satisfy the property.
- So if the input does not have the property then the tester must find a PROOF/WITNESS with high probability.

Typical 1-sided-error tester

1-sided-error algorithm

Query some bits of the input. The bits to be queried can be either uniformly chosen or chosen in a cleaver co-related fashion.

- If the answers of the queried bits contains a WITNESS that the input is not in the property then REJECT
- Else ACCEPT

Goal is to use some nice structure for the property for making the queries, like

- the Szemeredi's Regularity Lemma for graphs,
- properties of Fourier coefficients for algebraic functions, etc

Usually, the proof of SOUNDNESS is the hard part.

So what is the success probability of the tester?

Say the tester uses the random string r and queries the bits in Q_r (also say $|Q_r| = q$). Then the probability of success is

$$\Pr_r[Q_r \text{ contains a WITNESS}].$$

So what is the success probability of the tester?

Say the tester uses the random string r and queries the bits in Q_r (also say $|Q_r| = q$). Then the probability of success is

$$\Pr_r[Q_r \text{ contains a WITNESS}].$$

Thus a 1-sided-error property tester can successfully test a property $\mathcal{P}$ with q queries only if, an input x is "far" from $\mathcal{P}$ implies there is a lots of WITNESS of size q hidden in x.

Lower bounds for 1-sided-error testing

Lower bounds for 1-sided-error testing

$\mathcal{D}_N$ be a distribution on the the set of inputs that are far from $\mathcal{P}$.

Lower bounds for 1-sided-error testing

$\mathcal{D}_N$ be a distribution on the the set of inputs that are far from $\mathcal{P}$.

The input x is chosen according to the distribution $\mathcal{D}_N$.

Lower bounds for 1-sided-error testing

$\mathcal{D}_N$ be a distribution on the the set of inputs that are far from $\mathcal{P}$.

The input x is chosen according to the distribution $\mathcal{D}_N$.

And now if one shows that any deterministic algorithms that makes q queries will catch a WITNESS with very low probability then we obtain a lower bound of q on the query complexity for testing $\mathcal{P}$.

Lower bounds for 1-sided-error testing

$\mathcal{D}_N$ be a distribution on the the set of inputs that are far from $\mathcal{P}$.

The input x is chosen according to the distribution $\mathcal{D}_N$.

And now if one shows that any deterministic algorithms that makes q queries will catch a WITNESS with very low probability then we obtain a lower bound of q on the query complexity for testing $\mathcal{P}$.

For example: Checking whether $f : [n] \to [n]$ is 1-to-1 or 2-to-1 requires at least $\sqrt{n}$ queries. (By Birthday Paradox)

2-sided-error property tester

2-sided-property tester

Let $\mathcal{P}$ be a property. A 2-sided-error tester for $\mathcal{P}$ is a *randomized* algorithm $\mathcal{A}$ with black box access to an input x and satisfies:

- (Completeness) If $x \in \mathcal{P} \Rightarrow \Pr[\mathcal{A} \text{ accepts}] \geq 2/3$.
- (Soundness) If x is ϵ-far from $\mathcal{P} \Rightarrow \Pr[\mathcal{A} \text{ rejects}] \geq 2/3$.

We allow the algorithm to be *adaptive* (queries may depend on the outcome of previous queries).

2-sided-error tester

2-sided-error tester

- The tester does not have to find a PROOF/WITNESS to REJECT or ACCEPT.

2-sided-error tester

- The tester does not have to find a PROOF/WITNESS to REJECT or ACCEPT.
- The tester can use estimation/approximation as a tool.

For example: Distinguishing whether a string $x \in \{0,1\}^n$ has $n/4$ $1's$ OR $n/3$ $1's$ can be done using CONSTANT number of queries.

In general 2-sided-error algorithms can be very complicated.

Lower bounds for 2-sided-error testing

Lower bounds for 2-sided-error testing

Let $\mathcal{D}_N$ be a distribution on the the set of inputs that are far from $\mathcal{P}$ and $\mathcal{D}_Y$ be a distribution on the the set of inputs that satisfy $\mathcal{P}$.

Lower bounds for 2-sided-error testing

Let $\mathcal{D}_N$ be a distribution on the the set of inputs that are far from $\mathcal{P}$ and $\mathcal{D}_Y$ be a distribution on the the set of inputs that satisfy $\mathcal{P}$. The input x is chosen in the following manner:

Lower bounds for 2-sided-error testing

Let $\mathcal{D}_N$ be a distribution on the the set of inputs that are far from $\mathcal{P}$ and $\mathcal{D}_Y$ be a distribution on the the set of inputs that satisfy $\mathcal{P}$. The input x is chosen in the following manner:

- With probability $1/2$ the input x is chosen according to the distribution $\mathcal{D}_Y$

Lower bounds for 2-sided-error testing

Let $\mathcal{D}_N$ be a distribution on the the set of inputs that are far from $\mathcal{P}$ and $\mathcal{D}_Y$ be a distribution on the the set of inputs that satisfy $\mathcal{P}$. The input x is chosen in the following manner:

- With probability $1/2$ the input x is chosen according to the distribution $\mathcal{D}_Y$
- With the other $1/2$ probability the input x is chosen according to the distribution $\mathcal{D}_N$.

Lower bounds for 2-sided-error testing

Let $\mathcal{D}_N$ be a distribution on the the set of inputs that are far from $\mathcal{P}$ and $\mathcal{D}_Y$ be a distribution on the the set of inputs that satisfy $\mathcal{P}$. The input x is chosen in the following manner:

- With probability $1/2$ the input x is chosen according to the distribution $\mathcal{D}_Y$
- With the other $1/2$ probability the input x is chosen according to the distribution $\mathcal{D}_N$.

And now if one shows that any deterministic algorithms that makes q queries cannot distinguish the two kind of inputs then by Yao's Lemma we obtain a lower bound of q on the query complexity for testing $\mathcal{P}$.

Lower bounds for 2-sided-error testing

Let $\mathcal{D}_N$ be a distribution on the the set of inputs that are far from $\mathcal{P}$ and $\mathcal{D}_Y$ be a distribution on the the set of inputs that satisfy $\mathcal{P}$. The input x is chosen in the following manner:

- With probability $1/2$ the input x is chosen according to the distribution $\mathcal{D}_Y$
- With the other $1/2$ probability the input x is chosen according to the distribution $\mathcal{D}_N$.

And now if one shows that any deterministic algorithms that makes q queries cannot distinguish the two kind of inputs then by Yao's Lemma we obtain a lower bound of q on the query complexity for testing $\mathcal{P}$.

So, if the distribution of answers to the queries are similar when the input is drawn according to $\mathcal{D}_N$ and when it is drawn according to $\mathcal{D}_Y$ then the query complexity is $\geq q$.

Outline

Testing of Function Properties

- The property $\mathcal{P}$ is a set of functions from $\Sigma^n \to \Sigma$. For example: Linear functions, functions that are 1-to-1, functions accepted by a constant width read-once branching program etc.

Testing of Function Properties

- The property $\mathcal{P}$ is a set of functions from $\Sigma^n \to \Sigma$. For example: Linear functions, functions that are 1-to-1, functions accepted by a constant width read-once branching program etc.
- The input is a truth-table of a function $f : \Sigma^n \to \Sigma$.

Testing of Function Properties

- The property $\mathcal{P}$ is a set of functions from $\Sigma^n \to \Sigma$. For example: Linear functions, functions that are 1-to-1, functions accepted by a constant width read-once branching program etc.
- The input is a truth-table of a function $f : \Sigma^n \to \Sigma$.
- Queries are of form: $x \in \Sigma^n \longrightarrow f(x)$.

Property Tester for $\mathcal{P}$

A 1-sided-error tester for $\mathcal{P}$ is a *randomized* algorithm $\mathcal{A}$ that given query access to a truth-table of a function f does the following:

- If $f \in \mathcal{P} \Rightarrow \Pr[\mathcal{A} \text{ accepts}] = 1$.
- If for at least $\epsilon|\Sigma|^n$ number of strings in Σ^n the value of f has to be changed so that the property $\mathcal{P}$ is satisfied then $\Pr[\mathcal{A} \text{ rejects}] \geq 2/3$.

Testing of Linearity

Linearity testing

Given query access to a Boolean function $f : \{0,1\}^n \to \{0,1\}$ test if f is linear, that is, if for all $x, y \in \{0,1\}^n$, $f(x) \oplus f(y) = f(x \oplus y)$.

Testing of Linearity

Linearity testing

Given query access to a Boolean function $f : \{0,1\}^n \to \{0,1\}$ test if f is linear, that is, if for all $x, y \in \{0,1\}^n$, $f(x) \oplus f(y) = f(x \oplus y)$.

The obvious test is the following: pick two random $x, y \in \{0,1\}^n$ and if $f(x) \oplus f(y) \neq f(x \oplus y)$ then REJECT else ACCEPT.

Testing of Linearity

Linearity testing

Given query access to a Boolean function $f : \{0,1\}^n \to \{0,1\}$ test if f is linear, that is, if for all $x, y \in \{0,1\}^n$, $f(x) \oplus f(y) = f(x \oplus y)$.

The obvious test is the following: pick two random $x, y \in \{0,1\}^n$ and if $f(x) \oplus f(y) \neq f(x \oplus y)$ then REJECT else ACCEPT.

Linearity Testing [Blum-Luby-Rubinfeld]

The above tester has the following properties:

- If f is linear then the tester always ACCEPTS.
- If f is ϵ-far from linear then the tester REJECTS with high probability. (Proof using Fourier Analysis).

Generalization of Linearity Testing

Given query access to a function $f : \mathbb{F}^n \to \mathbb{F}$ test if f is a degree d polynomial.

Generalization of Linearity Testing

Given query access to a function $f : \mathbb{F}^n \to \mathbb{F}$ test if f is a degree d polynomial.

Low-degree testing [Babai-Fortnow-Lund, Rubinfeld-Sudan]

The query complexity for testing degree d polynomials is a function of $|\mathbb{F}|$ and d. When $|\mathbb{F}| = 2$ then the query complexity is 2^d and when $|\mathbb{F}|$ is around d then the query complexity is $poly(d)$.

Generalization of Linearity Testing

Given query access to a function $f : \mathbb{F}^n \to \mathbb{F}$ test if f is a degree d polynomial.

Low-degree testing [Babai-Fortnow-Lund, Rubinfeld-Sudan]

The query complexity for testing degree d polynomials is a function of $|\mathbb{F}|$ and d. When $|\mathbb{F}| = 2$ then the query complexity is 2^d and when $|\mathbb{F}|$ is around d then the query complexity is $poly(d)$.

This tester in also used in Probabilistically Checkable Proofs (PCP) [Arora-Safra, Arora-Lund-Motwani-Sudan-Szegedy]

Degree d tester, when $\mathbb{F} > d$.

Algorithm (For $|\mathbb{F}| > d$)

- *Pick a random* $x \in \mathbb{F}^n$
- *Pick a random line through* x. *Pick a random* $y \in \mathbb{F}^n$ *and consider all points of form* $x + \lambda y$.
- *Query at all the* $|\mathbb{F}|$ *points.*
- *If* f *is a degree* d *polynomial then restricted to this line it is a degree* d *univariate polynomial in variable* λ.
- *Use the points* $f(x + \lambda y)$, *when* $\lambda \neq 0$ *to fit a degree* d *polynomial.*
- *If the polynomial evaluated at* $\lambda = 0$ *is equal to* $f(x)$ *then ACCEPT else REJECT.*

Testing Parity

Given access to the truthtable of a function $f : \{0,1\}^n \to \{0,1\}$ test if the function f is a parity of some k variables.

- Can you show a upper bound independent of n?

Testing Parity

Given access to the truthtable of a function $f : \{0,1\}^n \to \{0,1\}$ test if the function f is a parity of some k variables.

- Can you show a upper bound independent of n?
- Can you show a upper bound of $O(klogk)$?

Testing Parity

Given access to the truthtable of a function $f : \{0,1\}^n \to \{0,1\}$ test if the function f is a parity of some k variables.

- Can you show a upper bound independent of n?
- Can you show a upper bound of $O(klogk)$?
- Can you show a lower bound of $\Omega(k)$?

Outline

Testing of graph property

- A property $\mathcal{P}$ is a set of graphs. For example: all bipartite graphs, all graphs that is isomorphic to a particular graph, all graphs where there exists a path from vertex 1 to vertex 2, ...

Testing of graph property

- A property $\mathcal{P}$ is a set of graphs. For example: all bipartite graphs, all graphs that is isomorphic to a particular graph, all graphs where there exists a path from vertex 1 to vertex 2, ...
- How is the graph given as input?

Dense Graph Model

A property $\mathcal{P}$ is a set of graphs. For example: all bipartite graphs, all graphs that is isomorphic to a particular graph, all graphs where there exists a path from vertex 1 to vertex 2, ...

Dense Graph Model

A property $\mathcal{P}$ is a set of graphs. For example: all bipartite graphs, all graphs that is isomorphic to a particular graph, all graphs where there exists a path from vertex 1 to vertex 2, ... The graph is given as an adjacency matrix. The input size is $\binom{|V|}{2}$.

Dense Graph Model

A property $\mathcal{P}$ is a set of graphs. For example: all bipartite graphs, all graphs that is isomorphic to a particular graph, all graphs where there exists a path from vertex 1 to vertex 2, ... The graph is given as an adjacency matrix. The input size is $\binom{|V|}{2}$.

A query is of form: Is there an edge between vertex i and j?

Dense Graph Model

A property $\mathcal{P}$ is a set of graphs. For example: all bipartite graphs, all graphs that is isomorphic to a particular graph, all graphs where there exists a path from vertex 1 to vertex 2, ... The graph is given as an adjacency matrix. The input size is $\binom{|V|}{2}$.

A query is of form: Is there an edge between vertex i and j?

Definition

A 1-sided-error tester for $\mathcal{P}$ is a *randomized* algorithm $\mathcal{A}$ that given query access to a graph G does the following:

- If $G \in \mathcal{P} \Rightarrow \Pr[\mathcal{A} \text{ accepts}] = 1$.
- If at least $\epsilon\binom{|V|}{2}$ number of entries of the adjacency matrix has to be changed so that the property $\mathcal{P}$ is satisfied then $\Pr[\mathcal{A} \text{ rejects}] \geq 2/3$.

Testing of Bipartiteness in the dense graph model

Given query access to the adjacency matrix of a graph G, test if G is bipartite of one has to remove $\epsilon\binom{|V|}{2}$ edges to make it bipartite.

Testing of Bipartiteness in the dense graph model

Given query access to the adjacency matrix of a graph G, test if G is bipartite of one has to remove $\epsilon\binom{|V|}{2}$ edges to make it bipartite.

Algorithm [Goldreich-Goldwasser-Ron]

- Pick $O(1/\epsilon^2 \log(1/\epsilon))$ number of vertices at random.
- Query all the pairs of selected vertices.
- If the induced graph is not bipartite REJECT else ACCEPT

Proof: If the graph is bipartite the algorithm always accept.

So now we have to prove that if G is ϵ-far from being bipartite then the induced graph is not bipartite with high probability.

Proof of Soundness of the Algorithm for Testing Bipartiteness

Since it is a 1-sided-error algorithm for every possible bipartition of the vertex set we should catch a violating edge, that is edges within the same part.

Proof of Soundness of the Algorithm for Testing Bipartiteness

Since it is a 1-sided-error algorithm for every possible bipartition of the vertex set we should catch a violating edge, that is edges within the same part.

If the graph is ϵ-far from being bipartite then any bipartition of the vertex set will have at least $\epsilon|V|^2$ violating edges.

Proof of Soundness of the Algorithm for Testing Bipartiteness

Since it is a 1-sided-error algorithm for every possible bipartition of the vertex set we should catch a violating edge, that is edges within the same part.

If the graph is ϵ-far from being bipartite then any bipartition of the vertex set will have at least $\epsilon|V|^2$ violating edges.

Note that given a particular bipartition by randomly sampling of $O(1/\epsilon^2)$ edges we would catch a violation for that bipartition with high probability. But we have to catch for all the bipartitions with high probability. Unfortunately, simple union bound does not give the math as the number of such bipartitions is $2^{|V|}$.

Proof of Soundness of the Algorithm for Testing Bipartiteness (contd...)

So we think of the selected vertices as two sets V_A and V_B. Vertices V_A induces the subgraph G_A.

After we have queried the subgraph G_A we show only a "small" number of partitions survive with high probability.

And then we can say, using union bound, that the second set V_B helps to catch the violations for the small number of surviving bipartitions.

Various other Graph Properties

- k-colorability of graphs $--$ $O(k/\epsilon)$.

Various other Graph Properties

- k-colorability of graphs $--$ $O(k/\epsilon)$.
- Is there a clique of size ρn $--$ $O(1/\epsilon)$ number of queries. (2-sided-error)

Various other Graph Properties

- k-colorability of graphs $--$ $O(k/\epsilon)$.
- Is there a clique of size ρn $--$ $O(1/\epsilon)$ number of queries. (2-sided-error)
- Triangle free-ness $--$ $tower(1/\epsilon)$. (Using Regularity Lemma)

What all can be tested?

Can we characterize all the graph properties that can be tested by a 1-sided-error tester using constant number of queries.

What all can be tested?

Can we characterize all the graph properties that can be tested by a 1-sided-error tester using constant number of queries.

Theorem (Alon-Shapira)

A graph property is called monotone if it is closed under removal of edges and vertices. Every monotone graph property is testable with constant number of queries.

What all can be tested?

Can we characterize all the graph properties that can be tested by a 1-sided-error tester using constant number of queries.

Theorem (Alon-Shapira)

A graph property is called monotone if it is closed under removal of edges and vertices. Every monotone graph property is testable with constant number of queries.

The proof uses Szemeredi's Regularity Lemma.

What all can be tested?

Can we characterize all the graph properties that can be tested by a 1-sided-error tester using constant number of queries.

Theorem (Alon-Shapira)

A graph property is called monotone if it is closed under removal of edges and vertices. Every monotone graph property is testable with constant number of queries.

The proof uses Szemeredi's Regularity Lemma.
The proof roughly based on the idea that testing monotone graph properties can be reduced to testing whether the graph has a regular-partition with certain parameters.

What all can be tested?

Can we characterize all the graph properties that can be tested by a 1-sided-error tester using constant number of queries.

Theorem (Alon-Shapira)

A graph property is called monotone if it is closed under removal of edges and vertices. Every monotone graph property is testable with constant number of queries.

The proof uses Szemeredi's Regularity Lemma.
The proof roughly based on the idea that testing monotone graph properties can be reduced to testing whether the graph has a regular-partition with certain parameters.
And testing whether a graph has a regular-partition can be tested with constant number of queries.

In the dense-graph-model its all about regularity

Theorem (Alon-Fischer-Newman-Shapira)

A graph property P can be tested with a constant number of queries if and only if testing P can be reduced to testing the property of satisfying one of finitely many Szemeredi-partitions.

Testing of Connectivity

Problem

Can we test whether in a graph is connected?

Testing of Connectivity

Problem

Can we test whether in a graph is connected?

Actually, ... this problem does not make much sense - in the dense graph model.
All graphs are just $|V|$ changes away from being connected and hence all graphs are ϵ-close to being connected.

Testing of Connectivity

Problem

Can we test whether in a graph is connected?

Actually, ... this problem does not make much sense - in the dense graph model.
All graphs are just $|V|$ changes away from being connected and hence all graphs are ϵ-close to being connected.

So we need some other models for sparse-graph-properties.

Sparse Graph Model

- The input is a graph with m edges.

Sparse Graph Model

- The input is a graph with m edges.
- Queries are of the form: What is the ith neighbor of vertex v?

Sparse Graph Model

- The input is a graph with m edges.
- Queries are of the form: What is the ith neighbor of vertex v?
- If the degree of v is less than i then the answer to query is "NONE".

Sparse Graph Model

- The input is a graph with m edges.
- Queries are of the form: What is the ith neighbor of vertex v?
- If the degree of v is less than i then the answer to query is "NONE".

Property Tester for Bounded Degree Model

A 1-sided-error tester for $\mathcal{P}$ is a *randomized* algorithm $\mathcal{A}$ that given query access to a graph G does the following:

- If $G \in \mathcal{P} \Rightarrow \Pr[\mathcal{A} \text{ accepts}] = 1$.
- If at least ϵm number of edges has to be added or removed so that the property $\mathcal{P}$ is satisfied then $\Pr[\mathcal{A} \text{ rejects}] \geq 2/3$.

Testing Connectivity in Sparse Graph Model

Observation

If a graph G is ϵ-far (in the sparse-graph-model) from being connected then it has more than $\epsilon m + 1$ connected components. And thus it must have at least $(\epsilon/2)m$ number of components of size at most $2n/\epsilon m$.

Testing Connectivity in Sparse Graph Model

Observation

If a graph G is ϵ-far (in the sparse-graph-model) from being connected then it has more than $\epsilon m + 1$ connected components. And thus it must have at least $(\epsilon/2)m$ number of components of size at most $2n/\epsilon m$.

Algorithm

- *Randomly pick $4n/\epsilon m$ vertices.*

Testing Connectivity in Sparse Graph Model

Observation

If a graph G is ϵ-far (in the sparse-graph-model) from being connected then it has more than $\epsilon m + 1$ connected components. And thus it must have at least $(\epsilon/2)m$ number of components of size at most $2n/\epsilon m$.

Algorithm

- *Randomly pick $4n/\epsilon m$ vertices.*
- *Do a BFS from each of the selected vertices till you find $2n/\epsilon m$ vertices.*

Testing Connectivity in Sparse Graph Model

Observation

If a graph G is ϵ-far (in the sparse-graph-model) from being connected then it has more than $\epsilon m + 1$ connected components. And thus it must have at least $(\epsilon/2)m$ number of components of size at most $2n/\epsilon m$.

Algorithm

- *Randomly pick $4n/\epsilon m$ vertices.*
- *Do a BFS from each of the selected vertices till you find $2n/\epsilon m$ vertices.*
- *If you find a component of size less than $2n/\epsilon m$ then REJECT, else ACCEPT.*

Other problems that has constant query complexity in the sparse graph model

- Cycle-freeness,
- Eulerianess,
- subgraph freeness

All the above has similar algorithms to connectivity testing.

Testing of *st*-connectedness

Problem

Can we test whether in a graph there is a path from a given vertex s to another given vertex t?

Testing of *st*-connectedness

Problem

Can we test whether in a graph there is a path from a given vertex s to another given vertex t?

Actually, ... this problem does not make much sense - in the sparse graph model also.
All graphs are just 1 change away from having *st*-connectivity and hence all graphs are ϵ-close to being *st*-connected.

Testing of *st*-connectedness

Problem

Can we test whether in a graph there is a path from a given vertex s to another given vertex t?

Actually, ... this problem does not make much sense - in the sparse graph model also.
All graphs are just 1 change away from having *st*-connectivity and hence all graphs are ϵ-close to being *st*-connected.

So we need some other models for this.

Orientation Model

- The input graph is a directed graph.

Orientation Model

- The input graph is a directed graph.
- The underlying un-directed graph is known in advance.

Orientation Model

- The input graph is a directed graph.
- The underlying un-directed graph is known in advance.
- Queries are of the form: What is the orientation of the edge e?

Orientation Model

- The input graph is a directed graph.
- The underlying un-directed graph is known in advance.
- Queries are of the form: What is the orientation of the edge e?

Property Tester for Orientation Model

A 1-sided-error tester for $\mathcal{P}$ is a *randomized* algorithm $\mathcal{A}$ that given query access to a graph G does the following:

- If $G \in \mathcal{P} \Rightarrow \Pr[\mathcal{A} \text{ accepts}] = 1$.
- If at least ϵm number of edges has to be re-oriented so that the property $\mathcal{P}$ is satisfied then $\Pr[\mathcal{A} \text{ rejects}] \geq 2/3$.

Testing in the orientation model

st-connectivity [C-Fischer-Lachish-Matsliah-Newman]

There is a 1-sided-error tester that makes $2^{2^{2^{O(1/\epsilon)}}}$ number of queries and tests for st-connectivity in the orientation model (ϵ is the distance parameter).

Testing in the orientation model

st-connectivity [C-Fischer-Lachish-Matsliah-Newman]

There is a 1-sided-error tester that makes $2^{2^{2^{O(1/\epsilon)}}}$ number of queries and tests for st-connectivity in the orientation model (ϵ is the distance parameter).

- Other properties like Eulerianness has also been studied in this model. But their query complexity is not constant.
- Not many properties are known to have constant query complexity in the orientation model.
- Even proving that a constant size witness exist is also hard. For example: If G is ϵ far from being s-to-all connected then does there exist a constant size witness?

Characterization in the Sparse-Graph-Model and Orientation-Model

Characterization in the Sparse-Graph-Model and Orientation-Model

Characterization of properties that can be tested using constant number of queries in the Sparse-Graph-Model. −− OPEN

Characterization in the Sparse-Graph-Model and Orientation-Model

Characterization of properties that can be tested using constant number of queries in the Sparse-Graph-Model. −− OPEN

Characterization of properties that can be tested using constant number of queries in the Orientation-Model. −− OPEN

What we saw till now...

What we saw till now...

- Function Property Testing
  - Linearity, low degree, constant-width-read-once-BP, k-juntas have constant query complexity
  - Testing of k parity.
- Distribution Testing
  - Uniformity testing has query complexity $\tilde{\Theta}(\sqrt{|Range|})$.

What we saw till now...

- Function Property Testing
  - Linearity, low degree, constant-width-read-once-BP, k-juntas have constant query complexity
  - Testing of k parity.
- Distribution Testing
  - Uniformity testing has query complexity $\tilde{\Theta}(\sqrt{|Range|})$.
- Graph Property Testing
  - k-colorability in dense graph model is testable with $O(k)$ queries,
  - Dense Graph Model - Testing is all about regularity,
  - Sparse Graph Model - testing of connectivity
  - Orientation Model - testing of s-connectivity

Current directions....

- Get tight query complexity for testing various properties.
- Classify Boolean function properties that can be tested using constant number of queries.
- Connection to communication complexity: like connection to gap-Hamming problem.
- Get lower bounds on the query complexity (dependence on ϵ) for graph properties: connection to additive combinatorics.
- Connection to LDC/PIR.
- Connection to learning theory.

Outline

Testing of Graph Isomorphism

Let H be a fixed graph. Then given query access to the adjacency matrix of a graph G test if G is isomorphic to H or if G is ϵ-far from being isomorphic to H.

Testing of Graph Isomorphism

Let H be a fixed graph. Then given query access to the adjacency matrix of a graph G test if G is isomorphic to H or if G is ϵ-far from being isomorphic to H.

Graph Isomorphism Testing [Fischer-Matsliah]

The 1-sided-query complexity for testing isomorphism to a fixed graph is $\tilde{\Theta}(|V|)$, whereas the 2-sided-error query complexity is $\tilde{\Theta}(\sqrt{|V|})$.

Testing of Graph Isomorphism

Let H be a fixed graph. Then given query access to the adjacency matrix of a graph G test if G is isomorphic to H or if G is ϵ-far from being isomorphic to H.

Graph Isomorphism Testing [Fischer-Matsliah]

The 1-sided-query complexity for testing isomorphism to a fixed graph is $\tilde{\Theta}(|V|)$, whereas the 2-sided-error query complexity is $\tilde{\Theta}(\sqrt{|V|})$.

GI Testing with constant number of queries [Fischer]

The query complexity for testing isomorphism to a fixed graph is constant iff the given graph is close to a graph that is generated by a constant number of cliques.

Hyper-graph isomorphism testing and its generalizations

Hyper-Graph Isomorphism testing: Let H be a fixed d-refular-hypergraph. Then given query access to the adjacency matrix of a d-regular-hypergraph G test if G is isomorphic to H or if G is ϵ-far from being isomorphic to H.

Hyper-graph isomorphism testing and its generalizations

Hyper-Graph Isomorphism testing: Let H be a fixed d-refular-hypergraph. Then given query access to the adjacency matrix of a d-regular-hypergraph G test if G is isomorphic to H or if G is ϵ-far from being isomorphic to H.

Testing Isomorphism under Group Operations: Let $\mathcal{G}$ be a primitive subgroup of S_n. Let $x \in \{0,1\}^n$ be a fixed string. Then given a string $y \in \{0,1\}$ test if x is isomorphic to y under permutation of the indices by elements of the group $\mathcal{G}$, that is, is there a $\pi \in \mathcal{G}$ such that for all i, $x_i = y_{\pi(i)}$, OR for all $\pi \in (G)$ for at least ϵn indices i, $x_i \neq y_{\pi(i)}$.

Testing Isomorphism under Group Operations: Let $\mathcal{G}$ be a primitive subgroup of S_n. Let $x \in \{0,1\}^n$ be a fixed string. Then given a string $y \in \{0,1\}$ test if x is isomorphic to y under permutation of the indices by elements of the group $\mathcal{G}$, that is, is there a $\pi \in \mathcal{G}$ such that for all i, $x_i = y_{\pi(i)}$, OR for all $\pi \in (G)$ for at least ϵn indices i, $x_i \neq y_{\pi(i)}$.

Testing Isomorphism under Group Operations [Babai-C]

The query complexity for test isomorphism under primitive group operation is $\tilde{\Theta}(\log |G|)$. This implies the query complexity for testing d-regular hypergraph isomorphism is $\tilde{\Theta}(|V|)$.
For 2-sided-error the bounds are $\tilde{\Theta}(\sqrt{\log |G|})$ and $\tilde{\Theta}(\sqrt{|V|})$ respectively.

Boolean Function Isomorphism Testing

Let $f : \{0,1\}^n \to \{0,1\}$ be a fixed function. Then given query access to the truth-table of a function g test if g is isomorphic to f upto a permutation of its variable, that is, does there exist a permutation $\pi \in S_n$ such that for all x, $f(x^\pi) = g(x)$, where $x_i^\pi = x_{\pi(i)}$.

Boolean Function Isomorphism Testing

Let $f : \{0,1\}^n \to \{0,1\}$ be a fixed function. Then given query access to the truth-table of a function g test if g is isomorphic to f upto a permutation of its variable, that is, does there exist a permutation $\pi \in S_n$ such that for all x, $f(x^\pi) = g(x)$, where $x_i^\pi = x_{\pi(i)}$.
For example:

- Is the function g a dictator function? −− Constant query complexity.
- Is the function a parity on k variable? −− Query complexity $O(k \log k)$ and $\Omega(k)$
- Is the function isomorphic to Majority? −− Constant Query Complexity.

Boolean Function Isomorphism Testing

Let $f : \{0,1\}^n \to \{0,1\}$ be a fixed function. Then given query access to the truth-table of a function g test if g is isomorphic to f upto a permutation of its variable, that is, does there exist a permutation $\pi \in S_n$ such that for all x, $f(x^\pi) = g(x)$, where $x_i^\pi = x_{\pi(i)}$.

Boolean FI testing [Alon-Blais, C-Garcia-Soriano-Matsliah]

The 1-sided-error query complexity for testing isomorphism to a k-junta is $\Theta(k \log n)$ where as the 2-sided-error query complexity is $O(k \log k)$.

Characterization?

Just like in case of graph isomorphism Fischer proved that query complexity is constant iff the graph is generated by a constant number of cliques, can we say something like that for function isomorphism.

Characterization?

Just like in case of graph isomorphism Fischer proved that query complexity is constant iff the graph is generated by a constant number of cliques, can we say something like that for function isomorphism.

We know isomorphism to k-junta takes only $k \log k$ queries. Also isomorphism to a symmetric function takes constant number of queries. Can we combine to say something like -

Characterization?

Just like in case of graph isomorphism Fischer proved that query complexity is constant iff the graph is generated by a constant number of cliques, can we say something like that for function isomorphism.

We know isomorphism to k-junta takes only $k \log k$ queries. Also isomorphism to a symmetric function takes constant number of queries. Can we combine to say something like -

Conjecture

If $f(x)$ depends on $|X|$ and at most k indices then the query complexity for testing isomorphism to f is $O(k \log k)$ and $\Omega(\log k)$.

Conclusion

Conclusion

- Function Property Testing
  - Linearity, low degree, constant-width-read-once-BP, k-juntas have constant query complexity
  - Monotonicity - query complexity is $\Omega(n)$ and $O(n^2)$
  - Testing distribution - uniformity testing has query complexity $\tilde{\Theta}(\sqrt{|Range|})$.

Conclusion

- Function Property Testing
  - Linearity, low degree, constant-width-read-once-BP, k-juntas have constant query complexity
  - Monotonicity - query complexity is $\Omega(n)$ and $O(n^2)$
  - Testing distribution - uniformity testing has query complexity $\tilde{\Theta}(\sqrt{|Range|})$.
- Graph Property Testing
  - k-colorability in dense graph model is testable with $O(k)$ queries,
  - Dense Graph Model - Testing is all about regularity,
  - Sparse Graph Model - testing of connectivity
  - Orientation Model - testing of s-connectivity

Conclusion

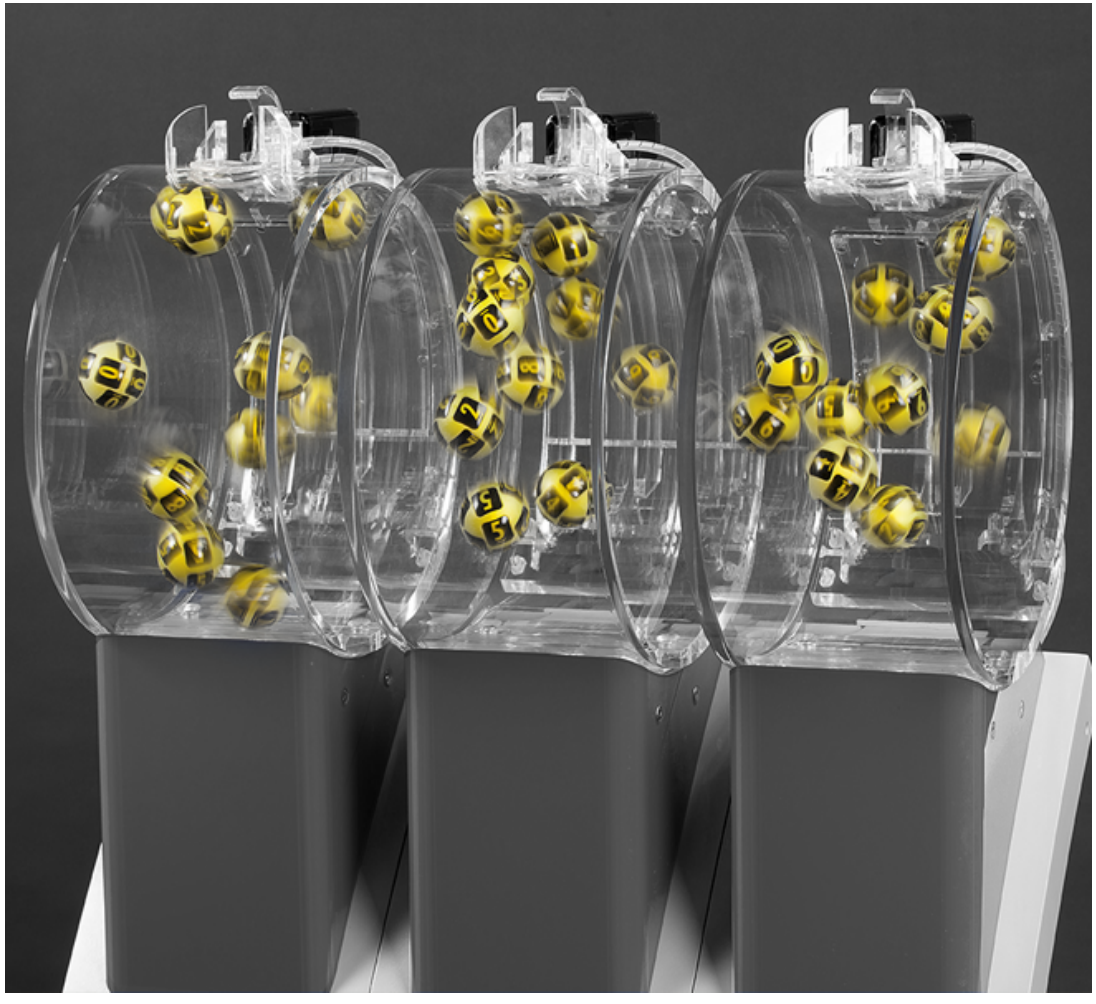
- Function Property Testing
  - Linearity, low degree, constant-width-read-once-BP, k-juntas have constant query complexity
  - Monotonicity - query complexity is $\Omega(n)$ and $O(n^2)$
  - Testing distribution - uniformity testing has query complexity $\tilde{\Theta}(\sqrt{|Range|})$.
- Graph Property Testing
  - k-colorability in dense graph model is testable with $O(k)$ queries,
  - Dense Graph Model - Testing is all about regularity,
  - Sparse Graph Model - testing of connectivity
  - Orientation Model - testing of s-connectivity
- Isomorphism Testing
  - Generalization of GI testing
  - Isomorphism to k-junta can be tested with $O(k \log k)$ queries.

Property Testing of Properties of Distributions

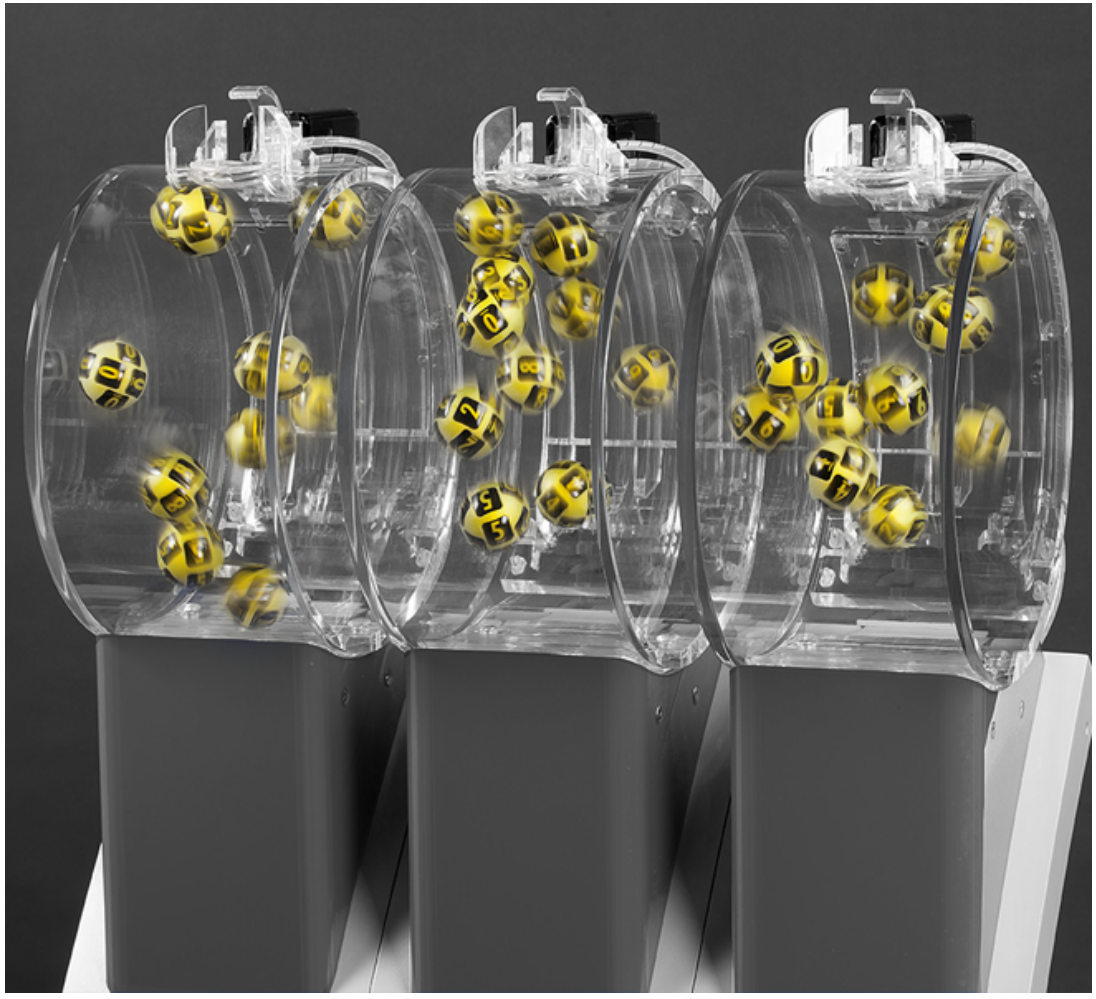
Sourav Chakraborty

Indian Statistical Institute
Kolkata, India

A teaser: What is this?



Guiding Problem



How can you test if the machine is truely random?

Testing of Distribution Properties

Given access to a distribution on domain of size n how many queries/samples one need to test if the distribution has a certain property or is "far" from having the property.

Testing of Distribution Properties

Given access to a distribution on domain of size n how many queries/samples one need to test if the distribution has a certain property or is "far" from having the property.

Eg:, the simplest and the most fundamental property to test is

"Is a distribution on $\{1, \ldots, n\}$ uniform on the domain?"

Testing of Distribution Properties

Given access to a distribution on domain of size n how many queries/samples one need to test if the distribution has a certain property or is "far" from having the property.

Eg:, the simplest and the most fundamental property to test is

"Is a distribution on $\{1, \ldots, n\}$ uniform on the domain?"

- The distribution may be implicitly given.
- Usually a query means drawing a random sample according to the distribution.
- Usually "far" means far in the variation distance.

HOW MANY QUERIES/SAMPLES ARE NECESSARY AND SUFFICIENT?

Importance of Distribution Testing

- Distribution Testing is often the central problem in many algorithms and protocols.

Importance of Distribution Testing

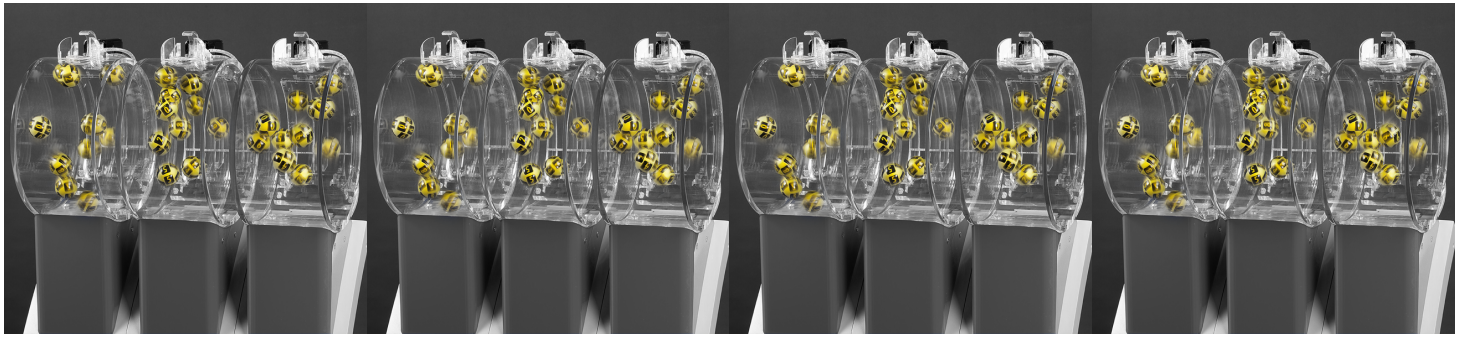
- Distribution Testing is often the central problem in many algorithms and protocols.
- The real life application is huge.

Importance of Distribution Testing

- Distribution Testing is often the central problem in many algorithms and protocols.
- The real life application is huge.
  1. Given a CNF formula producing a satisfying assignments uniformly at random from the set of all satisfying assignments is a crucial problem in the SAT-solver community.
     How to check if the output of a "claimed algorithm" is according to the uniform distribution?

Importance of Distribution Testing

- Distribution Testing is often the central problem in many algorithms and protocols.
- The real life application is huge.
  1. Given a CNF formula producing a satisfying assignments uniformly at random from the set of all satisfying assignments is a crucial problem in the SAT-solver community.
     How to check if the output of a "claimed algorithm" is according to the uniform distribution?
  2. Checking if a random number generator is correct.



Classical Sampling

- The queries are sample drawn according to the distribution
- "far" means total variation distance or the ℓ_1 distance.

Our Goal

To design an algorithm that, given access to random samples drawn according to a distribution μ on $\{1, \ldots, n\}$, will

- ACCEPT, with probability $2/3$, if μ is the uniform distribution, and
- REJECT, with probability $2/3$, if μ is ϵ-far from being the uniform distribution, that is, if the ℓ_1 distance from the uniform distribution is at least ϵ.

Testing Uniformity using Classical Samples

Uniformity Testing [Batu-Fortnow-Rubinfeld-Smith-White]

2-side-error query complexity for testing uniformity is $\tilde{\Theta}(\sqrt{k})$.

Sketch of Proof for Testing Uniformity

Uniformity Testing [Batu-Fortnow-Rubinfeld-Smith-White]

2-side-error query complexity for testing uniformity is $\tilde{\Theta}(\sqrt{k})$.

Sketch of Proof for Testing Uniformity

Uniformity Testing [Batu-Fortnow-Rubinfeld-Smith-White]

2-side-error query complexity for testing uniformity is $\tilde{\Theta}(\sqrt{k})$.

Proof.

Upper bound: Take random $\sqrt{k}$ samples and check if they fall in different buckets. If they all fall on distinct buckets estimate the fraction of elements that fall in these buckets.

Sketch of Proof for Testing Uniformity

Uniformity Testing [Batu-Fortnow-Rubinfeld-Smith-White]

2-side-error query complexity for testing uniformity is $\tilde{\Theta}(\sqrt{k})$.

Proof.

Upper bound: Take random $\sqrt{k}$ samples and check if they fall in different buckets. If they all fall on distinct buckets estimate the fraction of elements that fall in these buckets.

Lower bound: Distinguishing whether μ is uniform with support size k from μ is uniform with support size $k/2$ requires $\sqrt{k}$ queries. Just like distinguishing 1-to-1 function from 2-to-1 functions. □

Uniformity Testing Using Classical Sampling

- The queries are sample drawn according to the distribution
- "far" means total variation distance or the ℓ_1 distance.

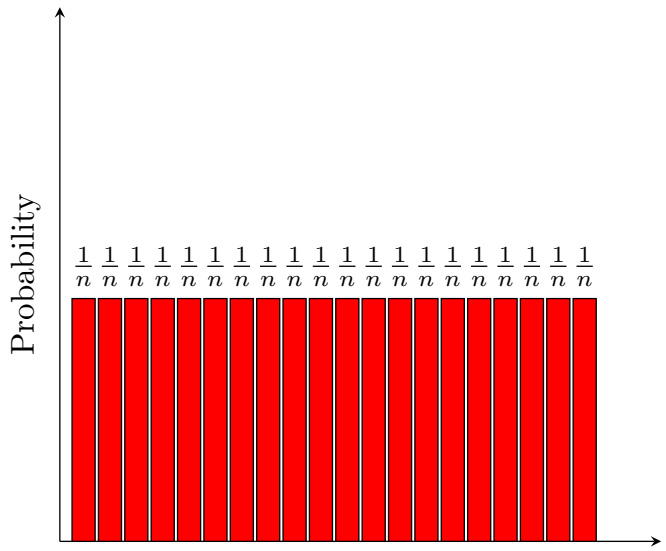


Figure: Uniform Distribution

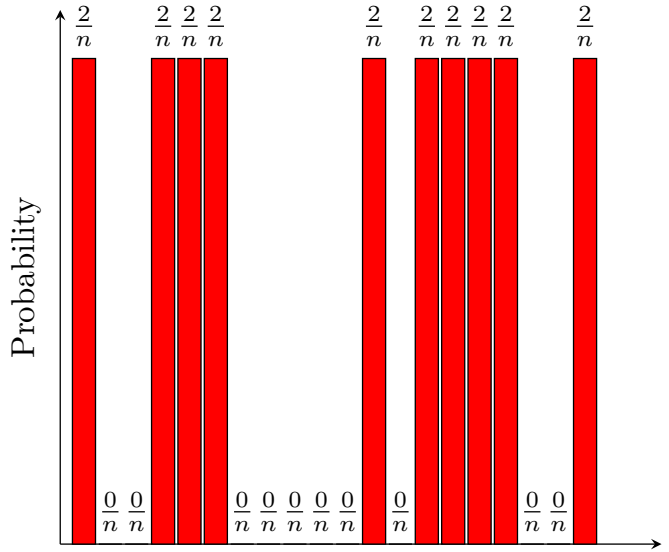


Figure: 1/2-far from uniform

Uniformity Testing Using Classical Sampling

- The queries are sample drawn according to the distribution
- "far" means total variation distance or the ℓ_1 distance.

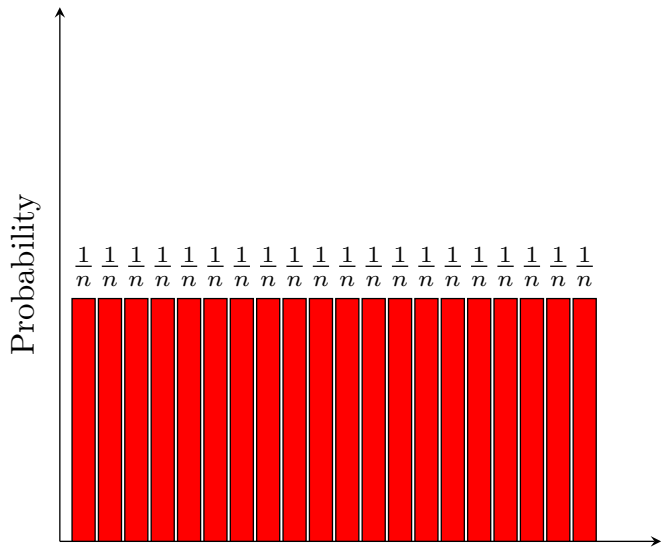


Figure: Uniform Distribution

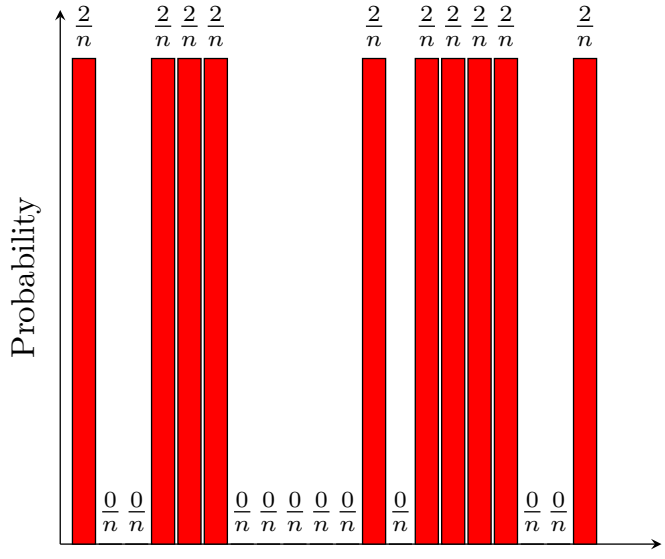


Figure: 1/2-far from uniform

- If $< \sqrt{n}/100$ samples are drawn then with high probability you see only distinct samples from either distribution.

Although sublinear, not practical

Theorem (Batu-Fortnow-Rubinfeld-Smith-White (JACM 2013))

Testing whether a distribution is ϵ-close to uniform has query complexity $\Theta(\sqrt{n}/\epsilon^2)$. [Paninski (Trans. Inf. Theory 2008)]

Although sublinear, not practical

Theorem (Batu-Fortnow-Rubinfeld-Smith-White (JACM 2013))

Testing whether a distribution is ϵ-close to uniform has query complexity $\Theta(\sqrt{n}/\epsilon^2)$. [Paninski (Trans. Inf. Theory 2008)]

1. Given a CNF formula producing a satisfying assignments uniformly at random from the set of all satisfying assignments is a crucial problem in the SAT-solver community.
   How to check if the output of a "claimed algorithm" is according to the uniform distribution?

Although sublinear, not practical

Theorem (Batu-Fortnow-Rubinfeld-Smith-White (JACM 2013))

Testing whether a distribution is ϵ-close to uniform has query complexity $\Theta(\sqrt{n}/\epsilon^2)$. [Paninski (Trans. Inf. Theory 2008)]

1. Given a CNF formula producing a satisfying assignments uniformly at random from the set of all satisfying assignments is a crucial problem in the SAT-solver community.
   How to check if the output of a "claimed algorithm" is according to the uniform distribution?
   Typically, number of variables in the formula ≈ 1000 and the number of satisfying assignments $\approx 2^{70}$.

Although sublinear, not practical

Theorem (Batu-Fortnow-Rubinfeld-Smith-White (JACM 2013))

Testing whether a distribution is ϵ-close to uniform has query complexity $\Theta(\sqrt{n}/\epsilon^2)$. [Paninski (Trans. Inf. Theory 2008)]

1. Given a CNF formula producing a satisfying assignments uniformly at random from the set of all satisfying assignments is a crucial problem in the SAT-solver community.
   How to check if the output of a "claimed algorithm" is according to the uniform distribution?
   Typically, number of variables in the formula ≈ 1000 and the number of satisfying assignments $\approx 2^{70}$.

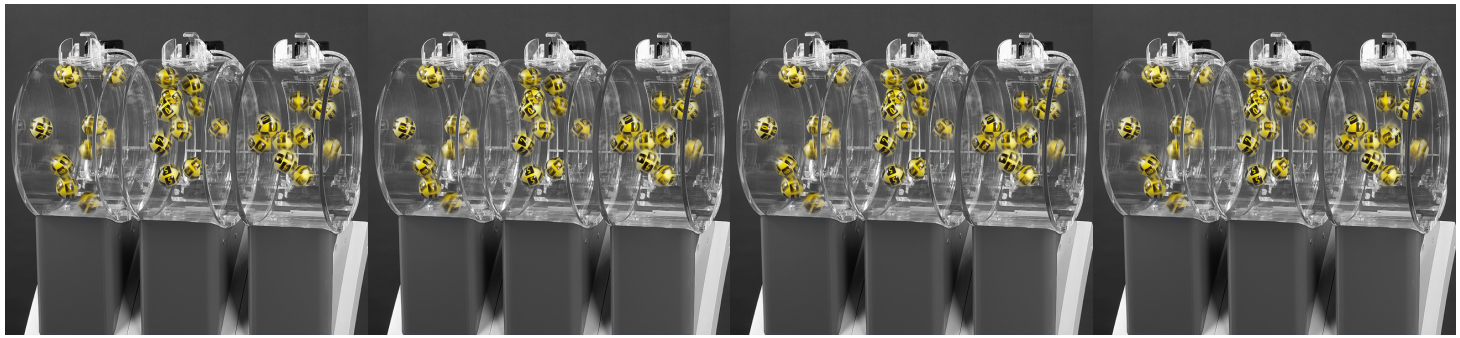
Number of queries needed is around 2^{35}.

Although sublinear, not practical

Theorem (Batu-Fortnow-Rubinfeld-Smith-White (JACM 2013))

Testing whether a distribution is ϵ-close to uniform has query complexity $\Theta(\sqrt{n}/\epsilon^2)$. [Paninski (Trans. Inf. Theory 2008)]

2. Checking if a random number generator is correct.

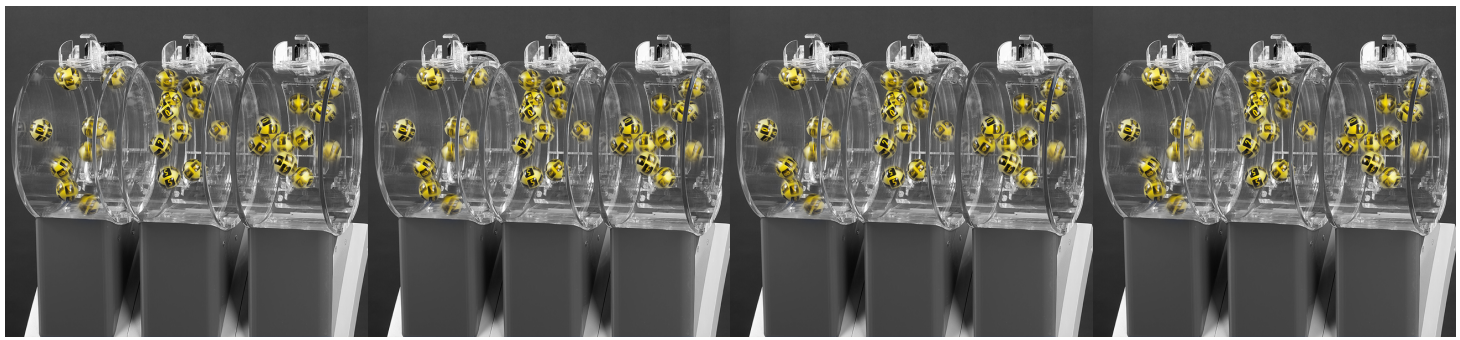


Although sublinear, not practical

Theorem (Batu-Fortnow-Rubinfeld-Smith-White (JACM 2013))

Testing whether a distribution is ϵ-close to uniform has query complexity $\Theta(\sqrt{n}/\epsilon^2)$. [Paninski (Trans. Inf. Theory 2008)]

2. Checking if a random number generator is correct.



If the machine outputs 12 digit numbers then

Number of times the machine has to be run is 10^6 times.

More sophisticated models of sampling

In the literature more sophisticated models of querying has been studied...

More sophisticated models of sampling

In the literature more sophisticated models of querying has been studied...

- *Quantum queries*

Theorem (C-Fischer-Matsliah-Wolf (2011))

Testing if a distribution on domain size n is uniform has quantum query complexity $\widetilde{\Theta}(n^{1/3})$.

More sophisticated models of sampling

In the literature more sophisticated models of querying has been studied...

- *Quantum queries*

Theorem (C-Fischer-Matsliah-Wolf (2011))

Testing if a distribution on domain size n is uniform has quantum query complexity $\widetilde{\Theta}(n^{1/3})$.

- Many more has been studied ...

More sophisticated models of sampling

In the literature more sophisticated models of querying has been studied...

- *Quantum queries*

Theorem (C-Fischer-Matsliah-Wolf (2011))

Testing if a distribution on domain size n is uniform has quantum query complexity $\widetilde{\Theta}(n^{1/3})$.

- Many more has been studied ...

- *Conditional Sampling*
  Introduced independently by C-Fischer-Matsliah-Goldhirsh (SICOMP 2016) and Cannone-Ron-Servedio (SICOMP 2015).

Conditional Sampling

Definition (Conditional Sampling)

Given a distribution $\mathcal{D}$ on a domain D one can

- *Specify a set $S \subseteq D$,*
- *Draw samples according to the distribution $\mathcal{D}|_S$, that is, $\mathcal{D}$ under the condition that the samples belong to S.*

Conditional Sampling

Definition (Conditional Sampling)

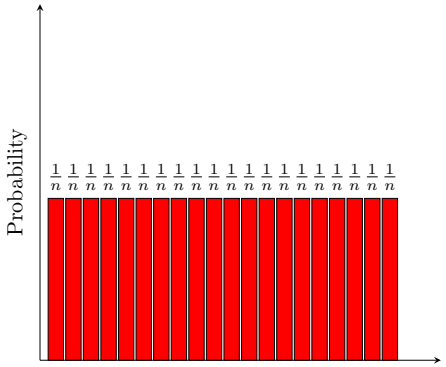
Given a distribution $\mathcal{D}$ on a domain D one can

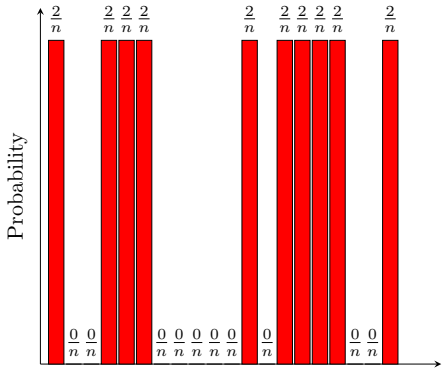
- *Specify a set $S \subseteq D$,*
- *Draw samples according to the distribution $\mathcal{D}|_S$, that is, $\mathcal{D}$ under the condition that the samples belong to S.*

Clearly such a sampling is at least as powerful as drawing normal samples.

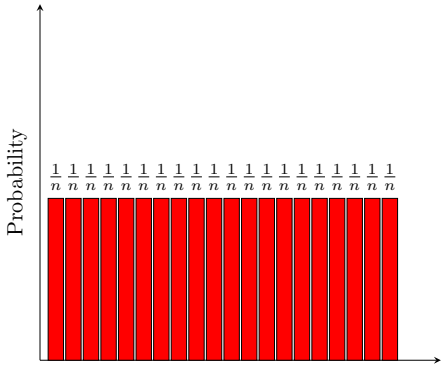
But how much powerful is it?

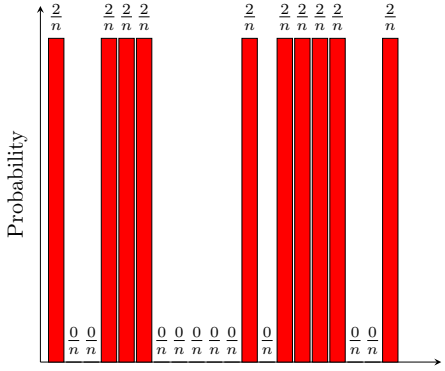
Testing Uniformity Using Conditional Sampling





Testing Uniformity Using Conditional Sampling

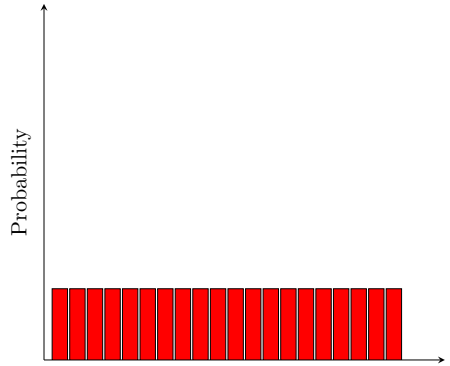


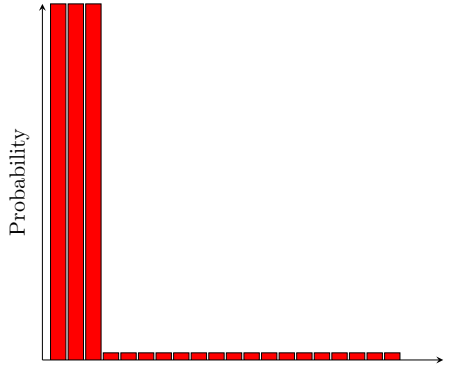


An algorithm for testing uniformity using conditional sampling:

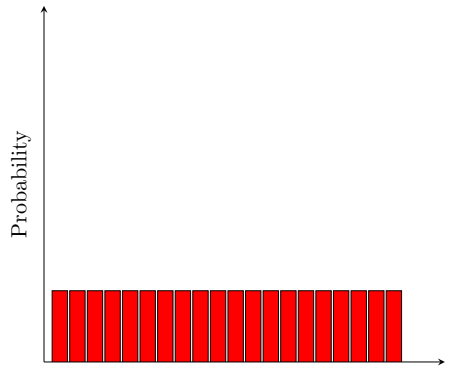
1. Draw two elements x and y uniformly at random from the domain. Let $S = \{x, y\}$.
2. In the case of the "far" distribution, with probability $1/2$, one of the two elements will have probability 0, and the other probability non-zero.
3. Now a constant number of conditional samples drawn from $\mathcal{D}|_S$ is enough to identify that it is not uniform.

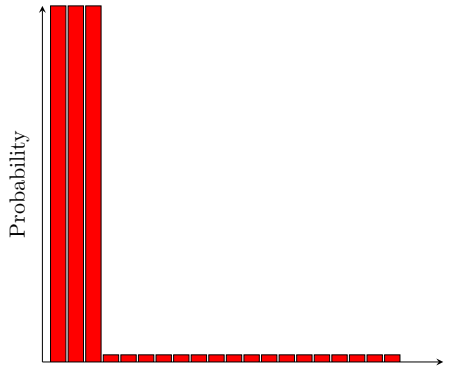
How about other distributions?





How about other distributions?

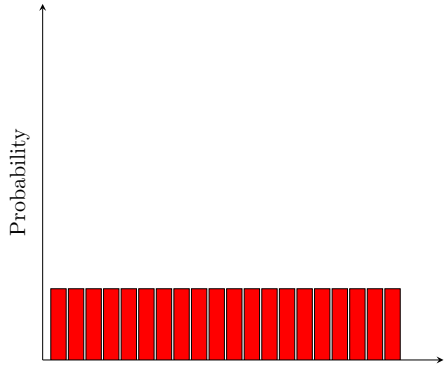


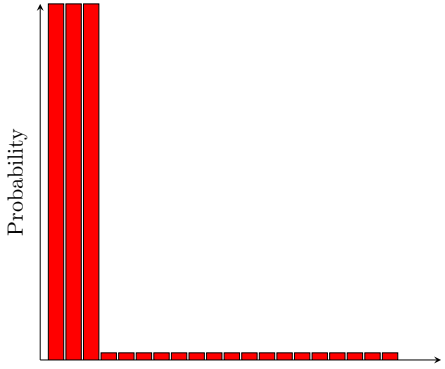


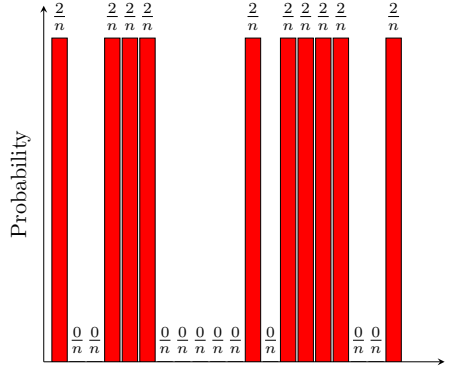
Previous algorithm fails in this case:

1. Draw two elements x and y uniformly at random from the domain. Let $S = \{x, y\}$.
2. In the case of the "far" distribution, with probability almost 1, both the two elements will have probability same, namely ϵ.
3. Probability that we will be able to distinguish the far distribution from the uniform distribution is very low.

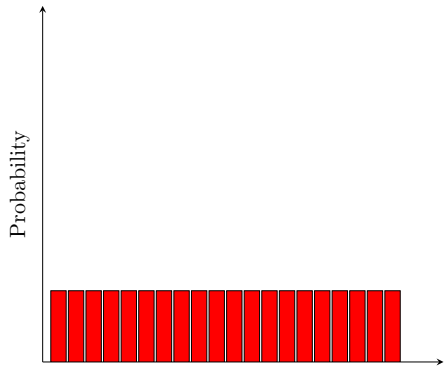
Testing Uniformity Using Conditional Sampling

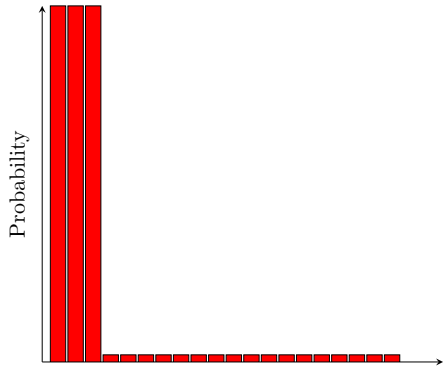


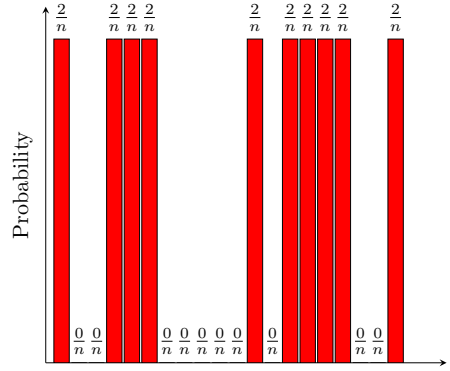




Testing Uniformity Using Conditional Sampling







A quick fix:

1. Draw x uniformly at random from the domain and draw y according to the distribution $\mathcal{D}$. Let $S = \{x, y\}$.
2. In the case of the "far" distribution, with constant probability, x will have "low" probability and y will have "high" probibility.
3. We will be able to distinguish the far distribution from the uniform distribution using constant number of conditional samples from $\mathcal{D}|_S$.
4. The constant depend on the farness parameter.

Power of Conditional Samples

Theorem (C-Fischer-Matsliah-Goldhirsh (SICOMP 2016))

For testing if a distribution is uniform one needs only poly$(1/\epsilon)$ number of conditional samples.

Power of Conditional Samples

Theorem (C-Fischer-Matsliah-Goldhirsh (SICOMP 2016))

For testing if a distribution is uniform one needs only poly$(1/\epsilon)$ number of conditional samples.

Theorem (Canonne-Ron-Servedio (SICOMP 2015))

For testing if a distribution is uniform one needs only $\tilde{\theta}(1/\epsilon^2)$ number of conditional samples.

Uniformity Tester using Conditional Sampling

Given μ over $\{1, \ldots, n\}$ and the distance parameter ϵ.

Uniformity Tester

1. Let S be a set of $10/\epsilon$ samples drawn according to the distribution;
2. Let T be a set of $10/\epsilon$ samples drawn according to the uniform distribution over $\{1, \ldots, n\}$;
3. Use classical uniformity tester to test if $\mu|_{S \cup T}$ is uniform (with distance parameter $1/40\epsilon^2$.

Uniformity Tester using Conditional Sampling

Given μ over $\{1, \ldots, n\}$ and the distance parameter ϵ.

Uniformity Tester

1. Let S be a set of $10/\epsilon$ samples drawn according to the distribution;
2. Let T be a set of $10/\epsilon$ samples drawn according to the uniform distribution over $\{1, \ldots, n\}$;
3. Use classical uniformity tester to test if $\mu|_{S \cup T}$ is uniform (with distance parameter $1/40\epsilon^2$.

Proof of Correctness

- Fact 1:With probability $\geq 8/9$, $\exists\, i \in S$ with $\mu(i) \geq \frac{1}{n}(1 + \frac{\epsilon}{2})$
- Fact 2:With probability $\geq 8/9$, $\exists\, i \in T$ with $\mu(i) \leq \frac{1}{n}$

Power of Conditional Samples

Theorem (C-Fischer-Matsliah-Goldhirsh (SICOMP 2016))

For testing if a distribution is uniform one needs only poly$(1/\epsilon)$ number of conditional samples.

Power of Conditional Samples

Theorem (C-Fischer-Matsliah-Goldhirsh (SICOMP 2016))

For testing if a distribution is uniform one needs only $poly(1/\epsilon)$ *number of conditional samples.*

Theorem (Canonne-Ron-Servedio (SICOMP 2015))

For testing if a distribution is uniform one needs only $\tilde{\theta}(1/\epsilon^2)$ *number of conditional samples.*

Power of Conditional Samples

Theorem (C-Fischer-Matsliah-Goldhirsh (SICOMP 2016))

For testing if a distribution is uniform one needs only $poly(1/\epsilon)$ *number of conditional samples.*

Theorem (Canonne-Ron-Servedio (SICOMP 2015))

For testing if a distribution is uniform one needs only $\tilde{\theta}(1/\epsilon^2)$ *number of conditional samples.*

Theorem (C-Fischer-Matsliah-Goldhirsh (SICOMP 2016))

If $\mathcal{P}$ *is any label invariant property then testing whether a distribution satisfy the property* $\mathcal{P}$ *requires* $poly(\log n)$ *number of condtional samples.*

Application to Real Life Problems

The main challenge for using conditional sampling in real life is:
Is conditional sampling implementable?

Application to Real Life Problems

The main challenge for using conditional sampling in real life is:
Is conditional sampling implementable?

Recall the random satisfying assignment problem:

1. Given a CNF formula producing a satisfying assignments uniformly at random from the set of all satisfying assignments is a crucial problem in the SAT-solver community.
   How to check if the output of a "claimed algorithm" is according to the uniform distribution?

Application to Real Life Problems

The main challenge for using conditional sampling in real life is:
Is conditional sampling implementable?

Recall the random satisfying assignment problem:

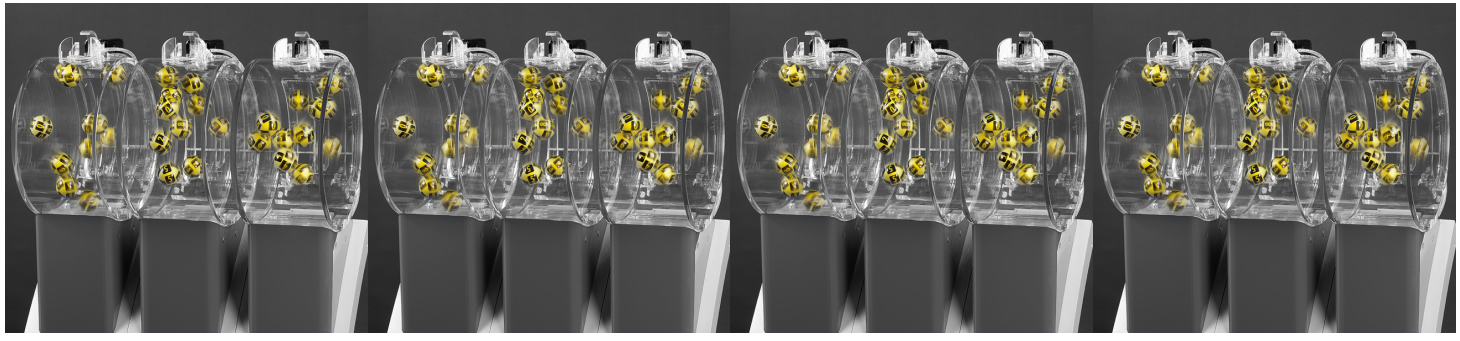
1. Given a CNF formula producing a satisfying assignments uniformly at random from the set of all satisfying assignments is a crucial problem in the SAT-solver community.
   How to check if the output of a "claimed algorithm" is according to the uniform distribution?

Theorem (C-Meel (AAAI 2019))

Using conditional sampling we can design a "practical" algorithm that given black box access to an algorithm can test if the algorithm indeed performs the task properly.

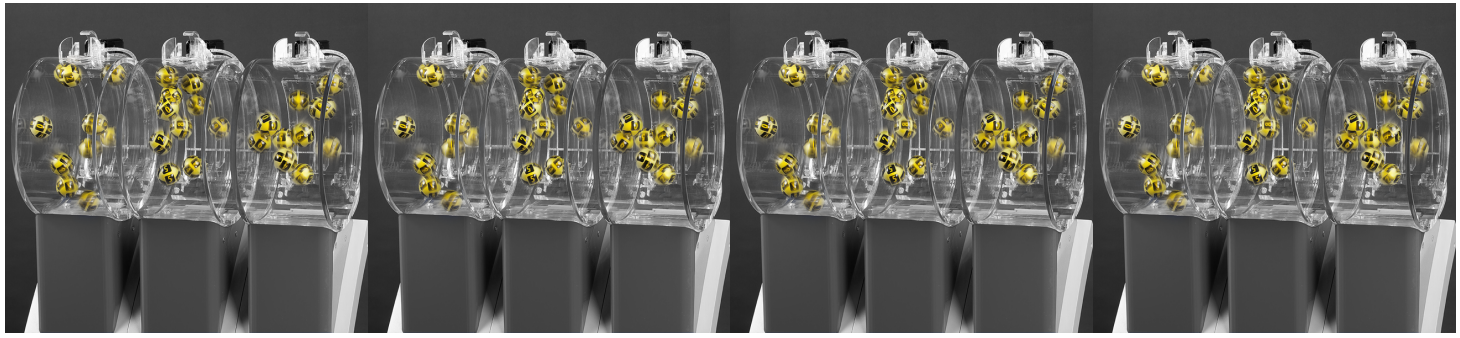
How about the Lottery Machine Problem

2. Checking if a random number generator is correct.



How about the Lottery Machine Problem

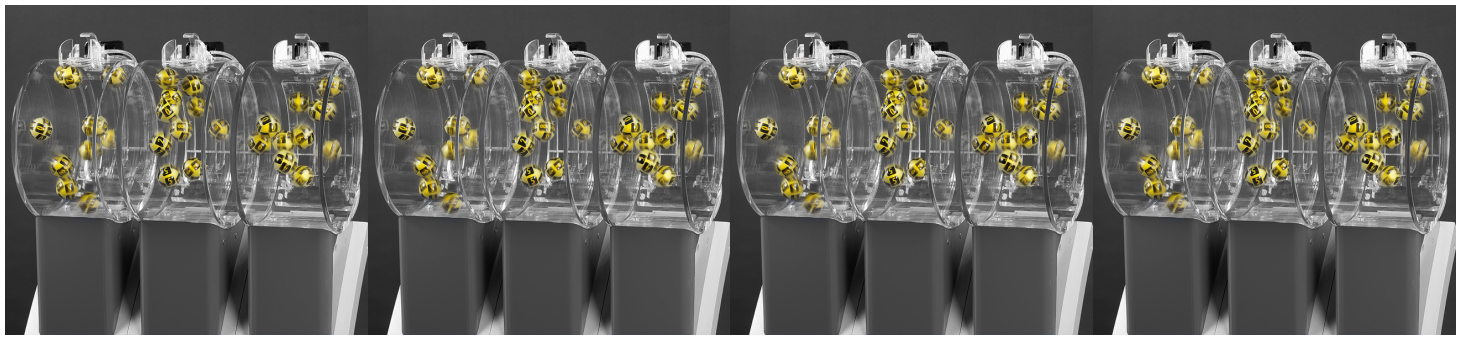
2. Checking if a random number generator is correct.



- In Lottery Machine we can easily draw conditional samples when the set S is of the form $S_1 \times S_2 \times \ldots$, where the $S_i \subseteq \{0, 1, \ldots, 9\}$.

How about the Lottery Machine Problem

2. Checking if a random number generator is correct.



- In Lottery Machine we can easily draw conditional samples when the set S is of the form $S_1 \times S_2 \times \ldots$, where the $S_i \subseteq \{0, 1, \ldots, 9\}$.
- But recall that for our algorithm for testing uniformity the set S can be any two elements in the domain and not necessarily of the special structure for which we can execute the conditional samples.

Conditional Sampling on Structured Domain

Theorem (Bhattacharyya-C (ToCT 2018))

When the domain is of the form $D_1 \times D_2 \times \cdots \times D_m$ and conditional sample can be drawn for sets of form $S_1 \times \cdots \times S_m$, where $S_i \subseteq D_i$, then $\Omega(m)$ number of conditional samples are necessary and $O(m^2)$ number of samples are sufficient.

Further Works related to Conditional Testing

Many different extension and directions are being investigated. For example:

- *Quantum Conditional Sampling* (Sardharwalla-Strelchuk-Jozsa (QIC 2016)).
- *Big data* (Canonne-Rubinfeld (ICALP 2014))
- *Learning using Conditional Sampling* (Aliakbarpour-Blais-Rubinfeld (COLT 2016))
- *New Computational Model* (Gouleakis-Tzamos-Zampetakis (SODA 2018))

A good survey for this area is
"A Chasm Between Identity and Equivalence Testing with Conditional Queries" by Jayadev Acharya, Clément L. Canonne and Gautam Kamath.

Latest set of works related to Distribution Testing

- New proofs of old theorems
- Tolerant Testers
- Testing with Noise
- Learning of Distributions
- Different distance parameters
- Connection to PCPP

A nice collection of talks/surveys on this topic can be found in the webpage of "Frontiers in Distribution Testing" workshop (held at FOCS 2017). The link can be found from the webpage of Clement Canonne.

Future Direction

Future Direction

- One of the most important problem of the present and future is testing if an algorithm is correct. Property testing has a big role to play.

Future Direction

- One of the most important problem of the present and future is testing if an algorithm is correct. Property testing has a big role to play.
- Major challenge: To model the problems properly.

Future Direction

- One of the most important problem of the present and future is testing if an algorithm is correct. Property testing has a big role to play.
- Major challenge: To model the problems properly.
- Interesting progress has been made in recent times on applying property testing successfully to real life problems.

Future Direction

- One of the most important problem of the present and future is testing if an algorithm is correct. Property testing has a big role to play.
- Major challenge: To model the problems properly.
- Interesting progress has been made in recent times on applying property testing successfully to real life problems.
- Lots of interesting and challenging theoretical problems arise.