Testing of Index-Invariant Properties in the Huge Object Model

Arijit Ghosh

Indian Statistical Institute

Joint work with

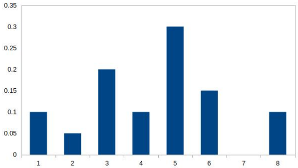
Sourav Chakraborty (Indian Statistical Institute)
Eldar Fischer (Technion)
Gopinath Mishra (University of Warwick)
Sayantan Sen (National University of Singapore)

Contents

Distribution Testing

Definition (Probability Distribution)

A probability distribution D over a universe $\{0,1\}^n$ is a non-negative function $D : \{0,1\}^n \to [0,1]$ such that $\sum_{\mathbf{x} \in \{0,1\}^n} D(\mathbf{x}) = 1$.



Definition (Distribution Property)

A distribution property $\mathcal{P}$ is a collection of distributions over $\{0,1\}^n$.

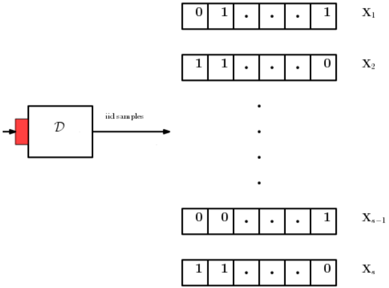
How are the distributions given?

Huge Object Model

- Consider distributions defined over $\{0,1\}^n$.
- For large n, even reading a few samples is infeasible.
- To address this, Goldreich and Ron [ITCS 2021] defined the *huge object model*.
- Samples may only be queried in a few places.
- Goal is to minimize sample and query complexities.

Sampling & Query Model

- Take s iid samples $\{X_1, \ldots, X_s\}$ from D.
- At each step, query some index i, $i \in [n]$ from some X_j, $j \in [s]$.



Motivation

- Testing properties of high dimensional distribution has several applications. For example clusterability testing has applications to computer vision.
- Understanding the properties of CNF-samplers is another important problem with wide applications.
- Testing properties of the distribution over the satisfying assignments of the CNF-formula such as uniformity, or high entropy.
- When a formula's size is large, reading the full assignment of the variables is very costly in practice.
- Ideally we would want to read the input only in few places.

Property Testing in Huge Object Model

Given a property $\mathcal{P}$, design an algorithm $\mathcal{A}$ such that

Input: A distribution D accessible via iid samples and queries to samples, and two parameters ε_1 and ε_2 with $0 \leq \varepsilon_1 < \varepsilon_2 \leq 1$.

Output: With probability at least $\frac{2}{3}$, output:

- Yes if D is ε_1-close to $\mathcal{P}$.
- No if D is ε_2-far from $\mathcal{P}$.

- D is ε_1-close to $\mathcal{P}$ if $\min_{D' \in \mathcal{P}} d_{EM}(D, D') \leq \varepsilon_1$.

Earth Mover Distance (EMD)

Let D_1 and D_2 be two probability distributions over $\{0,1\}^n$. The EMD between D_1 and D_2 is denoted by $d_{EM}(D_1, D_2)$, and is defined as the solution to the following linear program:

$$\textbf{Minimize} \sum_{\mathbf{X},\mathbf{Y} \in \{0,1\}^n} f_{\mathbf{XY}} d_H(\mathbf{X}, \mathbf{Y})$$

$$\textbf{Subject to} \sum_{\mathbf{Y} \in \{0,1\}^n} f_{\mathbf{XY}} = D_1(\mathbf{X}) \ \forall \mathbf{Y} \in \{0,1\}^n, \quad \sum_{\mathbf{X} \in \{0,1\}^n} f_{\mathbf{XY}} = D_2(\mathbf{Y}) \ \forall \mathbf{X} \in \{0,1\}^n$$

$$0 \leq f_{\mathbf{XY}} \leq 1, \forall \mathbf{X}, \mathbf{Y} \in \{0,1\}^n$$

Index-Invariant Property

Definition (Index-Invariant Distribution Property)

A property $\mathcal{P}$ is index-invariant if for all $D \in \mathcal{P}$ and all permutation $\sigma : [n] \to [n]$, $D_\sigma \in \mathcal{P}$, where

$$D_\sigma(w_{\sigma(1)}, \ldots, w_{\sigma(n)}) = D(w_1, \ldots, w_n) \quad \forall (w_1, \ldots, w_n) \in \{0,1\}^n$$

Index-Invariant Property

Definition (Index-Invariant Distribution Property)

A property $\mathcal{P}$ is index-invariant if for all $D \in \mathcal{P}$ and all permutation $\sigma : [n] \to [n]$, $D_\sigma \in \mathcal{P}$, where

$$D_\sigma(w_{\sigma(1)}, \ldots, w_{\sigma(n)}) = D(w_1, \ldots, w_n) \quad \forall (w_1, \ldots, w_n) \in \{0,1\}^n$$

- Property $\textsc{Monotone}$: $D \in \textsc{Monotone}$ property if

$$\mathbf{X} \preceq \mathbf{Y} \text{ implies } D(\mathbf{X}) \leq D(\mathbf{Y}), \text{ for any } \mathbf{X}, \mathbf{Y} \in \{0,1\}^n,$$

- Property $\textsc{Low-VC-dimension}$: $D \in \textsc{Low-VC-dimension}$ if the support of D has VC-dimension at most d.

Index-Invariant Property

Definition (Index-Invariant Distribution Property)

A property $\mathcal{P}$ is index-invariant if for all $D \in \mathcal{P}$ and all permutation $\sigma : [n] \to [n]$, $D_\sigma \in \mathcal{P}$, where

$$D_\sigma(w_{\sigma(1)}, \ldots, w_{\sigma(n)}) = D(w_1, \ldots, w_n) \quad \forall (w_1, \ldots, w_n) \in \{0,1\}^n$$

- Property $\textsc{Monotone}$: $D \in \textsc{Monotone}$ property if

$$\mathbf{X} \preceq \mathbf{Y} \text{ implies } D(\mathbf{X}) \leq D(\mathbf{Y}), \text{ for any } \mathbf{X}, \mathbf{Y} \in \{0,1\}^n,$$

- Property $\textsc{Low-VC-dimension}$: $D \in \textsc{Low-VC-dimension}$ if the support of D has VC-dimension at most d.

Similarly, we can define non-index-invariant properties.

Index-Invariant Property

Definition (Index-Invariant Distribution Property)

A property $\mathcal{P}$ is index-invariant if for all $D \in \mathcal{P}$ and all permutation $\sigma : [n] \to [n]$, $D_\sigma \in \mathcal{P}$, where

$$D_\sigma(w_{\sigma(1)}, \ldots, w_{\sigma(n)}) = D(w_1, \ldots, w_n) \quad \forall (w_1, \ldots, w_n) \in \{0,1\}^n$$

- Property $\textsc{Monotone}$: $D \in \textsc{Monotone}$ property if

$$\mathbf{X} \preceq \mathbf{Y} \text{ implies } D(\mathbf{X}) \leq D(\mathbf{Y}), \text{ for any } \mathbf{X}, \mathbf{Y} \in \{0,1\}^n,$$

- Property $\textsc{Low-VC-dimension}$: $D \in \textsc{Low-VC-dimension}$ if the support of D has VC-dimension at most d.

Similarly, we can define non-index-invariant properties.

- Identity with a fixed distribution.

Not the same as Label-invariant properties that consider all permutations $\tau : \{0,1\}^n \to \{0,1\}^n$!

Contents

Testing via Learning

If we can learn D, we can also test for any property $\mathcal{P}$.

Definition (Learning a Distribution)

Given sample and query accesses to an unknown distribution D over $\{0,1\}^n$, and a parameter $\varepsilon \in (0,1)$, construct a distribution $\widetilde{D}$ such that $d_{EM}(D, \widetilde{D}) \leq \varepsilon$.

Goal is to minimize query complexity.

Theorem (Folklore)

For any distribution D over $\{0,1\}^n$, $\widetilde{\mathcal{O}}(2^n)$ queries are sufficient to construct $\widetilde{D}$.

Can we learn D with better query complexity ?

Our Result: Learning Clusterable distributions

Definition (Clusterable distribution)

A distribution D over $\{0,1\}^n$ is called *(ζ, δ, r)-clusterable* if there is a partition $\mathcal{C}_0, \ldots, \mathcal{C}_s$ of $\{0,1\}^n$ such that $D(\mathcal{C}_0) \leq \zeta$, $s \leq r$, and for every $1 \leq i \leq s$, $d_H(\mathbf{U}, \mathbf{V}) \leq \delta$ for any $\mathbf{U}, \mathbf{V} \in \mathcal{C}_i$.

Our Result: Learning Clusterable distributions

Definition (Clusterable distribution)

A distribution D over $\{0,1\}^n$ is called *(ζ, δ, r)-clusterable* if there is a partition $\mathcal{C}_0, \ldots, \mathcal{C}_s$ of $\{0,1\}^n$ such that $D(\mathcal{C}_0) \leq \zeta$, $s \leq r$, and for every $1 \leq i \leq s$, $d_H(\mathbf{U}, \mathbf{V}) \leq \delta$ for any $\mathbf{U}, \mathbf{V} \in \mathcal{C}_i$.

Clusterable distributions can be learnt easily.

Theorem

Clusterability $\Rightarrow$ Distribution Learning with Constant Queries.

Overview of Learning Algorithm

- Take two sets of samples $\mathcal{S} = \{\mathbf{X}_1, \dots, \mathbf{X}_{t_1}\}$ and $\mathcal{T} = \{\mathbf{Y}_1, \dots, \mathbf{Y}_{t_2}\}$ from D. Also sample a set of indices $R \subseteq [n]$ u.a.r.
- Project $\mathcal{S}$ and $\mathcal{T}$ to R to obtain $\mathcal{S}_R = \{x_1, \dots, x_{t_1}\}$ and $\mathcal{T}_R = \{y_1, \dots, y_{t_2}\}$.
- For every $y_j \in \mathcal{T}_R$, if there exists $x_i \in \mathcal{S}_R$, if $d_H(x_i, y_j) \leq 2\delta$, assign y_j to x_i. If no x_i found, keep it unassigned.
- If total number of unassigned vectors is more than 3ζ, REJECT.

Learning Algorithm Overview Contd.

- Estimate the relative weight w_i of every $x_1, \dots, x_{t_1} \in \mathcal{S}_R$.
- Construct new vectors $\mathbf{Z}_1, \dots, \mathbf{Z}_{t_1}$ such that $d_H(\mathbf{Z}_i, \mathbf{X}_i) \leq \delta/10$.
- Define D': $D'(\mathbf{Z}_i) = w_i$ for every $i \in [t_1]$ and $D'(\ell) = 0$ for $\ell \in \{0,1\}^n \setminus \{\mathbf{Z}_1, \dots, \mathbf{Z}_{t_1}\}$.
- Output D'.

Results from VC Theory

Definition (VC Dimension)

For a set of vectors $V \subseteq \{0,1\}^n$ and a sequence of indices $I = (i_1, \ldots, i_k)$, with $i_j \in [n]$, let $V\mid_I$ denote the set of *projections* of V onto I, i.e.

$$V\mid_I = \{(v_{i_1}, \ldots, v_{i_k}) : (v_1, \ldots, v_n) \in V\}.$$

If $V\mid_I = \{0,1\}^k$, then we say that V *shatters* I. The *VC-dimension* of V is the size of the largest index sequence I that is shattered by V.

Definition (α-packing number)

For a set of vectors $V \subset \{0,1\}^n$ and $\alpha \in (0,1)$, the *α-packing number* $\mathcal{M}(\alpha, V)$ of V is the cardinality of the largest subset $W \subseteq V$ such that $\forall\, \mathbf{X}, \mathbf{Y} \in V$, $d_H(\mathbf{X}, \mathbf{Y}) \geq \alpha$.

Small packing number implies clusterability!

Our Result 2: Learning distributions of bounded VC-dimension

Theorem (Haussler's Packing Theorem)

If the VC-dimension of a set of vectors V is d, then the α-packing number of V is

$$\mathcal{M}(\alpha, V) \leq e(d+1)\left(\frac{2e}{\alpha}\right)^d$$

Our Result 2: Learning distributions of bounded VC-dimension

Theorem (Haussler's Packing Theorem)

If the VC-dimension of a set of vectors V is d, then the α-packing number of V is

$$\mathcal{M}(\alpha, V) \leq e(d+1)\left(\frac{2e}{\alpha}\right)^d$$

Theorem

If the support of D has VC-dimension at most d, then D can be learned using constant number of queries.

- Bounded VC-dimension implies clusterability by Haussler's Packing theorem.
- Call the algorithm for learning clusterable distributions.

Our Result 3: Tester for Bounded VC-dimension Properties

Theorem

Let $\mathcal{P}$ be an index-invariant property such that any $D \in \mathcal{P}$ has VC-dimension at most d. There exists a tester that can distinguish whether $D \in \mathcal{P}$ or D is ε-far from $\mathcal{P}$ using $\mathrm{poly}(\frac{1}{\varepsilon})$ queries.

Our Result 3: Tester for Bounded VC-dimension Properties

Theorem

Let $\mathcal{P}$ be an index-invariant property such that any $D \in \mathcal{P}$ has VC-dimension at most d. There exists a tester that can distinguish whether $D \in \mathcal{P}$ or D is ε-far from $\mathcal{P}$ using $\mathrm{poly}(\frac{1}{\varepsilon})$ queries.

- Follows from the learning result.
- Sample and query complexities of the tester are exponential and doubly-exponential in d respectively.

Are these dependencies necessary?

Our Result 4: Tightness of Bounded VC-dimension Property Tester

Theorem

There exists an index-invariant property $\mathcal{P}_{vc}$ with VC-dimension at most d such that any tester for $\mathcal{P}_{vc}$ requires $2^{\Omega(d)}$ samples and $2^{2^{d-\mathcal{O}(1)}}$ queries.

- Follows from Yao's lemma.
- Take a matrix A of dimension $k \times \ell$ such that $d_H(A_{\cdot,j}, A_{\cdot,t}) \geq 1/3$ with $\ell = 2^{2^{d-10}}$.
- Construct $\mathbf{V}_1, \ldots, \mathbf{V}_k$ where $\mathbf{V}_i$ is the n/ℓ times "blow-up" of the i-th row of A.
- Define $D_A(\mathbf{V}_i) = \frac{1}{k} = \frac{1}{2^d}$ for every $i \in [k]$.

D_{yes}: Choose a permutation $\sigma : [n] \to [n]$ u.a.r and pick D_A^σ.

Tightness of VC Tester Contd.

- Choose $\ell' = 2^{2^{d-20}}$ many column vectors uniformly at random from A to construct the matrix B of dimension $k \times \ell'$.
- Construct $\mathbf{W}_1, \ldots \mathbf{W}_k$ where $\mathbf{W}_i$ is the n/ℓ' times blow-up of the i-th row of B.
- Define $D_B(\mathbf{W}_i) = \frac{1}{k} = \frac{1}{2^d}$ for every $i \in [k]$.

D_{no}: Choose a permutation $\sigma : [n] \to [n]$ u.a.r and pick D_B^σ.

Lemma

$$d_{EM}(D_A^\sigma, D_B^\sigma) \geq 1/8.$$

Contents

Adaptive vs. Non-adaptive Testers

- Adaptive testers can query depending upon the answers to previous queries.
- Non-adaptive testers' queries are oblivious to answers to previous queries.
- Adaptive testers are more powerful.

Adaptive vs. Non-adaptive Testers

- Adaptive testers can query depending upon the answers to previous queries.
- Non-adaptive testers' queries are oblivious to answers to previous queries.
- Adaptive testers are more powerful.
- For dense graphs, there is a tight quadratic gap ([Goldreich-Trevisan'03 & Goldreich-Wigderson'21]).
- For functions and sparse graphs, this gap is exponential ([Ron-Servedio'15, Goldreich-Ron'97]).

What about huge object model?

Our Result 4: Exponential Gap for General Properties

Theorem

Any non-index-invariant property that can be adaptively tested using q queries, can be non-adaptively tested using at most 2^q queries.

Our Result 4: Exponential Gap for General Properties

Theorem

Any non-index-invariant property that can be adaptively tested using q queries, can be non-adaptively tested using at most 2^q queries.

- Overall idea is to follow the decision tree $\mathcal{T}$ of the adaptive tester.
- Since $\mathcal{T}$ has depth q, we can first non-adaptively make all $2^q - 1$ "potential queries" inside $\mathcal{T}$, and then follow the correct root to leaf path.

Our Result 4: Exponential Gap for General Properties

Theorem

Any non-index-invariant property that can be adaptively tested using q queries, can be non-adaptively tested using at most 2^q queries.

- Overall idea is to follow the decision tree $\mathcal{T}$ of the adaptive tester.
- Since $\mathcal{T}$ has depth q, we can first non-adaptively make all $2^q - 1$ "potential queries" inside $\mathcal{T}$, and then follow the correct root to leaf path.

Is this gap is tight?

Tightness of Exponential Gap

$\mathcal{P}_{Pal}$: $\mathbf{S} \in \mathcal{P}_{Pal}$ if $|\mathbf{S}| = n$ & $\mathbf{S} = \mathbf{v}\mathbf{v}^{\mathbf{R}}\mathbf{w}\mathbf{w}^{\mathbf{R}}$, where $\mathbf{v}\mathbf{v}^{\mathbf{R}}$ is over the alphabet $\{0, 1\}$, and $\mathbf{w}\mathbf{w}^{\mathbf{R}}$ is over the alphabet $\{2, 3\}$.

Lemma

$\mathcal{P}_{Pal}$ can be tested using $\mathcal{O}(\log n)$ adaptive queries, but $\Omega(\sqrt{n})$ non-adaptive queries are necessary.

- Upper bound follows from binary search.
- Lower bound follows from result of Alon-Krivelevich-Newman-Szegedy '99.

Exponential Tightness Proof Contd.

$1_{\mathcal{P}_{Pal}}$: For any $D \in 1_{\mathcal{P}_{Pal}}$, $|Supp(D)| = 1$, and for $x \in Supp(D)$, $x \in \mathcal{P}_{Pal}$.

Theorem

$1_{\mathcal{P}_{Pal}}$ can be tested adaptively using $\mathcal{O}(\log n)$ queries, but $\Omega(\sqrt{n})$ queries are necessary for any non-adaptive tester.

- First test if the support size of D is 1 $\Rightarrow \widetilde{\mathcal{O}}(1/\varepsilon)$ queries are enough.
- If the above test passes, test for $\mathcal{P}_{Pal}$.
- Lower bound follows from testing $\mathcal{P}_{Pal}$.

Our Result 5: Quadratic Gap for Index-Invariant Properties

Theorem

Any index-invariant property that can be adaptively tested using q queries, can be non-adaptively tested using at most q^2 queries.

Our Result 5: Quadratic Gap for Index-Invariant Properties

Theorem

Any index-invariant property that can be adaptively tested using q queries, can be non-adaptively tested using at most q^2 queries.

- Consider an adaptive tester $\mathcal{A}$ with sample complexity s and query complexity q.
- Simulate a *semi-adaptive* tester $\mathcal{A}'$ that queries q indices from each of the s samples.
- Apply a uniformly random permutation σ over $[n]$ and run $\mathcal{A}'$ over D_σ. Sample Complexity s and Query Complexity qs.

Our Result 5: Quadratic Gap for Index-Invariant Properties

Theorem

Any index-invariant property that can be adaptively tested using q queries, can be non-adaptively tested using at most q^2 queries.

- Consider an adaptive tester $\mathcal{A}$ with sample complexity s and query complexity q.
- Simulate a *semi-adaptive* tester $\mathcal{A}'$ that queries q indices from each of the s samples.
- Apply a uniformly random permutation σ over $[n]$ and run $\mathcal{A}'$ over D_σ. Sample Complexity s and Query Complexity qs.

Is this gap is tight?

Tightness of Quadratic Gap

Theorem

There exists an index-invariant property $\mathcal{P}_{\mathrm{Gap}}$ that can be tested adaptively using $\widetilde{\mathcal{O}}(n)$ queries, but requires $\widetilde{\Omega}(n^2)$ non-adaptive queries.

Lemma (Valiant-Valiant'11)

Given an unknown distribution D over $[2n]$, accessed via iid samples and a parameter $\varepsilon \in (0, 1/8)$, to distinguish whether D has support size at most n or D has at least $(1+\varepsilon)n$ elements in the support, $\Theta(\frac{n}{\log n})$ samples from D are necessary and sufficient.

Encode hard distributions from Valiant-Valiant's result using a secret sharing code.

Contents

Conclusion

- This is a very recent model with lots of potential applications.
- We proved that distributions whose support has bounded VC-dimension can be learned in constant number queries.
- For index-invariant properties, there is a tight quadratic gap between adaptive vs. non-adaptive testers.
- This is in contrast with a tight exponential gap for general properties.
- Recently Adar-Fischer [AF'23] studied various notions of adaptivity in this model.
- It would be interesting to see new notions!

Conclusion

- This is a very recent model with lots of potential applications.
- We proved that distributions whose support has bounded VC-dimension can be learned in constant number queries.
- For index-invariant properties, there is a tight quadratic gap between adaptive vs. non-adaptive testers.
- This is in contrast with a tight exponential gap for general properties.
- Recently Adar-Fischer [AF'23] studied various notions of adaptivity in this model.
- It would be interesting to see new notions!

THANK YOU!!