Testing shape restriction properties of probability distributions in a unified way

by

Themistoklis Gouleakis

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of
Master of Science in Computer Science and Engineering
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June 2015

© Massachusetts Institute of Technology 2015. All rights reserved.

Signature redacted
Author
Department of Electrical Engineering and Computer Science
May 20, 2015

Signature redacted
Certified by
Ronitt Rubinfeld
Professor
Thesis Supervisor

Signature redacted
Accepted by
Professor Leslie A. Kolodziej
Chair, Department Committee on Graduate Theses
Testing shape restriction properties of probability distributions in
a unified way

by

Themistoklis Gouleakis

Submitted to the Department of Electrical Engineering and Computer Science
on May 20, 2015, in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science and Engineering

Abstract

We study the question of testing structured properties of discrete distributions. Specifically, given sample access to an arbitrary distribution $D$ over $[n]$ and a property $P$, the goal is to distinguish between $D \in P$ and $\ell_1(D, P) > \varepsilon$. Building on a result of [9], we develop a general algorithm for this question, which applies to a large range of “shape-constrained” properties, including monotone, log-concave, $t$-modal and Poisson Binomial distributions. Our generic property tester works for properties that exclusively contain distributions which can be well approximated by $L$-histograms for a small (usually logarithmic in the domain size) value of $L$. The sample complexity of this generic approach is $\tilde{O}(L \cdot \sqrt{n})$. Moreover, for all cases considered, our algorithm has near-optimal sample complexity. Finally, we also describe a generic method to prove lower bounds for this problem, and use it to derive strong converses to our algorithmic results. More specifically, we use the following reduction technique: we compose the property tester for a class $C$ of distributions with an agnostic learner for that same class to get a tester for subset $C_{\text{HARD}} \subseteq C$ for which a lower bound is known.

Thesis Supervisor: Ronitt Rubinfeld
Title: Professor
Acknowledgments

I would like to primarily thank my advisor Ronitt Rubinfeld for introducing me to the field of property testing and constantly giving me new ideas for problems and very helpful advice throughout these years.

I would also like to thank my undergrad advisor Dimitris Achioptas for the close collaboration and guidance in my first research steps.

Many thanks to all my research collaborators: Maryam Aliakbarpour, Amartya Shankha Biswas, Clement Cannone, Ilias Diakonikolas, John Peebles, Anak Yodpinyanee and all my other members of the MIT theory group for creating a very friendly environment. To mention a few, I would like to thank my officemates and my friends Arturs Backurs, Cheng Chen, Gautam Kamath, Madalina Persu, Ilya Razenshteyn, Adam Sealfon, Katerina Sotiraki, Christos Tzamos, Manolis Zampetakis for their help and support at various times.

Lastly, I would like to thank Paris Kanellakis’ family for supporting me with a fellowship in the first year of my studies.
Contents

1 Introduction 13
   1.1 Property testing ........................................... 13
   1.2 Motivation and related work ............................... 20
       1.2.1 Testing properties of probability distributions in a unified way. .... 20

2 Preliminaries 23
   2.1 Definitions .............................................. 23
       2.1.1 Testing and learning .................................. 23
       2.1.2 Properties of interest ............................... 26

3 Testing distributions in a unified way 29
   3.1 Structural results ...................................... 30
   3.2 Computing the distances .................................. 34
   3.3 Algorithm ............................................... 36
       3.3.1 The algorithm ....................................... 36
       3.3.2 Proof of Theorem 3.3.2 ............................ 38
       3.3.3 Sample complexity. ................................. 38
       3.3.4 Correctness. ....................................... 38
   3.4 Considering the effective support size only ............ 42
       3.4.1 Proof of Theorem 3.4.4 ............................ 44
       3.4.2 Application: Testing Poisson Binomial Distributions ......... 45
   3.5 Lower bounds ............................................. 48
A Summary of results

B Ommited proofs

B.1 Proof of Lemma 1.1.7 ......................................................... 55
B.2 Proof of Theorem 3.1.7 ......................................................... 58
List of Figures
List of Tables

1.1 Summary of distribution testing results. ............................... 20

A.1 Summary of our results. ..................................................... 53
Chapter 1

Introduction

1.1 Property testing

Property testing is a framework where one has to come up with an algorithm that solves relaxed decision problem in sublinear time. Namely, it has to distinguish between a “YES’ instance and a “NO’ instance whose distance from any “YES’ instance is at least $\varepsilon$ for some distance measure on the set of possible instances. The hope is that this relaxed notion will in some cases enable us to solve the problem in time sublinear in the length of the input. This seems quite surprising since it implies that the algorithm will only look at an infinitesimal fraction of the input. As we are going to see later, randomness (i.e. looking at random places of the input) will always play a significant role.

In particular, for the special case of testing properties of probability distributions, the problem can be phrased as follows: given a class (property) of distributions $\mathcal{P}$ and sample access to an arbitrary distribution $D$, one must distinguish between the case that (a) $D \in \mathcal{P}$, versus (b) $\|D - D’\|_1 > \varepsilon$ for all $D’ \in \mathcal{P}$ (i.e., $D$ is either in the class, or far from it). The algorithm has access to either samples from the distribution, or is able to query a point in the probability density function (pdf) or the cumulative distribution function (CDF) of the input distribution. We will call sample (query) complexity of such an algorithm the number of samples (resp. queries) needed in order for it to satisfy the aforementioned guarantees.
Note that, the size of the input here is considered to be the size of the domain of the input distribution. We are only interested in sublinear sample or query complexity since otherwise the algorithm could just learn the distribution and and computationally check if it belongs to the class we are interested in (i.e as an ordinary decision problem).

**An example: Testing uniformity**

As a motivating example for testing properties of distributions, we will consider the problem of testing if a distribution is **uniform**. That is, we want to find an algorithm that will be taking a sublinear number of samples and will have the following performance guarantees:

- if \( p \) is uniform, it outputs ACCEPT with probability at least \( \frac{2}{3} \);
- if \( d_{TV}(p, U) \geq \varepsilon \), it outputs REJECT with probability at least \( \frac{2}{3} \).

At this point we observe that there exists a statistic for which the uniform distribution is extremal. Namely, given \( s \) samples, the uniform distribution has the minimum expected number of collisions between them (i.e pairs of identical samples). Indeed, let \( S = x_1, \ldots, x_s \) be the set of independent samples from \( p \). Let \( E_{ij} \) be the event that \( x_i = x_j \). By linearity of expectation on the indicator random variables \( X_{ij} \) for the events \( E_{ij} \), we have:

\[
\text{coll}(S) = E[\# \text{collisions}] = \sum_{\{i,j\}} \Pr[x_i = x_j] = \sum_{\{i,j\} \in S^2} \sum_{k=1}^{n} p_k^2 = \left(\frac{s}{2}\right) \cdot \|p\|_2^2
\]

Since \( \sum_{k=1}^{s} p_k = 1 \), the expected number of collisions is minimized when the distribution is uniform, in which case we have: \( \text{coll}(S) = \frac{s}{n} \).

From the above we conclude that the collision probability (i.e the expected fraction of collisions among the sample pairs) is equal to the \( l_2 \) norm of the distribution \( p \). Another useful observation is that we can estimate the \( l_2 \) distance of \( p \) from the uniform distribution using the \( l_2 \) norm of \( p \) as follows:

---

1Note that this type algorithm is called a property tester as we define more generally in section Section 2.1.1
\[ \|p - U\|_2^2 = \sum_{i \in [n]} (p_i - \frac{1}{n^2}) = \sum_{i \in [n]} p_i^2 - \frac{2}{n} \sum_{i \in [n]} p_i + \sum_{i \in [n]} \frac{1}{n^2} = \|p\|_2^2 - \frac{1}{n} \]

So, it is natural to use this statistic in order to decide whether to test whether a distribution is uniform or far from it.

The following lemma shows that the observed collision frequency is also concentrated around its mean:

**Lemma 1.1.1** ([9]). Let \( I \) be \( n \) interval and \( q \) be the conditional distribution of \( p \) on \( I \). Then,

\[
\left( \|q\|_2^2 - \frac{\varepsilon^2}{32|I|} \right) \leq \frac{\text{coll}(S_I)}{|S_I|} \leq \left( \|q\|_2^2 + \frac{\varepsilon^2}{32|I|} \right)
\]

with probability at least \( 1 - O(\log^{-3} n) \) provided that \(|S_I| = \Omega(\varepsilon^{-4} \sqrt{|I| \log \log n})\).

By setting \( I = [n] \), we get that \( \frac{\text{coll}(S_I)}{|S_I|} \) is with high probability an \((\pm \frac{\varepsilon^2}{32n})\)-additive estimate of \( \|p\|_2^2 \) and thus an \((\pm \frac{\varepsilon^2}{32n})\)-additive estimate of \( \|p - U\|_2^2 \).

We are now ready to check if the algorithm that accepts if and only if the estimate for \( \|p - U\|_2 \) is at most \( \varepsilon \) has the desired properties:

- If \( \|p - u\|_1 = 0 \Rightarrow \|p - u\|_2 = 0 \), then with probability at least 2/3 we will have: \( \frac{\text{coll}(S_I)}{|S_I|} < \frac{1}{n} + \frac{\varepsilon^2}{32n} \). So the algorithm returns ACCEPT. This property of the algorithm is called completeness.

- If \( \|p - u\|_1 > \varepsilon \Rightarrow \|p - u\|_2 > \frac{\varepsilon}{\sqrt{n}} \Rightarrow \|p - u\|_2 > \frac{\varepsilon^2}{n} \), then with probability at least 2/3 we will have: \( \frac{\text{coll}(S_I)}{|S_I|} > \frac{1}{n} + \frac{\varepsilon^2}{32n} \). So the algorithm returns REJECT. This property of the algorithm is called soundness.

In fact, the algorithm has even stronger guarantees in the sense that it will also return ACCEPT when the distribution is sufficiently close to being monotone. This is actually the case whenever a distance approximation algorithm for a given property exists as stated below:

**Theorem 1.1.2.** If there exists an algorithm that estimates the distance to a property \( \mathcal{P} \) up to additive error \( \varepsilon \), then there exists a algorithm that is able to distinguish between \( d(p, \mathcal{P}) < \varepsilon' \) and \( d(p, \mathcal{P}) > \varepsilon' + \varepsilon \).
This algorithm is called a tolerant tester for \( P \) and will be defined formally in Section 2.1.1.

**Lower bound** We now turn to showing that the algorithm described above is almost optimal in its dependence on \( n \) by showing an \( \Omega(\sqrt{n}) \) lower bound.\(^2\) We observe that the property of uniformity that we are interested in falls under the following definition:

**Definition 1.1.3.** A property \( P \) is called symmetric if it is closed under permutations of the domain. More formally, let \( \pi: [n] \rightarrow [n] \) be a permutation on \( n \) elements. Then \((p_1, p_2, \ldots, p_n) \in P \Rightarrow (p_{\pi(1)}, p_{\pi(2)}, \ldots, p_{\pi(n)}) \in P.\)

When we want to test for a symmetric property, then the specific labels of domain elements that we see in the sample are not important. The following definition is exactly the statistic of the set of samples that contains all the information we might need from it.

**Definition 1.1.4.** Given a multiset of samples \( S \) of size \( s \) from a distribution \( p \), we define the fingerprint of the sample as follows: Let \( C_i \), for \( 0 \leq i \leq s \) be the number of elements that appear exactly \( i \) times in \( S \). The multiset of values \( \{C_i\}_{0 \leq i \leq s} \) is called the fingerprint of the set \( S \).

**Lemma 1.1.5.** Let \( P \) be a symmetric property. If there exists a property tester for \( P \) the uses a multiset of samples \( S \), then there exists a property tester that gets as input only the fingerprint of \( S \).

**Proof.** Given that there is a property testing algorithm \( A \) that tests a symmetric property \( P \), we are going to build an algorithm that only takes the the fingerprint of the sample \( s \) as the input and also, tests for \( P \) with the same guarantees.

Our algorithm will read the input fingerprint, construct an "artificial" sample from it and feed it to the algorithm \( A \). More specifically, the algorithm \( A' \) is defined as follows:

- Read the fingerprint \( C_0, C_1, \ldots, C_s \)
- \( \text{cursor} = 1; \)
- **For** \( i = 1 \) to \( s \) **do**

\(^2\)Note that the logarithmic factor in lemma 1.1.1 goes away if we only ask for constant success probability.
- read $C_i$ and put the next $C_i$ elements (starting from the "cursor") in the sample $S$, $i$ times each element (in increasing order).
- cursor = cursor + $C_i$

- Run the algorithm $A$ using the "artificial" sample $S'$ and give the same output.

Example: if the fingerprint is $(C_0, C_1, ..., C_s) = (2, 3, 2, 1, 0, 0, 0, 0, 0, 0, 0, 0)$, $D = \{1..8\}$ then the algorithm creates the artificial sample: $S' = (1, 2, 3, 4, 4, 5, 5, 6, 6, 6)$.

We now have to argue that whatever the actual sample $S$ might have been, if we feed the algorithm $A$ with $S'$ instead, we will get the same output distribution. Indeed, since $P$ is symmetric, the fact that the actual labels of the elements might have been different does not affect the output distribution of the algorithm (i.e. if in fact number '7' had appeared 3 times in our example instead of number '6', this would not alter the output of the algorithm). Also, due to the fact that we take independent samples from the distribution, any permutation in the order in which the samples are taken has the same probability as all the rest of the permutations. So, the particular order of the samples does not give the algorithm any extra information about the distribution. This is why the output distribution of $A$ is not affected if we feed it with the "artificial" sample $S'$ instead of the original sample $S$.

□

**Theorem 1.1.6.** Any property tester for uniformity has sample complexity $\Omega(\sqrt{n})$.

**Proof.** Since uniformity is a symmetric property, the above lemma applies. We are now ready to show that for every possible $\epsilon$, there exists a distribution $p$ such that $\|p - U_D\|_1 > \epsilon$ which is indistinguishable from $U_D$ with probability $1 - o(1)$ if we use $o(\sqrt{n})$ samples. For that, it suffices to show that both $p$ and $U_D$ will produce exactly the same fingerprint with probability $1 - o(1)$. If that happens, then algorithm will have to accept PASS on that fingerprint with probability $\alpha$ and FAIL with probability $1 - \alpha$ for some $\alpha \in [0, 1]$. Since this happens with probability $1 - o(1)$, the overall probability of "passing" distribution $U_D$ will be $(1 - o(1)) \alpha + o(1) = \alpha + o(1)$ and the probability of failing $p$ will be $(1 - o(1)) \cdot (1 - \alpha) + o(1) = 1 - \alpha + o(1)$. The algorithm is required to PASS $U_D$ with probability at least $2/3$ and FAIL $p$ with probability at least $2/3$, which means that $\alpha + o(1) + 1 - \alpha + o(1) > 4/3 \Leftrightarrow 1 + o(1) > 4/3$. Contradiction!
So, it remains to show that if we use \( o(\sqrt{n}) \) samples, then both \( p \) and \( U_D \) produce the same fingerprint with probability \( 1 - o(1) \) and this will imply the desired lower bound.

Consider the following family of distributions \( p_c \) (where \( c \in (0, 1] \) is a parameter):

- For a uniformly random subset \( B \subseteq [n] : |B| \leq c \cdot n \): \( p_c(i) = \frac{1}{c \cdot n} \)
- Otherwise, \( p_c(i) = 0 \)

We can now calculate the distance between \( p_c \) and \( U_D \):

\[
\|p_c - U_D\|_1 = c \cdot n \cdot (\frac{1}{c \cdot n} - \frac{1}{n}) + \frac{1}{n} \cdot (1 - c) \cdot n = 2 \cdot (1 - c)
\]

So,

\[
\|p_c - U_D\|_1 > \epsilon \iff 2 \cdot (1 - c) > \epsilon \iff c < 1 - \frac{\epsilon}{2}
\]

So, for any constant \( \epsilon \), we can choose a constant \( c \) which should satisfy the above inequality and makes the distribution \( p_c \) \( \epsilon \)-far from \( U_D \).

Let \( s = o(\sqrt{n}) \) be the number of samples that we take. In order to find the most possible fingerprint for both distributions, we are going to use a birthday paradox argument.

So, we want to estimate the probability of collisions (i.e the probability of getting at least a pair of samples with the same label among the \( s \) samples we take from each distribution).

The probability that we see a particular element at least twice when is sample from \( p_c \) is the following:

\[
\Pr[\text{element '1' appears at least 2 times}] \leq \left( \binom{s}{2} \right) \left( \frac{1}{c \cdot n} \right)^2
\]

So, by union bound, we get that

\[
\Pr[\exists i: \text{ element 'i' appears at least 2 times}] \leq n \cdot \left( \binom{s}{2} \right) \left( \frac{1}{c \cdot n} \right)^2 = \Theta(\frac{s^2}{n}) = o(1)
\]

since \( s = o(\sqrt{n}) \).

So, for every distribution \( p_c, c \in (0, 1] \), there all the \( s = o(\sqrt{n}) \) samples are distinct with probability at least \( 1 - o(1) \). Note that \( U_D = p_1 \), so the same holds for the uniform distribution.
as well. Also, the fact that the subset $B$ is selected uniformly at random ensures that the samples have no bias in favour of some element of $[n]$. Thus, the fingerprint of any function $p_c$ in the family we have described has the fingerprint: $C_0 = n - s, C_1 = s, \forall i > 1 : C_i = 0$ with probability at $1 - o(1)$. This makes any $p_c$ indistinguishable from $p_1 = U_D$, which finishes the proof. $\square$

**Remark:** The gap for the dependence on $\varepsilon$ for this problem was closed by Paninski in [27] where the sample complexity for uniformity was shown to be $\Theta(\frac{\sqrt{n}}{\varepsilon^2})$.

In [19], Diakonikolas et al also used a different approach also used a different approach based on a modified $\chi^2$-test to prove the following slightly more general result:

**Lemma 1.1.7** (Adapted from [19, Theorem 11]). *There exists an algorithm $\text{CHECK-SMALL-} \ell_2$ which, given parameters $\varepsilon, \delta \in (0, 1)$ and $C \cdot \sqrt{|I|/\varepsilon^2 \log(1/\delta)}$ independent samples from a distribution $D$ over $I$ (for some absolute constant $C > 0$), outputs either yes or no, and satisfies the following.*

- If $\|D - U_I\|_2 > \varepsilon/\sqrt{|I|}$, then the algorithm outputs no with probability at least $1 - \delta$;
- If $\|D - U_I\|_2 \leq \varepsilon/2\sqrt{|I|}$, then the algorithm outputs yes with probability at least $1 - \delta$.

**Results on other properties**

Over the last couple decades, a number of properties of distributions that have been examined including identity, closeness, independence, monotonicity and $k$-modality. We will mention here some results about the above properties. The interested reader may also look at [10] for a more detailed survey of the field of distribution property testing.

We will now give the definitions and notable results for those properties:

- **Testing Identity:** Given an explicit description of a distribution $p$ and samples from an unknown distribution $q$, test if $q = p$.

- **Testing Closeness:** Given samples from two unknown distributions $p, q$, test whether $p = q$ or $\|p - q\|_1 > \varepsilon$. 

19
• **Testing Independence**: Given samples from a joint distribution over \([n] \times [m]\), test whether the two random variables are independent or there is no independent joint distribution that is \(\varepsilon\)-close to the input distribution.

• **Testing monotonicity**: Given samples from a distribution \(p\), test if \(p_1 > p_2 > \cdots > p_n\).

• **k-modality**: Given samples from a distribution \(p\), test if \(p\) has at most \(k\) modes.

Some of the known results for the above properties are summarized in the table below:

<table>
<thead>
<tr>
<th>Class</th>
<th>Upperbound</th>
<th>Lowerbound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monotone</td>
<td>(O(\frac{1}{\varepsilon^2})) ([9]), (\tilde{O}(\frac{1}{\varepsilon^2})) ((\text{Corollary 3.3.4}))</td>
<td>(\Omega(\frac{1}{\varepsilon^2})) ([9]), (\Omega(\frac{1}{\varepsilon^2})) ((\text{Corollary 3.5.2}))</td>
</tr>
<tr>
<td>Identity</td>
<td>(O(\frac{1}{\varepsilon^2})) ((\text{Corollary 3.3.4}))</td>
<td>(\Omega(\frac{1}{\varepsilon^2})) ((\text{uniformity lower bound: [27]})</td>
</tr>
<tr>
<td>Closeness</td>
<td>(O(\max(\frac{n^{2/3}}{\varepsilon^2}, \frac{1}{\varepsilon^2})) ((\text{[14]})</td>
<td>(\Omega(\max(\frac{n^{2/3}}{\varepsilon^2}, \frac{1}{\varepsilon^2})) ((\text{[14]})</td>
</tr>
<tr>
<td>(t)-modal</td>
<td>(\tilde{O}(t\varepsilon^2)) ((\text{Corollary 3.3.5}))</td>
<td>(\Omega(\frac{1}{\varepsilon^2})) ((\text{Corollary 3.5.2}))</td>
</tr>
<tr>
<td>Independence</td>
<td>(\tilde{O}(n^{2/3}m^{1/3})\ \text{poly}(\frac{1}{\varepsilon})) ((\text{[7]}))</td>
<td>(\Omega(n^{2/3}m^{1/3})) ((\text{[7]}))</td>
</tr>
</tbody>
</table>

Table 1.1: Summary of distribution testing results.

### 1.2 Motivation and related work

#### 1.2.1 Testing properties of probability distributions in a unified way.

Inferring information about the probability distribution that underlies a data sample is an essential question in Statistics, and one that has ramifications in every field of the natural sciences and quantitative research. In many situations, it is natural to assume that this data exhibits some simple structure because of known properties of the origin of the data, and in fact these assumptions are crucial in making the problem tractable.

As a result, the problem of deciding whether a distribution possesses such a structural property has been widely investigated both in theory and practice, in the context of *shape restricted inference* \([5, 32]\). Here, it is guaranteed or thought that the unknown distribution satisfies a shape constraint, such as having a monotone or log-concave probability density function \([31, 4, 37]\). From a different perspective, a recent line of work in Theoretical Computer
Science, originating from the papers of Batu et al. [8, 7, 21] has also been tackling similar questions in the setting of property testing (see [28, 29, 30] for surveys on this field). This very active area has seen a spate of results and breakthroughs over the past decade, culminating in very efficient (both sample and time-wise) algorithms for a wide range of distribution testing problems [6, 22, 2, 17, 14, 1, 19]. In many cases, this led to a tight characterization of the number of samples required for these tasks as well as the development of new tools and techniques, drawing connections to learning and information theory [33, 34, 35].

In this thesis, we will consider the following property testing problem: given a class (property) of distributions \( \mathcal{P} \) and sample access to an arbitrary distribution \( D \), one must distinguish between the case that (a) \( D \in \mathcal{P} \), versus (b) \( \|D - D'|\| > \varepsilon \) for all \( D' \in \mathcal{P} \) (i.e., \( D \) is either in the class, or far from it). While many of the previous works have focused on the testing of specific properties of distributions or obtained algorithms and lower bounds on a case-by-case basis, an emerging trend in distribution testing is to design general frameworks that can be applied to several distribution property testing problems [36, 19]. This direction, the testing analog of a similar movement in distribution learning [11, 13, 12], aims at abstracting the minimal assumptions that are shared by a large variety of problems, and giving algorithms that can be used for any of these problems.

One goal of this work is to provide a unified framework for the question of testing various properties of probability distributions.

Batu et al. [9] initiated the study of efficient property testers for monotonicity and obtained (nearly) matching upper and lower bounds for this problem; while [1] later considered testing the class of Poisson Binomial Distributions, and settled the complexity of this problem (up to the precise dependence on \( \varepsilon \)). Another body of work by [6], [9] and [17] shows how assumptions on the shape of the distributions can lead to significantly more efficient algorithms. They describe such improvements in the case of identity and closeness testing as well as for entropy estimation, under monotonicity or \( k \)-modality constraints. Specifically, Batu et al. show in [9] how to obtain a \( O(\log^3 n/\varepsilon^3) \)-sample tester for closeness in this setting, in stark contrast to the \( \Omega(n^{2/3}) \) general lower bound. Diakonikolas et al. [17] later gave
\( \hat{O}(\sqrt{\log n}) \) and \( \hat{O}(\log^{2/3} n) \)-sample testing algorithms for testing respectively identity and closeness of monotone distributions, and obtained similar results for k-modal distributions. Except for these two cases of monotonicity and PBDs nothing is known about testing the properties we study.

Moreover, for the specific problem of identity testing (i.e. given the explicit description of a distribution \( D^* \) and sample access to an unknown distribution \( D \), to decide whether \( D \) is equal to \( D^* \) or far from it), a recent result of [19] describes a general algorithm which applies to a large range of shape or structural constraints, and yields optimal identity testers for classes of distributions that satisfy them. We observe that while the question they answer can be cast as a specialized instance of membership testing, our results are incomparable to them, both because of the distinction above (testing with versus testing for structure) and as the structural assumptions they rely on are fundamentally different from ours.

One idea about how to tackle this problem is inspired by Batu et al. [9], which show that monotone distributions can be well-approximated by piecewise constant densities on a suitable partition of the domain; and leverage this fact to reduce monotonicity testing to uniformity testing on each interval of this partition. While their argument is tailored specifically for the setting of monotonicity setting, we will try to generalize this idea to other classes of probability distributions like k-modal, Log-concave, monotone hazard risk (MHR), Poisson binomial distributions (PBDs) etc. We will also consider mixtures of the above distributions.
Chapter 2

Preliminaries

2.1 Definitions

Hereafter, we write $[n]$ for the set $\{1, \ldots, n\}$, and log for the logarithm in base 2. A \textit{probability distribution} over a finite domain $\mathcal{X}$ is a non-negative function $D: \mathcal{X} \rightarrow [0, 1]$ such that $\sum_{x \in \mathcal{X}} D(x) = 1$; we denote by $\mathcal{U}_\mathcal{X}$ the uniform distribution on $\mathcal{X}$. Moreover, given a distribution $D$ over $\mathcal{X}$ and a set $S \subseteq \mathcal{X}$, we write $D(S)$ for the total probability weight $\sum_{x \in S} D(x)$ assigned to $S$ by $D$.

2.1.1 Testing and learning

We first define precisely the notions of \textit{testing, tolerant testing, learning} and \textit{proper learning} of distributions over a domain $[n]$.

\textbf{Definition 2.1.1 (Testing).} Fix any property $\mathcal{P}$ of distributions, and let ORACLE$_D$ be an oracle providing some type of access to $D$. A \textit{q-sample testing algorithm for} $\mathcal{P}$ \textit{is a randomized algorithm} $T$ \textit{which takes as input} $n$, $\varepsilon \in (0, 1]$, \textit{as well as access to ORACLE$_D$}. After making at most $q(\varepsilon, n)$ calls to the oracle, $T$ outputs either ACCEPT or REJECT, such that the following holds:

- if $D \in \mathcal{P}$, $T$ outputs ACCEPT with probability at least $2/3$;
• if \( d_{TV}(D, P) \geq \varepsilon \), \( T \) outputs REJECT with probability at least \( 2/3 \).

We shall also be interested in tolerant testers – roughly, algorithms robust to a relaxation of the first item above:

**Definition 2.1.2** (Tolerant testing). Fix property \( P \) and ORACLE\(_D\) as above. A \( q \)-sample tolerant testing algorithm for \( P \) is a randomized algorithm \( T \) which takes as input \( n \), \( 0 \leq \varepsilon_1 < \varepsilon_2 \leq 1 \), as well as access to ORACLE\(_D\). After making at most \( q(\varepsilon_1, \varepsilon_2, n) \) calls to the oracle, \( T \) outputs either ACCEPT or REJECT, such that the following holds:

- if \( d_{TV}(D, P) \leq \varepsilon_1 \), \( T \) outputs ACCEPT with probability at least \( 2/3 \);
- if \( d_{TV}(D, P) \geq \varepsilon_2 \), \( T \) outputs REJECT with probability at least \( 2/3 \).

Another class of algorithms we consider is that of (proper) learners. We give the precise definition below:

**Definition 2.1.3** (Learning). Let \( C \) be a class of probability distributions and \( D \in C \) be an unknown distribution. Let also \( \mathcal{H} \) be a hypothesis class of distributions. A \( q \)-sample learning algorithm for \( C \) is a randomized algorithm \( \mathcal{L} \) which takes as input \( n \), \( \varepsilon, \delta \in (0, 1) \), as well as access to SAMP\(_D\) and outputs the description of a distribution \( \hat{D} \in \mathcal{H} \) such that with probability at least \( 1 - \delta \) one has \( d_{TV}(D, \hat{D}) \leq \varepsilon \).

If in addition \( \mathcal{H} \subseteq C \), then we say \( \mathcal{L} \) is a proper learning algorithm.

The exact formalization of what learning a probability distribution means has been considered in Kearns et al. [24].

A more general definition of a type of learning algorithms follows:

**Definition 2.1.4** (Agnostic learning). Let \( C \) and \( \mathcal{H} \) be two classes of probability distributions over \( X \). A (semi-)agnostic learner for \( C \) (using hypothesis class \( \mathcal{H} \)) is a learning algorithm \( \mathcal{A} \) which, given sample access to an arbitrary distribution \( D \) and parameter \( \varepsilon \), outputs a hypothesis \( \hat{D} \in \mathcal{H} \) such that, with high probability, \( \hat{D} \) does “as much as well as the best approximation from \( C \):”

\[
d_{TV}(D, \hat{D}) \leq c \cdot \text{OPT}_{C,D} + O(\varepsilon)
\]
where $\text{OPT}_{C,D} \overset{\text{def}}{=} \inf_{D_C \in C} d_{TV}(D_C, D)$ and $c \geq 1$ is some absolute constant (if $c = 1$, the learner is said to be agnostic).

Note that a (semi-)agnostic learner with parameters $c, \varepsilon$ is also a learner with parameter $\varepsilon$. 
2.1.2 Properties of interest

We give here the formal descriptions of the classes of distributions involved in this work. Recall that a distribution $D$ over $[n]$ is monotone (non-increasing) if its probability mass function (pmf) satisfies $D(1) \geq D(2) \geq \ldots D(n)$. A natural generalization of the class $\mathcal{M}$ of monotone distributions is the set of $t$-modal distributions, i.e. distributions whose pmf can go “up and down” or “down and up” up to $t$ times:

**Definition 2.1.5** $(t$-modal$)$. Fix any distribution $D$ over $[n]$, and integer $t$. $D$ is said to have $t$ modes if there exists a sequence $i_0 < \cdots < i_{t+1}$ such that either $(-1)^j D(i_j) < (-1)^j D(i_{j+1})$ for all $0 \leq j \leq t$, or $(-1)^j D(i_j) > (-1)^j D(i_{j+1})$ for all $0 \leq j \leq t$. We call $D$ $t$-modal if it has at most $t$ modes, and write $\mathcal{M}_t$ for the class of all $t$-modal distributions (omitting the dependence on $n$). The particular case of $t = 1$ corresponds to the set $\mathcal{M}_1$ of unimodal distributions.

**Definition 2.1.6** (Log-Concave). A distribution $D$ over $[n]$ is said to be log-concave if it satisfies the following conditions: (i) for any $1 < i < j < k \leq n$ such that $D(i)D(k) > 0$, $D(j) > 0$; and (ii) for all $1 < k < n$, $D(k)^2 \geq D(k - 1)D(k + 1)$. We write $\mathcal{L}$ for the class of all log-concave distributions (omitting the dependence on $n$).

**Definition 2.1.7** (Concave and Convex). A distribution $D$ over $[n]$ is said to be concave if it satisfies the following conditions: (i) for any $1 \leq i < j < k \leq n$ such that $D(i)D(k) > 0$, $D(j) > 0$; and (ii) for all $1 < k < n$ such that $D(k - 1)D(k + 1) > 0$, $2D(k) \geq D(k - 1) + D(k + 1)$; it is convex if the reverse inequality holds in (ii). We write $\mathcal{K}^-$ (resp. $\mathcal{K}^+$) for the class of all concave (resp. convex) distributions (omitting the dependence on $n$).

It is not hard to see that convex and concave distributions are unimodal; moreover, every concave distribution is also log-concave, i.e. $\mathcal{K}^- \subseteq \mathcal{L}$. Note that in both definitions 2.1.6 and 2.1.7, condition (i) is equivalent to enforcing that the distribution be supported on an interval.
**Definition 2.1.8** (Monotone Hazard Rate). A distribution $D$ over $[n]$ is said to have monotonically hazard rate (MHR) if its hazard rate $H(i) \overset{\text{def}}{=} \frac{D(i)}{\sum_{j=1}^{i} D(j)}$ is a non-decreasing function. We write $\mathcal{MHR}$ for the class of all MHR distributions (omitting the dependence on $n$).

It is known that every log-concave distribution is both unimodal and MHR (see e.g. [3, Proposition 10]), and that monotone distributions are MHR. Finally, we recall the definition of the two following classes, which both extend the family of Binomial distributions $\mathcal{BLN}_n$: the first, by removing the need for each of the independent Bernoulli summands to share the same bias parameter.

**Definition 2.1.9.** A random variable $X$ is said to follow a **Poisson Binomial Distribution** (with parameter $n \in \mathbb{N}$) if it can be written as $X = \sum_{k=1}^{n} X_k$, where $X_1, \ldots, X_n$ are independent, non-necessarily identically distributed Bernoulli random variables. We denote by $\mathcal{PBD}_n$ the class of all such Poisson Binomial Distributions.

One can generalize even further, by allowing each random variable of the summation to be integer-valued:

**Definition 2.1.10.** Fix any $k \geq 0$. We say a random variable $X$ is a $k$-Sum of Independent Integer Random Variables ($k$-SIIRV) with parameter $n \in \mathbb{N}$ if it can be written as $X = \sum_{j=1}^{n} X_j$, where $X_1, \ldots, X_n$ are independent, non-necessarily identically distributed random variables taking value in $\{0, 1, \ldots, k-1\}$. We denote by $\mathcal{k-SIIRV}_n$ the class of all such $k$-SIIRVs.
Chapter 3

Testing distributions in a unified way

At the core of our approach is an idea of Batu et al. [9], which show that monotone distributions can be well-approximated by piecewise constant densities on a suitable partition of the domain; and leverage this fact to reduce monotonicity testing to uniformity testing on each interval of this partition. While their argument is tailored specifically for the setting of monotonicity setting, we are able to abstract the key ingredients, and generalize this idea to many classes of probability distributions. In more detail, we provide a generic testing algorithm which applies indifferently to any class of distributions which admit succinct decompositions – that is, which are well-approximated (in a strong $\ell_2$ sense) by piecewise constant densities on a small number of intervals (we hereafter refer to this approximation property as (Succinctness); and extend the notation to apply to any class $C$ of distributions for which all $D \in C$ satisfy this property). Crucially, the algorithm does not care about how these decompositions can be obtained: for the purpose of testing these structural properties we only need to establish their existence.
3.1 Structural results

We start by defining formally the "structural criterion" we shall rely on, before describing the algorithm at the heart of our result in Section 3.3.1. (We note that a modification of this algorithm will be described in Section 3.4, and will allow us to derive Corollary 3.3.6.)

Definition 3.1.1 (Decompositions). Let $\gamma > 0$ and $L = L(\gamma, n) \geq 1$. A class of distributions $C$ on $[n]$ is said to be $(\gamma, L)$-decomposable if for every $D \in C$ there exists $\ell \leq L$ and a partition $I(\gamma, D) = (I_1, \ldots, I_\ell)$ of $[n]$ such that, for all $j \in [\ell]$, one of the following holds:

(i) $D(I_j) \leq \frac{\gamma}{2}$; or

(ii) $\max_{i \in I_j} D(i) \leq (1 + \gamma) \cdot \min_{i \in I_j} D(i)$.

Further, if $I(\gamma, D)$ is dyadic (i.e., each $I_k$ is of the form $[j \cdot 2^t + 1, (j + 1) \cdot 2^t]$ for some integers $i, j$, corresponding to the leaves of a recursive bisection of $[n]$), then $C$ is said to be $(\gamma, L)$-splittable.

Lemma 3.1.2. If $C$ is $(\gamma, L)$-decomposable, then it is $(\gamma, O(L \log n))$-splittable.

Proof. We will begin by proving a claim that for every partition $I = \{I_1, I_2, \ldots, I_L\}$ of the interval $[1, n]$ into $L$ intervals, there exists a refinement of that partition which consists of at most $L \cdot \log n$ dyadic intervals. So, it suffices to prove that every interval $[a, b] \subseteq [1, n]$, can be partitioned in at most $O(\log n)$ dyadic intervals. Indeed, let $\ell$ be the largest integer such that $2^\ell \leq \frac{b-a}{2}$ and let $m$ be the smallest integer such that $m \cdot 2^\ell \geq a$. If follows that $m \cdot 2^\ell \leq a + \frac{b-a}{2} = \frac{a+b}{2}$ and $(m + 1) \cdot 2^\ell \leq b$. So, the interval $I = [m \cdot 2^\ell + 1, (m + 1) \cdot 2^\ell]$ is fully contained in $[a, b]$ and has size at least $\frac{b-a}{4}$.

In the following, we will also use the following fact:

$$m \cdot 2^\ell = m \cdot 2^{\ell - \ell'} \cdot 2^{\ell'} = m' \cdot 2^{\ell'}$$  \hspace{1cm} (3.1)

for every $\ell' \leq \ell$.

Now consider the following procedure: Starting from right (resp. left) side of the interval $I$, we add the largest interval which is adjacent to it and fully contained in $[a, b]$ and recurse...
until we cover the whole interval \([m+1) \cdot 2^d + 1, b]\) (resp. \([a, m \cdot 2^d]\)). Clearly, at the end of this procedure, the whole interval \([a, b]\) is covered by dyadic intervals. It remains to show that the procedure takes \(O(\log n)\) steps. Indeed, using Equation 3.1, we can see that at least half of the remaining left or right interval is covered in each step (except maybe for the first 2 steps where it is at least a quarter). Thus, the procedure will take at most \(2 \log n + 2 = O(\log n)\) steps in total. From the above, we can see that each of the \(L\) intervals of the partition \(\mathcal{I}\) can be covered with \(O(\log n)\) dyadic intervals, which completes the proof of the claim.

In order to complete the proof of the lemma, notice that the two conditions in Theorem 3.1.1 are closed under taking subsets. \(\Box\)

**Extension to mixtures.** Finally, it is immediate to see that if \(D_1, \ldots, D_k\) are \((\gamma, L)\)-decomposable, then any mixture \(\alpha_1 D_1 + \cdots + \alpha_k D_k\) is \((\gamma, kL)\)-decomposable.

In the following, we show that many "natural" classes of distributions are (succinctly) decomposable.

**Theorem 3.1.3 (Monotonicity).** For all \(\gamma > 0\), the class \(\mathcal{M}\) of monotone distributions on \([n]\) is \((\gamma, L)\)-splittable for \(L \equiv O\left(\frac{\log^2 n}{\gamma}\right)\).

Note that this proof can already be found in [9, Theorem 10], interwoven with the analysis of their algorithm. For the sake of being self-contained, we reproduce the structural part of their argument, removing its algorithmic aspects:

**Proof of Theorem 3.1.3.** We define the \(\mathcal{I}\) recursively as follows: \(\mathcal{I}^{(0)} = ([1, n])\), and for \(j \geq 0\) the partition \(\mathcal{I}^{(j+1)}\) is obtained from \(\mathcal{I}^{(j)} = (I^{(j)}_1, \ldots, I^{(j)}_\ell)\) by going over the \(I^{(j)}_i = [a^{(j)}_i, b^{(j)}_i]\) in order, and:

(a) if \(D(I^{(j)}_i) \leq \frac{2}{L}\), then \(I^{(j)}_i\) is added as element of \(\mathcal{I}^{(j+1)}\) ("marked as leaf");

(b) else, if \(D(b^{(j)}_i) \leq (1 + \gamma)D(a^{(j)}_i)\), then \(I^{(j)}_i\) is added as element of \(\mathcal{I}^{(j+1)}\) ("marked as leaves");

(c) otherwise, bisect \(I^{(j)}_i\) in \(I^{(j)}_{L_i} , I^{(j)}_{R_i}\) (with \(|I^{(j)}_{L_i}| = \left\lfloor |I^{(j)}|/2 \right\rfloor\) and add both \(I^{(j)}_{L_i}\) and \(I^{(j)}_{R_i}\) as elements of \(\mathcal{I}^{(j+1)}\).
and repeat until convergence (that is, whenever the last item is not applied for any of the intervals). Clearly, this process is well-defined, and will eventually terminate (as \((\ell_j)_j\) is a non-decreasing sequence of natural numbers, upper bounded by \(n\)). Let \(I = (I_1, \ldots, I_\ell)\) (with \(I_i = [a_i, a_{i+1})\)) be its outcome, so that the \(I_i\)'s are consecutive intervals all satisfying either (a) or (b). As (b) clearly implies (ii), we only need to show that \(\ell \leq L\); for this purpose, we shall leverage as in [9] the fact that \(D\) is monotone to bound the number of recursion steps.

The recursion above defines a complete binary tree (with the leaves being the intervals satisfying (a) or (b), and the internal nodes the other ones). Let \(t\) be the number of recursion steps the process goes through before converging to \(I\) (height of the tree); as mentioned above, we have \(t \leq \log n\) (as we start with an interval of size \(n\), and the length is halved at each step.). Observe further that if at any point an interval \(I_i^{(j)} = [a_i^{(j)}, b_i^{(j)}]\) has \(D(a_i^{(j)}) \leq \frac{\gamma}{nL}\), then it immediately (as well as all the \(I_k^{(j)}\)'s for \(k \geq i\) by monotonicity) satisfies (a) and is no longer split ("becomes a leaf"). So at any \(j \leq t\), the number of intervals \(i_j\) for which neither (a) nor (b) holds must satisfy

\[
1 \geq D(a_1^{(j)}) > (1 + \gamma)D(a_2^{(j)}) > (1 + \gamma)^2D(a_3^{(j)}) > \cdots > (1 + \gamma)^{i_j-1}D(a_{i_j}^{(j)}) \geq (1 + \gamma)^{i_j-1}\frac{\gamma}{nL}
\]

where \(a_k\) denotes the beginning of the \(k\)-th interval (again we use monotonicity to argue that the extrema were reached at the ends of each interval), so that \(i_j \leq 1 + \frac{\log \frac{nL}{\gamma}}{\log(1+\gamma)}\). In particular, the total number of internal nodes is then

\[
\sum_{i=1}^{t} i_j \leq t \cdot \left(1 + \frac{\log \frac{nL}{\gamma}}{\log(1+\gamma)}\right) = (1 + o(1)) \frac{\log^2 n}{\log(1+\gamma)} \leq L.
\]

This implies the same bound on the number of leaves \(\ell\). \(\square\)

Observe that if \(j\) satisfies the second item then \(\|D_{I_j} - U_{I_j}\|_1 \leq \gamma\) by ???. Thus, the
def distance between a monotone $D$ and its flattening $\bar{D} \overset{\text{def}}{=} \Phi(D, \mathcal{I}(D, \gamma))$ is at most

$$\|D - \bar{D}\|_1 \leq \sum_{j: D(I_j) \leq \frac{1}{L}} (D(I_j) + \bar{D}(I_j)) + \sum_{j: D(I_j) > \frac{1}{L}} D(I_j) \cdot \|D_{I_j} - \mathcal{U}_{I_j}\|_1$$

$$\leq L \cdot 2 \frac{\gamma}{L} + \gamma = 3\gamma.$$

**Corollary 3.1.4** (Unimodality). For all $\gamma > 0$, the class $\mathcal{M}_1$ of unimodal distributions on $[n]$ is $(\gamma, L)$-decomposable for $L \overset{\text{def}}{=} O\left(\frac{\log^2 n}{\gamma}\right)$.

**Proof.** For any $D \in \mathcal{M}_1$, $[n]$ can be partitioned in two intervals $I, J$ such that $D_I, D_J$ are either monotone non-increasing or non-decreasing. Applying Theorem 3.1.3 to $D_I$ and $D_J$ and taking the union of both partitions yields a (no longer necessarily dyadic) partition of $[n]$. □

The same argument yields an analogue statement for $t$-modal distributions:

**Corollary 3.1.5** ($t$-modality). For any $t \geq 1$ and all $\gamma > 0$, the class $\mathcal{M}_t$ of $t$-modal distributions on $[n]$ is $(\gamma, L)$-decomposable for $L \overset{\text{def}}{=} O\left(\frac{t \log^2 n}{\gamma}\right)$.

**Corollary 3.1.6** (Log-concavity, concavity and convexity). For all $\gamma > 0$, the classes $\mathcal{L}^-$ and $\mathcal{K}^+$ of log-concave, concave and convex distributions on $[n]$ are $(\gamma, L)$-decomposable for $L \overset{\text{def}}{=} O\left(\frac{\log^2 n}{\gamma}\right)$.

**Proof.** This is directly implied by Corollary 3.1.4, recalling that log-concave, concave and convex distributions are unimodal. □

**Theorem 3.1.7** (Monotone Hazard Rate). For all $\gamma > 0$, the class $\mathcal{MHR}$ of MHR distributions on $[n]$ is $(\gamma, L)$-decomposable for $L \overset{\text{def}}{=} O\left(\frac{\log n}{\gamma}\right)$.

**Proof.** This follows from adapting the proof of [11], which establishes that every MHR distribution can be approximated in $\ell_1$ distance by a $O(\log(n/\varepsilon)/\varepsilon)$-histogram. For completeness, we reproduce their argument, suitably modified to our purposes, in Section B.2. □
Theorem 3.1.8 (k-histogram). The class of k-histogram distributions on \([n]\) is \((0, k)\)-decomposable.

Proof. It is immediate from the definition.

3.2 Computing the distances

This section contains details of the distance estimation procedures for these classes, required in the last stage of Algorithm 1.

Lemma 3.2.1 (Monotonicity \([9, \text{Lemma 8}]\)). There exists a procedure \(\text{ESTIMATEDISTANCE}_M\) that, on input \(n\) as well as the full (succinct) specification of a \(\ell\)-histogram \(D\) on \([n]\), computes the distance \(\ell_1(D, M)\) in time \(\text{poly}(\ell)\).

A straightforward modification of the algorithm above (e.g., by adapting the underlying linear program to take as input the location \(m \in [\ell]\) of the mode of the distribution; then trying all \(\ell\) possibilities, running the subroutine \(\ell\) times and picking the minimum value) results in a similar claim for unimodal distributions:

Lemma 3.2.2 (Unimodality). There exists a procedure \(\text{ESTIMATEDISTANCE}_{M_1}\) that, on input \(n\) as well as the full (succinct) specification of a \(\ell\)-histogram \(D\) on \([n]\), computes the distance \(\ell_1(D, M_1)\) in time \(\text{poly}(\ell)\).

Lemma 3.2.3 (k-histogram). There exists a procedure \(\text{ESTIMATEDISTANCE}_H\) that, on input \(n\) as well as the full (succinct) specification of a \(\ell\)-histogram \(D\) on \([n]\), computes the distance \(\ell_1(D, H)\) in time \(\text{poly}(\ell)\).

The latter can be done by running \(\binom{n}{\ell}\) linear programs (i.e., one for each possible partition of the domain \([n]\) into \(\ell\) segments) and picking the smallest value at the end.

Similar statements hold for the classes of concave and convex distributions \(K^+, K^-\), also based on linear programming (specifically, on running \(O(n^2)\) different linear programs – one for each possible support \([a, b] \subset [n]\) – and taking the minimum over them).
Lemma 3.2.4 (MHR). There exists a (non-efficient) procedure \textsc{EstimateDistance}_{MHR} that, on input \( n \) as well as the full specification of a distribution \( D \) on \([n]\), computes (an \( \varepsilon \)-accurate approximation of) the distance \( \ell_1(D, MHR) \) in time \( 2^{\tilde{O}_e(n)} \).

Lemma 3.2.5 (Log-concavity). There exists a (non-efficient) procedure \textsc{EstimateDistance}_{L} that, on input \( n \) as well as the full specification of a distribution \( D \) on \([n]\), computes (an \( \varepsilon \)-accurate approximation of) the distance \( \ell_1(D, L) \) in time \( 2^{\tilde{O}_e(n)} \).

Lemma 3.2.4 and Lemma 3.2.5. We here give a naive algorithm for these two problems, based on an exhaustive search over a (huge) \( \varepsilon \)-cover \( S \) of distributions over \([n]\). Essentially, \( S \) contains all possible distributions whose probabilities \( p_1, \ldots, p_n \) are of the form \( j\varepsilon/n \), for \( j \in \{0, \ldots, n/\varepsilon\} \) (so that \( |S| = \tilde{O}((n/\varepsilon)^n) \)). It is not hard to see that this indeed defines an \( \varepsilon \)-cover of the set of all distributions, and moreover that it can be computed in time \( \text{poly}(|S|) \).

To approximate the distance from an explicit distribution \( D \) to the class \( C \) (either \( MHR \) or \( L \)), it is enough to go over every element \( S \) of \( S \), checking (this time, efficiently) if \( S \) satisfies the constraints for \( C \), and if so computing the values \( ||S - D||_1 \). At the end of this enumeration, the procedure outputs the minimum value obtained. \( \Box \)
3.3 Algorithm

In this section, we obtain our main result, restated below:

Theorem 3.3.1. There exists an algorithm TESTSPLITTABLE which, given sampling access to an unknown distribution $D$ over $[n]$ and parameter $\varepsilon \in (0, 1]$, can distinguish with probability $2/3$ between (a) $D \in \mathcal{P}$ versus (b) $\ell_1(D, \mathcal{P}) > \varepsilon$, for any property $\mathcal{P}$ that satisfies the above natural structural criterion (Succinctness). Moreover, for many such properties this algorithm is efficient, and its sample complexity is optimal (up to logarithmic factors and the exact dependence on $\varepsilon$).

3.3.1 The algorithm

Theorem 3.3.1, and with it Corollary 3.3.4 and Corollary 3.3.5 will follow from the theorem below, combined with the structural theorems from Section 3.1:

Theorem 3.3.2. Let $\mathcal{C}$ be a class of distributions over $[n]$ for which the following holds.

1. $\mathcal{C}$ is $(\gamma, L)$-splittable;

2. there exists a procedure ESTIMATEDISTANCE$_\mathcal{C}$ which, given as input the explicit description of a distribution $D$ over $[n]$, computes its distance $\ell_1(D, \mathcal{C})$ to $\mathcal{C}$.

Then, the algorithm TESTSPLITTABLE (Algorithm 1) is a $O(\max(\sqrt{n}L \log L/\varepsilon^3, L^2/\varepsilon^2))$-sample tester for $\mathcal{C}$, for $L = L(\varepsilon, n)$. (Moreover, if ESTIMATEDISTANCE$_\mathcal{C}$ is efficient, then so is TESTSPLITTABLE.)
Algorithm 1 TESTSPLITTABLE

**Require:** Domain $I$ (interval), sample access to $D$ over $I$; subroutine $\text{ESTIMATEDISTANCE}_C$

**Input:** Parameters $\varepsilon$ and function $L_C(\cdot, \cdot)$

1: **SETTING UP**
2: Define $\gamma \overset{\text{def}}{=} \frac{\varepsilon}{20}$, $L \overset{\text{def}}{=} L_C(\gamma, |I|)$, $\kappa \overset{\text{def}}{=} \frac{32L}{\varepsilon}$
3: Set $m \overset{\text{def}}{=} C \cdot \max \left( \kappa^2, \sqrt{|I|} \kappa \log L \right) = \tilde{O} \left( \max \left( \frac{\sqrt{|I|}}{\varepsilon^3} L, \frac{L^2}{\varepsilon^2} \right) \right) \triangleright C$ is an absolute constant.
4: 
5: **DECOMPOSITION**
6: Obtain a sequence $s$ of $m$ independent samples from $D$.
7: Let $\hat{D}$ be the empirical estimate of $D$ they induce.
8: while $\hat{D}(I) \geq \frac{1}{\kappa}$ and at most $L$ splits have been performed do \triangleright Implies $m_I \geq \frac{m}{\kappa}$.
9: Run $\text{CHECK-SMALL-}\ell_2$ (from Lemma 1.1.6) with parameters $\frac{\varepsilon}{8}$ and $\delta = \frac{1}{10L}$, using the samples of $s$ belonging to $I$.
10: if $\text{CHECK-SMALL-}\ell_2$ outputs no then
11: Bisect $I$, and recurse on both halves (using the same samples).
12: end if
13: end while
14: if more than $L$ splits have been performed then
15: return REJECT
16: else
17: Let $I \overset{\text{def}}{=} (I_1, \ldots, I_\ell)$ be the partition of $[n]$ from the leaves of the recursion. \triangleright $\ell \leq L$.
18: end if
19: 
20: **APPROXIMATION**
21: Learn the flattening $\Phi(D, \mathcal{I})$ of $D$ to $\ell_1$ error $\frac{\varepsilon}{4}$ (with probability $1/10$), using $O(\ell/\varepsilon^2)$ new samples. Let $\tilde{D}$ be the resulting hypothesis. \triangleright $\tilde{D}$ is a $\ell$-histogram.
22: 
23: **OFFLINE CHECK**
24: return ACCEPT if and only if $\text{ESTIMATEDISTANCE}_C(\tilde{D}) \leq \frac{\varepsilon}{2}$. \triangleright No sample needed.
3.3.2 Proof of Theorem 3.3.2

We now give the proof of our main result (Theorem 3.3.2), first analyzing the sample complexity of Algorithm 1 before arguing its correctness. For the latter, we will need a classical result from Probability, the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality, restated below:

**Theorem 3.3.3 ([20, 26]).** Let \( D \) be a distribution over \([n]\). Given \( m \) independent samples \( x_1, \ldots, x_m \) from \( D \), define the empirical distribution \( \hat{D} \) as follows:

\[
\hat{D}(i) \overset{\text{def}}{=} \frac{|\{ j \in [m] : x_j = i \}|}{m}, \quad i \in [n].
\]

Then, for all \( \varepsilon > 0 \),
\[
\Pr[D_K(D, \hat{D}) > \varepsilon] \leq 2e^{-2mn^2},
\]
where \( d_K(\cdot, \cdot) \) denotes the Kolmogorov distance (i.e., the \( \ell_\infty \) distance between cumulative distribution functions).

In particular, this implies that \( O(1/\varepsilon^2) \) samples suffice to learn a distribution up to \( \varepsilon \) in Kolmogorov distance.

3.3.3 Sample complexity.

The sample complexity is immediate, and comes from Steps 6 and 21. The total number of samples is

\[
m + O\left(\frac{\ell}{\varepsilon^2}\right) = O\left(\max\left(\frac{\sqrt{|I|}}{\varepsilon^3} L \log L, \frac{L^2}{\varepsilon^2}\right) + \frac{L}{\varepsilon^2}\right) = O\left(\max\left(\frac{\sqrt{|I|}}{\varepsilon^3} L \log L, \frac{L^2}{\varepsilon^2}\right)\right).
\]

3.3.4 Correctness.

Say an interval \( I \) is heavy if \( D(I) \geq \frac{2}{\kappa} \), and light if \( D(I) \leq \frac{1}{2\kappa} \). By choice of \( m \), the DKW inequality ensures that with probability at least 9/10, for every interval \( J \) we consider \( |\hat{D}(J) - D(J)| \leq 1/(2\kappa) \). In particular, every heavy interval \( I \) will have \( \hat{D}(I) \geq \frac{1}{\kappa} \) and every light interval \( I \) will have \( \hat{D}(I) < \frac{1}{\kappa} \). We hereafter condition on this.
We first argue that if the algorithm does not reject in Step 14, then with probability at least $9/10$ we have $\|D - \Phi(D,I)\|_1 \leq \varepsilon/4$. Indeed, we can write

$$\|D - \Phi(D,I)\|_1 = \sum_{k: \hat{D}(I_k) < 1/\kappa} D(I_k) \cdot \|D_{I_k} - U_{I_k}\|_1 + \sum_{k: \hat{D}(I_k) \geq 1/\kappa} D(I_k) \cdot \|D_{I_k} - U_{I_k}\|_1$$

$$\leq 2 \sum_{k: \hat{D}(I_k) < 1/\kappa} D(I_k) + \sum_{k: \hat{D}(I_k) \geq 1/\kappa} D(I_k) \cdot \|D_{I_k} - U_{I_k}\|_1.$$

Let us bound the two terms separately.

- If $\hat{D}(I) \geq 1/\kappa$, then this means at least $m/\kappa$ samples fell into $I$, and by our choice of constants we can apply Lemma 1.1.6 with $\delta = \frac{1}{10L}$; conditioning on all of the (at most $L$) events happening, which overall fails with probability at most $1/10$ by a union bound, we get

$$\|D_I\|_2^2 = \|D_I - U_I\|_2^2 + \frac{1}{|I|} \leq \left(1 + \frac{\varepsilon^2}{64}\right) \frac{1}{|I|}$$

as CHECK-SMALL-$\ell_2$ returned yes; and by ?? this implies $\|D_I - U_I\|_1 \leq \varepsilon/8$.

- If $\hat{D}(I_k) < 1/\kappa$, we know that $I$ cannot be heavy, and therefore $D(I) < \frac{2}{\kappa} = \frac{\varepsilon}{16L}$.

Putting it together, this yields

$$\|D - \Phi(D,I)\|_1 \leq 2 \sum_{k: \hat{D}(I_k) < 1/\kappa} \frac{\varepsilon}{16L} + \frac{\varepsilon}{8} \sum_{k: \hat{D}(I_k) \geq 1/\kappa} D(I_k) \leq \varepsilon/8 + \varepsilon/8 = \varepsilon/4.$$

**Soundness** By contrapositive, we argue that if the test returns ACCEPT, then (with probability at least $2/3$) $D$ is $\varepsilon$-close to $C$. Indeed, conditioning on $\hat{D}$ being $(\varepsilon/4)$-close to $\Phi(D,I)$, we get by the triangle inequality that

$$\|D - C\|_1 \leq \|D - \Phi(D,I)\|_1 + \|\Phi(D,I) - \hat{D}\|_1 + \text{dist}(\hat{D}, C)$$

$$\leq \varepsilon/4 + \varepsilon/4 + \varepsilon/2 = \varepsilon.$$

Overall, this happens except with probability at most $1/10 + 1/10 + 1/10 < 1/3$.

**Completeness** Assume $D \in C$. Then the choice of of $\gamma$ and $L$ ensures the existence of a
good dyadic partition \( \mathcal{I}(\gamma, D) \) in the sense of Theorem 3.1.1. For any \( I \) in this partition for which (i) holds \( (D(I) \leq \frac{\varepsilon}{L} < \frac{1}{2\kappa}) \), \( I \) will have \( \hat{D}(I) < \frac{1}{\kappa} \) and be kept as a “light leaf”. For the other ones, (ii) holds: let \( I \) be one of these (at most \( L \) intervals.

- If \( \hat{D}(I) < \frac{1}{\kappa} \), then \( I \) is kept as “light leaf”.
- If \( \hat{D}(I) \geq \frac{1}{\kappa} \), then by our choice of constants we can apply Lemma 1.1.6 with \( \delta = \frac{1}{10\kappa} \); conditioning on all of the (at most \( L \)) events happening, which overall fails with probability at most \( 1/10 \) by a union bound, CHECK-SMALL-\( \ell_2 \) will output yes, as

\[
\|D_I - U_I\|_2^2 = \|D_I\|_2^2 - \frac{1}{|I|} \leq \left(1 + \frac{\varepsilon^2}{400}\right) \frac{1}{|I|} - \frac{1}{|I|} = \frac{\varepsilon^2}{400 |I|} \leq \frac{\varepsilon^2}{256 |I|}
\]

and \( I \) is kept as “flat leaf”.

Therefore, as \( \mathcal{I}(\gamma, D) \) is dyadic the DECOMPOSITION stage is guaranteed to stop within at most \( L \) splits (in the worst case, it goes on until \( \mathcal{I}(\gamma, D) \) is considered, at which point it succeeds).\(^1\) Thus Step 14 passes, and the algorithm reaches the APPROXIMATION stage. By the foregoing discussion, this implies \( \Phi(D, \mathcal{I}) \) is \( \varepsilon/4 \)-close to \( D \) (and hence to \( C \)); \( \tilde{D} \) is then (except with probability at most \( 1/10 \)) \( (\varepsilon/4 + \varepsilon/4 = \varepsilon/2) \)-close to \( C \), and the algorithm returns ACCEPT.

We can now combine the structural results from section Section 3.1 and Theorem 3.3.2 to get the following corollaries:

**Corollary 3.3.4.** The algorithm TESTSPLITTABLE can test the classes of monotone, unimodal, log-concave, concave, convex, and monotone hazard rate (MHR) distributions, with \( \tilde{O}(\sqrt{n}/\varepsilon^4) \) samples.

\(^1\)In more detail, we want to argue that if \( D \) is in the class, then a decomposition with at most \( L \) pieces is found by the algorithm. Since there is a dyadic decomposition with at most \( L \) pieces (namely, \( \mathcal{I}(\gamma, D) = (I_1, \ldots, I_L) \)), it suffices to argue that the algorithm will never split one of the \( I_j \)'s (as every single \( I_j \) will eventually be considered by the recursive binary splitting, unless the algorithm stopped recursing in this “path” before even considering \( I_j \), which is even better). But this is the case by the above argument, which ensures each such \( I_j \) will be recognized as satisfying one of the two conditions for “good decomposition” (being either close to uniform in \( \ell_2 \), or having very little mass).
Corollary 3.3.5. The algorithm TESTSPLITTABLE can test the class of $t$-modal distributions, with $\tilde{O}(t\sqrt{n}/\varepsilon^4)$ samples.

Corollary 3.3.6. The algorithm TESTSPLITTABLE can test the classes of Binomial and Poisson Binomial Distributions, with $\tilde{O}(n^{1/4}/\varepsilon^4)$ samples.

Corollary 3.3.7. The algorithm TESTSPLITTABLE can test the class of $k$-histogram Distributions, with $\tilde{O}(k \cdot \sqrt{n})/\varepsilon^3$ samples.

Extension to mixtures. We finally note that, in contrast to the case-specific analogous results of [9, 23], Corollary 3.3.4, Corollary 3.3.5 and Corollary 3.3.6 easily extend to mixtures of distributions from these respective classes. More specifically, since the mixture of any $k$ $(\gamma, L)$-decomposable distributions is $(\gamma, k \cdot L)$-decomposable, our results imply the following:

Corollary 3.3.8. Let $C_k$ be the set of mixtures of (at most) $k$ distributions from $C$, where $C$ is any union of the classes of monotone, unimodal, log-concave and monotone hazard rate (MHR) distributions. Then, $C_k$ can be tested with $\tilde{O}(k\sqrt{n}/\varepsilon^4)$ samples. Similarly, the class $M_{t,k}$ of mixtures of (at most) $k$ $t$-modal distributions can be tested with $\tilde{O}(kt\sqrt{n}/\varepsilon^4)$ samples.
3.4 Considering the effective support size only

The general approach we have been following so far gives, out-of-the-box, an efficient testing algorithm with sample complexity $O(V^3)$ for a large range of properties. However, this sample complexity can for some classes $P$ be brought down a lot more, by taking advantage in a preprocessing step of good concentration guarantees of distributions in $P$.

As a motivating example, consider the class of Poisson Binomial Distributions (PBD). It is well-known (see e.g. [25, Section 2]) that PBDs are unimodal, and more specifically that $\mathcal{PBD}_n \subseteq \mathcal{L} \subseteq \mathcal{M}_1$. Therefore, using our generic framework we can test Poisson Binomial Distributions with $O(\sqrt{n})$ samples. This is, however, far from optimal: as shown in [23], a sample complexity of $\Theta(n^{1/4})$ is both necessary and sufficient. The reason our general algorithm ends up making quadratically too many queries can be explained as follows. PBDs are tightly concentrated around their expectation, so that they “morally” live on a support of size $m = O(\sqrt{n})$. Yet, instead of testing them on this very small support, in the above we still consider the entire range $[n]$, and thus end up paying a dependence $\sqrt{n}$ – instead of $\sqrt{m}$.

If we could use that observation to first reduce the domain to the effective support of the distribution, then we could call our testing algorithm on this reduced domain of size $O(\sqrt{n})$. In the rest of this section, we formalize and develop this idea, and in Section 3.4.2 will obtain as a direct application a $O(n^{1/4})$-query testing algorithm for $\mathcal{PBD}_n$.

**Definition 3.4.1.** Given $\varepsilon > 0$, the $\varepsilon$-effective support of a distribution $D$ is the smallest interval $I$ such that $D(I) \geq 1 - \varepsilon$.

The last definition we shall require is of the conditioned distributions of a class $C$:

**Definition 3.4.2.** For any class of distributions $C$ over $[n]$, define the set of conditioned distributions of $C$ (with respect to $\varepsilon > 0$ and interval $I \subseteq [n]$) as $C^{\varepsilon,I} \triangleq \{ D_I : D \in C, D(I) \geq 1 - \varepsilon \}$.

Finally, we will require the following simple result:

**Lemma 3.4.3.** Let $D$ be a distribution over $[n]$, and $I \subseteq [n]$ an interval such that $D(I) \geq 1 - \frac{\varepsilon}{10}$. Then,
• If \( D \in C \), then \( D \in C_{\frac{\ell_1}{10}} \);

• If \( \ell_1(D, C) > \varepsilon \), then \( \ell_1(D, C_{\frac{\ell_1}{10}}) > \frac{7\varepsilon}{10} \).

Proof. The first item is obvious. As for the second, let \( P \in C \) be any distribution with \( P(I) \geq 1 - \frac{\varepsilon}{10} \). By assumption, \( \| D - P \|_1 > \varepsilon \): but we have, writing \( \alpha = 1/10 \),

\[
\| D_I - P_I \|_1 = \sum_{i \in I} \left| \frac{D(i)}{D(I)} - \frac{P(i)}{P(I)} \right| = \frac{1}{D(I)} \sum_{i \in I} \left| D(i) - P(i) \right| \left( 1 - \frac{D(I)}{P(I)} \right)
\]

\[
\geq \frac{1}{D(I)} \left( \sum_{i \in I} |D(i) - P(i)| \right) \geq \frac{1}{D(I)} \left( \sum_{i \in I} |D(i) - P(i)| - \alpha \varepsilon \right)
\]

\[
\geq \frac{1}{D(I)} \left( \| D - P \|_1 - \sum_{i \in I} |D(i) - P(i)| - \alpha \varepsilon \right) \geq \frac{1}{D(I)} \left( \| D - P \|_1 - 3\alpha \varepsilon \right)
\]

\[
> (1 - 3\alpha)\varepsilon = \frac{7}{10} \varepsilon.
\]

We now proceed to state and prove our result – namely, efficient testing of structured classes of distributions with nice concentration properties.

**Theorem 3.4.4.** Let \( C \) be a class of distributions over \([n]\) for which the following holds.

1. there is a function \( M(\cdot, \cdot) \) such that each \( D \in C \) has \( \varepsilon \)-effective support of size at most \( M(n, \varepsilon) \);

2. for every \( \varepsilon \in [0, 1] \) and interval \( I \subseteq [n] \), \( C_{\varepsilon, I} \) is \((\gamma, L)\)-splittable;

3. there exists an efficient procedure \( \text{ESTIMATEDISTANCE}_{C_{\varepsilon, I}} \) which, given as input the explicit description of a distribution \( D \) over \([n]\) and interval \( I \subseteq [n] \), computes the distance \( \ell_1(D_I, C_{\varepsilon, I}) \).

Then, the algorithm \( \text{TESTEFFECTIVESPLITTABLE} \) (Algorithm 2) is a \( O \left( \max \left( \frac{1}{\varepsilon^3} \sqrt{m\ell} \log \ell, \frac{\ell^2}{\varepsilon^3} \right) \right) \)-sample tester for \( C \), where \( m = M(n, \frac{\varepsilon}{60}) \) and \( \ell = L \left( \frac{\varepsilon}{1200}, m \right) \).
Algorithm 2 TestEffectiveSplittable

Require: Domain $\mathcal{X}$ (interval of size $n$), sample access to $D$ over $\mathcal{X}$; subroutine EstimateDistance$_{C,e,i}$

Input: Parameters $\varepsilon \in (0, 1]$, function $L(\cdot, \cdot)$, and upper bound function $M(\cdot, \cdot)$ for the effective support of the class $C$.

1: Set $m \overset{\text{def}}{=} O(1/\varepsilon^2)$, $\tau \overset{\text{def}}{=} M(n, \frac{\varepsilon}{60})$.

2: Effective Support

3: Compute $\hat{D}$, an empirical estimate of $D$, by drawing $m$ independent samples from $D$.

4: Let $J$ be the largest interval of the form $\{1, \ldots, j\}$ such that $\hat{D}(J) \leq \frac{\varepsilon}{30}$.

5: Let $K$ be the largest interval of the form $\{k, \ldots, n\}$ such that $\hat{D}(K) \leq \frac{\varepsilon}{30}$.

6: Set $I \leftarrow [n] \setminus (J \cup K)$.

7: if $|I| > \tau$ then return REJECT

8: end if

9:

10: Testing

11: Call TestSplittable with $I$ (providing simulated access to $D_I$ by rejection sampling, returning FAIL if the number of samples $q$ from $D_I$ required by the subroutine is not obtained after $O(q)$ samples from $D$), EstimateDistance$_{C,e,i}$, parameters $\varepsilon' \overset{\text{def}}{=} \frac{\tau}{10}$ and $L(\cdot, \cdot)$.

12: return ACCEPT if TestSplittable accepts, REJECT otherwise.

3.4.1 Proof of Theorem 3.4.4

By the choice of $m$ and the DKW inequality, with probability at least $23/24$ the estimate $\hat{D}$ satisfies $d_K(D, \hat{D}) \leq \frac{\varepsilon}{60}$. Conditioning on that from now on, we get that $D(I) \geq \hat{D}(I) - \frac{\varepsilon}{30} \geq 1 - \frac{\varepsilon}{10}$. Furthermore, denoting by $j$ and $k$ the two inner endpoints of $J$ and $K$ in Steps 4 and 5, we have $D(J \cup \{j + 1\}) \geq \hat{D}(J \cup \{j + 1\}) - \frac{\varepsilon}{60} > \frac{\varepsilon}{60}$ (similarly for $D(K \cup \{k - 1\})$), so that $I$ has size at most $\sigma + 1$, where $\sigma$ is the $\frac{\varepsilon}{60}$-effective support size of $D$. 

44
Finally, note that since $D(I) = \Omega(1)$ by our conditioning, the simulation of samples by rejection sampling will succeed with probability at least $23/24$ and the algorithm will not output FAIL.

**Sample complexity.** The sample complexity is the sum of the $O(1/\varepsilon^2)$ in Step 3 and the $O(q)$ in Step 11. From Theorem 3.3.1 and the choice of $I$, this latter quantity is $O\left(\max\left(\frac{1}{\varepsilon^3} \sqrt{M(n, \frac{\varepsilon}{60})\ell \log \ell, \frac{\varepsilon^2}{\ell^2}}\right)\right)$ where $\ell = L\left(\frac{\varepsilon}{1200}, M(n, \frac{\varepsilon}{60})\right)$.

**Correctness.** If $D \in \mathcal{C}$, then by the setting of $\tau$ (set to be an upper bound on the $\frac{\varepsilon}{60}$-effective support size of any distribution in $\mathcal{C}$) the algorithm will go beyond Step 6. The call to TESTSPLITTABLE will then end up in the algorithm returning ACCEPT in Step 12, with probability at least $2/3$ by Lemma 3.4.3, Theorem 3.3.1 and our choice of parameters.

Similarly, if $D$ is $\varepsilon$-far from $\mathcal{C}$, then either its effective support is too large (and then the test on Step 6 fails), or the main tester will detect that its conditional distribution on $I$ is $\frac{\tau\varepsilon}{10}$-far from $\mathcal{C}$ and output REJECT in Step 12.

Overall, in either case the algorithm is correct except with probability at most $1/24 + 1/24 + 1/3 = 5/12$ (by a union bound). Repeating constantly many times and outputting the majority vote brings the probability of failure down to $1/3$. \qed

### 3.4.2 Application: Testing Poisson Binomial Distributions

In this section, we illustrate the use of our generic two-stage approach to test the class of Poisson Binomial Distributions. Specifically, we prove the following result:

**Corollary 3.4.5.** The class of Poisson Binomial Distributions can be tested with $\tilde{O}(n^{1/4}/\varepsilon^4)$ samples, using Algorithm 2.

This is a direct consequence of Theorem 3.4.4 and the lemmas below. The first one states that, indeed, PBDs have small effective support:

**Fact 3.4.6.** For any $\varepsilon > 0$, a PBD has $\varepsilon$-effective support of size $O\left(\sqrt{n\log(1/\varepsilon)}\right)$. 45
Proof. By an additive Chernoff Bound, any random variable \(X\) following a Poisson Binomial Distribution has \(\Pr[|X - EX| > \gamma n] \leq 2e^{-2\gamma n}\). Taking \(\gamma \overset{\text{def}}{=} \sqrt{\frac{1}{2n} \ln \frac{2}{\epsilon}}\), we get that \(\Pr[X \in I] \geq 1 - \epsilon\), where \(I \overset{\text{def}}{=} [EX - \sqrt{\frac{1}{2} \ln \frac{2}{\epsilon}} \cdot EX + \sqrt{\frac{1}{2} \ln \frac{2}{\epsilon}}]\).

It is clear that if \(D \in \mathcal{PBD}_n\) (and therefore is unimodal), then for any interval \(I \subseteq [n]\) the conditional distribution \(D_I\) is still unimodal, and thus the class of conditioned PBDs \(\mathcal{PBD}_n^{\epsilon, I} \overset{\text{def}}{=} \{D_I : D \in \mathcal{PBD}_n, D(I) \geq 1 - \epsilon\}\) falls under Corollary 3.1.4. The last piece we need to apply our generic testing framework is the existence of an efficient algorithm to compute the distance between an (explicit) distribution and the class of conditioned PBDs. This is provided by our next lemma:

Claim 3.4.7. There exists a procedure \(\text{ESTIMATEDISTANCE}_{\mathcal{PBD}_n^{\epsilon, I}}\) that, on input \(n\) and \(\epsilon, \in [0, 1]\), \(I \subseteq [n]\) as well as the full specification of a distribution \(D\) on \([n]\), computes a value \(\tau\) such that \(\tau \in [1 \pm 2\epsilon]f_1(D, \mathcal{PBD}_n^{\epsilon, I}) \pm \frac{\epsilon}{100}\), in time \(\text{poly}(n) \cdot (1/\epsilon)^{O(\log^2 1/\epsilon)}\).

Proof. The goal is to find a \(\gamma\)-approximation of the minimum value of \(\sum_{i \in I} \left| \frac{P(i)}{P(I)} - \frac{D(i)}{D(I)} \right|\), subject to \(P(I) = \sum_{i \in I} P(i) \geq 1 - \epsilon\) and \(P \in \mathcal{PBD}_n\). We first note that, given the parameters \(n \in \mathbb{N}\) and \(p_1, \ldots, p_n \in [0, 1]\) of a PBD \(P\), the vector of \((n + 1)\) probabilities \(P(0), \ldots, P(n)\) can be obtained in time \(O(n^2)\), e.g. by dynamic programming. Therefore, computing the \(\ell_1\) distance between \(D\) and any PBD with known parameters can be done efficiently. To conclude, we invoke a result of Daskalakis and Papadimitriou, that guarantees the existence of a succinct (proper) cover of \(\mathcal{PBD}_n\):

Theorem 3.4.8 ([18, Theorem 1]). For all \(n, \gamma > 0\), there exists a set \(S_\gamma \subseteq \mathcal{PBD}_n\) such that:

(i) \(S_\gamma\) is a \(\gamma\)-cover of \(\mathcal{PBD}_n\) in \(\ell_1\); that is, for all \(D \in \mathcal{PBD}_n\), there exists some \(D' \in S_\gamma\) such that \(\|D - D'\|_1 \leq \gamma\)

(ii) \(|S_\gamma| \leq n^2 + n \cdot (1/\gamma)^{O(\log^2 1/\gamma)}\)

(iii) \(S_\gamma\) can be computed in time \(O(n^2 \log n) + O(n \log n) \cdot (1/\gamma)^{O(\log^2 1/\gamma)}\).

and each \(D \in S_\gamma\) is explicitly described by its set of parameters.

Set \(\gamma \overset{\text{def}}{=} \frac{\epsilon}{250}\). Fix \(P \in \mathcal{PBD}_n\) such that \(P(I) \geq 1 - \epsilon\), and \(Q \in S_\gamma\) such that \(\|P - Q\|_1 \leq \gamma\).

In particular, it is easy to see via the correspondence between \(\ell_1\) and total variation distance.
that \(|P(I) - Q(I)| \leq \gamma/2\). By a calculation analogue as in Lemma 3.4.3, we have

\[
\|P_I - Q_I\|_1 = \sum_{i \in I} \left| \frac{P(i)}{P(I)} - \frac{Q(i)}{Q(I)} \right| = \sum_{i \in I} \left| \frac{P(i)}{P(I)} - \frac{Q(i)}{Q(I)} \right| + Q(i) \left( \frac{1}{P(I)} - \frac{1}{Q(I)} \right)
\]

\[
= \frac{1}{P(I)} \left( \sum_{i \in I} |P(i) - Q(i)| \pm \frac{\gamma}{2} \right) = \frac{1}{P(I)} \left( \|P - Q\|_1 \pm \frac{5\gamma}{2} \right)
\]

where we used the fact that \(\sum_{i \in I} |P(i) - Q(i)| = 2 \left( \sum_{i \in I: P(i) > Q(i)} (P(i) - Q(i)) \right) = Q(I) - P(I) \in [-2\gamma, 2\gamma]\). By the triangle inequality, this implies that the minimum of \(\|P_I - D_I\|_1\) over the distributions \(P\) of \(S_e\) with \(P(I) \geq 1 - (e + \gamma/2)\) will be within an additive \(O(\varepsilon)\) of \(\ell_1(D, PBD_n^{e,f})\). The fact that the former can be done in time \(\text{poly}(n) \cdot (1/\varepsilon)^{O(\log^2 1/\varepsilon)}\) concludes the proof.

It is not hard to see that the fact this only guarantees an approximation of \(\ell_1(D, PBD_n^{e,f})\) instead of the exact value is sufficient for the purpose of Algorithm 1.

Proof of Corollary 3.4.5. Combining the above, we invoke Theorem 3.4.4 with \(M(n, \varepsilon) = O(\sqrt{n \log(1/\varepsilon)})\) (Fact 3.4.6) and \(L(m, \gamma) = O\left(\frac{\log^2 m}{\gamma}\right)\) (Corollary 3.1.4). This yields the claimed sample complexity; finally, the efficiency is a direct consequence of Claim 3.4.7. \(\square\)
3.5 Lower bounds

We now turn to proving converses to our positive results – namely, that many of the upper bounds we obtain cannot be significantly improved upon. As in our algorithmic approach, we describe for this purpose a generic framework for obtaining lower bounds.

In order to state our results, we will require the usual definition of agnostic learning. Recall that an algorithm is said to be a semi-agnostic learner for a class $C$ if it satisfies the following. Given sample access to an arbitrary distribution $D$ and parameter $\varepsilon$, it outputs a hypothesis $\hat{D}$ which (with high probability) does “almost as well as it gets”:

$$\|D - \hat{D}\|_1 \leq c \cdot \text{OPT}_{C,D} + O(\varepsilon)$$

where $\text{OPT}_{C,D} \overset{\text{def}}{=} \inf_{D' \in C} \ell_1(D', D)$, and $c \geq 1$ is some absolute constant (if $c = 1$, the learner is said to be agnostic).

**High-level idea.** The motivation for our result is the observation of [9] that “monotonicity is at least as hard as uniformity”. Unfortunately, their specific argument does not generalize easily to other classes of distribution, making it impossible to extend it readily. The starting point of our approach is to observe that while uniformity testing is hard in general, it becomes very easy under the promise that the distribution is monotone, or even only close to monotone (namely, $O(1/\varepsilon^2)$ samples suffice). This can give an alternate proof of the lower bound for monotonicity testing, via a different reduction: first, test if the unknown distribution is monotone; if it is, test whether it is uniform, now assuming closeness to monotone.

More generally, this idea applies to any class $C$ which (a) contains the uniform distribution, and (b) for which we have a $o(\sqrt{n})$-sample agnostic learner $L$, as follows. Assuming we have a tester $T$ for $C$ with sample complexity $o(\sqrt{n})$, define a uniformity tester as below.

- test if $D \in C$ using $T$; if not, reject (as $\mathcal{U} \in C$, $D$ cannot be uniform);
- otherwise, agnostically learn $D$ with $L$ (since $D$ is close to $C$), and obtain hypothesis $\hat{D}$;
- check offline if $\hat{D}$ is close to uniform.
By assumption, $T$ and $L$ each use $o(\sqrt{n})$ samples, so does the whole process; but this contradicts the lower bound of $[21, 27]$ on uniformity testing. Hence, $T$ must use $\Omega(\sqrt{n})$ samples.

This “testing-by-narrowing” reduction argument can be further extended to other reductions than to uniformity, as we show below:

**Theorem 3.5.1.** Let $C$ be a class of distributions over $[n]$ for which the following holds:

1. there exists a semi-agnostic learner $L$ for $C$, with sample complexity $q_L(n, \varepsilon, \delta)$ and “agnostic constant” $c$;
2. there exists a subclass $C_{\text{Hard}} \subseteq C$ such that testing $C_{\text{Hard}}$ requires $q_H(n, \varepsilon)$ samples.

Suppose further that $q_L(n, \varepsilon, 1/10) = o(q_H(n, \varepsilon))$. Then, any tester for $C$ must use $\Omega(q_H(n, \varepsilon))$ samples.

*Proof.* The above theorem relies on the reduction outlined above, which we rigorously detail here. Assuming $C$, $C_{\text{Hard}}$, $L$ as above (with semi-agnostic constant $c \geq 1$), and a tester $T$ for $C$ with sample complexity $q_T(n, \varepsilon)$, we define a tester $T_{\text{Hard}}$ for $C_{\text{Hard}}$. On input $\varepsilon \in (0, 1]$ and given sample access to a distribution $D$ on $[n]$, $T_{\text{Hard}}$ acts as follows:

- call $T$ with parameters $n$, $\varepsilon'/c$ (where $\varepsilon' \overset{\text{def}}{=} \frac{\varepsilon}{c}$) and failure probability $1/6$, to $\varepsilon'/c$-test if $D \in C$. If not, reject.
- otherwise, agnostically learn a hypothesis $\hat{D}$ for $D$, with $L$ called with parameters $n$, $\varepsilon'$ and failure probability $1/6$;
- check offline if $\hat{D}$ is $\varepsilon'$-close to $C_{\text{Hard}}$, accept if and only if this is the case.

We condition on both calls (to $T$ and $L$) to be successful, which overall happens with probability at least $2/3$ by a union bound. The completeness is immediate: if $D \in C_{\text{Hard}} \subseteq C$, $T$ accepts, and the hypothesis $\hat{D}$ satisfies $\|\hat{D} - D\|_1 \leq \varepsilon'$. Therefore, $\ell_1(\hat{D}, C_{\text{Hard}}) \leq \varepsilon'$, and $T_{\text{Hard}}$ accepts.

For the soundness, we proceed by contrapositive. Suppose $T_{\text{Hard}}$ accepts; it means that each step was successful. In particular, $\ell_1(\hat{D}, C) \leq \varepsilon'/c$; so that the hypothesis outputted by the agnostic learner satisfies $\|\hat{D} - D\|_1 \leq c \cdot \text{OPT} + \varepsilon' \leq 2\varepsilon'$. In turn, since the last step passed
and by a triangle inequality we get, as claimed, $\ell_1(D, C_{\text{Hard}}) \leq 2\varepsilon' + \ell_1(\hat{D}, C_{\text{Hard}}) \leq 3\varepsilon' = \varepsilon$.

Observing that the overall sample complexity is $q_T(n, \frac{\varepsilon'}{c}) + q_L(n, \varepsilon', \frac{1}{10}) = q_T(n, \varepsilon') + o(q_H(n, \varepsilon'))$ concludes the proof.

Taking $C_{\text{Hard}}$ to be the singleton consisting of the uniform distribution, and from the semi-agnostic learners of [11, 12] (each with sample complexity either poly$(1/\varepsilon)$ or poly$(\log n, 1/\varepsilon)$), we obtain the following:²

**Corollary 3.5.2.** Testing log-concavity, convexity, concavity, MHR, unimodality and t-modality each require $\Omega\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ samples (the latter for $t = o(\sqrt{n})$).

Similarly, we can use another result of [16] which shows how to agnostically learn Poisson Binomial Distributions with $\tilde{O}(1/\varepsilon^2)$ samples.³ Taking $C_{\text{Hard}}$ to be the single Bin$(n, 1/2)$ distribution (along with the testing lower bound of [35]), this yields the following:

**Corollary 3.5.3.** Testing $\mathcal{PBD}_n$ and $\mathcal{BIN}_n$ each require $\Omega\left(\frac{n^{1/4}}{\varepsilon^2}\right)$ samples.

Finally, we derive a lower bound on testing $k$-SIIRVs from the agnostic learner of [15] (which has sample complexity poly$(k, 1/\varepsilon)$ samples, independent of $n$):

**Corollary 3.5.4.** There exists an absolute constant $c > 0$ such that testing $k$-SIIRV$_n$ requires $\Omega\left(k^{1/2}n^{1/4}\right)$ samples, for any $k = o(n^c)$.

**Corollary 3.5.4.** To prove this result, it is enough by Theorem 3.5.1 to exhibit a particular $k$-SIIRV $S$ such that testing identity to $S$ requires this many samples. Moreover, from [35] this last part amounts to proving that the (truncated) 2/3-norm $\|S_{-\varepsilon_0}^{\text{max}}\|_{2/3}$ of $S$ is $\Omega\left(k^{1/2}n^{1/4}\right)$ (for some small $\varepsilon_0 > 0$). Our hard instance $S$ will be defined as follows: it is defined as the distribution of $X_1 + \cdots + X_n$, where the $X_i$'s are independent integer random variables uniform on $\{0, \ldots, k - 1\}$ (in particular, for $k = 2$ we get a Bin$(n, 1/2)$ distribution). It is straightforward to verify that $\mathbb{E}S = \frac{n(k-1)}{2}$ and $\sigma^2 \overset{\text{def}}{=} \text{Var}S = \frac{(k^2-1)n}{12} = \Theta(k^2n)$; moreover,

²Specifically, these lower bounds hold as long as $\varepsilon = \Omega(1/n^\alpha)$ for some absolute constant $\alpha > 0$ (so that the sample complexity of the agnostic learner is indeed negligible in front of $\sqrt{n}/\varepsilon^2$).

³Note the quasi-quadratic dependence on $\varepsilon$ of the learner, which allows us to get $\varepsilon$ into our lower bound for $n \gg \text{poly log}(1/\varepsilon)$. 

50
$S$ is log-concave (as the convolution of $n$ uniform distributions). From this last point, we get that (i) the maximum probability of $S$, attained at its mode, is $\|S\|_\infty = \Theta(1/\sigma)$; and (ii) for every $j$ in an interval $I$ of length $2\sigma$ centered at this mode, $S(j) \geq \Omega(\|S\|_\infty)$. Putting this together, we get that the $2/3$-norm (and similarly the truncated $2/3$-norm) of $S$ is lower bounded by

$$\left( \sum_{j \in I} S(j)^{2/3} \right)^{3/2} \geq \left( 2\sigma \cdot \Omega(1/\sigma)^{2/3} \right)^{3/2} = \Omega\left( \sigma^{1/2} \right) = \Omega\left( k^{1/2} n^{1/4} \right)$$

which concludes the proof. \qed
Appendix A

Summary of results

<table>
<thead>
<tr>
<th>Class</th>
<th>Upperbound</th>
<th>Lowerbound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monotone</td>
<td>$\tilde{O}(\frac{d^2}{e^6})$ [9], $\tilde{O}(\frac{d^2}{e^2})$ (Corollary 3.3.4)</td>
<td>$\Omega(\frac{d^2}{e^2})$ [9], $\Omega(\frac{d^2}{e^2})$ (Corollary 3.5.2)</td>
</tr>
<tr>
<td>Unimodal</td>
<td>$\tilde{O}(\frac{d^2}{e^2})$ (Corollary 3.3.4)</td>
<td>$\Omega(\frac{d^2}{e^2})$ (Corollary 3.5.2)</td>
</tr>
<tr>
<td>$k$-mixtures of monotone and/or unimodal</td>
<td>$\tilde{O}(k\frac{d^2}{e^2})$ (Corollary 3.3.4)</td>
<td>$\Omega(\frac{d^2}{e^2})$ (Corollary 3.5.2)</td>
</tr>
<tr>
<td>$t$-modal</td>
<td>$\tilde{O}(\frac{t^2}{e^2})$ (Corollary 3.3.5)</td>
<td>$\Omega(\frac{d^2}{e^2})$ (Corollary 3.5.2)</td>
</tr>
<tr>
<td>$k$-mixtures of $t$-modal</td>
<td>$\tilde{O}(\frac{k^2n^2}{e^2})$ (Corollary 3.3.4)</td>
<td>$\Omega(\frac{d^2}{e^2})$ (Corollary 3.5.2)</td>
</tr>
<tr>
<td>Log-concave, concave, convex</td>
<td>$\tilde{O}(\frac{d^2}{e^2})$ (Corollary 3.3.4)</td>
<td>$\Omega(\frac{d^2}{e^2})$ (Corollary 3.5.2)</td>
</tr>
<tr>
<td>$k$-mixtures of log-concave, concave and/or convex</td>
<td>$\tilde{O}(\frac{k^2n^2}{e^2})$ (Corollary 3.3.4)</td>
<td>$\Omega(\frac{d^2}{e^2})$ (Corollary 3.5.2)</td>
</tr>
<tr>
<td>Monotone Hazard Rate (MHR)</td>
<td>$\tilde{O}(\frac{n^{1/2}}{e^2})$ (Corollary 3.3.4)</td>
<td>$\Omega(\frac{n^{1/2}}{e^2})$ (Corollary 3.5.2)</td>
</tr>
<tr>
<td>$k$-mixtures of MHR</td>
<td>$\tilde{O}(\frac{k^{1/2}n^{1/2}}{e^2})$ (Corollary 3.3.4)</td>
<td>$\Omega(\frac{n^{1/2}}{e^2})$ (Corollary 3.5.3)</td>
</tr>
<tr>
<td>Binomial, Poissson Binomial (PBD)</td>
<td>$\tilde{O}(\frac{n^{1/2}}{e^2})$ [23] (1), $\tilde{O}(\frac{n^{1/2}}{e^2})$ (Corollary 3.3.6)</td>
<td>$\Omega(\frac{n^{1/2}}{e^2})$ [23], Corollary 3.5.3</td>
</tr>
<tr>
<td>$k$-mixtures of Binomials and/or PBD</td>
<td>$\tilde{O}(\frac{k^{1/2}n^{1/2}}{e^2})$ (Corollary 3.3.6)</td>
<td>$\Omega(k^{1/2}n^{1/2})$ (Corollary 3.5.4)</td>
</tr>
</tbody>
</table>

Table A.1: Summary of our results.

We stress that although two of these results are not new (the $\tilde{O}(\sqrt{n}/\varepsilon^6)$ upper and $\Omega(\sqrt{n}/\varepsilon^2)$ lower bounds on testing monotonicity can be found in [9], while the $\Theta(n^{1/4})$ sample complexity of testing PBDs is pinned down$^1$ in [23]), the crux here is that we are able to derive them in a unified way, by applying the same generic algorithm to these different tasks. As an addition to its generality, we emphasize that our framework yields a much cleaner and simpler proof of the results from [23], for both the upper and lower bounds.
Moreover, prior to our work nothing was known on testing membership to the other classes we consider.

\[ \text{Specifically, [23] obtain a } n^{1/4} \cdot \tilde{O}(1/\varepsilon^2) + \tilde{O}(1/\varepsilon^6) \text{ sample complexity, to be compared with our } \tilde{O}(n^{1/4} / \varepsilon^4) \text{ upper bound;} \text{ as well as an } \Omega(n^{1/4} / \varepsilon^6) \text{ lower bound.} \]
Appendix B

Omitted proofs

B.1 Proof of Lemma 1.1.6

We now give the proof of Lemma 1.1.6, restated below:

Proof. To do so, we first describe an algorithm that distinguishes between \( \|D - \mathcal{U}\|^2_2 \geq \epsilon^2/n \) and \( \|D - \mathcal{U}\|^2_2 < \epsilon^2/(2n) \) with probability at least 2/3, using \( C \cdot \sqrt{n} \) samples. Boosting the success probability to \( 1 - \delta \) at the price of a multiplicative \( \log \frac{1}{\delta} \) factor can then be achieved by standard techniques.

Similarly as in the proof of Theorem 11 (whose algorithm we use, but with a threshold \( \tau \equiv \frac{3m^2\epsilon^2}{4n} \) instead of \( \frac{3m}{\sqrt{n}} \)), define the quantities

\[
Z_k \equiv \left( X_k - \frac{m}{n} \right)^2 - X_k, \quad k \in [n]
\]

and \( Z \equiv \sum_{k=1}^n Z_k \), where the \( X_k \)'s (and thus the \( Z_k \)'s) are independent by Poissonization, and \( X_k \sim \text{Poisson}(mD(k)) \). It is not hard to see that \( \mathbb{E}Z_k = \Delta_k^2 \), where \( \Delta_k \equiv \left( \frac{1}{n} - D(k) \right) \), so that \( \mathbb{E}Z = m^2\|D - \mathcal{U}\|^2_2 \). Furthermore, we also get

\[
\text{Var} Z_k = 2m^2 \left( \frac{1}{n} - \Delta_k \right)^2 + 4m^3 \left( \frac{1}{n} - \Delta_k \right) \Delta_k
\]
so that
\[ \text{Var } Z = 2m^2 \left( \sum_{k=1}^{n} \Delta_k^2 + \frac{1}{n} - 2m \sum_{k=1}^{n} \Delta_k^3 \right) \]  \hspace{1cm} (B.1)
(after expanding and since \( \sum_{k=1}^{n} \Delta_k = 0 \)).

**Soundness.** Almost straight from [19], but the threshold has changed. Assume \( \Delta^2 \overset{\text{def}}{=} \|D - U\|_2^2 \geq \varepsilon^2 / n \); we will show that \( \Pr[Z < \tau] \leq 1/3 \). By Chebyshev’s inequality, it is sufficient to show that \( \tau \leq \mathbb{E}[Z] - \sqrt{3\text{Var } Z} \), as
\[
\Pr\left[ \mathbb{E}[Z] - Z > \sqrt{3\text{Var } Z} \right] \leq 1/3 .
\]
As \( \tau < \frac{3}{4} \mathbb{E}[Z] \), arguing that \( \sqrt{3\text{Var } Z} \leq \frac{1}{4} \mathbb{E}[Z] \) is enough, i.e. that \( 48 \text{Var } Z \leq (\mathbb{E}[Z])^2 \). From (B.1), this is equivalent to showing
\[
\Delta^2 + \frac{1}{n} - 2m \sum_{k=1}^{n} \Delta_k^3 \leq \frac{m^2 \Delta^4}{96} .
\]
We bound the LHS term by term.

- As \( \Delta^2 \geq \frac{\varepsilon^2}{n} \), we get \( m^2 \Delta^2 \geq \frac{C^2}{\varepsilon^2} \), and thus \( \frac{m^2 \Delta^4}{288} \geq \frac{C^2}{288 \varepsilon^2} \Delta^2 \geq \Delta^2 \) (as \( C \geq 17 \) and \( \varepsilon \leq 1 \)).
- Similarly, \( \frac{m^2 \Delta^4}{288} \geq \frac{C^2}{288 \varepsilon^2} \cdot \frac{\varepsilon^2}{n} \geq \frac{1}{n} \).
- Finally, recalling that\(^1\)
\[
\sum_{k=1}^{n} |\Delta_k|^3 \leq \left( \sum_{k=1}^{n} |\Delta_k|^2 \right)^{3/2} = \Delta^3
\]
we get that \( 2m \sum_{k=1}^{n} |\Delta_k|^3 \leq 2m \Delta^3 = \frac{m^2 \Delta^4}{288} \cdot \frac{288}{m \Delta} \leq \frac{m^2 \Delta^4}{288} \), using the fact that \( \frac{m \Delta}{288} \geq \frac{C}{576} \geq 1 \) (by choice of \( C \geq 576 \)).

Overall, the LHS is at most \( 3 \cdot \frac{m^2 \Delta^4}{288} = \frac{m^2 \Delta^4}{96} \), as claimed.

---

\(^1\) For any sequence \( x = (x_1, \ldots, x_n) \in \mathbb{R}^n \), \( p > 0 \) \( \iff \|x\|_p \) is non-increasing. In particular, for \( 0 < p \leq q < \infty \),
\[
\left( \sum_{i} |x_i|^q \right)^{1/q} = \|x\|_q \leq \|x\|_p = \left( \sum_{i} |x_i|^p \right)^{1/p} .
\]

56
Completeness. Assume $\Delta^2 = \|D - \mathcal{U}\|_2^2 < \varepsilon^2/(4n)$. We need to show that $\Pr[Z \geq \tau] \leq 1/3$. Chebyshev's inequality implies

$$\Pr[Z - EZ > \sqrt{3}\sqrt{\text{Var} Z}] \leq 1/3$$

and therefore it is sufficient to show that

$$\tau \geq EZ + \sqrt{3}\sqrt{\text{Var} Z}$$

Recalling the expressions of $EZ$ and $\text{Var} Z$ from (B.1), this is tantamount to showing

$$\frac{3m^2\varepsilon^2}{4n} \geq m^2\Delta^2 + \sqrt{6m\Delta^2 + \frac{1}{n} - 2m \sum_{k=1}^{n}\Delta_k^3}$$

or equivalently

$$\frac{3m^2\varepsilon^2}{4\sqrt{n}} \geq m\sqrt{n}\Delta^2 + \sqrt{6\left[1 + n\Delta^2 - 2nm\sum_{k=1}^{n}\Delta_k^3\right]}.$$

Since $\sqrt{1 + n\Delta^2 - 2nm\sum_{k=1}^{n}\Delta_k^3} \leq \sqrt{1 + n\Delta^2} \leq \sqrt{1 + \varepsilon^2/4} \leq \sqrt{5}/4$, we get that the second term is at most $\sqrt{30}/4 < 3$. All that remains is to show that $m\sqrt{n}\Delta^2 \geq 3m\frac{\varepsilon^2}{4\sqrt{n}} - 3$. But as $\Delta^2 < \varepsilon^2/(4n)$, $m\sqrt{n}\Delta^2 \leq m\frac{\varepsilon^2}{4\sqrt{n}}$; and our choice of $m \geq C \cdot \frac{\sqrt{n}}{\varepsilon^2}$ for some absolute constant $C \geq 6$ ensures this holds. \hfill \Box

---

To see why, one can easily prove that if $\|x\|_p = 1$, then $\|x\|_q^q \leq 1$ (bounding each term $|x_i|^q \leq |x_i|^p$), and therefore $\|x\|_q \leq 1 = \|x\|_p$. Next, for the general case, apply this to $y = x/\|x\|_p$, which has unit $\ell_p$ norm, and conclude by homogeneity of the norm.
B.2 Proof of Theorem 3.1.7

In this section, we prove our structural result for MHR distributions, Theorem 3.1.7:

**Theorem 3.1.8** (k-histogram). The class of k-histogram distributions on $[n]$ is $(0,k)$-decomposable.

**Proof.** We reproduce and adapt the argument of [11, Section 5.1] to meet our definition of decomposability (which, albeit related, is incomparable to theirs). First, we modify the algorithm at the core of their constructive proof, in Algorithm 3: note that the only two changes are in Steps 2 and 3, where we use parameters respectively $\frac{3}{n}$ and $\frac{2}{n^2}$. Following

```
Algorithm 3 DECOMPOSE-MHR'(D, γ)

Require: explicit description of MHR distribution D over [n]; accuracy parameter γ > 0
1: Set J ← [n] and Q ← ∅.
2: Let I ← RIGHT-INTERVAL(D, J, $\frac{3}{n}$) and I' ← RIGHT-INTERVAL(D, J \ I, $\frac{2}{n^2}$). Set J ← J \ (I ∪ I').
3: Set i ∈ J to be the smallest integer such that $D(i) \geq \frac{2}{n^2}$. If no such i exists, let $I''$ ← J and go to Step 9. Otherwise, let $I''$ ← \{1, ..., i − 1\} and J ← J \ $I''$.
4: while J ≠ ∅ do
5:   Let j ∈ J be the smallest integer such that $D(j) \notin \left[\frac{1}{1+γ}, 1 + γ\right]D(i)$. If no such j exists, let $I''$ ← J; otherwise let $I'''$ ← \{i, ..., j − 1\}.
6:   Add $I'''$ to Q and set J ← J \ $I'''$.
7:   Let i ← j.
8: end while
9: Return Q ∪ \{I, I', I''\}
```

the structure of their proof, we write $Q = \{I_1, ..., I_{|Q|}\}$ with $I_i = [a_i, b_i]$, and define $Q' = \{I_i \in Q : D(a_i) > D(a_{i+1})\}$, $Q'' = \{I_i \in Q : D(a_i) \leq D(a_{i+1})\}$.

We immediately obtain the analogues of their Lemmas 5.2 and 5.3:

**Lemma B.2.1.** We have $\prod_{I_i \in Q'}$ $\frac{D(a_i)}{D(a_{i+1})} \leq \frac{n}{γ}$.

**Lemma B.2.2.** Step 4 of Algorithm 3 adds at most $O\left(\frac{1}{ε} \log \frac{n}{ε}\right)$ intervals to Q.
Sketch. This derives from observing that now $D(I \cup I') \geq \gamma/n$, which as in [11, Lemma 5.3] in turn implies

$$1 \geq \frac{\gamma}{n} (1 + \gamma)^{|Q'| - 1}$$

so that $|Q'| = O\left(\frac{1}{\varepsilon} \log \frac{n}{\varepsilon}\right)$.

Again following their argument, we also get

$$\frac{D(a_{iQ+1})}{D(a_i)} = \prod_{i \in Q'} \frac{D(a_{i+1})}{D(a_i)}, \prod_{i \in Q'} \frac{D(a_{i+1})}{D(a_i)}$$

by combining Lemma B.2.1 with the fact that $D(a_{iQ+1}) \leq 1$ and that by construction $D(a_i) \geq \gamma/n^2$, we get

$$\prod_{i \in Q'} \frac{D(a_{i+1})}{D(a_i)} \leq \frac{n}{\gamma} \cdot \frac{n^2}{\gamma} = \frac{n^3}{\gamma}.$$ 

But since each term in the product is at least $(1 + \gamma)$ (by construction of $Q$ and the definition of $Q''$), this leads to

$$(1 + \gamma)^{|Q'|} \leq \frac{n^3}{\gamma}$$

and thus $|Q''| = O\left(\frac{1}{\varepsilon} \log \frac{n}{\varepsilon}\right)$ as well. 

It remains to show that $Q \cup \{I, I', I''\}$ is indeed a good decomposition of $[n]$ for $D$, as per Theorem 3.1.1. Since by construction every interval in $Q$ satisfies item (ii), we only are left with the case of $I, I'$ and $I''$. For the first two, as they were returned by RIGHT-INTERVAL either (a) they are singletons, in which case item (ii) trivially holds; or (b) they have at least two elements, in which case they have probability mass at most $\frac{2}{n}$ (by the choice of parameters for RIGHT-INTERVAL) and thus item (i) is satisfied. Finally, it is immediate to see that by construction $D(I'') \leq n \cdot \gamma/n^2 = \gamma/n$, and item (i) holds as well. 

$\square$
Bibliography


61


[18] Constantinos Daskalakis and Christos Papadimitriou. Sparse covers for sums of indicators. *ArXiv*, June 2013. 3.4.8

[19] Ilias Diakonikolas, Daniel M. Kane, and Vladimir Nikishkin. Testing Identity of Structured Distributions. In *Proceedings of SODA*, 2015. 1.1.6, 1.2.1, B.1


