Testing Shape Restrictions of Discrete Distributions

Clément L. Canonne* Ilias Diakonikolas† Themis Gouleakis‡ Ronitt Rubinfeld§

July 30, 2015

Abstract

We study the question of testing *structured* properties (classes) of discrete distributions. Specifically, given sample access to an arbitrary distribution D over $[n]$ and a property $\mathcal{P}$, the goal is to distinguish between $D \in \mathcal{P}$ and $\ell_1(D, \mathcal{P}) > \varepsilon$. We develop a general algorithm for this question, which applies to a large range of "shape-constrained" properties, including monotone, log-concave, t-modal, piecewise-polynomial, and Poisson Binomial distributions. Moreover, for all cases considered, our algorithm has near-optimal sample complexity with regard to the domain size and is computationally efficient. For most of these classes, we provide the first non-trivial tester in the literature. In addition, we also describe a generic method to prove lower bounds for this problem, and use it to show our upper bounds are nearly tight. Finally, we extend some of our techniques to tolerant testing, deriving nearly–tight upper and lower bounds for the corresponding questions.

1 Introduction

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. Such assumptions translate as constraints on the probability distribution – e.g., it is supposed to be Gaussian, or to meet a smoothness or "fat tail" condition (see e.g., [Man63, Hou86, TLSM95]).

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* [BDBB72, SS01] and *model selection* [MP07]. 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 [SN99, BB05, Wal09]. From a different perspective, a recent line of work in

*Columbia University. Email: ccanonne@cs.columbia.edu. Research supported by NSF CCF-1115703 and NSF CCF-1319788.

†University of Edinburgh. Email: ilias.d@ed.ac.uk. Research supported by EPSRC grant EP/L021749/1, a Marie Curie Career Integration Grant, and a SICSA grant. This work was performed in part while visiting CSAIL, MIT.

‡CSAIL, MIT. Email: tgoule@mit.edu.

§CSAIL, MIT and the Blavatnik School of Computer Science, Tel Aviv University. Email: ronitt@csail.mit.edu.

Theoretical Computer Science, originating from the papers of Batu et al. [BFR+00, BFF+01, GR00] has also been tackling similar questions in the setting of property testing (see [Ron08, Ron10, Rub12] 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 [BDKR05, GMV06, AAK+07, DDS+13, CDVV14, AD15, DKN15b]. 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 [VV10, VV11a, VV14].

In this paper, we focus on the following general 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) $\|D - D'\|_1 > \varepsilon$ 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* property testing problems [Val11, VV11a, DKN15b, DKN15a]. This direction, the testing analog of a similar movement in distribution learning [CDSS13, CDSS14b, CDSS14a, ADLS15], 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. In this work, we make significant progress in this direction by providing a unified framework for the question of testing various properties of probability distributions. More specifically, we describe a generic technique to obtain upper bounds on the sample complexity of this question, which applies to a broad range of structured classes. Our technique yields sample near-optimal and computationally efficient testers for a wide range of distribution families. Conversely, we also develop a general approach to prove lower bounds on these sample complexities, and use it to derive tight or nearly tight bounds for many of these classes.

Related work. Batu et al. [BKR04] initiated the study of efficient property testers for monotonicity and obtained (nearly) matching upper and lower bounds for this problem; while [AD15] later considered testing the class of Poisson Binomial Distributions, and settled the sample complexity of this problem (up to the precise dependence on ε). Indyk, Levi, and Rubinfeld [ILR12], focusing on distributions that are piecewise constant on t intervals ("t-histograms") described a $\tilde{O}(\sqrt{tn}/\varepsilon^5)$-sample algorithm for testing membership to this class. Another body of work by [BDKR05], [BKR04], and [DDS+13] 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 [BKR04] 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. Daskalakis et al. [DDS+13] later gave $\tilde{O}(\sqrt{\log n})$ and $\tilde{O}(\log^{2/3} n)$-sample testing algorithms for testing respectively identity and closeness of monotone distributions, and obtained similar results for k-modal distributions. Finally, we briefly mention two related results, due respectively to [BDKR05] and [DDS12a]. The first one states that for the task of getting a multiplicative *estimate* of the entropy of a distribution, assuming monotonicity enables exponential savings in sample complexity – $O(\log^6 n)$, instead of $\Omega(n^c)$ for the general case. The second describes how to test if an unknown k-modal distribution is in fact monotone, using only $O(k/\varepsilon^2)$ samples. Note that the latter line of work differs from ours in that it *presupposes* the distributions satisfy some structural property, and uses this knowledge to test something else about the distribution; while we are given *a priori* arbitrary distributions, and must

check whether the structural property holds. Except for the properties of monotonicity and being a PBD, nothing was previously known on testing the shape restricted properties that we study. Independently and concurrently to this work, Acharya, Daskalakis, and Kamath obtained a sample near-optimal efficient algorithm for testing log-concavity.[1]

Moreover, for the specific problems of identity and closeness testing,[2] recent results of [DKN15b, DKN15a] describe 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 theirs, 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.

1.1 Results and Techniques

Upper Bounds. A natural way to tackle our membership testing problem would be to first learn the unknown distribution D *as if* it satisfied the property, before checking if the hypothesis obtained is indeed both close to the original distribution and to the property. Taking advantage of the purported structure, the first step could presumably be conducted with a small number of samples; things break down, however, in the second step. Indeed, most approximation results leading to the improved learning algorithms one would apply in the first stage only provide very weak guarantees, in the ℓ_1 sense. For this reason, they lack the robustness that would be required for the second part, where it becomes necessary to perform *tolerant* testing between the hypothesis and D – a task that would then entail a number of samples almost linear in the domain size. To overcome this difficulty, we need to move away from these global ℓ_1 closeness results and instead work with stronger requirements, this time in ℓ_2 norm.

At the core of our approach is an idea of Batu et al. [BKR04], which show that monotone distributions can be well-approximated (in a certain technical sense) by piecewise constant densities on a suitable interval partition of the domain; and leverage this fact to reduce monotonicity testing to uniformity testing on each interval of this partition. While the argument of [BKR04] is tailored specifically for the setting of monotonicity testing, we are able to abstract the key ingredients, and obtain a generic membership tester that applies to a wide range of distribution families. In more detail, we provide a testing algorithm which applies to any class of distributions which admits succinct approximate decompositions – that is, each distribution in the class can be well-approximated (in a strong ℓ_2 sense) by piecewise constant densities on a small number of intervals (we hereafter refer to this approximation property, formally defined in Definition 3.1, as (Succinctness); and extend the notation to apply to any *class* $\mathcal{C}$ of distributions for which all $D \in \mathcal{C}$ satisfy (Succinctness)). 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*. Specific examples are given in the corollaries below. Informally, our main algorithmic result, informally stated (see Theorem 3.3 for a detailed formal statement), is as follows:

[1] Following the communication of a preliminary version of this paper (February 2015), we were informed that [ADK15] subsequently obtained near-optimal testers for some of the classes we consider. To the best of our knowledge, their work builds on ideas from [AD15] and their techniques are orthogonal to ours.

[2] Recall that the identity testing problem asks, 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; while in closeness testing both distributions to compare are unknown.

Theorem 1.1 (Main Theorem)**.** *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 computationally efficient, and its sample complexity is optimal (up to logarithmic factors and the exact dependence on ε).*

We then instantiate this result to obtain "out-of-the-box" *computationally efficient* testers for several classes of distributions, by showing that they satisfy the premise of our theorem (the definition of these classes is given in Section 2.1):

Corollary 1.2. *The algorithm* TestSplittable *can test the classes of monotone, unimodal, log-concave, concave, convex, and monotone hazard rate (MHR) distributions, with $\tilde{O}\left(\sqrt{n}/\varepsilon^{7/2}\right)$ samples.*

Corollary 1.3. *The algorithm* TestSplittable *can test the class of t-modal distributions, with $\tilde{O}\left(\sqrt{tn}/\varepsilon^{7/2}\right)$ samples.*

Corollary 1.4. *The algorithm* TestSplittable *can test the classes of t-histograms and t-piecewise degree-d distributions, with $\tilde{O}\left(\sqrt{tn}/\varepsilon^{3}\right)$ and $\tilde{O}\left(\sqrt{t(d+1)n}/\varepsilon^{7/2} + t(d+1)/\varepsilon^{3}\right)$ samples respectively.*

Corollary 1.5. *The algorithm* TestSplittable *can test the classes of Binomial and Poisson Binomial Distributions, with $\tilde{O}\left(n^{1/4}/\varepsilon^{7/2}\right)$ samples.*

We remark that the aforementioned sample upper bounds are information-theoretically near-optimal in the domain size n (up to logarithmic factors). See Table 1 and the following subsection for the corresponding lower bounds. We did not attempt to optimize the dependence on the parameter ε, though a more careful analysis can lead to such improvements.

We stress that prior to our work, no non-trivial testing bound was known for most of these classes – specifically, our nearly-tight bounds for t-modal with $t > 1$, log-concave, concave, convex, MHR, and piecewise polynomial distributions are new. Moreover, although a few of our applications were known in the literature (the $\tilde{O}(\sqrt{n}/\varepsilon^6)$ upper and $\Omega(\sqrt{n}/\varepsilon^2)$ lower bounds on testing monotonicity can be found in [BKR04], while the $\Theta(n^{1/4})$ sample complexity of testing PBDs was recently given[3] in [AD15], and the task of testing t-histograms is considered in [ILR12]), the crux here is that we are able to derive them in a *unified* way, by applying the same generic algorithm to all these different distribution families. We note that our upper bound for t-histograms (Corollary 1.4) also improves on the previous $\tilde{O}\left(\sqrt{tn}/\varepsilon^5\right)$-sample tester, as long as $t = \tilde{O}\left(n^{1/3}/\varepsilon^2\right)$. In addition to its generality, our framework yields much cleaner and conceptually simpler proofs of the upper and lower bounds from [AD15].

Extension to Mixtures of Structured Distributions. We finally note that, in contrast to the case-specific analogous results of [BKR04, AD15], Corollary 1.2, Corollary 1.3, Corollary 1.4,

[3]For the sample complexity of testing monotonicity, [BKR04] originally states an $\tilde{O}\left(\sqrt{n}/\varepsilon^4\right)$ upper bound, but the proof seems to only result in an $\tilde{O}\left(\sqrt{n}/\varepsilon^6\right)$ bound. Regarding the class of PBDs, [AD15] obtain an $n^{1/4} \cdot \tilde{O}\left(1/\varepsilon^2\right) + \tilde{O}\left(1/\varepsilon^6\right)$ sample complexity, to be compared with our $\tilde{O}\left(n^{1/4}/\varepsilon^{7/2}\right) + O\left(\log^4 n/\varepsilon^4\right)$ upper bound; as well as an $\Omega\left(n^{1/4}/\varepsilon^2\right)$ lower bound.

Class	Upperbound	Lowerbound
Monotone	$\tilde{O}\left(\frac{\sqrt{n}}{\varepsilon^6}\right)$ [BKR04], $\tilde{O}\left(\frac{\sqrt{n}}{\varepsilon^{7/2}}\right)$ (Corollary 1.2)	$\Omega\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ [BKR04], $\Omega\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ (Corollary 1.7)
Unimodal	$\tilde{O}\left(\frac{\sqrt{n}}{\varepsilon^{7/2}}\right)$ (Corollary 1.2)	$\Omega\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ (Corollary 1.7)
t-modal	$\tilde{O}\left(\frac{\sqrt{tn}}{\varepsilon^{7/2}}\right)$ (Corollary 1.3)	$\Omega\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ (Corollary 1.7)
Log-concave, concave, convex	$\tilde{O}\left(\frac{\sqrt{n}}{\varepsilon^{7/2}}\right)$ (Corollary 1.2)	$\Omega\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ (Corollary 1.7)
Monotone Hazard Rate (MHR)	$\tilde{O}\left(\frac{\sqrt{n}}{\varepsilon^{7/2}}\right)$ (Corollary 1.2)	$\Omega\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ (Corollary 1.7)
Binomial, Poisson Binomial (PBD)	$\tilde{O}\left(\frac{n^{1/4}}{\varepsilon^2}+\frac{1}{\varepsilon^6}\right)$ [AD15], $\tilde{O}\left(\frac{n^{1/4}}{\varepsilon^{7/2}}\right)$ (Corollary 1.5)	$\Omega\left(\frac{n^{1/4}}{\varepsilon^2}\right)$ ([AD15], Corollary 1.8)
k-mixtures of monotone, unimodal, log-concave, concave, convex, and/or MHR	$\tilde{O}\left(\frac{\sqrt{kn}}{\varepsilon^{7/2}}+\frac{k}{\varepsilon^3}\right)$ (Corollary 1.6)	
k-mixtures of t-modal	$\tilde{O}\left(\frac{\sqrt{ktn}}{\varepsilon^{7/2}}+\frac{kt}{\varepsilon^3}\right)$ (Corollary 1.6)	
k-mixtures of Binomials and/or PBD	$\tilde{O}\left(\frac{\sqrt{kn}}{\varepsilon^{7/2}}+\frac{k}{\varepsilon^3}\right)$ (Corollary 1.5)	
t-histograms	$\tilde{O}\left(\frac{\sqrt{tn}}{\varepsilon^5}\right)$ [ILR12], $\tilde{O}\left(\frac{\sqrt{tn}}{\varepsilon^3}\right)$ (Corollary 1.4)	$\Omega(\sqrt{tn})$ for $t \le \frac{1}{\varepsilon}$ [ILR12], $\Omega\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ (Corollary 1.7)
t-piecewise degree-d	$\tilde{O}\left(\frac{\sqrt{t(d+1)n}}{\varepsilon^{7/2}}+\frac{t(d+1)}{\varepsilon^3}\right)$ (Corollary 1.4)	$\Omega\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ (Corollary 1.7)
k-SIIRV		$\Omega(k^{1/2}n^{1/4})$ (Corollary 1.9)

Table 1: Summary of results.

and Corollary 1.5 easily extend to *mixtures* of distributions from the respective classes, incurring a mild dependence on the number of components. Specifically, our main theorem implies the following:

Corollary 1.6. *Let $\mathcal{C}_k$ be the set of mixtures of (at most) k distributions from $\mathcal{C}$, where $\mathcal{C}$ is any union of the classes of monotone, unimodal, log-concave, concave, convex, and monotone hazard rate (MHR) distributions. Then, $\mathcal{C}_k$ can be tested with $\tilde{O}(\sqrt{kn}/\varepsilon^{7/2}+k/\varepsilon^3)$ samples. Similarly, the class $\mathcal{M}_{t,k}$ of mixtures of (at most) k t-modal distributions can be tested with $\tilde{O}(\sqrt{ktn}/\varepsilon^{7/2}+kt/\varepsilon^3)$ samples.*

Lower Bounds. To complement our upper bounds, we give a generic framework for proving lower bounds against testing classes of distributions. In more detail, we describe how to *reduce* – under a mild assumption on the property $\mathcal{C}$ – the problem of testing *membership to* $\mathcal{C}$ ("does $D \in \mathcal{C}$?") to testing *identity to* D^* ("does $D = D^*$?"), for any explicit distribution D^* in $\mathcal{C}$. While these two problems need not in general be related,[4] we show that our reduction-based approach applies to a large number of natural properties, and obtain lower bounds that nearly match our upper bounds for all of them. Moreover, this lets us derive a simple proof of the lower bound of [AD15] on testing the class of PBDs. The reader is referred to Theorem 6.1 for the formal statement of our

[4] As a simple example, consider the class $\mathcal{C}$ of *all* distributions, for which testing membership is trivial.

reduction-based lower bound theorem. In this section, we state the concrete corollaries we obtain for specific structured distribution families:

Corollary 1.7. *Testing log-concavity, convexity, concavity, MHR, unimodality, t-modality, t-histograms, and t-piecewise degree-d distributions each require $\Omega(\sqrt{n}/\varepsilon^2)$ samples (the last three for $t = o(\sqrt{n})$ and $t(d+1) = o(\sqrt{n})$, respectively), for any $\varepsilon \geq 1/n^{O(1)}$.*

Corollary 1.8. *Testing the classes of Binomial and Poisson Binomial Distributions each require $\Omega\Big(n^{1/4}/\varepsilon^2\Big)$ samples, for any $\varepsilon \geq 1/n^{O(1)}$.*

Corollary 1.9. *There exist absolute constants $c > 0$ and $\varepsilon_0 > 0$ such that testing the class of k-SIIRV distributions requires $\Omega(k^{1/2}n^{1/4})$ samples, for any $k = o(n^c)$ and $\varepsilon \leq \varepsilon_0$.*

Tolerant Testing. Using our techniques, we also establish nearly–tight upper and lower bounds on tolerant testing for shape restrictions. Similarly, our upper and lower bounds are matching as a function of the domain size. More specifically, we give a simple generic upper bound approach (namely, a learning followed by tolerant testing algorithm). Our tolerant testing lower bounds follow the same reduction-based approach as in the non-tolerant case. In more detail, our results are as follows (see Section 6 and Section 7):

Corollary 1.10. *Tolerant testing of log-concavity, convexity, concavity, MHR, unimodality, and t-modality can be performed with $O(\frac{1}{(\varepsilon_2-\varepsilon_1)^2}\frac{n}{\log n})$ samples, for $\varepsilon_2 \geq C\varepsilon_1$ (where $C > 2$ is an absolute constant).*

Corollary 1.11. *Tolerant testing of the classes of Binomial and Poisson Binomial Distributions can be performed with $O(\frac{1}{(\varepsilon_2-\varepsilon_1)^2}\frac{\sqrt{n\log(1/\varepsilon_1)}}{\log n})$ samples, for $\varepsilon_2 \geq C\varepsilon_1$ (where $C > 2$ is an absolute constant).*

Corollary 1.12. *Tolerant testing of log-concavity, convexity, concavity, MHR, unimodality, and t-modality each require $\Omega\Big(\frac{1}{(\varepsilon_2-\varepsilon_1)}\frac{n}{\log n}\Big)$ samples (the latter for $t = o(n)$).*

Corollary 1.13. *Tolerant testing of the classes of Binomial and Poisson Binomial Distributions each require $\Omega\Big(\frac{1}{(\varepsilon_2-\varepsilon_1)}\frac{\sqrt{n}}{\log n}\Big)$ samples.*

On the scope of our results. We point out that our main theorem is likely to apply to many other classes of structured distributions, due to the mild structural assumptions it requires. However, we did not attempt here to be comprehensive; but rather to illustrate the generality of our approach. Moreover, for all properties considered in this paper the generic upper and lower bounds we derive through our methods turn out to be optimal up to at most polylogarithmic factors (with regard to the support size). The reader is referred to Table 1 for a summary of our results and related work.

[4] *Tolerant testing* of a property $\mathcal{P}$ is defined as follows: given $0 \leq \varepsilon_1 < \varepsilon_2 \leq 1$, one must distinguish between (a) $\ell_1(D,\mathcal{P}) \leq \varepsilon_1$ and (b) $\ell_1(D,\mathcal{P}) \geq \varepsilon_2$. This turns out to be, in general, a much harder task than that of "regular" testing (where we take $\varepsilon_1 = 0$).

1.2 Organization of the Paper

We start by giving the necessary background and definitions in Section 2, before turning to our main result, the proof of Theorem 1.1 (our general testing algorithm) in Section 3. In Section 4, we establish the necessary structural theorems for each classes of distributions considered, enabling us to derive the upper bounds of Table 1. Section 5 introduces a slight modification of our algorithm which yields stronger testing results for classes of distributions with small effective support, and use it to derive Corollary 1.5, our upper bound for Poisson Binomial distributions. Second, Section 6 contains the details of our lower bound methodology, and of its applications to the classes of Table 1. Finally, Section 6.2 is concerned with the extension of this methodology to *tolerant* testing, of which Section 7 describes a generic upper bound counterpart.

2 Notation and Preliminaries

2.1 Definitions

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 $D(1) \geq D(2) \geq \dots D(n)$. 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:[5]

Definition 2.1 (t-modal)**.** Fix any distribution D over $[n]$, and integer t. D is said to have t *modes* if there exists a sequence $i_0 < \dots < i_{t+1}$ 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 2.2 (Log-Concave)**.** A distribution D over $[n]$ is said to be *log-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 $1 < k < n$, $D(k)^2 \geq D(k-1)D(k+1)$. We write $\mathcal{L}$ for the class of all log-concave distributions (omitting the dependence on n).

Definition 2.3 (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 Definition 2.2 and Definition 2.3, condition (i) is equivalent to enforcing that the distribution be supported on an interval.

Definition 2.4 (Monotone Hazard Rate)**.** A distribution D over $[n]$ is said to have *monotone hazard rate* (MHR) if its *hazard rate* $H(i) \stackrel{\text{def}}{=} \frac{D(i)}{\sum_{j=i}^{n} D(j)}$ is a non-decreasing function. We write $\mathcal{MHR}$ for the class of all MHR distributions (omitting the dependence on n).

[5] Note that this slightly deviates from the Statistics literature, where only the peaks are counted as modes (so that what is usually referred to as a bimodal distribution is, according to our definition, 3-modal).

It is known that every log-concave distribution is both unimodal and MHR (see e.g. [An96, Proposition 10]), and that monotone distributions are MHR. Two other classes of distributions have elicited significant interest in the context of density estimation, that of *histograms* (piecewise constant) and *piecewise polynomial densities*:

Definition 2.5 (Piecewise Polynomials [CDSS14a])**.** A distribution D over $[n]$ is said to be a *t-piecewise degree-d distribution* if there is a partition of $[n]$ into t disjoint intervals $I_1, \ldots, I_t$ such that $D(i) = p_j(i)$ for all $i \in I_j$, where each $p_1, \ldots p_t$ is a univariate polynomial of degree at most d. We write $\mathcal{P}_{t,d}$ for the class of all t-piecewise degree-d distributions (omitting the dependence on n). (We note that t-piecewise degree-0 distributions are also commonly referred to as *t-histograms*, and write $\mathcal{H}_t$ for $\mathcal{P}_{t,0}$.)

Finally, we recall the definition of the two following classes, which both extend the family of Binomial distributions $\mathcal{BIN}_n$: the first, by removing the need for each of the independent Bernoulli summands to share the same bias parameter.

Definition 2.6. 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.

It is not hard to show that Poisson Binomial Distributions are in particular log-concave. One can generalize even further, by allowing each random variable of the summation to be integer-valued:

Definition 2.7. 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 k-$\mathcal{SIIRV}_n$ the class of all such k-SIIRVs.

2.2 Tools from previous work

We first restate a result of Batu et al. relating closeness to uniformity in ℓ_2 and ℓ_1 norms to "overall flatness" of the probability mass function, and which will be one of the ingredients of the proof of Theorem 1.1:

Lemma 2.8 ([BFR+00, BFF+01])**.** *Let D be a distribution on a domain S. (a) If $\max_{i \in S} D(i) \leq (1+\varepsilon) \min_{i \in S} D(i)$, then $\|D\|_2^2 \leq (1+\varepsilon^2)/|S|$. (b) If $\|D\|_2^2 \leq (1+\varepsilon^2)/|S|$, then $\|D - \mathcal{U}_S\|_1 \leq \varepsilon$.*

To check condition (b) above we shall rely on the following, which one can derive from the techniques in [DKN15b] and whose proof we defer to Appendix A:

Lemma 2.9 (Adapted from [DKN15b, Theorem 11])**.** *There exists an algorithm* Check-Small-ℓ_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 $\|D - \mathcal{U}_I\|_2 > \varepsilon/\sqrt{|I|}$, then the algorithm outputs* no *with probability at least $1 - \delta$;*
- *If $\|D - \mathcal{U}_I\|_2 \leq \varepsilon/2\sqrt{|I|}$, then the algorithm outputs* yes *with probability at least $1 - \delta$.*

Finally, we will also rely on a classical result from Probability, the *Dvoretzky–Kiefer–Wolfowitz* (DKW) inequality, restated below:

Theorem 2.10 ([DKW56, Mas90])**.** *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:*

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

Then, for all $\varepsilon > 0$, $\Pr\Big[\, \|D - \hat{D}\|_{\text{Kol}} > \varepsilon \,\Big] \leq 2e^{-2m\varepsilon^2}$, where $\|\cdot - \cdot\|_{\text{Kol}}$ denotes the Kolmogorov distance (i.e., the ℓ_∞ distance between cumulative distribution functions).

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

3 The General Algorithm

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

Theorem 1.1 (Main Theorem)**.** *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 computationally efficient, and its sample complexity is optimal (up to logarithmic factors and the exact dependence on ε).*

Intuition. Before diving into the proof of this theorem, we first provide a high-level description of the argument. The algorithm proceeds in 3 stages: the first, the *decomposition step*, attempts to recursively construct a partition of the domain in a small number of intervals, with a very strong guarantee. If the decomposition succeeds, then the unknown distribution D will be close (in ℓ_1 distance) to its "flattening" on the partition; while if it fails (too many intervals have to be created), this serves as evidence that D does not belong to the class and we can reject. The second stage, the *approximation step*, then learns this flattening of the distribution – which can be done with few samples since by construction we do not have many intervals. The last stage is purely computational, the *projection step*: where we verify that the flattening we have learned is indeed close to the class $\mathcal{C}$. If all three stages succeed, then by the triangle inequality it must be the case that D is close to $\mathcal{C}$; and by the structural assumption on the class, if $D \in \mathcal{C}$ then it will admit succinct enough partitions, and all three stages will go through.

Turning to the proof, we start by defining formally the "structural criterion" we shall rely on, before describing the algorithm at the heart of our result in Section 3.1. (We note that a modification of this algorithm will be described in Section 5, and will allow us to derive Corollary 1.5.)

Definition 3.1 (Decompositions)**.** Let $\gamma > 0$ and $L = L(\gamma, n) \geq 1$. A class of distributions $\mathcal{C}$ on $[n]$ is said to be (γ, L)-*decomposable* if for every $D \in \mathcal{C}$ there exists $\ell \leq L$ and a partition $\mathcal{I}(\gamma, D) = (I_1, \ldots, I_\ell)$ of $[n]$ such that, for all $j \in [\ell]$, one of the following holds:

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

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

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

Lemma 3.2. *If $\mathcal{C}$ is (γ, L)-decomposable, then it is $(\gamma, O(L \log n))$-splittable.*

Proof. We will begin by proving a claim that for every partition $\mathcal{I} = \{I_1, I_2, ...I_L\}$ 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 ℓ 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 $\frac{b-a}{4}$.

We will also use the fact that, for every $\ell' \leq \ell$,

$$m \cdot 2^\ell = m \cdot 2^{\ell - \ell'} \cdot 2^{\ell'} = m' \cdot 2^{\ell'} \tag{1}$$

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 $[(m+1) \cdot 2^\ell + 1, b]$ (resp. $[a, m \cdot 2^\ell]$). 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 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 Definition 3.1 are closed under taking subsets. □

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

3.1 The algorithm

Theorem 1.1, and with it Corollary 1.2 and Corollary 1.3 will follow from the theorem below, combined with the structural theorems from Section 4:

Theorem 3.3. *Let $\mathcal{C}$ be a class of distributions over $[n]$ for which the following holds.*

1. *$\mathcal{C}$ is $(\gamma, L(\gamma, n))$-splittable;*
2. *there exists a procedure* ProjectionDist$_\mathcal{C}$ *which, given as input a parameter $\alpha \in (0, 1)$ and the explicit description of a distribution D over $[n]$, returns* yes *if the distance $\ell_1(D, \mathcal{C})$ to $\mathcal{C}$ is at most $\alpha/10$, and* no *if $\ell_1(D, \mathcal{C}) \geq 9\alpha/10$ (and either* yes *or* no *otherwise).*

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

3.2 Proof of Theorem 3.3

We now give the proof of our main result (Theorem 3.3), first analyzing the sample complexity of Algorithm 1 before arguing its correctness. For the latter, we will need the following simple lemma from [ILR12], restated below:

Algorithm 1 TestSplittable

Require: Domain I (interval), sample access to D over I; subroutine $\textsc{ProjectionDist}_{\mathcal{C}}$
Input: Parameters ε and function $L_{\mathcal{C}}(\cdot,\cdot)$.

Setting Up
  Define $\gamma \stackrel{\text{def}}{=} \frac{\varepsilon}{80}$, $L \stackrel{\text{def}}{=} L_{\mathcal{C}}(\gamma, |I|)$, $\kappa \stackrel{\text{def}}{=} \frac{\varepsilon}{160L}$, $\delta \stackrel{\text{def}}{=} \frac{1}{10L}$; and $c > 0$ be as in Lemma 2.9.
  Set $m \stackrel{\text{def}}{=} C \cdot \max\left(\frac{1}{\kappa}, \frac{\sqrt{L|I|}}{\varepsilon^3}\right) \cdot \log|I| = \tilde{O}\left(\frac{\sqrt{L|I|}}{\varepsilon^3} + \frac{L}{\varepsilon}\right)$ ▷ C is an absolute constant.
  Obtain a sequence $\mathbf{s}$ of m independent samples from D. ▷ For any $J \subseteq I$, let m_J be the number of samples falling in J.

Decomposition
  while $m_I \geq \max\left(c \cdot \frac{\sqrt{|I|}}{\varepsilon^2} \log\frac{1}{\delta}, \kappa m\right)$ and at most L splits have been performed **do**
    Run Check-Small-ℓ_2 (from Lemma 2.9) with parameters $\frac{\varepsilon}{40}$ and δ, using the samples of $\mathbf{s}$ belonging to I.
    if Check-Small-ℓ_2 outputs no **then**
      Bisect I, and recurse on both halves (using the same samples).
    end if
  end while
  if more than L splits have been performed **then**
    return REJECT
  else
    Let $\mathcal{I} \stackrel{\text{def}}{=} (I_1, \ldots, I_\ell)$ be the partition of $[n]$ from the leaves of the recursion. ▷ $\ell \leq L$.
  end if

Approximation
  Learn the flattening $\Phi(D, \mathcal{I})$ of D to ℓ_1 error $\frac{\varepsilon}{20}$ (with probability 1/10), using $O(\ell/\varepsilon^2)$ new samples. Let $\tilde{D}$ be the resulting hypothesis. ▷ $\tilde{D}$ is a ℓ-histogram.

Offline Check
  return ACCEPT if and only if $\textsc{ProjectionDist}_{\mathcal{C}}(\varepsilon, \tilde{D})$ returns yes. ▷ No sample needed.

Fact 3.4 ([ILR12, Fact 1]). *Let D be a distribution over $[n]$, and $\delta \in (0,1]$. Given $m \geq C \cdot \frac{\log\frac{n}{\delta}}{\eta}$ independent samples from D (for some absolute constant $C > 0$), with probability at least $1-\delta$ we have that, for every interval $I \subseteq [n]$:*

(i) if $D(I) \geq \frac{\eta}{4}$, then $\frac{D(I)}{2} \leq \frac{m_I}{m} \leq \frac{3D(I)}{2}$;
(ii) if $\frac{m_I}{m} \geq \frac{\eta}{2}$, then $D(I) > \frac{\eta}{4}$;
(iii) if $\frac{m_I}{m} < \frac{\eta}{2}$, then $D(I) < \eta$;

where $m_I \stackrel{\text{def}}{=} |\{\, j \in [m] : x_j \in I \,\}|$ is the number of the samples falling into I.

3.3 Sample complexity.

The sample complexity is immediate, and comes from Steps 4 and 20. The total number of samples is

$$m + O\left(\frac{\ell}{\varepsilon^2}\right) = O\left(\frac{\sqrt{|I| \cdot L}}{\varepsilon^3} \log |I| + \frac{L}{\varepsilon} \log |I| + \frac{L}{\varepsilon^2}\right) = O\left(\frac{\sqrt{|I| \cdot L}}{\varepsilon^3} \log |I| + \frac{L}{\varepsilon^2}\right).$$

3.4 Correctness.

Say an interval I considered during the execution of the "Decomposition" step is *heavy* if m_I is big enough on Step 7, and *light* otherwise; and let $\mathcal{H}$ and $\mathcal{L}$ denote the sets of heavy and light intervals respectively. By choice of m and a union bound over all $|I|^2$ possible intervals, we can assume on one hand that with probability at least $9/10$ the guarantees of Fact 3.4 hold simultaneously for all intervals considered. We hereafter condition on this event.

We first argue that if the algorithm does not reject in Step 13, then with probability at least $9/10$ we have $\|D - \Phi(D, \mathcal{I})\|_1 \le \varepsilon/20$. Indeed, we can write

$$\begin{aligned}\|D - \Phi(D,\mathcal{I})\|_1 &= \sum_{k\colon I_k \in \mathcal{L}} D(I_k) \cdot \|D_{I_k} - \mathcal{U}_{I_k}\|_1 + \sum_{k\colon I_k \in \mathcal{H}} D(I_k) \cdot \|D_{I_k} - \mathcal{U}_{I_k}\|_1 \\ &\le 2 \sum_{k\colon I_k \in \mathcal{L}} D(I_k) + \sum_{k\colon I_k \in \mathcal{H}} D(I_k) \cdot \|D_{I_k} - \mathcal{U}_{I_k}\|_1 .\end{aligned}$$

Let us bound the two terms separately.

- If $I' \in \mathcal{H}$, then by our choice of threshold we can apply Lemma 2.9 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

  $$\|D_{I'}\|_2^2 = \|D_{I'} - \mathcal{U}_{I'}\|_2^2 + \frac{1}{|I'|} \le \left(1 + \frac{\varepsilon^2}{1600}\right) \frac{1}{|I'|}$$

  as Check-Small-ℓ_2 returned yes; and by Lemma 2.8 this implies $\|D_{I'} - \mathcal{U}_{I'}\|_1 \le \varepsilon/40$.

- If $I' \in \mathcal{L}$, then we claim that $D(I') \le \max(\kappa, 2c \cdot \frac{\sqrt{|I'|}}{m\varepsilon^2} \log \frac{1}{\delta})$. Clearly, this is true if $D(I') \le \kappa$, so it only remains to show that $D(I') \le 2c \cdot \frac{\sqrt{|I'|}}{m\varepsilon^2} \log \frac{1}{\delta}$. But this follows from Fact 3.4 (i), as if we had $D(I') > 2c \cdot \frac{\sqrt{|I'|}}{m\varepsilon^2} \log \frac{1}{\delta}$ then $m_{I'}$ would have been big enough, and $I' \notin \mathcal{L}$. Overall,

  $$\sum_{I' \in \mathcal{L}} D(I') \le \sum_{I' \in \mathcal{L}} \left(\kappa + 2c \cdot \frac{\sqrt{|I'|}}{m\varepsilon^2} \log \frac{1}{\delta}\right) \le L\kappa + 2 \sum_{I' \in \mathcal{L}} c \cdot \frac{\sqrt{|I'|}}{m\varepsilon^2} \log \frac{1}{\delta} \le \frac{\varepsilon}{160} \left(1 + \sum_{I' \in \mathcal{L}} \sqrt{\frac{|I'|}{|I|\, L}}\right) \le \frac{\varepsilon}{80}$$

  for a sufficiently big choice of constant $C > 0$ in the definition of m; where we first used that $|\mathcal{L}| \le L$, and then that $\sum_{I' \in \mathcal{L}} \sqrt{\frac{|I'|}{|I|}} \le \sqrt{L}$ by Jensen's inequality.

Putting it together, this yields

$$\|D - \Phi(D,\mathcal{I})\|_1 \le 2 \cdot \frac{\varepsilon}{80} + \frac{\varepsilon}{40} \sum_{I' \in \mathcal{H}} D(I_k) \le \varepsilon/40 + \varepsilon/40 = \varepsilon/20.$$

Soundness. By contrapositive, we argue that if the test returns ACCEPT, then (with probability at least 2/3) D is ε-close to $\mathcal{C}$. Indeed, conditioning on $\tilde{D}$ being $\varepsilon/20$-close to $\Phi(D,\mathcal{I})$, we get by the triangle inequality that

$$\begin{aligned}\|D-\mathcal{C}\|_1 &\le \|D-\Phi(D,\mathcal{I})\|_1 + \|\Phi(D,\mathcal{I})-\tilde{D}\|_1 + \operatorname{dist}\Big(\tilde{D},\mathcal{C}\Big)\\ &\le \frac{\varepsilon}{20}+\frac{\varepsilon}{20}+\frac{9\varepsilon}{10}=\varepsilon.\end{aligned}$$

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

Completeness. Assume $D\in\mathcal{C}$. Then the choice of of γ and L ensures the existence of a good dyadic partition $\mathcal{I}(\gamma,D)$ in the sense of Definition 3.1. For any I in this partition for which (i) holds ($D(I)\le\frac{\gamma}{L}<\frac{\kappa}{2}$), I will have $\frac{m_I}{m}<\kappa$ and be kept as a "light leaf" (this by contrapositive of Fact 3.4 (ii)). For the other ones, (ii) holds: let I be one of these (at most L) intervals.

- If m_I is too small on Step 7, then I is kept as "light leaf."
- Otherwise, then by our choice of constants we can use Lemma 2.8 and apply Lemma 2.9 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, CHECK-SMALL-ℓ_2 will output yes, as

$$\|D_I-\mathcal{U}_I\|_2^2=\|D_I\|_2^2-\frac{1}{|I|}\le\left(1+\frac{\varepsilon^2}{6400}\right)\frac{1}{|I|}-\frac{1}{|I|}=\frac{\varepsilon^2}{6400\,|I|}$$

  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).[6] Thus Step 13 passes, and the algorithm reaches the APPROXIMATION stage. By the foregoing discussion, this implies $\Phi(D,\mathcal{I})$ is $\varepsilon/20$-close to D (and hence to $\mathcal{C}$); $\tilde{D}$ is then (except with probability at most $1/10$) ($\frac{\varepsilon}{20}+\frac{\varepsilon}{20}=\frac{\varepsilon}{10}$)-close to $\mathcal{C}$, and the algorithm returns ACCEPT.

4 Structural Theorems

In this section, we show that a wide range of natural distribution families are succinctly decomposable, and provide efficient projection algorithms for each class.

4.1 Existence of Structural Decompositions

Theorem 4.1 (Monotonicity). *For all $\gamma>0$, the class $\mathcal{M}$ of monotone distributions on $[n]$ is (γ,L)-splittable for $L\stackrel{\text{def}}{=}O\Big(\frac{\log^2 n}{\gamma}\Big)$.*

[6] 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, $\mathcal{I}(\gamma,D)=(I_1,\dots,I_t)$), 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 ℓ_2, or having very little mass).

Note that this proof can already be found in [BKR04, 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 4.1. 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 $\mathcal{I}^{(j)} = (I_1^{(j)}, \dots, I_{\ell_j}^{(j)})$ by going over the $I_i^{(j)} = [a_i^{(j)}, b_i^{(j)}]$ in order, and:

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

(b) else, if $D(b_i^{(j)}) \leq (1+\gamma) D(a_i^{(j)})$, then $I_i^{(j)}$ is added as element of $\mathcal{I}^{(j+1)}$ ("marked as leaf");

(c) otherwise, bisect $I^{(j)}$ in $I_{\mathrm{L}}^{(j)}$, $I_{\mathrm{R}}^{(j)}$ (with $\left|I_{\mathrm{L}}^{(j)}\right| = \left\lceil \left|I^{(j)}\right| /2 \right\rceil$) and add both $I_{\mathrm{L}}^{(j)}$ and $I_{\mathrm{R}}^{(j)}$ 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 $\mathcal{I} = (I_1, \dots, I_\ell)$ (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 [BKR04] 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 $\mathcal{I}$ (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

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

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 ℓ. □

Corollary 4.2 (Unimodality)**.** *For all $\gamma > 0$, the class $\mathcal{M}_1$ of unimodal distributions on $[n]$ is (γ, L)-decomposable for $L \stackrel{\text{def}}{=} O\left(\frac{\log^2 n}{\gamma}\right)$.*

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 4.1 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 4.3 (t-modality)**.** *For any $t \geq 1$ and all $\gamma > 0$, the class $\mathcal{M}_t$ of t-modal distributions on $[n]$ is (γ, L)-decomposable for $L \stackrel{\text{def}}{=} O\left(\frac{t \log^2 n}{\gamma}\right)$.*

Corollary 4.4 (Log-concavity, concavity and convexity)**.** *For all $\gamma > 0$, the classes $\mathcal{L}$, $\mathcal{K}^-$ and $\mathcal{K}^+$ of log-concave, concave and convex distributions on $[n]$ are (γ, L)-decomposable for $L \stackrel{\text{def}}{=} O\left(\frac{\log^2 n}{\gamma}\right)$.*

Proof. This is directly implied by Corollary 4.2, recalling that log-concave, concave and convex distributions are unimodal. □

Theorem 4.5 (Monotone Hazard Rate)**.** *For all $\gamma > 0$, the class $\mathcal{MHR}$ of MHR distributions on $[n]$ is (γ, L)-decomposable for $L \stackrel{\text{def}}{=} O\left(\frac{\log n}{\gamma}\right)$.*

Proof. This follows from adapting the proof of [CDSS13], which establishes that every MHR distribution can be approximated in ℓ_1 distance by a $O(\log(n/\varepsilon)/\varepsilon)$-histogram. For completeness, we reproduce their argument, suitably modified to our purposes, in Appendix B. □

Theorem 4.6 (Piecewise Polynomials)**.** *For all $\gamma > 0$, $t, d \geq 0$, the class $\mathcal{P}_{t,d}$ of t-piecewise degree-d distributions on $[n]$ is (γ, L)-decomposable for $L \stackrel{\text{def}}{=} O\left(\frac{t(d+1)}{\gamma} \log^2 n\right)$. (Moreover, for the class of t-histograms $\mathcal{H}_t$ ($d = 0$) one can take $L = t$.)*

Proof. The last part of the statement is obvious, so we focus on the first claim. Observing that each of the t pieces of a distribution $D \in \mathcal{P}_{t,d}$ can be subdivided in at most $d+1$ intervals on which D is monotone (being degree-d polynomial on each such pieces), we obtain a partition of $[n]$ into at most $t(d+1)$ intervals. D being monotone on each of them, we can apply an argument almost identical to that of Theorem 4.1 to argue that each interval can be further split into $O(\log^2 n/\gamma)$ subintervals, yielding a good decomposition with $O(t(d+1)\log^2 n/\gamma)$ pieces. □

4.2 Projection Step: computing the distances

This section contains details of the distance estimation procedures for these classes, required in the last stage of Algorithm 1. (Note that some of these results are phrased in terms of distance approximation, as estimating the distance $\ell_1(D, \mathcal{C})$ to sufficient accuracy in particular yields an algorithm for this stage.)

We focus in this section on achieving the sample complexities stated in Corollary 1.2, Corollary 1.3, and Corollary 1.4. While almost all the distance estimation procedures we give in this section are efficient, running in time polynomial in all the parameters or even with only a polylogarithmic dependence on n, there are two exceptions – namely, the procedures for monotone hazard rate (Lemma 4.9) and log-concave (Lemma 4.10) distributions. We *do* describe computationally efficient procedures for these two cases as well in Section 4.2.1, at a modest additive cost in the sample complexity.

Lemma 4.7 (Monotonicity [BKR04, Lemma 8])**.** *There exists a procedure* ProjectionDist$_{\mathcal{M}}$ *that, on input n as well as the full (succinct) specification of a ℓ-histogram D on $[n]$, computes the (exact) distance $\ell_1(D, \mathcal{M})$ in time* poly(ℓ).

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 ℓ possibilities, running the subroutine ℓ times and picking the minimum value) results in a similar claim for unimodal distributions:

Lemma 4.8 (Unimodality)**.** *There exists a procedure* ProjectionDist$_{\mathcal{M}_1}$ *that, on input n as well as the full (succinct) specification of a ℓ-histogram D on $[n]$, computes the (exact) distance $\ell_1(D, \mathcal{M}_1)$ in time* poly(ℓ).

A similar result can easily be obtained for the class of t-modal distributions as well, with a poly(ℓ, t)-time algorithm based on a combination of dynamic and linear programming. Analogous statements hold for the classes of concave and convex distributions $\mathcal{K}^+, \mathcal{K}^-$, also based on linear programming (specifically, on running $O(n^2)$ different linear programs – one for each possible support $[a, b] \subseteq [n]$ – and taking the minimum over them).

Lemma 4.9 (MHR)**.** *There exists a (non-efficient) procedure* ProjectionDist$_{\mathcal{MHR}}$ *that, on input n, ε, as well as the full specification of a distribution D on $[n]$, distinguishes between $\ell_1(D, \mathcal{MHR}) \leq \varepsilon$ and $\ell_1(D, \mathcal{MHR}) > 2\varepsilon$ in time $2^{\tilde{O}_\varepsilon(n)}$.*

Lemma 4.10 (Log-concavity)**.** *There exists a (non-efficient) procedure* ProjectionDist$_{\mathcal{L}}$ *that, on input n, ε, as well as the full specification of a distribution D on $[n]$, distinguishes between $\ell_1(D, \mathcal{L}) \leq \varepsilon$ and $\ell_1(D, \mathcal{L}) > 2\varepsilon$ in time $2^{\tilde{O}_\varepsilon(n)}$.*

Lemma 4.9 and Lemma 4.10. We here give a naive algorithm for these two problems, based on an exhaustive search over a (huge) ε-cover $\mathcal{S}$ of distributions over $[n]$. Essentially, $\mathcal{S}$ 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 $|\mathcal{S}| = O((n/\varepsilon)^n)$). It is not hard to see that this indeed defines an ε-cover of the set of all distributions, and moreover that it can be computed in time poly$(|\mathcal{S}|)$. To approximate the distance from an explicit distribution D to the class $\mathcal{C}$ (either $\mathcal{MHR}$ or $\mathcal{L}$), it is enough to go over every element S of $\mathcal{S}$, checking (this time, efficiently) if $\|S - D\|_1 \leq \varepsilon$ and if there is a distribution $P \in \mathcal{C}$ close to S (this time, pointwise, that is $|P(i) - S(i)| \leq \varepsilon/n$ for all i) – which also implies $\|S - P\|_1 \leq \varepsilon$ and thus $\|P - D\|_1 \leq 2\varepsilon$. The test for pointwise closeness can be done by checking feasibility of a linear program with variables corresponding to the logarithm of probabilities, i.e. $x_i \equiv \ln P(i)$. Indeed, this formulation allows to rephrase the log-concave and MHR constraints as linear constraints, and pointwise approximation is simply enforcing that $\ln(S(i) - \varepsilon/n) \leq x_i \leq \ln(S(i) + \varepsilon/n)$ for all i. At the end of this enumeration, the procedure accepts if and only if for some S both $\|S - D\|_1 \leq \varepsilon$ and the corresponding linear program was feasible. □

Lemma 4.11 (Piecewise Polynomials)**.** *There exists a procedure* ProjectionDist$_{\mathcal{P}_{t,d}}$ *that, on input n as well as the full specification of an ℓ-histogram D on $[n]$, computes an approximation Δ of the distance $\ell_1(D, \mathcal{P}_{t,d})$ such that $\ell_1(D, \mathcal{P}_{t,d}) \leq \Delta \leq 3\ell_1(D, \mathcal{P}_{t,d}) + \varepsilon$, and runs in time $O(n^3) \cdot$ poly$(\ell, t, d, \frac{1}{\varepsilon})$.*

Moreover, for the special case of t-histograms ($d = 0$) there exists a procedure ProjectionDist$_{\mathcal{H}_t}$*, which, given inputs as above, computes an approximation Δ of the distance $\ell_1(D, \mathcal{H}_t)$ such that $\ell_1(D, \mathcal{H}_t) \leq \Delta \leq 4\ell_1(D, \mathcal{H}_t) + \varepsilon$, and runs in time* poly$(\ell, t, \frac{1}{\varepsilon})$*, independent of n.*

Proof. We begin with ProjectionDist$_{\mathcal{H}_t}$. Fix any distribution D on $[n]$. Given any explicit partition of $[n]$ into intervals $\mathcal{I} = (I_1, \ldots, I_t)$, one can easily show that $\|D - \Phi(D, \mathcal{I})\|_1 \leq 2\text{OPT}_{\mathcal{I}}$, where OPT$_{\mathcal{I}}$ is the optimal distance of D to any histogram on $\mathcal{I}$. To get a 2-approximation of $\ell_1(D, \mathcal{H}_t)$, it thus suffices to find the minimum, over all possible partitionings $\mathcal{I}$ of $[n]$ into t intervals, of the quantity $\|D - \Phi(D, \mathcal{I})\|_1$ (which itself can be computed in time $T = O(\min(t\ell, n))$). By a simple dynamic programming approach, this can be performed in time $O(tn^2 \cdot T)$. The quadratic

dependence on n, which follows from allowing the endpoints of the t intervals to be at any point of the domain, is however far from optimal and can be reduced to $(t/\varepsilon)^2$, as we show below.

For $\eta > 0$, define an *η-granular decomposition* of a distribution D over $[n]$ to be a partition of $[n]$ into $s = O(1/\eta)$ intervals $J_1, \dots, J_s$ such that each interval J_i is either a singleton or satisfies $D(J_i) \le \eta$. (Note that if D is a known ℓ-histogram, one can compute an η-granular decomposition of D in time $O(\ell/\eta)$ in a greedy fashion.)

Claim 4.12. *Let D be a distribution over $[n]$, and $\mathcal{J} = (J_1, \dots, J_s)$ be an η-granular decomposition of D (with $s \ge t$). Then, there exists a partition of $[n]$ into t intervals $\mathcal{I} = (I_1, \dots, I_t)$ and a t-histogram H on $\mathcal{I}$ such that $\|D - H\|_1 \le 2\ell_1(D, \mathcal{H}_t) + 2t\eta$, and $\mathcal{I}$ is a coarsening of $\mathcal{J}$.*

Before proving it, we describe how this will enable us to get the desired time complexity for $\textsc{ProjectionDist}_{\mathcal{H}_t}$. Phrased differently, the claim above allows us to run our dynamic program using the $O(1/\eta)$ endpoints of the $O(1/\eta)$ instead of the n points of the domain, paying only an additive error $O(t\eta)$. Setting $\eta = \frac{\varepsilon}{4t}$, the guarantee for $\textsc{ProjectionDist}_{\mathcal{H}_t}$ follows.

Proof of Claim 4.12. Let $\mathcal{J} = (J_1, \dots, J_s)$ be an η-granular decomposition of D, and $H^* \in \mathcal{H}_t$ be a histogram achieving $\textsc{opt} = \ell_1(D, \mathcal{H}_t)$. Denote further by $\mathcal{I}^* = (I_1^*, \dots, I_t^*)$ the partition of $[n]$ corresponding to H^*. Consider now the $r \le t$ endpoints of the I_i^*'s that do not fall on one of the endpoints of the J_i's: let $J_{i_1}, \dots, J_{i_r}$ be the respective intervals in which they fall (in particular, these cannot be singleton intervals), and $S = \cup_{j=1}^r J_{i_j}$ their union. By definition of η-granularity, $D(S) \le t\eta$, and it follows that $H^*(S) \le t\eta + \frac{1}{2}\textsc{opt}$. We define H from H^* in two stages: first, we obtain a (sub)distribution H' by modifying H^* on S, setting for each $x \in J_{i_j}$ the value of H to be the minimum value (among the two options) that H^* takes on J_{i_j}. H' is thus a t-histogram, and the endpoints of its intervals are endpoints of $\mathcal{J}$ as wished; but it may not sum to one. However, by construction we have that $H'([n]) \ge 1 - H^*(S) \ge 1 - t\eta - \frac{1}{2}\textsc{opt}$. Using this, we can finally define our t-histogram distribution H as the renormalization of H'. It is easy to check that H is a valid t-histogram on a coarsening of $\mathcal{J}$, and

$$\|D - H\|_1 \le \|D - H'\|_1 + (1 - H'([n])) \le \|D - H^*\|_1 + \|H^* - H'\|_1 + t\eta + \frac{1}{2}\textsc{opt} \le 2\textsc{opt} + 2t\eta$$

as stated. □

Turning now to $\textsc{ProjectionDist}_{\mathcal{P}_{t,d}}$, we apply the same initial dynamic programming approach, which will result on a running time of $O(n^2 t \cdot T)$, where T is the time required to estimate (to sufficient accuracy) the distance of a given (sub)distribution over an interval I onto the space $\mathcal{P}_d$ of degree-d polynomials. Specifically, we will invoke the following result, adapted from [CDSS14a] to our setting:

Theorem 4.13. *Let p be a ℓ-histogram over $[-1, 1)$. There is an algorithm $\textsc{ProjectSinglePoly}(d, \eta)$ which runs in time $\mathrm{poly}(\ell, d+1, 1/\eta)$, and outputs a degree-$d$ polynomial q which defines a pdf over $[-1, 1)$ such that $\|p - q\|_1 \le 3\ell_1(p, \mathcal{P}_d) + O(\eta)$.*

The proof of this modification of [CDSS14a, Theorem 9] is deferred to Appendix C. Applying it as a blackbox with η set to $O(\varepsilon/t)$ and noting that computing the ℓ_1 distance to our explicit distribution on a given interval of the degree-d polynomial returned incurs an additional $O(n)$ factor, we obtain the claimed guarantee and running time. □

4.2.1 Computationally Efficient Procedures for Log-concave and MHR Distributions

We now describe how to obtain *efficient* testing for the classes $\mathcal{L}$ and $\mathcal{MHR}$ – that is, how to obtain polynomial-time distance estimation procedures for these two classes, unlike the ones described in the previous section. At a very high-level, the idea is in both case to write down a linear program on variables related *logarithmically* to the probabilities we are searching, as enforcing the log-concave and MHR constraints on these new variables can be done linearly. The catch now becomes the ℓ_1 objective function (and, to a lesser extent, the fact that the probabilities must sum to one), now highly non-linear.

The first insight is to leverage the structure of log-concave (resp. monotone hazard rate) distributions to express this objective as slightly stronger constraints, specifically pointwise $(1 \pm \varepsilon)$-multiplicative closeness, much easier to enforce in our "logarithmic formulation." Even so, doing this naively fails, essentially because of a too weak distance guarantee between our explicit histogram $\hat{D}$ and the unknown distribution we are trying to find: in the completeness case, we are only promised ε-closeness in ℓ_1, while we would also require good additive pointwise closeness of the order ε^2 or ε^3.

The second insight is thus to observe that we "almost" have this for free: indeed, if we do not reject in the first stage of the testing algorithm, we do obtain an explicit k-histogram $\hat{D}$ with the guarantee that D is ε-close to the distribution P to test. However, we *also* implicitly have another distribution $\hat{D}'$ that is $\sqrt{\varepsilon/k}$-close to P *in Kolmogorov distance*: as in the recursive descent we take enough samples to use the DKW inequality (Theorem 2.10) with this parameter, i.e. an additive overhead of $O(k/\varepsilon)$ samples (on top of the $\tilde{O}(\sqrt{kn}/\varepsilon^{7/2})$). If we are willing to increase this overhead by just a small amount, that is to take $\tilde{O}(\max(k/\varepsilon, 1/\varepsilon^4))$, we can guarantee that $\hat{D}'$ be also $\tilde{O}(\varepsilon^2)$-close to P in Kolmogorov distance.

Combining these ideas yield the following distance estimation lemmas:

Lemma 4.14 (Monotone Hazard Rate)**.** *There exists a procedure* $\textsc{ProjectionDist}^*_{\mathcal{MHR}}$ *that, on input n as well as the full specification of a k-histogram distribution D on $[n]$ and of a ℓ-histogram distribution D' on $[n]$, runs in time* $\mathrm{poly}(n, 1/\varepsilon)$*, and satisfies the following.*

- *If there is $P \in \mathcal{MHR}$ such that $\|D - P\|_1 \leq \varepsilon$ and $\|D' - P\|_{\mathrm{Kol}} \leq \varepsilon^3$, then the procedure returns* yes*;*
- *If $\ell_1(D, \mathcal{MHR}) > 100\varepsilon$, then the procedure returns* no*.*

Lemma 4.15 (Log-concavity)**.** *There exists a procedure* $\textsc{ProjectionDist}^*_{\mathcal{L}}$ *that, on input n as well as the full specifications of a k-histogram distribution D on $[n]$ and a ℓ-histogram distribution D' on $[n]$, runs in time* $\mathrm{poly}(n, k, \ell, 1/\varepsilon)$*, and satisfies the following.*

- *If there is $P \in \mathcal{L}$ such that $\|D - P\|_1 \leq \varepsilon$* and *$\|D' - P\|_{\mathrm{Kol}} \leq \frac{\varepsilon^2}{\log^2(1/\varepsilon)}$, then the procedure returns* yes*;*
- *If $\ell_1(D, \mathcal{L}) \geq 100\varepsilon$, then the procedure returns* no*.*

The proofs of these two lemmas are quite technical and deferred to Appendix C. With these in hand, a simple modification of our main algorithm (specifically, setting $m = \tilde{O}(\max(\sqrt{|I|}/\varepsilon^3 L, L^2/\varepsilon^2, 1/\varepsilon^c))$ for c either 4 or 6 instead of $\tilde{O}(\max(\sqrt{|I|}/\varepsilon^3 L, L^2/\varepsilon^2))$, to get the desired Kolmogorov distance guarantee; and providing the empirical histogram defined by these m samples along to the distance estimation procedure) suffices to obtain the following counterpart to Corollary 1.2:

Corollary 4.16. *The algorithm* TestSplittable*, after this modification, can* efficiently *test the classes of log-concave and monotone hazard rate (MHR) distributions, with respectively* $\tilde{O}(\sqrt{n}/\varepsilon^{7/2} + 1/\varepsilon^4)$ *and* $\tilde{O}(\sqrt{n}/\varepsilon^{7/2} + 1/\varepsilon^6)$ *samples.*

5 Going Further: Reducing the Support Size

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

As a motivating example, consider the class of Poisson Binomial Distributions (PBD). It is well-known (see e.g. [KG71, 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 $\tilde{O}(\sqrt{n})$ samples. This is, however, far from optimal: as shown in [AD15], a sample complexity of $\Theta\left(n^{1/4}\right)$ 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 5.2 will obtain as a direct application a $\tilde{O}\left(n^{1/4}\right)$-query testing algorithm for $\mathcal{PBD}_n$.

Definition 5.1. Given $\varepsilon > 0$, the *ε-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 $\mathcal{C}$:

Definition 5.2. For any class of distributions $\mathcal{C}$ over $[n]$, define the set of *conditioned distributions of $\mathcal{C}$* (with respect to $\varepsilon > 0$ and interval $I \subseteq [n]$) as $\mathcal{C}^{\varepsilon,I} \stackrel{\text{def}}{=} \{ D_I : D \in \mathcal{C}, D(I) \geq 1 - \varepsilon \}$.

Finally, we will require the following simple result:

Lemma 5.3. *Let D be a distribution over $[n]$, and $I \subseteq [n]$ an interval such that $D(I) \geq 1 - \frac{\varepsilon}{10}$. Then,*

- *If $D \in \mathcal{C}$, then $D_I \in \mathcal{C}^{\frac{\varepsilon}{10},I}$;*
- *If $\ell_1(D, \mathcal{C}) > \varepsilon$, then $\ell_1(D_I, \mathcal{C}^{\frac{\varepsilon}{10},I}) > \frac{7\varepsilon}{10}$.*

Proof. The first item is obvious. As for the second, let $P \in \mathcal{C}$ be any distribution with $P(I) \geq 1 - \frac{\varepsilon}{10}$.

By assumption, $\|D - P\|_1 > \varepsilon$: but we have, writing $\alpha = 1/10$,

$$\begin{aligned}
\|D_I - P_I\|_1 &= \sum_{i\in I}\left|\frac{D(i)}{D(I)} - \frac{P(i)}{P(I)}\right| = \frac{1}{D(I)}\sum_{i\in I}\left|D(i) - P(i) + P(i)\left(1 - \frac{D(I)}{P(I)}\right)\right| \\
&\geq \frac{1}{D(I)}\Big(\sum_{i\in I}|D(i) - P(i)| - \left|1 - \frac{D(I)}{P(I)}\right|\sum_{i\in I}P(i)\Big) \\
&= \frac{1}{D(I)}\Big(\sum_{i\in I}|D(i) - P(i)| - |P(I) - D(I)|\Big) \geq \frac{1}{D(I)}\Big(\sum_{i\in I}|D(i) - P(i)| - \alpha\varepsilon\Big) \\
&\geq \frac{1}{D(I)}\Big(\|D - P\|_1 - \sum_{i\notin I}|D(i) - P(i)| - \alpha\varepsilon\Big) \geq \frac{1}{D(I)}\Big(\|D - P\|_1 - 3\alpha\varepsilon\Big) \\
&> (1 - 3\alpha)\varepsilon = \frac{7}{10}\varepsilon.
\end{aligned}$$

□

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

Theorem 5.4. *Let $\mathcal{C}$ be a class of distributions over $[n]$ for which the following holds.*

1. *there is a function $M(\cdot,\cdot)$ such that each $D \in \mathcal{C}$ has ε-effective support of size at most $M(n,\varepsilon)$;*
2. *for every $\varepsilon \in [0,1]$ and interval $I \subseteq [n]$, $\mathcal{C}^{\varepsilon,I}$ is (γ, L)-splittable;*
3. *there exists an efficient procedure $\textsc{ProjectionDist}_{\mathcal{C}^{\varepsilon,I}}$ which, given as input the explicit description of a distribution D over $[n]$ and interval $I \subseteq [n]$, computes the distance $\ell_1(D_I, \mathcal{C}^{\varepsilon,I})$.*

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

Algorithm 2 TestEffectiveSplittable

Require: Domain Ω (interval of size n), sample access to D over Ω; subroutine $\textsc{ProjectionDist}_{\mathcal{C}^{\varepsilon,I}}$

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

1: Set $m \stackrel{\text{def}}{=} O(1/\varepsilon^2)$, $\tau \stackrel{\text{def}}{=} M(n, \frac{\varepsilon}{60})$.
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 $\{1,\dots,j\}$ such that $\hat{D}(J) \le \frac{\varepsilon}{30}$.
5: Let K be the largest interval of the form $\{k,\dots,n\}$ such that $\hat{D}(K) \le \frac{\varepsilon}{30}$.
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), $\textsc{ProjectionDist}_{\mathcal{C}^{\varepsilon,I}}$, parameters $\varepsilon' \stackrel{\text{def}}{=} \frac{7\varepsilon}{10}$ and $L(\cdot,\cdot)$.
12: **return** ACCEPT if TestSplittable accepts, REJECT otherwise.
13:

5.1 Proof of Theorem 5.4

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

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 1.1 and the choice of I, this latter quantity is $O\Big(\max\Big(\frac{1}{\varepsilon^3}\sqrt{M\ell}\log M, \frac{\ell^2}{\varepsilon^2}\Big)\Big)$ where $M = M(n, \frac{\varepsilon}{60})$ and $\ell = L(\frac{\varepsilon}{1200}, M(n, \frac{\varepsilon}{60}))$.

Correctness. If $D \in \mathcal{C}$, then by the setting of τ (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 5.3, Theorem 1.1 and our choice of parameters.

Similarly, if D is ε-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 $\frac{7\varepsilon}{10}$-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 $1/3$. □

5.2 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 5.5. *The class of Poisson Binomial Distributions can be tested with* $\tilde{O}\left(n^{1/4}/\varepsilon^{7/2}\right) + O\left(\log^4 n/\varepsilon^4\right)$ *samples, using Algorithm 2.*

This is a direct consequence of Theorem 5.4 and the lemmas below. The first one states that, indeed, PBDs have small effective support:

Fact 5.6. *For any* $\varepsilon > 0$*, a PBD has* ε*-effective support of size* $O\left(\sqrt{n \log(1/\varepsilon)}\right)$.

Proof. By an additive Chernoff Bound, any random variable X following a Poisson Binomial Distribution has $\Pr[\,|X - \mathbb{E}X| > \gamma n\,] \leq 2e^{-2\gamma^2 n}$. Taking $\gamma \stackrel{\text{def}}{=} \sqrt{\frac{1}{2n} \ln \frac{2}{\varepsilon}}$, we get that $\Pr[\,X \in I\,] \geq 1 - \varepsilon$, where $I \stackrel{\text{def}}{=} [\mathbb{E}X - \sqrt{\frac{1}{2} \ln \frac{2}{\varepsilon}}, \mathbb{E}X + \sqrt{\frac{1}{2} \ln \frac{2}{\varepsilon}}]$. □

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* $\mathcal{PBD}_n^{\varepsilon,I} \stackrel{\text{def}}{=} \{\, D_I : D \in \mathcal{PBD}_n, D(I) \geq 1 - \varepsilon \,\}$ falls under Corollary 4.2. The last piece we need to apply our generic testing framework is the existence of an algorithm to compute the distance between an (explicit) distribution and the class of conditioned PBDs. This is provided by our next lemma:

Claim 5.7. *There exists a procedure* $\textsc{ProjectionDist}_{\mathcal{PBD}_n^{\varepsilon,I}}$ *that, on input* n *and* $\varepsilon, \in [0,1]$, $I \subseteq [n]$ *as well as the full specification of a distribution* D *on* $[n]$*, computes a value* τ *such that* $\tau \in [1 \pm 2\varepsilon] \cdot \ell_1(D, \mathcal{PBD}_n^{\varepsilon,I}) \pm \frac{\varepsilon}{100}$*, in time* $n^2\,(1/\varepsilon)^{O(\log 1/\varepsilon)}$.

Proof. The goal is to find a $\gamma = \Theta(\varepsilon)$-approximation of the minimum value of $\sum_{i \in I} \left|\frac{P(i)}{P(I)} - \frac{D(i)}{D(I)}\right|$, subject to $P(I) = \sum_{i \in I} P(i) \geq 1 - \varepsilon$ 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 $O(n^2)$ by dynamic programming. Therefore, computing the ℓ_1 distance between D and any PBD with known parameters can be done efficiently. To conclude, we invoke a result of Diakonikolas, Kane, and Stewart, that guarantees the existence of a succinct (proper) cover of $\mathcal{PBD}_n$:

Theorem 5.8 ([DKS15, Theorem 14] (rephrased))**.** *For all* $n, \gamma > 0$*, there exists a set* $\mathcal{S}_\gamma \subseteq \mathcal{PBD}_n$ *such that:*

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

(ii) $|\mathcal{S}_\gamma| \leq n\,(1/\gamma)^{O(\log 1/\gamma)}$

(iii) $\mathcal{S}_\gamma$ *can be computed in time* $n\,(1/\gamma)^{O(\log 1/\gamma)}$

and each $D \in \mathcal{S}_\gamma$ *is explicitly described by its set of parameters.*

We further observe that the factor n in both the size of the cover and running time can be easily removed in our case, as we know a good approximation of the support size of the candidate PBDs. (That is, we only need to enumerate over a subset of the cover of [DKS15], that of the PBDs with effective support compatible with our distribution D.)

Set $\gamma \stackrel{\text{def}}{=} \frac{\varepsilon}{250}$. Fix $P \in \mathcal{PBD}_n$ such that $P(I) \geq 1-\varepsilon$, and $Q \in \mathcal{S}_\gamma$ such that $\|P-Q\|_1 \leq \gamma$. In particular, it is easy to see via the correspondence between ℓ_1 and total variation distance that $|P(I)-Q(I)| \leq \gamma/2$. By a calculation analogue as in Lemma 5.3, we have

$$\begin{aligned}
\|P_I - Q_I\|_1 &= \sum_{i\in I}\left|\frac{P(i)}{P(I)} - \frac{Q(i)}{Q(I)}\right| = \sum_{i\in I}\left|\frac{P(i)}{P(I)} - \frac{Q(i)}{P(I)} + Q(i)\left(\frac{1}{P(I)} - \frac{1}{Q(I)}\right)\right| \\
&= \sum_{i\in I}\left|\frac{P(i)}{P(I)} - \frac{Q(i)}{P(I)}\right| \pm \sum_{i\in I} Q(i)\left|\frac{1}{P(I)} - \frac{1}{Q(I)}\right| = \frac{1}{P(I)}\left(\sum_{i\in I}|P(i)-Q(i)| \pm |P(I)-Q(I)|\right) \\
&= \frac{1}{P(I)}\left(\sum_{i\in I}|P(i)-Q(i)| \pm \frac{\gamma}{2}\right) = \frac{1}{P(I)}\left(\|P-Q\|_1 \pm \frac{5\gamma}{2}\right) \\
&\in \left[\|P-Q\|_1 - 5\gamma/2, (1+2\varepsilon)\left(\|P-Q\|_1 + 5\gamma/2\right)\right]
\end{aligned}$$

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

As previously mentioned, this approximation guarantee for $\ell_1(D, \mathcal{PBD}_n^{\varepsilon,I})$ is sufficient for the purpose of Algorithm 1.

Proof of Corollary 5.5. Combining the above, we invoke Theorem 5.4 with $M(n,\varepsilon) = O(\sqrt{n\log(1/\varepsilon)})$ (Fact 5.6) and $L(m,\gamma) = O(\frac{\log^2 m}{\gamma})$ (Corollary 4.2). This yields the claimed sample complexity; finally, the efficiency is a direct consequence of Claim 5.7. □

6 Lower Bounds

6.1 Reduction-based Lower Bound Approach

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 $\mathcal{C}$ if it satisfies the following. Given sample access to an arbitrary distribution D and parameter ε, it outputs a hypothesis $\hat{D}$ which (with high probability) does "almost as well as it gets":

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

where $\mathrm{OPT}_{\mathcal{C},D} \stackrel{\text{def}}{=} \inf_{D'\in\mathcal{C}} \ell_1(D', D)$, 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 [BKR04] that "monotonicity is at least as hard as uniformity." Unfortunately, their specific argument does not generalize easily to other classes of distributions, 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 $\mathcal{C}$ which (a) contains the uniform distribution, and (b) for which we have a $o(\sqrt{n})$-sample agnostic learner $\mathcal{L}$, as follows. Assuming we have a tester $\mathcal{T}$ for $\mathcal{C}$ with sample complexity $o(\sqrt{n})$, define a uniformity tester as below.

- test if $D \in \mathcal{C}$ using $\mathcal{T}$; if not, reject (as $\mathcal{U} \in \mathcal{C}$, D cannot be uniform);
- otherwise, agnostically learn D with $\mathcal{L}$ (since D is close to $\mathcal{C}$), and obtain hypothesis $\hat{D}$;
- check offline if $\hat{D}$ is close to uniform.

By assumption, $\mathcal{T}$ and $\mathcal{L}$ each use $o(\sqrt{n})$ samples, so does the whole process; but this contradicts the lower bound of [BFR+00, Pan08] on uniformity testing. Hence, $\mathcal{T}$ must use $\Omega(\sqrt{n})$ samples.

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

Theorem 6.1. *Let $\mathcal{C}$ be a class of distributions over $[n]$ for which the following holds:*

(i) there exists a semi-agnostic learner $\mathcal{L}$ for $\mathcal{C}$, with sample complexity $q_L(n, \varepsilon, \delta)$ and "agnostic constant" c;

(ii) there exists a subclass $\mathcal{C}_{\text{Hard}} \subseteq \mathcal{C}$ such that testing $\mathcal{C}_{\text{Hard}}$ requires $q_H(n, \varepsilon)$ samples.

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

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

- call $\mathcal{T}$ with parameters n, $\frac{\varepsilon'}{c}$ (where $\varepsilon' \stackrel{\text{def}}{=} \frac{\varepsilon}{3}$) and failure probability $1/6$, to $\frac{\varepsilon'}{c}$-test if $D \in \mathcal{C}$. If not, reject.
- otherwise, agnostically learn a hypothesis $\hat{D}$ for D, with $\mathcal{L}$ called with parameters n, ε' and failure probability $1/6$;
- check offline if $\hat{D}$ is ε'-close to $\mathcal{C}_{\text{Hard}}$, accept if and only if this is the case.

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

For the soundness, we proceed by contrapositive. Suppose $\mathcal{T}_{\text{Hard}}$ accepts; it means that each step was successful. In particular, $\ell_1(\hat{D}, \mathcal{C}) \leq \varepsilon'/c$; 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, $\ell_1(D, \mathcal{C}_{\text{Hard}}) \leq 2\varepsilon' + \ell_1(\hat{D}, \mathcal{C}_{\text{Hard}}) \leq 3\varepsilon' = \varepsilon$.

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

Taking $\mathcal{C}_{\text{Hard}}$ to be the singleton consisting of the uniform distribution, and from the semi-agnostic learners of [CDSS13, CDSS14a] (each with sample complexity either $\text{poly}(1/\varepsilon)$ or $\text{poly}(\log n, 1/\varepsilon)$),

we obtain the following:[7]

Corollary 1.7. *Testing log-concavity, convexity, concavity, MHR, unimodality, t-modality, t-histograms, and t-piecewise degree-d distributions each require $\Omega(\sqrt{n}/\varepsilon^2)$ samples (the last three for $t = o(\sqrt{n})$ and $t(d+1) = o(\sqrt{n})$, respectively), for any $\varepsilon \geq 1/n^{O(1)}$.*

Similarly, we can use another result of [DDS12b] which shows how to agnostically learn Poisson Binomial Distributions with $\tilde{O}(1/\varepsilon^2)$ samples.[8] Taking $\mathcal{C}_{\text{Hard}}$ to be the single $\text{Bin}(n, 1/2)$ distribution (along with the testing lower bound of [VV14]), this yields the following:

Corollary 1.8. *Testing the classes of Binomial and Poisson Binomial Distributions each require $\Omega\left(n^{1/4}/\varepsilon^2\right)$ samples, for any $\varepsilon \geq 1/n^{O(1)}$.*

Finally, we derive a lower bound on testing k-SIIRVs from the agnostic learner of [DDO+13] (which has sample complexity $\text{poly}(k, 1/\varepsilon)$ samples, independent of n):

Corollary 1.9. *There exist absolute constants $c > 0$ and $\varepsilon_0 > 0$ such that testing the class of k-SIIRV distributions requires $\Omega(k^{1/2}n^{1/4})$ samples, for any $k = o(n^c)$ and $\varepsilon \leq \varepsilon_0$.*

Corollary 1.9. To prove this result, it is enough by Theorem 6.1 to exhibit a particular k-SIIRV S such that testing identity to S requires this many samples. Moreover, from [VV14] this last part amounts to proving that the (truncated) 2/3-norm $\|S^{-\max}_{-\varepsilon_0}\|_{2/3}$ 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 $\text{Bin}(n, 1/2)$ distribution). It is straightforward to verify that $\mathbb{E}S = \frac{n(k-1)}{2}$ and $\sigma^2 \stackrel{\text{def}}{=} \text{Var}\, S = \frac{(k^2-1)n}{12} = \Theta(k^2 n)$; moreover, 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 2σ 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

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

which concludes the proof. □

6.2 Tolerant Testing

This lower bound framework from the previous section carries to *tolerant* testing as well, resulting in this analogue to Theorem 6.1:

Theorem 6.2. *Let $\mathcal{C}$ be a class of distributions over $[n]$ for which the following holds:*

(i) there exists a semi-agnostic learner $\mathcal{L}$ for $\mathcal{C}$, with sample complexity $q_L(n, \varepsilon, \delta)$ and "agnostic constant" c;

[7] Specifically, 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 $\sqrt{n}/\varepsilon^2$).

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

(ii) there exists a subclass $\mathcal{C}_{\mathrm{Hard}} \subseteq \mathcal{C}$ *such that* tolerant *testing* $\mathcal{C}_{\mathrm{Hard}}$ *requires* $q_H(n, \varepsilon_1, \varepsilon_2)$ *samples for some parameters* $\varepsilon_2 > (4c+1)\varepsilon_1$.

Suppose further that $q_L(n, \varepsilon_2 - \varepsilon_1, 1/10) = o(q_H(n, \varepsilon_1, \varepsilon_2))$. *Then, any* tolerant *tester for* $\mathcal{C}$ *must use* $\Omega(q_H(n, \varepsilon_1, \varepsilon_2))$ *samples (for some explicit parameters* $\varepsilon_1', \varepsilon_2'$*).*

Proof. The argument follows the same ideas as for Theorem 6.1, up to the details of the parameters. Assuming $\mathcal{C}$, $\mathcal{C}_{\mathrm{Hard}}$, $\mathcal{L}$ as above (with semi-agnostic constant $c \geq 1$), and a tolerant tester $\mathcal{T}$ for $\mathcal{C}$ with sample complexity $q(n, \varepsilon_1, \varepsilon_2)$, we define a tolerant tester $\mathcal{T}_{\mathrm{Hard}}$ for $\mathcal{C}_{\mathrm{Hard}}$. On input $0 < \varepsilon_1 < \varepsilon_2 \leq 1$ with $\varepsilon_2 > (4c+1)\varepsilon_1$, and given sample access to a distribution D on $[n]$, $\mathcal{T}_{\mathrm{Hard}}$ acts as follows. After setting $\varepsilon_1' \stackrel{\mathrm{def}}{=} \frac{\varepsilon_2 - \varepsilon_1}{4}$, $\varepsilon_2' \stackrel{\mathrm{def}}{=} \frac{\varepsilon_2 - \varepsilon_1}{2}$, $\varepsilon' \stackrel{\mathrm{def}}{=} \frac{\varepsilon_2 - \varepsilon_1}{16}$ and $\tau \stackrel{\mathrm{def}}{=} \frac{6\varepsilon_2 + 10\varepsilon_1}{16}$,

- call $\mathcal{T}$ with parameters n, $\frac{\varepsilon_1'}{c}$, $\frac{\varepsilon_2'}{c}$ and failure probability $1/6$, to tolerantly test if $D \in \mathcal{C}$. If $\ell_1(D, \mathcal{C}) > \varepsilon_2'/c$, reject.
- otherwise, agnostically learn a hypothesis $\hat{D}$ for D, with $\mathcal{L}$ called with parameters n, ε' and failure probability $1/6$;
- check offline if $\hat{D}$ is τ-close to $\mathcal{C}_{\mathrm{Hard}}$, accept if and only if this is the case.

We condition on both calls (to $\mathcal{T}$ and $\mathcal{L}$) to be successful, which overall happens with probability at least $2/3$ by a union bound. We first argue completeness: assume $\ell_1(D, \mathcal{C}_{\mathrm{Hard}}) \leq \varepsilon_1$. This implies $\ell_1(D, \mathcal{C}) \leq \varepsilon_1$, so that $\mathcal{T}$ accepts as $\varepsilon_1 \leq \varepsilon_1'/c$ (which is the case because $\varepsilon_2 > (4c+1)\varepsilon_1$). Thus, the hypothesis $\hat{D}$ satisfies $\|\hat{D} - D\|_1 \leq c \cdot \varepsilon_1'/c + \varepsilon' = \varepsilon_1' + \varepsilon'$. Therefore, $\ell_1(\hat{D}, \mathcal{C}_{\mathrm{Hard}}) \leq \|\hat{D} - D\|_1 + \ell_1(D, \mathcal{C}_{\mathrm{Hard}}) \leq \varepsilon_1' + \varepsilon' + \varepsilon_1 < \tau$, and $\mathcal{T}_{\mathrm{Hard}}$ accepts.

For the soundness, we again proceed by contrapositive. Suppose $\mathcal{T}_{\mathrm{Hard}}$ accepts; it means that each step was successful. In particular, $\ell_1(\hat{D}, \mathcal{C}) \leq \varepsilon_2'/c$; so that the hypothesis outputted by the agnostic learner satisfies $\|\hat{D} - D\|_1 \leq c \cdot \mathrm{OPT} + \varepsilon' \leq \varepsilon_2' + \varepsilon'$. In turn, since the last step passed and by a triangle inequality we get, as claimed, $\ell_1(D, \mathcal{C}_{\mathrm{Hard}}) \leq \varepsilon_2' + \varepsilon' + \ell_1(\hat{D}, \mathcal{C}_{\mathrm{Hard}}) \leq \varepsilon_2' + \varepsilon' + \tau < \varepsilon_2$.

Observing that the overall sample complexity is $q_T(n, \frac{\varepsilon_1'}{c}, \frac{\varepsilon_2'}{c}) + q_L(n, \varepsilon', \frac{1}{10}) = q_T(n, \frac{\varepsilon'}{c}) + o(q_H(n, \varepsilon'))$ concludes the proof. □

As before, we instantiate the general theorem to obtain specific lower bounds for tolerant testing of the classes we covered in this paper. That is, taking $\mathcal{C}_{\mathrm{Hard}}$ to be the singleton consisting of the uniform distribution (combined with the tolerant testing lower bound of [VV10]), and again from the semi-agnostic learners of [CDSS13, CDSS14a] (each with sample complexity either $\mathrm{poly}(1/\varepsilon)$ or $\mathrm{poly}(\log n, 1/\varepsilon)$), we obtain the following:

Corollary 1.12. *Tolerant testing of log-concavity, convexity, concavity, MHR, unimodality, and t-modality each require* $\Omega\left(\frac{1}{(\varepsilon_2 - \varepsilon_1)} \frac{n}{\log n}\right)$ *samples (the latter for* $t = o(n)$*).*

Similarly, we again turn to the class of Poisson Binomial Distributions, for which we can invoke as before the $\tilde{O}(1/\varepsilon^2)$-sample agnostic learner of [DDS12b]. As before, we would like to choose for $\mathcal{C}_{\mathrm{Hard}}$ the single $\mathrm{Bin}(n, 1/2)$ distribution; however, as no tolerant testing lower bound for this distribution exists – to the best of our knowledge – in the literature, we first need to establish the lower bound we will rely upon:

Theorem 6.3. *There exists an absolute constant* $\varepsilon_0 > 0$ *such that the following holds. Any algorithm which, given sampling access to an unknown distribution* D *on* Ω *and parameter* $\varepsilon \in (0, \varepsilon_0)$*, distinguishes with probability at least* $2/3$ *between (i)* $\|D - \mathrm{Bin}(n, 1/2)\|_1 \leq \varepsilon$ *and (ii)* $\|D - \mathrm{Bin}(n, 1/2)\|_1 \geq 100\varepsilon$ *must use* $\Omega\left(\frac{1}{\varepsilon} \frac{\sqrt{n}}{\log n}\right)$ *samples.*

The proof relies on a reduction from tolerant testing of *uniformity*, drawing on a result of Valiant and Valiant [VV10]; for the sake of conciseness, the details are deferred to Appendix D. With Theorem 6.3 in hand, we can apply Theorem 6.2 to obtain the desired lower bound:

Corollary 1.13. *Tolerant testing of the classes of Binomial and Poisson Binomial Distributions each require* $\Omega\Big(\frac{1}{(\varepsilon_2-\varepsilon_1)}\frac{\sqrt{n}}{\log n}\Big)$ *samples.*

We observe that both Corollary 1.12 and Corollary 1.13 are tight (with regard to the dependence on n), as proven in the next section (Section 7).

7 A Generic Tolerant Testing Upper Bound

To conclude this work, we address the question of tolerant testing of distribution classes. In the same spirit as before, we focus on describing a generic approach to obtain such bounds, in a clean conceptual manner. The most general statement of the result we prove in this section is stated below, which we then instantiate to match the lower bounds from Section 6.2:

Theorem 7.1. *Let $\mathcal{C}$ be a class of distributions over $[n]$ for which the following holds:*

(i) there exists a semi-agnostic learner $\mathcal{L}$ for $\mathcal{C}$, with sample complexity $q_L(n,\varepsilon,\delta)$ and "agnostic constant" c;

(ii) for any $\eta\in[0,1]$, every distribution in $\mathcal{C}$ has η-effective support of size at most $M(n,\eta)$.

Then, there exists an algorithm that, for any fixed $\kappa>1$ and on input $\varepsilon_1,\varepsilon_2\in(0,1)$ such that $\varepsilon_2\geq C\varepsilon_1$, has the following guarantee (where $C>2$ depends on c and κ only). The algorithm takes $O\Big(\frac{1}{(\varepsilon_2-\varepsilon_1)^2}\frac{m}{\log m}\Big)+q_L(n,\frac{\varepsilon_2-\varepsilon_1}{\kappa},\frac{1}{10})$ samples (where $m=M(n,\varepsilon_1)$), and with probability at least $2/3$ distinguishes between (a) $\ell_1(D,\mathcal{C})\leq\varepsilon_1$ and (b) $\ell_1(D,\mathcal{C})>\varepsilon_2$. (Moreover, one can take $C=(1+(5c+6)\frac{\kappa}{\kappa-1})$.)

Corollary 1.10. *Tolerant testing of log-concavity, convexity, concavity, MHR, unimodality, and t-modality can be performed with $O(\frac{1}{(\varepsilon_2-\varepsilon_1)^2}\frac{n}{\log n})$ samples, for $\varepsilon_2\geq C\varepsilon_1$ (where $C>2$ is an absolute constant).*

Applying now the theorem with $M(n,\varepsilon)=\sqrt{n\log(1/\varepsilon)}$ (as per Corollary 5.5), we obtain an improved upper bound for Binomial and Poisson Binomial distributions:

Corollary 1.11. *Tolerant testing of the classes of Binomial and Poisson Binomial Distributions can be performed with $O(\frac{1}{(\varepsilon_2-\varepsilon_1)^2}\frac{\sqrt{n\log(1/\varepsilon_1)}}{\log n})$ samples, for $\varepsilon_2\geq C\varepsilon_1$ (where $C>2$ is an absolute constant).*

High-level idea. Somewhat similar to the lower bound framework developed in Section 6, the gist of the approach is to reduce the problem of tolerant testing membership of D to the *class* $\mathcal{C}$ to that of tolerant testing identity to a known *distribution* – namely, the distribution $\hat{D}$ obtained after trying to agnostically learn D. Intuitively, an agnostic learner for $\mathcal{C}$ should result in a good enough hypothesis $\hat{D}$ (i.e., $\hat{D}$ close enough to both D and $\mathcal{C}$) when D is ε_1-close to $\mathcal{C}$; but output a $\hat{D}$ that is significantly far from either D or $\mathcal{C}$ when D is ε_2-far from $\mathcal{C}$ – sufficiently for us to be able to tell. Besides the many technical details one has to control for the parameters to work out, one key element is the use of a tolerant testing algorithm for closeness of two distributions due

to [VV11b], whose (tight) sample complexity scales as $n/\log n$ for a domain of size n. In order to get the right dependence on the effective support (required in particular for Corollary 1.11), we have to perform a first test to identify the effective support of the distribution and check its size, in order to only call this tolerant closeness testing algorithm on this much smaller subset. (This additional preprocessing step itself has to be carefully done, and comes at the price of a slightly worse constant $C = C(c, \kappa)$ in the statement of the theorem.)

7.1 Proof of Theorem 7.1

As described in the preceding section, the algorithm will rely on the ability to perform tolerant testing of equivalence between two unknown distributions (over some known domain of size m). This is ensured by an algorithm of Valiant and Valiant, restated below:

Theorem 7.2 ([VV11b, Theorem 3 and 4])**.** *There exists an algorithm $\mathcal{E}$ which, given sampling access to two unknown distributions D_1, D_2 over $[m]$, satisfies the following. On input $\varepsilon \in (0, 1]$, it takes $O(\frac{1}{\varepsilon^2}\frac{m}{\log m})$ samples from D_1 and D_2, and outputs a value Δ such that $|\|D_1 - D_2\|_1 - \Delta| \leq \varepsilon$ with probability $1 - 1/\operatorname{poly}(m)$. (Furthermore, $\mathcal{E}$ runs in time $\operatorname{poly}(m)$.)*

For the proof, we will also need this fact, similar to Lemma 5.3, which relates the distance of two distributions to that of their conditional distributions on a subset of the domain:

Fact 7.3. *Let D and P be distributions over $[n]$, and $I \subseteq [n]$ an interval such that $D(I) \geq 1 - \alpha$ and $P(I) \geq 1 - \beta$. Then,*

- $\|D_I - P_I\|_1 \leq \frac{3}{2}\frac{\|D-P\|_1}{D(I)} \leq 3\|D - P\|_1$ *(the last inequality for $\alpha \leq \frac{1}{2}$); and*
- $\|D_I - P_I\|_1 \geq \|D - P\|_1 - 2(\alpha + \beta)$.

Proof. To establish the first item, write:

$$\begin{aligned}
\|D_I - P_I\|_1 &= \sum_{i\in I}\left|\frac{D(i)}{D(I)} - \frac{P(i)}{P(I)}\right| = \frac{1}{D(I)}\sum_{i\in I}\left|D(i) - P(i) + P(i)(1 - \frac{D(I)}{P(I)})\right| \\
&\leq \frac{1}{D(I)}\Big(\sum_{i\in I}|D(i) - P(i)| + \left|1 - \frac{D(I)}{P(I)}\right|\sum_{i\in I}P(i)\Big) \\
&= \frac{1}{D(I)}\Big(\sum_{i\in I}|D(i) - P(i)| + |P(I) - D(I)|\Big) \leq \frac{1}{D(I)}\Big(\sum_{i\in I}|D(i) - P(i)| + \frac{1}{2}\|D - P\|_1\Big) \\
&\leq \frac{1}{D(I)}\cdot\frac{3}{2}\|D - P\|_1
\end{aligned}$$

where we used the fact that $|P(I) - D(I)| \leq \mathrm{d}_{\mathrm{TV}}(D, P) = \frac{1}{2}\|D - P\|_1$. Turning now to the second item, we have:

$$\begin{aligned}
\|D_I - P_I\|_1 &= \frac{1}{D(I)}\sum_{i\in I}\left|D(i) - P(i) + P(i)\Big(1 - \frac{D(I)}{P(I)}\Big)\right| \geq \frac{1}{D(I)}\Big(\sum_{i\in I}|D(i) - P(i)| - \left|1 - \frac{D(I)}{P(I)}\right|\sum_{i\in I}P(i)\Big) \\
&= \frac{1}{D(I)}\Big(\sum_{i\in I}|D(i) - P(i)| - |P(I) - D(I)|\Big) \geq \frac{1}{D(I)}\Big(\sum_{i\in I}|D(i) - P(i)| - (\alpha + \beta)\Big) \\
&\geq \frac{1}{D(I)}\Big(\|D - P\|_1 - \sum_{i\notin I}|D(i) - P(i)| - (\alpha + \beta)\Big) \geq \frac{1}{D(I)}\Big(\|D - P\|_1 - 2(\alpha + \beta)\Big) \\
&\geq \|D - P\|_1 - 2(\alpha + \beta).
\end{aligned}$$

□

With these two ingredients, we are in position to establish our theorem:

Proof of Theorem 7.1. The algorithm proceeds as follows, where we set $\varepsilon \stackrel{\text{def}}{=} \frac{\varepsilon_2-\varepsilon_1}{17\kappa}$, $\theta \stackrel{\text{def}}{=} \varepsilon_2 - ((6+c)\varepsilon_1 + 11\varepsilon)$, and $\tau \stackrel{\text{def}}{=} 2\frac{(3+c)\varepsilon_1+5\varepsilon}{2}$:

(1) using $O(\frac{1}{\varepsilon^2})$ samples, get (with probability at least $1-1/10$, by Theorem 2.10) a distribution $\tilde{D}$ $\frac{\varepsilon}{2}$-close to D in Kolmogorov distance; and let $I \subseteq [n]$ be the smallest interval such that $\tilde{D}(I) > 1 - \frac{3}{2}\varepsilon_1 - \varepsilon$. Output REJECT if $|I| > M(n,\varepsilon_1)$.

(2) invoke $\mathcal{L}$ on D with parameters ε and failure probability $\frac{1}{10}$, to obtain a hypothesis $\hat{D}$;

(3) call $\mathcal{E}$ (from Theorem 7.2) on D_I, $\hat{D}_I$ with parameter $\frac{\varepsilon}{6}$ to get an estimate $\hat{\Delta}$ of $\|D_I - \hat{D}_I\|_1$;

(4) output REJECT if $\hat{D}(I) < 1 - \tau$;

(5) compute "offline" (an estimate accurate within ε of) $\ell_1(\hat{D},\mathcal{C})$, denoted Δ;

(6) output REJECT is $\Delta + \hat{\Delta} > \theta$, and output ACCEPT otherwise.

The claimed sample complexity is immediate from Steps (2) and (3), along with Theorem 7.2. Turning to correctness, we condition on both subroutines meeting their guarantee (i.e., $\|D - \hat{D}\|_1 \le c\cdot\mathrm{OPT}+\varepsilon$ and $\|D-\hat{D}\|_1 \in [\hat{\Delta}-\varepsilon, \hat{\Delta}+\varepsilon]$), which happens with probability at least $8/10 - 1/\operatorname{poly}(n) \ge 3/4$ by a union bound.

- Soundness: If $\ell_1(D,\mathcal{C}) \le \varepsilon_1$, then D is ε_1-close to some $P \in \mathcal{C}$, for which there exists an interval $J \subseteq [n]$ of size at most $M(n,\varepsilon_1)$ such that $P(J) \ge 1-\varepsilon_1$. It follows that $D(J) \ge 1 - \frac{3}{2}\varepsilon_1$ (since $|D(J) - P(J)| \le \frac{\varepsilon_1}{2}$) and $\tilde{D}(J) \ge 1 - \frac{3}{2}\varepsilon_1 - 2\cdot\frac{\varepsilon}{2}\varepsilon$; establishing existence of a good interval I to be found (and Step (1) does not end with REJECT). Additionally, $\|D-\hat{D}\|_1 \le c\cdot\varepsilon_1 + \varepsilon$ and by the triangle inequality this implies $\ell_1(\hat{D},\mathcal{C}) \le (1+c)\varepsilon_1 + \varepsilon$.
  Moreover, as $D(I) \ge \tilde{D}(I) - 2\cdot\frac{\varepsilon}{2} \ge 1 - \frac{3}{2}\varepsilon_1 - 2\varepsilon$ and $\left|\hat{D}(I) - D(I)\right| \le \frac{1}{2}\|D-\hat{D}\|_1$, we do have
  $$\hat{D}(I) \ge 1 - \frac{3}{2}\varepsilon_1 - 2\varepsilon - \frac{c\varepsilon_1}{2} - \frac{\varepsilon}{2} = 1 - \tau$$
  and the algorithm does not reject in Step (4). To conclude, one has by Fact 7.3 that
  $$\|D_I - \hat{D}_I\|_1 \le \frac{3}{2}\frac{\|D-\hat{D}\|_1}{D(I)} \le \frac{3}{2}\frac{(c\varepsilon_1+\varepsilon)}{1-\frac{3}{2}\varepsilon_1 - 2\varepsilon} \le 3(c\varepsilon_1+\varepsilon) \qquad (\text{for } \varepsilon_1 < 1/4, \text{ as } \varepsilon < 1/17)$$
  Therefore, $\Delta + \hat{\Delta} \le \ell_1(\hat{D},\mathcal{C}) + \varepsilon + \|D_I - \hat{D}_I\|_1 + \varepsilon \le (4c+1)\varepsilon_1 + 6\varepsilon \le \varepsilon_2 - ((6+c)\varepsilon_1 + 11\varepsilon) = \theta$ (the last inequality by the assumption on $\varepsilon_2, \varepsilon_1$), and the tester accepts.
- Completeness: If $\ell_1(D,\mathcal{C}) > \varepsilon_2$, then we must have $\|D-\hat{D}\|_1 + \ell_1(\hat{D},\mathcal{C}) > \varepsilon_2$. If the algorithm does not already reject in Step (4), then $\hat{D}(I) \ge 1-\tau$. But, by Fact 7.3,
  $$\begin{aligned}\|D_I - \hat{D}_I\|_1 &\ge \|D-\hat{D}\|_1 - 2(D(I^c) + \hat{D}(I^c)) \ge \|D_I - \hat{D}_I\|_1 - 2\Big(\frac{3}{2}\varepsilon_1 + 2\varepsilon + \tau\Big)\\ &= \|D-\hat{D}\|_1 - ((6+c)\varepsilon_1 + 9\varepsilon)\end{aligned}$$
  we then have $\|D_I - \hat{D}_I\|_1 + \ell_1(\hat{D},\mathcal{C}) > \varepsilon_2 - ((6+c)\varepsilon_1 + 9\varepsilon)$. This implies $\Delta + \hat{\Delta} > \varepsilon_2 - ((6+c)\varepsilon_1 + 9\varepsilon) - 2\varepsilon = \varepsilon_2 - ((6+c)\varepsilon_1 + 11\varepsilon) = \theta$, and the tester rejects.

Finally, the testing algorithm defined above is computationally efficient as long as both the learning algorithm (Step (2)) and the estimation procedure (Step (5)) are.

□

References

[AAK+07] Noga Alon, Alexandr Andoni, Tali Kaufman, Kevin Matulef, Ronitt Rubinfeld, and Ning Xie. Testing k-wise and almost k-wise independence. In *Proceedings of the 39th ACM Symposium on Theory of Computing, STOC 2007, San Diego, California, USA, June 11-13, 2007*, pages 496–505, New York, NY, USA, 2007.

[AD15] Jayadev Acharya and Constantinos Daskalakis. Testing Poisson Binomial Distributions. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015, San Diego, CA, USA, January 4-6, 2015*, pages 1829–1840, 2015.

[ADK15] Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. *CoRR*, abs/1507.05952, July 2015.

[ADLS15] Jayadev Acharya, Ilias Diakonikolas, Jerry Zheng Li, and Ludwig Schmidt. Sample-optimal density estimation in nearly-linear time. *CoRR*, abs/1506.00671, 2015.

[AK03] Sanjeev Arora and Subhash Khot. Fitting algebraic curves to noisy data. *Journal of Computer and System Sciences*, 67(2):325 – 340, 2003. Special Issue on STOC 2002.

[An96] Mark Y. An. Log-concave probability distributions: theory and statistical testing. Technical report, Centre for Labour Market and Social Research, Denmark, 1996.

[BB05] Mark Bagnoli and Ted Bergstrom. Log-concave probability and its applications. *Economic Theory*, 26(2):445–469, 2005.

[BDBB72] Richard E. Barlow, Bartholomew D.J, J.M Bremner, and H.D Brunk. *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*. Wiley Series in Probability and Mathematical Statistics. J. Wiley, London, New York, 1972.

[BDKR05] Tuğkan Batu, Sanjoy Dasgupta, Ravi Kumar, and Ronitt Rubinfeld. The complexity of approximating the entropy. *SIAM Journal on Computing*, 35(1):132–150, 2005.

[BFF+01] Tuğkan Batu, Eldar Fischer, Lance Fortnow, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. Testing random variables for independence and identity. In *42nd Annual IEEE Symposium on Foundations of Computer Science, FOCS 2001, Las Vegas, Nevada, USA, October 14-17 2001*, pages 442–451, 2001.

[BFR+00] Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing that distributions are close. In *41st Annual IEEE Symposium on Foundations of Computer Science, FOCS 2000, Redondo Beach, California, USA, November 12-14 2000*, pages 259–269, 2000.

[BKR04] Tuğkan Batu, Ravi Kumar, and Ronitt Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *Proceedings of the 36th ACM Symposium on Theory of Computing, STOC 2004, Chicago, IL, USA, June 13-16, 2004*, pages 381–390, New York, NY, USA, 2004. ACM.

[CDSS13] Siu-on Chan, Ilias Diakonikolas, Rocco A. Servedio, and Xiaorui Sun. Learning mixtures of structured distributions over discrete domains. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2013, New Orleans, Louisiana, USA, January 6-8, 2013*, pages 1380–1394, 2013.

[CDSS14a] Siu-on Chan, Ilias Diakonikolas, Rocco A. Servedio, and Xiaorui Sun. Efficient density estimation via piecewise polynomial approximation. In *Proceedings of the 45th ACM Symposium on Theory of Computing, STOC 2014, New York, NY, USA, May 31 - June 03, 2014*, pages 604–613. ACM, 2014.

[CDSS14b] Siu-on Chan, Ilias Diakonikolas, Rocco A. Servedio, and Xiaorui Sun. Near-optimal density estimation in near-linear time using variable-width histograms. In *Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1844–1852, 2014.

[CDVV14] Siu-on Chan, Ilias Diakonikolas, Gregory Valiant, and Paul Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2014, Portland, Oregon, USA, January 5-7, 2014*, pages 1193–1203, 2014.

[DDO+13] Constantinos Daskalakis, Ilias Diakonikolas, Ryan O'Donnell, Rocco A. Servedio, and Li-Yang Tan. Learning Sums of Independent Integer Random Variables. In *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2013, Berkeley, CA, USA, October 26-29, 2013*, pages 217–226. IEEE Computer Society, 2013.

[DDS12a] Constantinos Daskalakis, Ilias Diakonikolas, and Rocco A. Servedio. Learning k-modal distributions via testing. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2012, Kyoto, Japan, January 17-19, 2012*, pages 1371–1385. Society for Industrial and Applied Mathematics (SIAM), 2012.

[DDS12b] Constantinos Daskalakis, Ilias Diakonikolas, and Rocco A. Servedio. Learning Poisson Binomial Distributions. In *Proceedings of the 44th ACM Symposium on Theory of Computing, STOC 2012 Conference, New York, NY, USA, May 19 - 22, 2012*, STOC '12, pages 709–728, New York, NY, USA, 2012. ACM.

[DDS+13] Constantinos Daskalakis, Ilias Diakonikolas, Rocco A. Servedio, Gregory Valiant, and Paul Valiant. Testing k-modal distributions: Optimal algorithms via reductions. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2013, New Orleans, Louisiana, USA, January 6-8, 2013*, pages 1833–1852. Society for Industrial and Applied Mathematics (SIAM), 2013.

[DKN15a] Ilias Diakonikolas, Daniel M. Kane, and Vladimir Nikishkin. Optimal algorithms and lower bounds for testing closeness of structured distributions. In *56th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2015 (to appear)*, 2015.

[DKN15b] Ilias Diakonikolas, Daniel M. Kane, and Vladimir Nikishkin. Testing Identity of Structured Distributions. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015, San Diego, CA, USA, January 4-6, 2015*, 2015.

[DKS15] Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Nearly optimal learning and sparse covers for sums of independent integer random variables. *CoRR*, abs/1505.00662, 2015.

[DKW56] Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27(3):642–669, 1956.

[GMV06] Sudipto Guha, Andrew McGregor, and Suresh Venkatasubramanian. Streaming and sublinear approximation of entropy and information distances. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2006, Miami, Florida, USA, January 22-26, 2006*, pages 733–742, Philadelphia, PA, USA, 2006. Society for Industrial and Applied Mathematics (SIAM).

[GR00] Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. Technical Report TR00-020, Electronic Colloquium on Computational Complexity (ECCC), 2000.

[Hou86] Philip Hougaard. Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, 73:397–96, 1986.

[ILR12] Piotr Indyk, Reut Levi, and Ronitt Rubinfeld. Approximating and Testing k-Histogram Distributions in Sub-linear Time. In *Proceedings of PODS*, pages 15–22, 2012.

[KG71] J. Keilson and H. Gerber. Some results for discrete unimodality. *Journal of the American Statistical Association*, 66(334):pp. 386–389, 1971.

[Man63] Benoit Mandelbrot. New methods in statistical economics. *Journal of Political Economy*, 71(5):pp. 421–440, 1963.

[Mas90] Pascal Massart. The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990.

[MP07] Pascal Massart and Jean Picard. Concentration inequalities and model selection. Lecture Notes in Mathematics, 33, 2003, Saint-Flour, Cantal, 2007. Springer.

[Pan08] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008.

[Ron08] Dana Ron. Property Testing: A Learning Theory Perspective. *Foundations and Trends in Machine Learning*, 1(3):307–402, 2008.

[Ron10] Dana Ron. Algorithmic and analysis techniques in property testing. *Foundations and Trends in Theoretical Computer Science*, 5:73–205, 2010.

[Rub12] Ronitt Rubinfeld. Taming Big Probability Distributions. *XRDS*, 19(1):24–28, September 2012.

[SN99] Debasis Sengupta and Asok K. Nanda. Log-concave and concave distributions in reliability. *Naval Research Logistics (NRL)*, 46(4):419–433, 1999.

[SS01] Mervyn J. Silvapulle and Pranab K. Sen. *Constrained Statistical Inference*. John Wiley & Sons, Inc., 2001.

[TLSM95] Constantino Tsallis, Silvio V. F. Levy, André M. C. Souza, and Roger Maynard. Statistical-mechanical foundation of the ubiquity of Lévy distributions in nature. *Phys. Rev. Lett.*, 75:3589–3593, Nov 1995.

[Val11] Paul Valiant. Testing symmetric properties of distributions. *SIAM Journal on Computing*, 40(6):1927–1968, 2011.

[VV10] Gregory Valiant and Paul Valiant. A CLT and tight lower bounds for estimating entropy. *Electronic Colloquium on Computational Complexity (ECCC)*, 17:179, 2010.

[VV11a] Gregory Valiant and Paul Valiant. Estimating the unseen: An $n/\log n$-sample estimator for entropy and support size, shown optimal via new clts. In *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011*, pages 685–694, 2011.

[VV11b] Gregory Valiant and Paul Valiant. The power of linear estimators. In *52nd Annual IEEE Symposium on Foundations of Computer Science, FOCS 2011, Palm Springs, CA, USA, October 22-25, 2011*, pages 403–412, 2011.

[VV14] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. In *55th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2014, Philadelphia, PA, USA, October 18-21, 2014*, 2014.

[Wal09] Guenther Walther. Inference and modeling with log-concave distributions. *Statistical Science*, 24(3):319–327, 2009.

A Proof of Lemma 2.9

We now give the proof of Lemma 2.9, restated below:

Lemma 2.9 (Adapted from [DKN15b, Theorem 11])**.** *There exists an algorithm* Check-Small-ℓ_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 $\|D - \mathcal{U}_I\|_2 > \varepsilon/\sqrt{|I|}$, then the algorithm outputs* no *with probability at least $1 - \delta$;*
- *If $\|D - \mathcal{U}_I\|_2 \leq \varepsilon/2\sqrt{|I|}$, then the algorithm outputs* yes *with probability at least $1 - \delta$.*

Proof. To do so, we first describe an algorithm that distinguishes between $\|D - \mathcal{U}\|_2^2 \geq \varepsilon^2/n$ and $\|D - \mathcal{U}\|_2^2 < \varepsilon^2/(2n)$ with probability at least $2/3$, using $C \cdot \frac{\sqrt{n}}{\varepsilon^2}$ 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 $\tau \stackrel{\text{def}}{=} \frac{3}{4}\frac{m^2\varepsilon^2}{n}$ instead of $\frac{4m}{\sqrt{n}}$), define the quantities

$$Z_k \stackrel{\text{def}}{=} \left(X_k - \frac{m}{n}\right)^2 - X_k, \qquad k \in [n]$$

and $Z \stackrel{\text{def}}{=} \sum_{k=1}^n Z_k$, 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 $\Delta_k \stackrel{\text{def}}{=} (\frac{1}{n} - D(k))$, 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

$$\text{Var}\, Z = 2m^2\left(\sum_{k=1}^n \Delta_k^2 + \frac{1}{n} - 2m\sum_{k=1}^n \Delta_k^3\right) \tag{2}$$

(after expanding and since $\sum_{k=1}^n \Delta_k = 0$).

Soundness. Almost *straight from [DKN15b], but the threshold has changed.* Assume $\Delta^2 \stackrel{\text{def}}{=} \|D - \mathcal{U}\|_2^2 \geq \varepsilon^2/n$; we will show that $\Pr[\, Z < \tau\,] \leq 1/3$. By Chebyshev's inequality, it is sufficient to show that $\tau \leq \mathbb{E}Z - \sqrt{3}\sqrt{\text{Var}\, Z}$, as

$$\Pr\Big[\, \mathbb{E}Z - Z > \sqrt{3}\sqrt{\text{Var}\, Z}\,\Big] \leq 1/3\,.$$

As $\tau < \frac{3}{4}\mathbb{E}Z$, arguing that $\sqrt{3}\sqrt{\text{Var}\, Z} \leq \frac{1}{4}\mathbb{E}Z$ is enough, i.e. that $48\,\text{Var}\, Z \leq (\mathbb{E}Z)^2$. From (2), 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[9]

  $$\sum_{k=1}^n |\Delta_k|^3 \leq \left(\sum_{k=1}^n |\Delta_k|^2\right)^{3/2} = \Delta^3$$

  we get that $\left|2m\sum_{k=1}^n |\Delta_k|^3\right| \leq 2m\Delta^3 = \frac{m^2\Delta^4}{288} \cdot \frac{2\cdot 288}{m\Delta} \leq \frac{m^2\Delta^4}{288}$, using the fact that $\frac{m\Delta}{2\cdot 288} \geq \frac{C}{576\varepsilon} \geq 1$ (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.

[9] For any sequence $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, $p > 0 \mapsto \|x\|_p$ is non-increasing. In particular, for $0 < p \leq q < \infty$,

$$\left(\sum_i |x_i|^q\right)^{1/q} = \|x\|_q \leq \|x\|_p = \left(\sum_i |x_i|^p\right)^{1/p}.$$

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 ℓ_p norm, and conclude by homogeneity of the norm.

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

$$\Pr\Big[\, Z - \mathbb{E}Z > \sqrt{3}\sqrt{\operatorname{Var} Z}\,\Big] \leq 1/3$$

and therefore it is sufficient to show that

$$\tau \geq \mathbb{E}Z + \sqrt{3}\sqrt{\operatorname{Var} Z}$$

Recalling the expressions of $\mathbb{E}Z$ and $\operatorname{Var} Z$ from (2), this is tantamount to showing

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

or equivalently

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

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. □

B Proof of Theorem 4.5

In this section, we prove our structural result for MHR distributions, Theorem 4.5:

Theorem 4.5 (Monotone Hazard Rate)**.** *For all $\gamma > 0$, the class $\mathcal{MHR}$ of MHR distributions on $[n]$ is (γ, L)-decomposable for $L \stackrel{\text{def}}{=} O\Big(\frac{\log n}{\gamma}\Big)$.*

Proof. We reproduce and adapt the argument of [CDSS13, 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 $\frac{\gamma}{n}$ and $\frac{\gamma}{n^2}$. Following the structure of their proof, we write $\mathcal{Q} = \{I_1, \ldots, I_{|\mathcal{Q}|}\}$ with $I_i = [a_i, b_i]$, and define $\mathcal{Q}' = \{\, I_i \in \mathcal{Q} :\ D(a_i) > D(a_{i+1}) \,\}$, $\mathcal{Q}'' = \{\, I_i \in \mathcal{Q} :\ D(a_i) \leq D(a_{i+1}) \,\}$.
We immediately obtain the analogues of their Lemmas 5.2 and 5.3:

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

Lemma B.2. *Step 4 of Algorithm 3 adds at most $O\Big(\frac{1}{\varepsilon}\log\frac{n}{\varepsilon}\Big)$ intervals to $\mathcal{Q}$.*

Sketch. This derives from observing that now $D(I \cup I') \geq \gamma/n$, which as in [CDSS13, Lemma 5.3] in turn implies

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

Algorithm 3 DECOMPOSE-MHR$'(D, \gamma)$

Require: explicit description of MHR distribution D over $[n]$; accuracy parameter $\gamma > 0$

1: Set $J \leftarrow [n]$ and $\mathcal{Q} \leftarrow \emptyset$.
2: Let $I \leftarrow$ RIGHT-INTERVAL$(D, J, \frac{\gamma}{n})$ and $I' \leftarrow$ RIGHT-INTERVAL$(D, J \setminus I, \frac{\gamma}{n})$. Set $J \leftarrow J \setminus (I \cup I')$.
3: Set $i \in J$ to be the smallest integer such that $D(i) \geq \frac{\gamma}{n^2}$. If no such i exists, let $I'' \leftarrow J$ and go to Step 9. Otherwise, let $I'' \leftarrow \{1, \ldots, i-1\}$ and $J \leftarrow J \setminus I''$.
4: **while** $J \neq \emptyset$ **do**
5: Let $j \in J$ bet the smallest integer such that $D(j) \notin [\frac{1}{1+\gamma}, 1+\gamma]D(i)$. If no such j exists, let $I''' \leftarrow J$; otherwise let $I''' \leftarrow \{i, \ldots, j-1\}$.
6: Add I''' to $\mathcal{Q}$ and set $J \leftarrow J \setminus I'''$.
7: Let $i \leftarrow j$.
8: **end while**
9: Return $\mathcal{Q} \cup \{I, I', I''\}$

.

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

Again following their argument, we also get

$$\frac{D(a_{|\mathcal{Q}|+1})}{D(a_1)} = \prod_{I_i \in \mathcal{Q}''} \frac{D(a_{i+1})}{D(a_i)} \cdot \prod_{I_i \in \mathcal{Q}'} \frac{D(a_{i+1})}{D(a_i)}$$

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

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

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

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

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

It remains to show that $\mathcal{Q} \cup \{I, I', I''\}$ is indeed a good decomposition of $[n]$ for D, as per Definition 3.1. Since by construction every interval in $\mathcal{Q}$ 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 $\frac{\gamma}{n}$ (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 in this case as well. □

C Proofs from Section 4

This section contains the proofs omitted from Section 4, namely the distance estimation procedures for t-piecewise degree-d (Theorem 4.13), monotone hazard rate (Lemma 4.14), and log-concave

distributions (Lemma 4.15).

C.1 Proof of Theorem 4.13

In this section, we prove the following:

Theorem C.1. *Let p be a ℓ-histogram over $[-1,1)$. There is an algorithm* ProjectSinglePoly(d,ε) *which runs in time* $\mathrm{poly}(\ell, d+1, 1/\varepsilon)$*, and outputs a degree-$d$ polynomial q which defines a pdf over $[-1,1)$ such that $\|p-q\|_1 \le 3\ell_1(p,\mathcal{P}_d)+O(\varepsilon)$.*

As mentioned in Section 4, the proof of this statement is a rather straightforward adaption of the proof of [CDSS14a, Theorem 9], with two differences: first, in our setting there is no uncertainty nor probabilistic argument due to sampling, as we are provided with an explicit description of the histogram p. Second, Chan et al. require some "well-behavedness" assumption on the distribution p (for technical reasons essentially due to the sampling access), that we remove here. Besides these two points, the proof is almost identical to theirs, and we only reproduce (our modification of) it here for the sake of completeness. (Any error introduced in the process, however, is solely our responsibility.)

Proof. Some preliminary definitions will be helpful:

Definition C.2 (Uniform partition)**.** Let p be a subdistribution on an interval $I \subseteq [-1,1)$. A partition $\mathcal{I} = \{I_1,\dots,I_\ell\}$ of I is (p,η)-*uniform* if $p(I_j) \le \eta$ for all $1 \le j \le \ell$.

We will also use the following notation: For this subsection, let $I = [-1,1)$ (I will denote a subinterval of $[-1,1)$ when the results are applied in the next subsection). We write $\|f\|_1^{(I)}$ to denote $\int_I |f(x)|dx$, and we write $\mathrm{d_{TV}}^{(I)}(p,q)$ to denote $\|p-q\|_1^{(I)}/2$. We write $\mathrm{OPT}_{1,d}^{(I)}$ to denote the infimum of the distance $\|p-g\|_1^{(I)}$ between p and any degree-d subdistribution g on I that satisfies $g(I) = p(I)$.

The key step of ProjectSinglePoly is Step 2 where it calls the FindSinglePoly procedure. In this procedure $T_i(x)$ denotes the degree-i Chebychev polynomial of the first kind. The function FindSinglePoly should be thought of as the CDF of a "quasi-distribution" f; we say that $f = F'$ is a "quasi-distribution" and not a *bona fide* probability distribution because it is not guaranteed to be non-negative everywhere on $[-1,1)$. Step 2 of FindSinglePoly processes f slightly to obtain a polynomial q which is an actual distribution over $[-1,1)$.

Algorithm 4 ProjectSinglePoly

Require: parameters d,ε; and the full description of a ℓ-histogram p over $[-1,1)$.
Ensure: a degree-d distribution q such that $\mathrm{d_{TV}}(p,q) \le 3\cdot \mathrm{OPT}_{1,d} + O(\varepsilon)$

1: Partition $[-1,1)$ into $z = \Theta((d+1)/\varepsilon)$ intervals $I_0 = [i_0,i_1),\dots,I_{z-1} = [i_{z-1},i_z)$, where $i_0 = -1$ and $i_z = 1$, such that for each $j \in \{1,\dots,z\}$ we have $p(I_j) = \Theta(\varepsilon/(d+1))$ or ($|I_j| = 1$ and $p(I_j) = \Omega(\varepsilon/(d+1))$).
2: Call FindSinglePoly$(d,\ \varepsilon,\ \eta := \Theta(\varepsilon/(d+1)),\ \{I_0,\dots,I_{z-1}\},\ p$ and output the hypothesis q that it returns.

The rest of this subsection gives the proof of Theorem C.1. The claimed running time bound is obvious (the computation is dominated by solving the $\mathrm{poly}(d,1/\varepsilon)$-size LP in ProjectSinglePoly,

Algorithm 5 FindSinglePoly

Require: degree parameter d; error parameter ε; parameter η; (p,η)-uniform partition $\mathcal{I}_I = \{I_1,\ldots,I_z\}$ of interval I into z intervals such that $\sqrt{\varepsilon z}\cdot\eta \le \varepsilon/2$; a subdistribution p on I

Ensure: a number τ and a degree-d subdistribution q on I such that $q(I) = p(I)$,

$$\textsc{opt}_{1,d}^{(I)} \le \|p-q\|_1^{(I)} \le 3\textsc{opt}_{1,d}^{(I)} + \sqrt{\varepsilon z(d+1)}\cdot\eta + \text{error},$$

$0 \le \tau \le \textsc{opt}_{1,d}^{(I)}$ and error $= O((d+1)\eta)$.

1: Let τ be the solution to the following LP:

minimize τ subject to the following constraints:

(Below $F(x) = \sum_{i=0}^{d+1} c_i T_i(x)$ where $T_i(x)$ is the degree-i Chebychev polynomial of the first kind, and $f(x) = F'(x) = \sum_{i=0}^{d+1} c_i T_i'(x)$.)

(a) $F(-1) = 0$ and $F(1) = p(I)$;

(b) For each $0 \le j < k \le z$,

$$\left|\left(p([i_j,i_k)) + \sum_{j\le\ell<k} w_\ell\right) - (F(i_k) - F(i_j))\right| \le \sqrt{\varepsilon\cdot(k-j)}\cdot\eta; \tag{3}$$

(c)

$$\sum_{0\le\ell<z} w_\ell = 0, \tag{4}$$

$$-y_\ell \le w_\ell \le y_\ell \qquad \text{for all } 0\le\ell<z, \tag{5}$$

$$\sum_{0\le\ell<z} y_\ell \le \tau; \tag{6}$$

(d) The constraints $|c_i| \le \sqrt{2}$ for $i = 0,\ldots,d+1$;

(e) The constraints

$$0 \le F(z) \le 1 \quad \text{for all } z \in J,$$

where J is a set of $O\big((d+1)^6\big)$ equally spaced points across $[-1,1)$;

(f) The constraints

$$\sum_{i=0}^{d} c_i T_i'(x) \ge 0 \quad \text{for all } x \in K,$$

where K is a set of $O((d+1)^2/\varepsilon)$ equally spaced points across $[-1,1)$.

2: Define $q(x) = \varepsilon f(I)/|I| + (1-\varepsilon)f(x)$. Output q as the hypothesis pdf.

with an additional term linear in ℓ when partitioning $[-1,1)$ in the initial first step), so it suffices to prove correctness.

Before launching into the proof we give some intuition for the linear program. Intuitively $F(x)$ represents the cdf of a degree-d polynomial distribution f where $f = F'$. Constraint (a) captures the endpoint constraints that any cdf must obey if it has the same total weight as p. Intuitively, constraint (b) ensures that for each interval $[i_j, i_k)$, the value $F(i_k) - F(i_j)$ (which we may alternately write as $f([i_j, i_k))$) is close to the weight $p([i_j, i_k))$ that the distribution puts on the interval. Recall that by assumption p is $\mathrm{OPT}_{1,d}$-close to some degree-d polynomial r. Intuitively the variable w_ℓ represents $\int_{[i_\ell, i_{\ell+1})}(r-p)$ (note that these values sum to zero by constraint (c)(4), and y_ℓ represents the absolute value of w_ℓ (see constraint (c)(5)). The value τ, which by constraint (c)(6) is at least the sum of the y_ℓ's, represents a lower bound on $\mathrm{OPT}_{1,d}$. The constraints in (d) and (e) reflect the fact that as a cdf, F should be bounded between 0 and 1 (more on this below), and the (f) constraints reflect the fact that the pdf $f = F'$ should be everywhere nonnegative (again more on this below).

We begin by observing that ProjectSinglePoly calls FindSinglePoly with input parameters that satisfy FindSinglePoly's input requirements:

(I) the non-singleton intervals $I_0, \ldots, I_{z-1}$ are (p, η)-uniform; and

(II) the singleton intervals each have weight at least $\frac{\eta}{10}$.

We then proceed to show that, from there, FindSinglePoly's LP is feasible and has a high-quality optimal solution.

Lemma C.3. *Suppose p is an ℓ-histogram over $[-1,1)$, so that conditions (I) and (II) above hold; then the LP defined in Step 1 of* FindSinglePoly *is feasible; and the optimal solution τ is at most $\mathrm{OPT}_{1,d}$.*

Proof. As above, let r be a degree-d polynomial pdf such that $\mathrm{OPT}_{1,d} = \|p - r\|_1$ and $r(I) = p(I)$.We exhibit a feasible solution as follows: take F to be the cdf of r (a degree d polynomial). Take w_ℓ to be $\int_{[i_\ell, i_{\ell+1})}(r-p)$, and take y_ℓ to be $|w_\ell|$. Finally, take τ to be $\sum_{0 \le \ell < z} y_\ell$.

We first argue feasibility of the above solution. We first take care of the easy constraints: since F is the cdf of a subdistribution over I it is clear that constraints (a) and (e) are satisfied, and since both r and p are pdfs with the same total weight it is clear that constraints (c)(4) and (f) are both satisfied. Constraints (c)(5) and (c)(6) also hold. So it remains to argue constraints (b) and (d).

Note that constraint (b) is equivalent to $p + (r-p) = r$ and r satisfying $(\mathcal{I}, \varepsilon/(d+1), \varepsilon)$-inequalities, therefore this constraint is satisfied.

To see that constraint (d) is satisfied we recall some of the analysis of Arora and Khot [AK03, Section 3]. This analysis shows that since F is a cumulative distribution function (and in particular a function bounded between 0 and 1 on I) each of its Chebychev coefficients is at most $\sqrt{2}$ in magnitude.

To conclude the proof of the lemma we need to argue that $\tau \le \mathrm{OPT}_{1,d}$. Since $w_\ell = \int_{[i_\ell, i_{\ell+1})}(r-p)$ it is easy to see that $\tau = \sum_{0 \le \ell < z} y_\ell = \sum_{0 \le \ell < z} |w_\ell| \le \|p - r\|_1$, and hence indeed $\tau \le \mathrm{OPT}_{1,d}$ as required. □

Having established that with high probability the LP is indeed feasible, henceforth we let τ denote the optimal solution to the LP and F, f, w_ℓ, c_i, y_ℓ denote the values in the optimal solution. A simple argument (see e.g. the proof of [AK03, Theorem 8]) gives that $\|F\|_\infty \le 2$. Given this

bound on $\|F\|_\infty$, the Bernstein–Markov inequality implies that $\|f\|_\infty = \|F'\|_\infty \le O((d+1)^2)$. Together with (f) this implies that $f(z) \ge -\varepsilon/2$ for all $z \in [-1,1)$. Consequently $q(z) \ge 0$ for all $z \in [-1,1)$, and

$$\int_{-1}^{1} q(x)dx = \varepsilon + (1-\varepsilon)\int_{-1}^{1} f(x)dx = \varepsilon + (1-\varepsilon)(F(1) - F(-1)) = 1.$$

So $q(x)$ is indeed a degree-d pdf. To prove Theorem C.1 it remains to show that $\|p-q\|_1 \le 3\mathrm{OPT}_{1,d} + O(\varepsilon)$.

We sketch the argument that we shall use to bound $\|p-q\|_1$. A key step in achieving this bound is to bound the $\|\cdot\|_{\mathcal{A}}$ distance between f and $\widehat{p}_m + w$ where $\mathcal{A} = \mathcal{A}_{d+1}$ is the class of all unions of $d+1$ intervals and w is a function based on the w_ℓ values (see (9) below). If we can bound $\|(p+w) - f\|_{\mathcal{A}} \le O(\varepsilon)$ then it will not be difficult to show that $\|r - f\|_{\mathcal{A}} \le \mathrm{OPT}_{1,d} + O(\varepsilon)$.. Since r and f are both degree-d polynomials we have $\|r-f\|_1 = 2\|r-f\|_{\mathcal{A}} \le 2\mathrm{OPT}_{1,d} + O(\varepsilon)$, so the triangle inequality (recalling that $\|p-r\|_1 = \mathrm{OPT}_{1,d}$) gives $\|p-f\|_1 \le 3\mathrm{OPT}_{1,d} + O(\varepsilon)$. From this point a simple argument (Proposition C.5) gives that $\|p-q\|_1 \le \|p-f\|_1 + O(\varepsilon)$, which gives the theorem.

We will use the following lemma that translates $(\mathcal{I},\eta,\varepsilon)$-inequalities into a bound on $\mathcal{A}_{d+1}$ distance.

Lemma C.4. *Let $\mathcal{I} = \{I_0 = [i_0, i_1), \ldots, I_{z-1} = [i_{z-1}, i_z)\}$ be a (p,η)-uniform partition of I, possibly augmented with singleton intervals. If $h\colon I \to \mathbb{R}$ and p satisfy the $(\mathcal{I},\eta,\varepsilon)$-inequalities, then*

$$\|p-h\|^{(I)}_{\mathcal{A}_{d+1}} \le \sqrt{\varepsilon z(d+1)} \cdot \eta + \text{error},$$

where error $= O((d+1)\eta)$.

Proof. To analyze $\|p-h\|_{\mathcal{A}_{d+1}}$, consider any union of $d+1$ disjoint non-overlapping intervals $S = J_1 \cup \cdots \cup J_{d+1}$. We will bound $\|p-h\|_{\mathcal{A}_{d+1}}$ by bounding $|p(S) - h(S)|$.

We lengthen intervals in S slightly to obtain $T = J'_1 \cup \cdots \cup J'_{d+1}$ so that each J'_j is a union of intervals of the form $[i_\ell, i_{\ell+1})$. Formally, if $J_j = [a,b)$, then $J'_j = [a',b')$, where $a' = \max_\ell \{\, i_\ell : i_\ell \le a \,\}$ and $b' = \min_\ell \{\, i_\ell : i_\ell \ge b \,\}$. We claim that

$$|p(S) - h(S)| \le O((d+1)\eta) + |p(T) - f(T)|\,. \tag{7}$$

Indeed, consider any interval of the form $J = [i_\ell, i_{\ell+1})$ such that $J \cap S \ne J \cap T$ (in particular, such an interval cannot be one of the singletons). We have

$$|p(J\cap S) - p(J \cap T)| \le p(J) \le O(\eta), \tag{8}$$

where the first inequality uses non-negativity of p and the second inequality follows from the bound $p([i_\ell, i_{\ell+1})) \le \eta$. The $(\mathcal{I},\eta,\varepsilon)$-inequalities (between h and p) implies that the inequalities in (8) also hold with h in place of p. Now (7) follows by adding (8) across all $J = [i_\ell, i_{\ell+1})$ such that $J \cap S \ne J \cap T$ (there are at most $2(d+1)$ such intervals J), since each interval J_j in S can change at most two such J's when lengthened.

Now rewrite T as a disjoint union of $s \le d+1$ intervals $[i_{L_1}, i_{R_1}) \cup \cdots \cup [i_{L_s}, i_{R_s})$. We have

$$|p(T) - h(T)| \le \sum_{j=1}^{s} \sqrt{R_j - L_j} \cdot \sqrt{\varepsilon}\eta$$

by $(\mathcal{I},\eta,\varepsilon)$-inequalities between p and h. Now observing that that $0 \le L_1 \le R_1 \cdots \le L_s \le R_s \le t = O((d+1)/\varepsilon)$, we get that the largest possible value of $\sum_{j=1}^{s}\sqrt{R_j - L_j}$ is $\sqrt{sz} \le \sqrt{(d+1)z}$, so the RHS of (7) is at most $O((d+1)\eta) + \sqrt{(d+1)z\varepsilon}\eta$, as desired. □

Recall from above that F, f, w_ℓ, c_i, y_ℓ, τ denote the values in the optimal solution. We claim that

$$\|(p+w) - f\|_{\mathcal{A}} = O(\varepsilon), \tag{9}$$

where w is the subdistribution which is constant on each $[i_\ell, i_{\ell+1})$ and has weight w_ℓ there, so in particular $\|w\|_1 \le \tau \le \mathrm{OPT}_{1,d}$. Indeed, this equality follows by applying Lemma C.4 with $h = f - w$. The lemma requires h and p to satisfy $(\mathcal{I},\eta,\varepsilon)$-inequalities, which follows from constraint (b) ($(\mathcal{I},\eta,\varepsilon)$-inequalities between $p+w$ and f) and observing that $(p+w) - f = p - (f - w)$. We have also used $\eta = \Theta(\varepsilon/(d+1))$ to bound the error term of the lemma by $O(\varepsilon)$.

Next, by the triangle inequality we have (writing $\mathcal{A}$ for $\mathcal{A}_{d+1}$)

$$\|r - f\|_{\mathcal{A}} \le \|r - (p+w)\|_{\mathcal{A}} + \|(p+w) - f\|_{\mathcal{A}}.$$

The last term on the RHS has just been shown to be $O(\varepsilon)$. The first term is bounded by

$$\|r - (p+w)\|_{\mathcal{A}} \le \frac{1}{2}\|r - (p+w)\|_1 \le \frac{1}{2}(\|r-p\|_1 + \|w\|_1) \le \mathrm{OPT}_{1,d}.$$

Altogether, we get that $\|r - f\|_{\mathcal{A}} \le \mathrm{OPT}_{1,d} + O(\varepsilon)$.

Since r and f are degree d polynomials, $\|r - f\|_1 = 2\|r - f\|_{\mathcal{A}} \le 2\mathrm{OPT}_{1,d} + O(\varepsilon)$. This implies $\|p - f\|_1 \le \|p - r\|_1 + \|r - f\|_1 \le 3\mathrm{OPT}_{1,d} + O(\varepsilon)$. Finally, we turn our quasidistribution f which has value $\ge -\varepsilon/2$ everywhere into a distribution q (which is nonnegative), by redistributing the weight. The following simple proposition bounds the error incurred.

Proposition C.5. *Let f and p be any sub-quasidistribution on I. If $q = \varepsilon f(I)/|I| + (1-\varepsilon)f$, then $\|q - p\|_1 \le \|f - p\|_1 + \varepsilon(f(I) + p(I))$.*

Proof. We have

$$q - p = \varepsilon(f(I)/|I| - p) + (1-\varepsilon)(f - p).$$

Therefore

$$\|q - p\|_1 \le \varepsilon\|f(I)/|I| - p\|_1 + (1-\varepsilon)\|f - p\|_1 \le \varepsilon(f(I) + p(I)) + \|f - p\|_1.$$ □

We now have $\|p - q\|_1 \le \|p - f\|_1 + O(\varepsilon)$ by Proposition C.5, concluding the proof of Theorem C.1. □

C.2 Proof of Lemma 4.14

Lemma 4.14 (Monotone Hazard Rate)**.** *There exists a procedure* $\textsc{ProjectionDist}^*_{\mathcal{MHR}}$ *that, on input n as well as the full specification of a k-histogram distribution D on $[n]$ and of a ℓ-histogram distribution D' on $[n]$, runs in time* $\mathrm{poly}(n, 1/\varepsilon)$*, and satisfies the following.*

- *If there is $P \in \mathcal{MHR}$ such that $\|D - P\|_1 \le \varepsilon$ and $\|D' - P\|_{\mathrm{Kol}} \le \varepsilon^3$, then the procedure returns* yes*;*
- *If $\ell_1(D, \mathcal{MHR}) > 100\varepsilon$, then the procedure returns* no*.*

Proof. For convenience, let $\alpha \stackrel{\text{def}}{=} \varepsilon^3$; we also write $[i,j]$ instead of $\{i,\dots,j\}$.

First, we note that it is easy to reduce our problem to the case where, in the completeness case, we have $P \in \mathcal{MHR}$ such that $\|D-P\|_1 \le 2\varepsilon$ and $\|D-P\|_{\text{Kol}} \le 2\alpha$; while in the soundness case $\ell_1(D,\mathcal{MHR}) \ge 99\varepsilon$. Indeed, this can be done with a linear program on $\text{poly}(k,\ell)$ variables, asking to find a $(k+\ell)$-histogram D'' on a refinement of D and D' minimizing the ℓ_1 distance to D, under the constraint that the Kolmogorov distance to D' be bounded by ε. (In the completeness case, clearly a feasible solution exists, as P is one.) We therefore follow with this new formulation: either

(a) D is ε-close to a monotone hazard rate distribution P (in ℓ_1 distance) *and* D is α-close to P (in Kolmogorov distance); and

(b) D is 32ε-far from monotone hazard rate

where D is a $(k+\ell)$-histogram.

We then proceed by observing the following easy fact: suppose P is a MHR distribution on $[n]$, i.e. such that the quantity $h_i \stackrel{\text{def}}{=} \frac{P(i)}{\sum_{j=i}^n P(i)}$, $i \in [n]$ is non-increasing. Then, we have

$$P(i) = h_i \prod_{j=1}^{i-1}(1-h_j), \qquad i \in [n]. \tag{10}$$

and there is a bijective correspondence between P and $(h_i)_{i\in[n]}$.

We will write a linear program with variables $y_1,\dots,y_n$, with the correspondence $y_i \stackrel{\text{def}}{=} \ln(1-h_i)$. Note that with this parameterization, we get that if the $(y_i)_{i\in[n]}$ correspond to a MHR distribution P, then for $i \in [n]$

$$P([i,n]) = \prod_{j=1}^{i-1} e^{y_j} = e^{\sum_{j=1}^{i-1} y_j}$$

and asking that $\ln(1-\varepsilon) \le \sum_{j=1}^{i-1} y_j - \ln D([i,n]) \le \ln(1+\varepsilon)$ amounts to requiring

$$P([i,n]) \in [1\pm\varepsilon]D([i,n]).$$

We focus first on the completeness case, to provide intuition for the linear program. Suppose there exists $P \in \mathcal{MHR}$ such $P \in \mathcal{MHR}$ such that $\|D-P\|_1 \le \varepsilon$ and $\|D'-P\|_{\text{Kol}} \le \alpha$. This implies that for all $i \in [n]$, $|P([i,n]) - D([i,n])| \le 2\alpha$. Define $I = \{b+1,\dots,n\}$ to be the longest interval such that $D(\{b+1,\dots,n\}) \le \frac{\varepsilon}{2}$. It follows that for every $i \in [n]\setminus I$,

$$\frac{P([i,n])}{D([i,n])} \le \frac{D([i,n])+2\alpha}{D([i,n])} \le 1 + \frac{2\alpha}{\varepsilon/2} = 1+4\varepsilon^2 \le 1+\varepsilon \tag{11}$$

and similarly $\frac{P([i,n])}{D([i,n])} \ge \frac{D([i,n])-2\alpha}{D[i,n]} \ge 1-\varepsilon$. This means that for the points i in $[n]\setminus I$, we can write constraints asking for multiplicative closeness (within $1\pm\varepsilon$) between $e^{\sum_{j=1}^{i-1} y_j}$ and $D([i,n])$, which is very easy to write down as linear constraints on the y_i's.

The linear program. Let T and S be respectively the sets of "light" and "heavy" points, defined as $T = \{\, i \in \{1,\dots,b\} :\ D(i) \le \varepsilon^2 \,\}$ and $S = \{\, i \in \{1,\dots,b\} :\ D(i) > \varepsilon^2 \,\}$, where b is as above. (In particular, $|S| \le 1/\varepsilon^2$.)

Algorithm 6 Linear Program

Find $y_1, \ldots, y_b$

s.t.

$$y_i \le 0 \tag{12}$$

$$y_{i+1} \le y_i \qquad \forall i \in \{1, \ldots, b-1\} \tag{13}$$

$$\ln(1-\varepsilon) \le \sum_{j=1}^{i-1} y_j - \ln D([i,n]) \le \ln(1+\varepsilon) \qquad \forall i \in \{1, \ldots, b\} \tag{14}$$

$$\frac{D(i)-\varepsilon_i}{(1+\varepsilon)D[i,n]} \le -y_i \le (1+4\varepsilon)\frac{D(i)+\varepsilon_i}{(1-\varepsilon)D[i,n]} \qquad \forall i \in T \tag{15}$$

$$\sum_{i \in T} \varepsilon_i \le \varepsilon \tag{16}$$

$$0 \le \varepsilon_i \le 2\alpha \qquad \forall i \in T \tag{17}$$

$$\ln\left(1 - \frac{D(i)+2\alpha}{(1-\varepsilon)D[i,n]}\right) \le y_i \le \ln\left(1 - \frac{D(i)-2\alpha}{(1+\varepsilon)D[i,n]}\right) \qquad \forall i \in S \tag{18}$$

Given a solution to the linear program above, define $\tilde{P}$ (a non-normalized probability distribution) by setting $\tilde{P}(i) = (1-e^{y_i})e^{\sum_{j=1}^{i-1} y_j}$ for $i \in \{1, \ldots, b\}$, and $\tilde{P}(i) = 0$ for $i \in I = \{b+1, \ldots, n\}$. A MHR distribution is then obtained by normalizing $\tilde{P}$.

Completeness. Suppose $P \in \mathcal{MHR}$ is as promised. In particular, by the Kolmogorov distance assumption we know that every $i \in T$ has $P(i) \le \varepsilon^2 + 2\alpha < 2\varepsilon^2$.

- For any $i \in T$, we have that $\frac{P(i)}{P[i,n]} \le \frac{2\varepsilon^2}{(1-\varepsilon)\varepsilon} \le 4\varepsilon$, and

$$\frac{D(i)-\varepsilon_i}{(1+\varepsilon)D[i,n]} \le \frac{P(i)}{P[i,n]} \le \underbrace{-\ln\left(1-\frac{P(i)}{P[i,n]}\right)}_{-y_i} \le (1+4\varepsilon)\frac{P(i)}{P[i,n]} = (1+4\varepsilon)\frac{D(i)+\varepsilon_i}{P[i,n]} \le \frac{1+4\varepsilon}{1-\varepsilon}\frac{D(i)+\varepsilon_i}{D[i,n]} \tag{19}$$

  where we used Equation 11 for the two outer inequalities; and so (15), (16), and (17) would follow from setting $\varepsilon_i \stackrel{\text{def}}{=} |P(i) - D(i)|$ (along with the guarantees on ℓ_1 and Kolmogorov distances between P and D).
- For $i \in S$, Constraint (18) is also met, as $\frac{P(i)}{P([i,n])} \in \left[\frac{D(i)-2\alpha}{P([i,n])}, \frac{D(i)+2\alpha}{P([i,n))}\right] \subseteq \left[\frac{D(i)-2\alpha}{(1+\varepsilon)D([i,n])}, \frac{D(i)+2\alpha}{(1-\varepsilon)D([i,n])}\right]$.

Soundness. Assume a feasible solution to the linear program is found. We argue that this implies D is $O(\varepsilon)$-close to some MHR distribution, namely to the distribution obtained by renormalizing $\tilde{P}$.

In order to do so, we bound separately the ℓ_1 distance between D and $\tilde{P}$, from I, S, and T. First, $\sum_{i \in I} \left|D(i) - \tilde{P}(i)\right| = \sum_{i \in I} D(i) \le \frac{\varepsilon}{2}$ by construction. For $i \in T$, we have $\frac{D(i)}{D[i,n]} \le \varepsilon$, and

$$\tilde{P}(i) = (1-e^{y_i})e^{\sum_{j=1}^{i-1} y_j} \in [1 \pm \varepsilon]\,(1-e^{y_i})D([i,n]).$$

Now,

$$1-(1-\varepsilon)\frac{D(i)-\varepsilon_i}{(1+\varepsilon)D[i,n]} \geq e^{-\frac{D(i)-\varepsilon_i}{(1+\varepsilon)D[i,n]}} \geq e^{y_i} \geq e^{-(1+4\varepsilon)\frac{D(i)+\varepsilon_i}{(1-\varepsilon)D[i,n]}} \geq 1-(1+4\varepsilon)\frac{D(i)+\varepsilon_i}{(1-\varepsilon)D[i,n]}$$

so that

$$(1-\varepsilon)\frac{(1-\varepsilon)}{(1+\varepsilon)}(D(i)-\varepsilon_i) \leq \tilde{P}(i) \leq (1+4\varepsilon)\frac{(1+\varepsilon)}{(1-\varepsilon)}(D(i)+\varepsilon_i)$$

which implies

$$(1-10\varepsilon)(D(i)-\varepsilon_i) \leq \tilde{P}(i) \leq (1+10\varepsilon)(D(i)+\varepsilon_i)$$

so that $\sum_{i\in T}\left|D(i)-\tilde{P}(i)\right| \leq 10\varepsilon\sum_{i\in T}D(i)+(1+10\varepsilon)\sum_{i\in T}\varepsilon_i \leq 10\varepsilon+(1+10\varepsilon)\varepsilon \leq 20\varepsilon$ where the last inequality follows from Constraint (16).

To analyze the contribution from S, we observe that Constraint (18) implies that, for any $i \in S$,

$$\frac{D(i)-2\alpha}{(1+\varepsilon)D([i,n])} \leq \frac{\tilde{P}(i)}{\tilde{P}([i,n])} \leq \frac{D(i)+2\alpha}{(1-\varepsilon)D([i,n])}$$

which combined with Constraint (14) guarantees

$$\frac{D(i)-2\alpha}{(1+\varepsilon)^2\tilde{P}([i,n])} \leq \frac{\tilde{P}(i)}{\tilde{P}([i,n])} \leq \frac{D(i)+2\alpha}{(1-\varepsilon)^2\tilde{P}([i,n])}$$

which in turn implies that $\left|\tilde{P}(i)-D(i)\right| \leq 3\varepsilon\tilde{P}(i)+2\alpha$. Recalling that $|S| \leq \frac{1}{\varepsilon^2}$ and $\alpha=\varepsilon^3$, this yields $\sum_{i\in S}\left|D(i)-\tilde{P}(i)\right| \leq 3\varepsilon\sum_{i\in S}\tilde{P}(i)+2\varepsilon \leq 3\varepsilon(1+\varepsilon)+2\varepsilon \leq 8\varepsilon$. Summing up, we get $\sum_{i=1}^{n}\left|D(i)-\tilde{P}(i)\right| \leq 30\varepsilon$ which finally implies by the triangle inequality that the ℓ_1 distance between D and the normalized version of $\tilde{P}$ (a valid MHR distribution) is at most 32ε.

Running time. The running time is immediate, from executing the two linear programs on $\mathrm{poly}(n, 1/\varepsilon)$ variables and constraints. □

C.3 Proof of Lemma 4.15

Lemma 4.15 (Log-concavity)**.** *There exists a procedure* $\textsc{ProjectionDist}^*_{\mathcal{L}}$ *that, on input* n *as well as the full specifications of a* k*-histogram distribution* D *on* $[n]$ *and a* ℓ*-histogram distribution* D' *on* $[n]$*, runs in time* $\mathrm{poly}(n,k,\ell,1/\varepsilon)$*, and satisfies the following.*

- *If there is* $P \in \mathcal{L}$ *such that* $\|D-P\|_1 \leq \varepsilon$ and $\|D'-P\|_{\mathrm{Kol}} \leq \frac{\varepsilon^2}{\log^2(1/\varepsilon)}$*, then the procedure returns* yes*;*
- *If* $\ell_1(D,\mathcal{L}) \geq 100\varepsilon$*, then the procedure returns* no*.*

Proof. We set $\alpha \stackrel{\text{def}}{=} \frac{\varepsilon^2}{\log^2(1/\varepsilon)}$, $\beta \stackrel{\text{def}}{=} \frac{\varepsilon^2}{\log(1/\varepsilon)}$, and $\gamma \stackrel{\text{def}}{=} \frac{\varepsilon^2}{10}$ (so that $\alpha \ll \beta \ll \gamma \ll \varepsilon$),

Given the explicit description of a distribution D on $[n]$, which a k-histogram over a partition $\mathcal{I}=(I_1,\ldots,I_k)$ of $[n]$ with $k=\mathrm{poly}(\log n, 1/\varepsilon)$ and the explicit description of a distribution D' on $[n]$, one must *efficiently* distinguish between:

(a) D is ε-close to a log-concave P (in ℓ_1 distance) *and* D' is α-close to P (in Kolmogorov distance); and

(b) D is 100ε-far from log-concave.

If we are willing to pay an extra factor of $O(n)$, we can assume without loss of generality that we know the mode of the closest log-concave distribution (which is implicitly assumed in the following: the final algorithm will simply try all possible modes).

Outline. First, we argue that we can simplify to the case where D is unimodal. Then, reduce to the case where where D and D' are only one distribution, satisfying both requirements from the completeness case. Both can be done efficiently (Section C.3.1), and make the rest much easier. Then, perform some *ad hoc* partitioning of $[n]$, using our knowledge of D, into $\tilde{O}(1/\varepsilon^2)$ pieces such that each piece is either a "heavy" singleton, or an interval I with weight very close (multiplicatively) to $D(I)$ *under the target log-concave distribution, if it exists* (Section C.3.2). This in particular simplifies the type of log-concave distribution we are looking for: it is sufficient to look for distributions putting that very specific weight on each piece, up to a $(1 + o(1))$ factor. Then, in Section C.3.3, we write and solve a linear program to try and find such a "simplified" log-concave distribution, and reject if no feasible solution exists.

Note that the first two sections allow us to argue that instead of additive (in ℓ_1) closeness, we can enforce constraints on *multiplicative* (within a $(1+\varepsilon)$ factor) closeness between D and the target log-concave distribution. This is what enables a linear program with variables being the logarithm of the probabilities, which plays very nicely with the log-concavity constraints.

We will require the following result of Chan, Diakonikolas, Servedio, and Sun:

Theorem C.6 ([CDSS13, Lemma 4.1])**.** *Let D be a distribution over $[n]$, log-concave and non-decreasing over $\{1, \dots, b\} \subseteq [n]$. Let $a \leq b$ such that $\sigma = D(\{1, \dots, a-1\}) > 0$, and write $\tau = D(\{a, \dots, b\})$. Then $\frac{D(b)}{D(a)} \leq 1 + \frac{\tau}{\sigma}$.*

C.3.1 Step 1

Reducing to D unimodal. Using a linear program, find a closest *unimodal* distribution $\tilde{D}$ to D (also a k-histogram on $\mathcal{I}$) under the constraint that $\|D - P\|_{\mathrm{Kol}} \leq \alpha$: this can be done in time $\mathrm{poly}(k)$. If $\|D - \tilde{D}\|_1 > \varepsilon$, output REJECT.

- If D is ε-close to a log-concave distribution P as above, then it is in particular ε-close to unimodal and we do not reject. Moreover, by the triangle inequality $\|\tilde{D} - P\|_1 \leq 2\varepsilon$ and $\|\tilde{D} - P\|_{\mathrm{Kol}} \leq 2\alpha$.
- If D is 100ε-far from log-concave and we do not reject, then $\ell_1(\tilde{D}, \mathcal{L}) \geq 99\varepsilon$.

Reducing to $D = D'$. First, we note that it is easy to reduce our problem to the case where, in the completeness case, we have $P \in \mathcal{L}$ such that $\|D - P\|_1 \leq 4\varepsilon$ and $\|D - P\|_{\mathrm{Kol}} \leq 4\alpha$; while in the soundness case $\ell_1(D, \mathcal{L}) \geq 97\varepsilon$. Indeed, this can be done with a linear program on $\mathrm{poly}(k, \ell)$ variables and constraints, asking to find a $(k+\ell)$-histogram D'' on a refinement of D and D' minimizing the ℓ_1 distance to D, under the constraint that the Kolmogorov distance to D' be bounded by 2α. (In the completeness case, clearly a feasible solution exists, as (the flattening on this $(k+\ell)$-interval partition) of P is one.) We therefore follow with this new formulation: either

(a) D is 4ε-close to a log-concave P (in ℓ_1 distance) *and* D is 4α-close to P (in Kolmogorov distance); and

(b) D is 97ε-far from log-concave;

where D is a $(k+\ell)$-histogram.

This way, we have reduced the problem to a slightly more convenient one, that of Section C.3.2.

Reducing to knowing the support $[a,b]$. The next step is to compute a good approximation of the support of any target log-concave distribution. This is easily obtained in time $O(k)$ as the interval $\{a,\cdots,b\}$ such that

- $D(\{1,\dots,a-1\}) \le \alpha$ but $D(\{1,\dots,a\}) > \alpha$; and
- $D(\{b+1,\dots,\}n) \le \alpha$ but $D(\{b,\dots,n\}) > \alpha$.

Any log-concave distribution that is α-close to D must include $\{a,\cdots,b\}$ in its support, since otherwise the ℓ_1 distance between D and P is already greater than α. Conversely, if P is a log-concave distribution α-close to D, it is easy to see that the distribution obtained by setting P to be zero outside $\{a,\cdots,b\}$ and renormalizing the result is still log-concave, and $O(\alpha)$-close to D.

C.3.2 Step 2

Given the explicit description of a *unimodal* distribution D on $[n]$, which a k-histogram over a partition $\mathcal{I} = (I_1,\dots,I_k)$ of $[n]$ with $k = \mathrm{poly}(\log n, 1/\varepsilon)$, one must *efficiently* distinguish between:

(a) D is ε-close to a log-concave P (in ℓ_1 distance) and α-close to P (in Kolmogorov distance); and

(b) D is 24ε-far from log-concave,

assuming we know the mode of the closest log-concave distribution, which has support $[n]$.

In this stage, we compute a partition $\mathcal{J}$ of $[n]$ into $\tilde{O}(1/\varepsilon^2)$ intervals (here, we implicitly use the knowledge of the mode of the closest log-concave distribution, in order to apply Theorem C.6 differently on two intervals of the support, corresponding to the non-decreasing and non-increasing parts of the target log-concave distribution).

As D is unimodal, we can efficiently ($O(\log k)$) find the interval S of heavy points, that is

$$S \stackrel{\text{def}}{=} \{\, x \in [n] :\ D(x) \ge \beta \,\}.$$

Each point in S will form a singleton interval in our partition. Let $T \stackrel{\text{def}}{=} [n] \setminus S$ be its complement (T is the union of at most two intervals T_1, T_2 on which D is monotone, the head and tail of the distribution). For convenience, we focus on only one of these two intervals, without loss of generality the "head" T_1 (on which D is non-decreasing).

1. Greedily find $J = \{1,\dots,a\}$, the smallest prefix of the distribution satisfying $D(J) \in [\frac{\varepsilon}{10} - \beta, \frac{\varepsilon}{10}]$.
2. Similarly, partition $T_1 \setminus J$ into intervals $I'_1,\dots,I'_s$ (with $s = O(1/\gamma) = O(1/\varepsilon^2)$) such that $\frac{\gamma}{10} \le D(I'_j) \le \frac{9}{10}\gamma$ for all $1 \le j \le s-1$, and $\frac{\gamma}{10} \le D(I'_s) \le \gamma$. This is possible as all points not in S have weight less than β, and $\beta \ll \gamma$.

Discussion: why doing this? We focus on the completeness case: let $P \in \mathcal{L}$ be a log-concave distribution such that $\|D-P\|_1 \le \varepsilon$ and $\|D-P\|_{\mathrm{Kol}} \le \alpha$. Applying Theorem C.6 on J and the

I'_j's, we obtain (using the fact that $\left|P(I'_j) - D(I'_j)\right| \le 2\alpha$) that:

$$\frac{\max_{x\in I'_j} P(x)}{\min_{x\in I'_j} P(x)} \le 1 + \frac{D(I'_j)+2\alpha}{D(J)-2\alpha} \le 1 + \frac{\gamma+2\alpha}{\frac{\varepsilon}{10}-2\alpha} = 1+\varepsilon+O\left(\frac{\varepsilon^2}{\log^2(1/\varepsilon)}\right) \stackrel{\text{def}}{=} 1+\kappa.$$

Moreover, we also get that each resulting interval I'_j will satisfy

$$D(I'_j)(1-\kappa_j) = D(I'_j) - 2\alpha \le P(I'_j) \le D(I'_j) + 2\alpha = D(I'_j)(1+\kappa_j)$$

with $\kappa_j \stackrel{\text{def}}{=} \frac{2\alpha}{D(I'_j)} = \Theta\Big(1/\log^2(1/\varepsilon)\Big)$.

Summing up, we have a partition of $[n]$ into $|S|+2 = \tilde{O}(1/\varepsilon^2)$ intervals such that:

- The (at most) two end intervals have $D(J) \in \left[\frac{\varepsilon}{10}-\beta, \frac{\varepsilon}{10}\right]$, and thus $P(J) \in \left[\frac{\varepsilon}{10}-\beta-2\alpha, \frac{\varepsilon}{10}+2\alpha\right]$;
- the $\tilde{O}(1/\varepsilon^2)$ singleton-intervals from S are points x with $D(x) \ge \beta$, so that $P(x) \ge \beta - 2\alpha \ge \frac{\beta}{2}$;
- each other interval $I = I'_j$ satisfies

$$(1-\kappa_j)D(I) \le P(I) \le (1+\kappa_j)D(I) \tag{20}$$

with $\kappa_j = O\Big(1/\log^2(1/\varepsilon)\Big)$; and

$$\frac{\max_{x\in I} P(x)}{\min_{x\in I} P(x)} \le 1+\kappa < 1+\frac{3}{2}\varepsilon. \tag{21}$$

We will use in the constraints of the linear program the fact that $(1+\frac{3}{2}\varepsilon)(1+\kappa_j) \le 1+2\varepsilon$, and $\frac{1-\kappa_j}{1+\frac{3}{2}\varepsilon} \ge \frac{1}{1+2\varepsilon}$.

C.3.3 Step 3

We start by computing the partition $\mathcal{J} = (J_1, \dots, J_\ell)$ as in Section C.3.2; with $\ell = \tilde{O}(1/\varepsilon^2)$; and write $J_j = \{a_j, \dots, b_j\}$ for all $j \in [\ell]$. We further denote by S and T the set of heavy and light points, following the notations from Section C.3.2; and let $T' \stackrel{\text{def}}{=} T_1 \cup T_2$ be the set obtained by removing the two "end intervals" (called J in the previous section) from T.

Lemma C.7 (Soundness)**.** *If the linear program (Algorithm 7) has a feasible solution, then* $\ell_1(D, \mathcal{L}) \le O(\varepsilon)$.

Proof. A feasible solution to this linear program will define (setting $p_i = e^{x_i}$) a sequence $p = (p_1, \dots, p_n) \in (0,1]^n$ such that

- p takes values in $(0,1]$ (from (22));
- p is log-concave (from (23));
- p is "$(1+O(\varepsilon))$-multiplicatively constant" on each interval J_j (from (24));
- p puts roughly the right amount of weight on each J_i:
  - weight $(1 \pm O(\varepsilon))D(J)$ on every J from T (from (24)), so that the ℓ_1 distance between D and p coming from T' is at most $O(\varepsilon)$;

Algorithm 7 Linear Program

Find $x_1, \ldots, x_n, \varepsilon_1, \ldots, \varepsilon_{|S|}$

s.t.

$$x_i \le 0 \tag{22}$$

$$x_i - x_{i-1} \ge x_{i+1} - x_i \qquad \forall i \in [n] \tag{23}$$

$$-\ln(1+2\varepsilon) \le x_i - \mu_j \le \ln(1+2\varepsilon), \qquad \forall j \in T', \forall i \in J_j \tag{24}$$

$$-2\frac{\varepsilon_i}{D(i)} \le x_i - \ln D(i) \le \frac{\varepsilon_i}{D(i)}, \qquad \forall i \in S \tag{25}$$

$$\sum_{i \in S} \varepsilon_i \le \varepsilon \tag{26}$$

$$0 \le \varepsilon_i \le 2\alpha \qquad \forall i \in S \tag{27}$$

$$\tag{28}$$

where $\mu_j \stackrel{\text{def}}{=} \ln \frac{D(J_j)}{|J_j|}$ for $j \in T'$.

– it puts weight approximately $D(J)$ on every singleton J from S, i.e. such that $D(J) \ge \beta$. To see why, observe that each ε_i is in $[0, 2\alpha]$ by constraints (27). In particular, this means that $\frac{\varepsilon_i}{D(i)} \le 2\frac{\alpha}{\beta} \ll 1$, and we have

$$D(i) - 4\varepsilon_i \le D(i) \cdot e^{-4\frac{\varepsilon_i}{D(i)}} \le p_i = e^{x_i} \le D(i) \cdot e^{2\frac{\varepsilon_i}{D(i)}} \le D(i) + 4\varepsilon_i$$

and together with (26) this guarantees that the ℓ_1 distance between D and p coming from S is at most ε.

Note that the solution obtained this way may not sum to one – i.e., is not necessarily a probability distribution. However, it is easy to renormalize p to obtain a *bona fide* probability distribution $\tilde{P}$ as follows: set $\tilde{P} = \frac{p(i)}{\sum_{i \in S \cup T'} p(i)}$ for all $i \in S \cup T'$, and $p(i) = 0$ for $i \in T \setminus T'$.

Since by the above discussion we know that $p(S \cup T')$ is within $O(\varepsilon)$ of $D(S \cup T')$ (itself in $[1-\frac{9\varepsilon}{5}, 1+\frac{9\varepsilon}{5}]$ by construction of T'), $\tilde{P}$ is a log-concave distribution such that $\|\tilde{P} - D\|_1 = O(\varepsilon)$. □

Lemma C.8 (Completeness)**.** *If There is P in $\mathcal{L}$ such that $\|D - P\|_1 \le \varepsilon$ and $\|D - P\|_{\text{Kol}} \le \alpha$, then the linear program (Algorithm 7) has a feasible solution.*

Proof. Let $P \in \mathcal{L}$ such that $\|D - P\|_1 \le \varepsilon$ and $\|D - P\|_{\text{Kol}} \le \alpha$. Define $x_i \stackrel{\text{def}}{=} \ln P(i)$ for all $i \in [n]$. Constraints (22) and (23) are immediately satisfied, since P is log-concave. By the discussion from Section C.3.2 (more specifically, Eq. (20) and (21)), constraint (24) holds as well.

Letting $\varepsilon_i \stackrel{\text{def}}{=} |P(i) - D(i)|$ for $i \in S$, we also immediately have (26) and (27) (since $\|P - D\|_1 \le \varepsilon$ and $\|D - P\|_{\text{Kol}} \le \alpha$ by assumption). Finally, to see why (25) is satisfied, we rewrite

$$x_i - \ln D(i) = \ln \frac{P(i)}{D(i)} = \ln \frac{D(i) \pm \varepsilon_i}{D(i)} = \ln(1 \pm \frac{\varepsilon_i}{D(i)})$$

and use the fact that $\ln(1+x) \le x$ and $\ln(1-x) \ge -2x$ (the latter for $x < \frac{1}{2}$, along with $\frac{\varepsilon_i}{D(i)} \le \frac{2\alpha}{\beta} \ll 1$). □

C.3.4 Putting it all together: Proof of Lemma 4.10

The algorithm is as follows (keeping the notations from Section C.3.1 to Section C.3.3):

- Set α, β, γ as above.
- Follow Section C.3.1 to reduce it to the case where D is unimodal and satisfies the conditions for Kolmogorov and ℓ_1 distance; and a good $[a, b]$ approximation of the support is known
- For each of the $O(n)$ possible modes $c \in [a, b]$:
  - Run the linear program Algorithm 7, return ACCEPT if a feasible solution is found
- None of the linear programs was feasible: return REJECT.

The correctness comes from Lemma C.7 and Lemma C.8 and the discussions in Section C.3.1 to Section C.3.3; as for the claimed running time, it is immediate from the algorithm and the fact that the linear program executed each step has $\mathrm{poly}(n, 1/\varepsilon)$ constraints and variables.

□

D Proof of Theorem 6.3

In this section, we establish our lower bound for tolerant testing of the Binomial distribution, restated below:

Theorem 6.3. *There exists an absolute constant $\varepsilon_0 > 0$ such that the following holds. Any algorithm which, given sampling access to an unknown distribution D on Ω and parameter $\varepsilon \in (0, \varepsilon_0)$, distinguishes with probability at least $2/3$ between (i) $\|D - \mathrm{Bin}(n, 1/2)\|_1 \leq \varepsilon$ and (ii) $\|D - \mathrm{Bin}(n, 1/2)\|_1 \geq 100\varepsilon$ must use $\Omega\left(\frac{1}{\varepsilon}\frac{\sqrt{n}}{\log n}\right)$ samples.*

The theorem will be a consequence of the (slightly) more general result below:

Theorem D.1. *There exist absolute constants $\varepsilon_0 > 0$ and $\lambda > 0$ such that the following holds. Any algorithm which, given* SAMP *access to an unknown distribution D on Ω and parameter $\varepsilon \in (0, \varepsilon_0)$, distinguishes with probability at least $2/3$ between (i) $\|D - \mathrm{Bin}\left(n, \frac{1}{2}\right)\|_1 \leq \varepsilon$ and (ii) $\|D - \mathrm{Bin}\left(n, \frac{1}{2}\right)\|_1 \geq \lambda\varepsilon^{1/3} - \varepsilon$ must use $\Omega\left(\varepsilon\frac{\sqrt{n}}{\log(\varepsilon^{2/3}n)}\right)$ samples.*

By choosing a suitable ϕ and working out the corresponding parameters, this for instance enables us to derive the following:

Corollary D.2. *There exists an absolute constant $\varepsilon_0 \in (0, 1/1000)$ such that the following holds. Any algorithm which, given* SAMP *access to an unknown distribution D on Ω, distinguishes with probability at least $2/3$ between (i) $\|D - \mathrm{Bin}\left(n, \frac{1}{2}\right)\|_1 \leq \varepsilon_0$ and (ii) $\|D - \mathrm{Bin}\left(n, \frac{1}{2}\right)\|_1 \geq 100\varepsilon_0$ must use $\Omega\left(\frac{\sqrt{n}}{\log n}\right)$ samples.*

By standard techniques, this will in turn imply Theorem 6.3.

Proof of Theorem D.1. Hereafter, we write for convenience $B_n \stackrel{\text{def}}{=} \mathrm{Bin}\left(n, \frac{1}{2}\right)$. To prove this lower bound, we will rely on the following:

Theorem D.3 ([VV10, Theorem 1]). *For any constant $\phi \in (0, 1/4)$, following holds. Any algorithm which, given* SAMP *access to an unknown distribution D on Ω, distinguishes with probability at least $2/3$ between (i) $\|D - \mathcal{U}_n\|_1 \le \phi$ and (ii) $\|D - \mathcal{U}_n\|_1 \ge \frac{1}{2} - \phi$, must have sample complexity at least $\frac{\phi}{32}\frac{n}{\log n}$.*

Without loss of generality, assume n is even (so that B_n has only one mode located at $\frac{n}{2}$). For $c > 0$, we write $I_{n,c}$ for the interval $\{\frac{n}{2} - c\sqrt{n}, \dots, \frac{n}{2} + c\sqrt{n}\}$ and $J_{n,c} \stackrel{\text{def}}{=} \Omega \setminus I_{n,c}$.

Fact D.4. *For any $c > 0$,*

$$\frac{B_n(\frac{n}{2} + c\sqrt{n})}{B_n(n/2)}, \frac{B_n(\frac{n}{2} - c\sqrt{n})}{B_n(n/2)} \underset{n\to\infty}{\sim} e^{-2c^2}$$

and

$$B_n(I_{n,c}) \in (1 \pm o(1)) \cdot [e^{-2c^2}, 1] \cdot 2c\sqrt{\frac{2}{\pi}} = \Theta(c)\,.$$

The reduction proceeds as follows: given sampling access to D on $[n]$, we can simulate sampling access to a distribution D' on $[N]$ (where $N = \Theta(n^2)$) such that

- if $\|D - \mathcal{U}_n\|_1 \le \phi$, then $\|D' - B_N\|_1 < \varepsilon$;
- if $\|D - \mathcal{U}_n\|_1 \ge \frac{1}{2} - \phi$, then $\|D' - B_N\|_1 > \varepsilon' - \varepsilon$

for $\varepsilon \stackrel{\text{def}}{=} \Theta(\phi^{3/2})$ and $\varepsilon' \stackrel{\text{def}}{=} \Theta(\phi^{\frac{1}{2}})$; in a way that preserves the sample complexity.

More precisely, define $c \stackrel{\text{def}}{=} \sqrt{2\ln\frac{1}{1-\phi}} = \Theta(\sqrt{\phi})$ (so that $\phi = 1 - e^{-2c^2}$) and N such that $|I_{N,c}| = n$ (that is, $N = (n/(2c))^2 = \Theta(n^2/\phi)$). From now on, we can therefore identify $[n]$ to $I_{N,c}$ in the obvious way, and see a draw from D as an element in $I_{N,c}$.

Let $p \stackrel{\text{def}}{=} B_N(I_{N,c}) = \Theta(\sqrt{\phi})$, and $B_{N,c}$, $\bar{B}_{N,c}$ respectively denote the conditional distributions induced by B_N on $I_{N,c}$ and $J_{N,c}$. Intuitively, we want D to be mapped to the conditional distribution of D' on $I_{N,c}$, and the conditional distribution of D' on $J_{N,c}$ to be exactly $\bar{B}_{N,c}$. This is done as by defining D' by the process below:

- with probability p, we draw a sample from D (seen as an element of $I_{N,c}$;
- with probability $1 - p$, we draw a sample from $\bar{B}_{N,c}$.

Let $\tilde{B}_N$ be defined as the distribution which exactly matches B_N on $J_{n,c}$, but is uniform on $I_{n,c}$:

$$\tilde{B}_N(i) = \begin{cases} \frac{p}{|I_{n,c}|} & i \in I_{n,c} \\ B_N(i) & i \in J_{n,c} \end{cases}$$

From the above, we have that $\|D' - \tilde{B}_N\|_1 = p \cdot \|D - \mathcal{U}_n\|_1$. Furthermore, by Fact D.4, Lemma 2.8 and the definition of $I_{n,c}$, we get that $\|B_N - \tilde{B}_N\|_1 = p \cdot \|(B_N)_{I_{n,c}} - \mathcal{U}_{I_{n,c}}\|_1 \le p \cdot \phi$. Putting it all together,

- If $\|D - \mathcal{U}_n\|_1 \le \phi$, then by the triangle inequality $\|D' - B_N\|_1 \le p(\phi + \phi) = 2p\phi$;
- If $\|D - \mathcal{U}_n\|_1 \ge \frac{1}{2} - \phi$, then similarly $\|D' - B_N\|_1 \ge p(\frac{1}{2} - \phi - \phi) = \frac{p}{4} - 2p\phi$.

Recalling that $p = \Theta(\sqrt{\phi})$ and setting $\varepsilon \stackrel{\text{def}}{=} 2p\phi$ concludes the reduction. From Theorem D.3, we conclude that

$$\frac{\phi}{32}\frac{n}{\log n} = \Omega\left(\phi\frac{\sqrt{\phi N}}{\log(\phi N)}\right) = \Omega\left(\varepsilon\frac{\sqrt{N}}{\log(\varepsilon^{2/3}N)}\right)$$

samples are necessary.

□

Proof of Corollary D.2. The corollary follows from the proof of Corollary D.2, by taking $\phi = 1/1000$ and computing the corresponding ε and $\varepsilon' - \varepsilon$ to check that indeed $\lim_{n\to\infty} \frac{\varepsilon'-\varepsilon}{\varepsilon} > 100$. □