Electronic Colloquium on Computational Complexity, Revision 1 of Report No. 75 (2017)

Fourier-Based Testing for Families of Distributions

Clément L. Canonne*
Columbia University
ccanonne.d@cs.columbia.edu

Ilias Diakonikolas†
University of Southern California
diakonik@usc.edu

Alistair Stewart‡
University of Southern California
stewart.al@gmail.com

August 7, 2017

Abstract

We study the general problem of testing whether an unknown distribution belongs to a specified family of distributions. More specifically, given a distribution family $\mathcal{P}$ and sample access to an unknown discrete distribution $\mathbf{P}$, we want to distinguish (with high probability) between the case that $\mathbf{P} \in \mathcal{P}$ and the case that $\mathbf{P}$ is ϵ-far, in total variation distance, from every distribution in $\mathcal{P}$. This is the prototypical hypothesis testing problem that has received significant attention in statistics and, more recently, in theoretical computer science.

The sample complexity of this general inference task depends on the underlying family $\mathcal{P}$. The gold standard in distribution property testing is to design sample-optimal and computationally efficient algorithms for this task. The main contribution of this work is a simple and general testing technique that is applicable to all distribution families whose *Fourier spectrum* satisfies a certain approximate *sparsity* property. To the best of our knowledge, ours is the first use of the Fourier transform in the context of distribution testing.

We apply our Fourier-based framework to obtain near sample-optimal and computationally efficient testers for the following fundamental distribution families: Sums of Independent Integer Random Variables (SIIRVs), Poisson Multinomial Distributions (PMDs), and Discrete Log-Concave Distributions. For the first two, ours are the first non-trivial testers in the literature, vastly generalizing previous work on testing Poisson Binomial Distributions. For the third, our tester improves on prior work in both sample and time complexity.

*Supported by NSF grants CCF-1115703 and NSF CCF-1319788.
†Supported by NSF Award CCF-1652862 (CAREER) and a Sloan Research Fellowship.
‡Supported by a USC startup grant.

ISSN 1433-8092

1 Introduction

1.1 Background and Motivation The prototypical inference question in the area of *distribution property testing* [BFR$^+$00] is the following: Given a set of samples from a collection of probability distributions, can we determine whether these distributions satisfy a certain property? During the past two decades, this broad question – whose roots lie in statistical hypothesis testing [NP33, LR05] – has received considerable attention by the computer science community, see [Rub12, Can15] for two recent surveys. After two decades of study, for many properties of interest there exist sample-optimal testers (matched by information-theoretic lower bounds) [Pan08, CDVV14, VV14, DKN15, ADK15, DK16].

In this work, we focus on the problem of testing whether an unknown distribution belongs to a given family of discrete *structured* distributions. Let $\mathcal{P}$ be a family of discrete distributions over a total order (e.g., $[n]$) or a partial order (e.g., $[n]^k$). The problem of *membership testing for* $\mathcal{P}$ is the following: Given sample access to an unknown distribution $\mathbf{P}$ (effectively supported on the same domain as $\mathcal{P}$), we want to distinguish between the case that $\mathbf{P} \in \mathcal{P}$ versus $d_{\mathrm{TV}}(\mathbf{P}, \mathcal{P}) > \epsilon$. (Here, d_{TV} denotes the total variation distance between distributions.) The sample complexity of this problem depends on the underlying family $\mathcal{P}$. For example, if $\mathcal{P}$ contains a single distribution over a domain of size n, the sample complexity of the testing problem is $\Theta(n^{1/2}/\epsilon^2)$ [CDVV14, VV14, DKN15, ADK15].

In this work, we give a general technique to test membership in various distribution families over discrete domains, i.e., to solve the following task:

> $\mathfrak{T}(\mathcal{P}, \epsilon)$: given a family of discrete distributions $\mathcal{P}$ over some partially or totally ordered set, parameter $\epsilon \in (0, 1]$, and sample access to an unknown distribution $\mathbf{P}$ over the same domain, how many samples are required to distinguish, with probability $3/5$, between the case that $\mathbf{P} \in \mathcal{P}$ versus $d_{\mathrm{TV}}(\mathbf{P}, \mathcal{P}) > \epsilon$?

Before we state our results in full generality, we present concrete applications to a number of well-studied distribution families.

1.2 Our Results Our first result is a nearly sample-optimal testing algorithm for sums of independent integer random variables (SIIRVs). Formally, an (n, k)-SIIRV is a sum of n independent integer random variables each supported in $\{0, 1, \ldots, k-1\}$. We will denote the set of (n, k)-SIIRVs by $\mathcal{SIIRV}_{n,k}$. SIIRVs comprise a rich class of distributions that arise in many settings. The special case of $k = 2$ was first considered by Poisson [Poi37] as a non-trivial extension of the Binomial distribution, and is known as Poisson binomial distribution (PBD). In application domains, SIIRVs have many uses in research areas such as survey sampling, case-control studies, and survival analysis, see e.g., [CL97] for a survey of the many practical uses of these distributions. In addition to their practical applications, SIIRVs are of fundamental probabilistic interest and have been extensively studied in the theory of probability and statistics [Che52, Hoe63, DP09b, Pre83, Kru86, BHJ92, CL10, CGS11]. We prove:

Theorem 1.1 (Testing SIIRVs)**.** *Given parameters $k, n \in \mathbb{N}$ and sample access to a distribution over $\mathbb{N}$, there exists an algorithm (Algorithm 1) for $\mathfrak{T}(\mathcal{SIIRV}_{n,k}, \epsilon)$ which takes*

$$O\left(\frac{kn^{1/4}}{\epsilon^2}\log^{1/4}\frac{1}{\epsilon} + \frac{k^2}{\epsilon^2}\log^2\frac{k}{\epsilon}\right)$$

samples, and runs in time $n(k/\epsilon)^{O(k\log(k/\epsilon))}$.

Prior to our work, no non-trivial[1] tester was known for (n,k)-SIIRVs for any $k > 2$. [CDGR17] showed a sample lower bound of $\Omega\left(k^{1/2}n^{1/4}/\epsilon^2\right)$, but their techniques did not yield any non-trivial sample upper bound.

For the special case of PBDs ($k = 2$), Acharya and Daskalakis [AD15] gave a tester with sample complexity $O\left(\frac{n^{1/4}}{\epsilon^2}\sqrt{\log 1/\epsilon} + \frac{\log^{5/2} 1/\epsilon}{\epsilon^6}\right)$, running time $O\left(\frac{n^{1/4}}{\epsilon^2}\sqrt{\log 1/\epsilon} + (1/\epsilon)^{O(\log^2 1/\epsilon)}\right)$, and also showed a sample lower bound of $\Omega(n^{1/4}/\epsilon^2)$. The special case of our Theorem 1.1 for $k = 2$ yields an improvement over [AD15] in both sample size and runtime:

Theorem 1.2 (Testing PBDs)**.** *Given parameter $n \in \mathbb{N}$ and sample access to a distribution over $\mathbb{N}$, there exists an algorithm (Algorithm 1) for $\mathfrak{T}(\mathcal{PBD}_n, \epsilon)$ which takes*

$$O\left(\frac{n^{1/4}}{\epsilon^2}\log^{1/4}\frac{1}{\epsilon} + \frac{\log^2 1/\epsilon}{\epsilon^2}\right)$$

samples, and runs in time $n^{1/4}\cdot\tilde{O}\left(1/\epsilon^2\right) + (1/\epsilon)^{O(\log\log(1/\epsilon))}$.

Note that the sample complexity of our algorithm is $n^{1/4}\cdot\tilde{O}(1/\epsilon^2)$, matching the information-theoretic lower bound up to a logarithmic factor in $1/\epsilon$. In particular, our algorithm does not incur the extraneous $\Omega(1/\epsilon^6)$ term of [AD15]. Moreover, our runtime has a $(1/\epsilon)^{O(\log\log(1/\epsilon))}$ dependence, as opposed to $(1/\epsilon)^{O(\log^2 1/\epsilon)}$. The improved running time relies on a more efficient computational "projection step" in our general framework, which leverages the geometric structure of Poisson Binomial distributions.

We remark that the guarantees provided by the above two theorems are actually stronger than the usual property testing one. Namely, whenever the algorithm returns accept, then it also provides a (proper) hypothesis $\mathbf{H}$ such that $d_{\mathrm{TV}}(\mathbf{P},\mathbf{H}) \le \epsilon$ with probability at least $3/5$.

A broad generalization of PBDs to the high-dimensional setting is the family of Poisson Multinomial Distributions (PMDs). Formally, an (n,k)-PMD is any random variable of the form $X = \sum_{i=1}^n X_i$, where the X_i's are independent random vectors supported on the set $\{e_1, e_2, \ldots, e_k\}$ of standard basis vectors in $\mathbb{R}^k$. We will denote by $\mathcal{PMD}_{n,k}$ the set of (n,k)-PMDs. PMDs comprise a broad class of discrete distributions of fundamental importance in computer science, probability, and statistics. A large body of work in the probability and statistics literature has been devoted to the study of the behavior of PMDs under various structural conditions [Bar88, Loh92, BHJ92, Ben03, Roo99, Roo10]. PMDs generalize the familiar multinomial distribution, and describe many distributions commonly encountered in computer science (see, e.g., [DP07, DP08, Val08, VV11]). Recent years have witnessed a flurry of research activity on PMDs and related distributions, from several perspectives of theoretical computer science, including learning [DDS12, DDO$^+$13, DKS16b, DKT15, DKS16c], property testing [Val08, VV10, VV11], computational game theory [DP07, DP08, BCI$^+$08, DP09a, DP14, GT14, CDS17], and derandomization [GMRZ11, BDS12, De15, GKM15]. We prove the following:

Theorem 1.3 (Testing PMDs)**.** *Given parameters $k, n \in \mathbb{N}$ and sample access to a distribution over $\mathbb{N}^k$, there exists an algorithm (Algorithm 7) for $\mathfrak{T}(\mathcal{PMD}_{n,k}, \epsilon)$ which takes*

$$O\left(\frac{n^{(k-1)/4}k^{2k}}{\epsilon^2}\log(k/\epsilon)^k\right)$$

samples, and runs in time $n^{O(k^3)}\cdot(1/\epsilon)^{O(k^3\frac{\log(k/\epsilon)}{\log\log(k/\epsilon)})^{k-1}}$ or alternatively in time $n^{O(k)}\cdot 2^{O(k^{5k}\log(1/\epsilon)^{k+2})}$.

[1] By the term "non-trivial" here we refer to a testing algorithm that uses fewer samples than just learning the unknown distribution and then checking whether it is close to a distribution in the family.

For the sake of intuition, we note that Theorem 1.3 is particularly interesting in the regime that n is large and k is small. Indeed, the sample complexity of testing PMDs is inherently *exponential* in k: We prove a sample lower bound of $\Omega_k(n^{(k-1)/4}/\epsilon^2)$ (Theorem 8.1),[2] nearly-matching our upper bound for constant k.

Finally, we demonstrate the versatility of our techniques by obtaining (Section 7) a testing algorithm for discrete log-concavity. Log-concave distributions constitute a broad and flexible non-parametric family that is extensively used in modeling and inference [Wal09]. In the discrete setting, log-concave distributions encompass a range of fundamental types of discrete distributions, including binomial, negative binomial, geometric, hypergeometric, Poisson, Poisson Binomial, hyper-Poisson, Pólya-Eggenberger, and Skellam distributions. Log-concave distributions have been studied in a wide range of different contexts including economics [An95], statistics and probability theory (see [SW14] for a recent survey), theoretical computer science [LV07], and algebra, combinatorics and geometry [Sta89]. We will denote by $\mathcal{LCV}_n$ the class of log-concave distributions over $[n]$. We prove:

Theorem 1.4 (Testing Log-Concavity)**.** *Given a parameter $n \in \mathbb{N}$ and sample access to a distribution over $\mathbb{N}$, there exists an algorithm (Algorithm 8) for $\mathfrak{T}(\mathcal{LCV}_n, \epsilon)$ which takes*

$$O\left(\frac{\sqrt{n}}{\epsilon^2}\right) + \tilde{O}\left(\frac{1}{\epsilon^{5/2}}\right)$$

samples, and runs in time $O(\sqrt{n} \cdot \mathrm{poly}(1/\epsilon))$.

Our discrete log-concavity tester improves on previous work in terms of both sample and time complexity. Specifically, [ADK15] gave a log-concavity tester with sample complexity $O\left(\sqrt{n}/\epsilon^2 + 1/\epsilon^5\right)$, while [CDGR17] obtained a tester with sample complexity $\tilde{O}\left(\sqrt{n}/\epsilon^{7/2}\right)$. Our sample complexity dominates both these bounds, and is significantly better when ϵ is small. The algorithms in [ADK15, CDGR17] run in $\mathrm{poly}(n/\epsilon)$ time, as they involve solving a linear program of $\mathrm{poly}(n/\epsilon)$ size. In contrast, the running time of our algorithm is *sublinear* in n.

1.3 Our Techniques and Comparison to Previous Work All the testing algorithms in this paper follow from a simple and general technique that may be of broader interest. The common property of the underlying distribution families $\mathcal{P}$ that allows for our unified testing approach is the following: Let $\mathbf{P}$ be the probability mass function of any distribution in $\mathcal{P}$. Then, *the Fourier transform of $\mathbf{P}$ is approximately sparse*, in a well-defined sense.

For concreteness, we elaborate on our technique for the case of SIIRVs. The starting point of our approach is the observation from [DKS16b] that (n, k)-SIIRVs – in addition to having a relatively small effective support – also have an approximately sparse Fourier representation. Roughly speaking, most of their Fourier mass is concentrated on a small subset of Fourier coefficients, which can be computed efficiently.

This suggests the following natural approach to testing (n, k)-SIIRVs: first, identify the effective support I of the distribution $\mathbf{P}$ and check that it is appropriately small. If it is not, then reject. Then, compute the corresponding small subset S of the Fourier domain, and check that almost no Fourier mass of $\mathbf{P}$ lies outside S. Otherwise, one can safely reject, as this is a certificate that $\mathbf{P}$ is not an (n, k)-SIIRV. Combining the two steps, one can show that learning the Fourier transform of $\mathbf{P}$ (in L_2-norm) on this small subset S only, is sufficient to learn $\mathbf{P}$ itself in total variation distance. The former goal can be performed with relatively few samples, as S is sufficiently small.

At this point, we have obtained a distribution $\mathbf{H}$ – succinctly represented by its Fourier transform on S – such that $\mathbf{P}$ and $\mathbf{H}$ are close in total variation distance. It only remains to perform a computational

[2] Here, we use the notation $\Omega_k(\cdot)$, $O_k(\cdot)$ to indicate that the parameter k is seen as a constant.

"projection step" to verify that $\mathbf{H}$ itself is close to some (n,k)-SIIRV. This will clearly be the case if indeed $\mathbf{P} \in \mathcal{SIIRV}_{n,k}$.

Although the aforementioned approach forms the core of our SIIRV testing algorithm (Algorithm 3), the actual tester has to address separately the case where $\mathbf{P}$ has small variance, which can be handled by a testing-via-learning approach. Our main contribution is thus to describe how to efficiently perform the second step, i.e., the Fourier sparsity testing. This is done in Theorem 3.1, which describes a simple algorithm to perform this step. The algorithm proceeds by essentially considering the Fourier coefficients of the empirical distribution (obtained by taking a small number of samples). Interestingly, the main idea underlying Theorem 3.1 is to avoid analyzing directly the behavior of these Fourier coefficients – which would naively require too high a time complexity. Instead, we rely on Plancherel's identity and reduce the problem to the analysis of a different task: that of the sample complexity of L_2 *identity testing* (Proposition 3.2). By a tight analysis of this L_2 tester, we get as a byproduct that several Fourier quantities of interest (of our empirical distribution) simultaneously enjoy good concentration – while arguing concentration of each of these terms separately would yield a suboptimal time complexity.

A nearly identical method works for PMDs as well. Moreover, our approach can be abstracted to yield a general testing framework, as we explain in Section 5. It is interesting to remark that the Fourier transform has been used to learn PMDs and SIIRVs [DKS16b, DKT15, DKS16c, DDKT16], and therefore it may not be entirely surprising that it has applications to testing as well. Notably, our Fourier testing technique gives an improved and nearly-optimal algorithms for log-concavity, for which no Fourier learning algorithm was known. More generally, testing membership to a class using the Fourier transform is significantly more challenging than learning. A fundamental difference is that in the testing setting we need to handle distributions that do *not* belong to the class (e.g., SIIRVs, PMDs), but are far from the class in an arbitrary way. In contrast, learning algorithms work under the promise that the distribution is in the underlying class, and thus can leverage the specific structure.

Testing via the Fourier Transform: the Advantage One may wonder how the detour via the Fourier transform enables us to obtain better sample complexity than an approach purely based on L_2 testing. Indeed, all distributions in the classes we consider, crucially, have small L_2 norm. For testing identity to such a distribution $\mathbf{P}$, the standard L_2 identity tester (see, e.g., [CDVV14] or Proposition 3.2), which works by checking how large the L_2-distance between the empirical and the hypothesis distribution is, will be optimal. We can thus test membership of a class of such distributions by (i) learning $\mathbf{P}$ assuming it belongs to the class, and then (ii) test whether what we learned is indeed close to $\mathbf{P}$ using the L_2 identity tester. The catch is that, in order to get guarantees in L_1-distance using this approach, would require us to learn to very small L_2 distance (because of the Cauchy–Schwarz inequality). In particular, if the unknown distribution $\mathbf{P}$ has support size N, we would have to learn to L_2 distance $\epsilon/\sqrt{N}$ in (i), and then in (ii) test that we are within L_2-distance $\epsilon/\sqrt{N}$ of the learned hypothesis.

However, if a distribution $\mathbf{P}$ has a sparse discrete Fourier transform (whose effective support is known), then suffices h to estimate only these few Fourier coefficients [DKS16b, DKS16d]. This step enables us to learn $\mathbf{P}$ in (i) not just to within L_1-distance ϵ, but indeed (crucially) within L_2-distance $\frac{\epsilon}{\sqrt{N}}$ with good sample complexity. Additionally, the identity testing algorithm can be put into a simpler form for a hypothesis with sparse Fourier transform, as previously mentioned. Now, the tester has higher sample complexity, roughly $\sqrt{N}/\epsilon^2$; but if it accepts, then we have learned the distribution $\mathbf{P}$ to within ϵ total variation distance, with much fewer samples than the $\Omega\left(N/\epsilon^2\right)$ required for arbitrary distributions over support size N. Lastly, we note that we can replace the support size N in the above description by the size of the *effective support*, i.e., the smallest set that contains $1-O(\epsilon)$ fraction of the mass. Doing so for the case of (n,k)-SIIRVs leads to a sample complexity proportional to $n^{1/4}$, instead of $n^{1/2}$.

1.4 Organization The rest of the paper is organized as follows: In Section 2, we set up notations and provide definitions as well as standard results relevant to our purposes. Section 3 contains the details of one of the main subroutines our testers rely on, namely for *Fourier sparsity testing*. We then give and analyze in Section 4 our Fourier-based tester for SIIRVs. In Section 5, we abstract and generalize this approach to obtain a general tester applicable to any class of distributions which enjoys good Fourier sparsity. Section 6 then contains our tester for Poisson Multinomial Distributions, which we get by extending our general technique to higher dimensions (this tester is complemented in Section 8 by our sample complexity lower bound on testing PMDs). Finally, we focus in Section 7 on the class of log-concave distributions, leveraging our Fourier-based tools to obtain a tester for this class.

All omitted proofs can be found in Appendix A. In Appendix B, we analyze the sample complexity of learning discrete log-concave distributions via the Maximum Likelihood Estimator, a result that we use for our log-concavity tester and which may be of independent interest.

2 Preliminaries

We begin with some standard notations and definitions, as well as basics of Fourier analysis and results from Probability that we shall use throughout the paper. We also state two structural results on SIIRVs, which will be useful to us in Section 4. For $m \in \mathbb{N}$, we write $[m]$ for the set $\{0, 1, \ldots, m-1\}$, and $\log$ (resp. $\ln$) for the binary logarithm (resp. the natural logarithm).

Distributions and Metrics A probability distribution over (discrete) domain Ω is a function $\mathbf{P} \colon \Omega \to [0,1]$ such that $\|\mathbf{P}\|_1 \stackrel{\text{def}}{=} \sum_{\omega \in \Omega} \mathbf{P}(\omega) = 1$; we denote by $\Delta(\Omega)$ the set of all probability distributions over domain Ω. Recall that for two probability distributions $\mathbf{P}, \mathbf{Q} \in \Delta(\Omega)$, their *total variation distance* (or statistical distance) is defined as $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) \stackrel{\text{def}}{=} \sup_{S \subseteq \Omega} (\mathbf{P}(S) - \mathbf{Q}(S)) = \frac{1}{2} \sum_{\omega \in \Omega} |\mathbf{P}(\omega) - \mathbf{Q}(\omega)|$, i.e. $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) = \frac{1}{2} \|\mathbf{P} - \mathbf{Q}\|_1$. Given a subset $\mathcal{P} \subseteq \Delta(\Omega)$ of distributions, the *distance from* $\mathbf{P}$ *to* $\mathcal{P}$ is then defined as $d_{\mathrm{TV}}(\mathbf{P}, \mathcal{P}) \stackrel{\text{def}}{=} \inf_{\mathbf{Q} \in \mathcal{P}} d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q})$. If $d_{\mathrm{TV}}(\mathbf{P}, \mathcal{P}) > \epsilon$, we say that $\mathbf{P}$ is ϵ*-far* from $\mathcal{P}$; otherwise, it is ϵ*-close*.

Property Testing We work in the standard setting of distribution testing: a *testing algorithm for a property* $\mathcal{P} \subseteq \Delta(\Omega)$ is an algorithm which, granted access to independent samples from an unknown distribution $\mathbf{P} \in \Delta(\Omega)$ as well as distance parameter $\epsilon \in (0, 1]$, outputs either accept or reject, with the following guarantees.

- if $\mathbf{P} \in \mathcal{P}$, then it outputs accept with probability at least $3/5$;
- if $d_{\mathrm{TV}}(\mathbf{P}, \mathcal{P}) > \epsilon$, then it outputs reject with probability at least $3/5$.

The two measures of interest here are the *sample complexity* of the algorithm (i.e., the number of samples from the distribution it takes in the worst case), and its running time.

Classes (Properties) of Distributions We now recall the definition of the three classes of discrete distributions central to this work, which all extend the family of Binomial distributions: the first two, by allowing each summand to be non-identically distributed:

Definition 2.1. Fix any $k \geq 2$. We say a random variable X is a (n,k)*-Sum of Independent Integer Random Variables (*(n,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 $[k] = \{0, 1, \ldots, k-1\}$. We denote by $\mathcal{SIIRV}_{n,k}$ the class of all such (n, k)-SIIRVs.

(The class of *Poisson Binomial Distributions*, denoted $\mathcal{PBD}_n$, corresponds to the case $k = 2$, that is 2-SIIRVS. Equivalently, this is the generalization of Binomials where each Bernoulli summand is allowed to have its own parameter). A different type of generalization is that of Poisson Multinomial Distributions, where each summand is a random variable supported on the k vectors of the standard basis of $\mathbb{R}^k$, instead of $[k]$:

Definition 2.2. Fix any $k \geq 2$. We say a random variable X is a (n, k)-*Poisson Multinomial Distribution ((n, k)-PMD)* 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 $\{e_1, \ldots, e_k\}$ (where $(e_i)_{i\in[k]}$ is the canonical basis of $\mathbb{R}^k$). We denote by $\mathcal{PMD}_{n,k}$ the class of all such (n, k)-PMDs.

Lastly, we recall the definition of discrete log-concavity.

Definition 2.3. A distribution $\mathbf{P}$ over $\mathbb{Z}$ is said to be *log-concave* if it satisfies the following conditions: (i) for any $i < j < k$ such that $\mathbf{P}(i)\mathbf{P}(k) > 0$, $\mathbf{P}(j) > 0$; and (ii) for all $k \in \mathbb{Z}$, $\mathbf{P}(k)^2 \geq \mathbf{P}(k-1)\mathbf{P}(k+1)$. We write $\mathcal{LCV}$ for the class of all log-concave distributions over $\mathbb{Z}$, and $\mathcal{LCV}_n \subseteq \mathcal{LCV}$ for that of all log-concave distributions over $[n]$.

Discrete Fourier Transform For our SIIRV testing algorithm, we will need the following definition of the Fourier transform.

Definition 2.4 (Discrete Fourier Transform)**.** For $x \in \mathbb{R}$, we let $e(x) \stackrel{\text{def}}{=} \exp(-2i\pi x)$. The *Discrete Fourier Transform (DFT) modulo M* of a function $F\colon [n] \to \mathbb{C}$ is the function $\widehat{F}\colon [M] \to \mathbb{C}$ defined as

$$\widehat{F}(\xi) = \sum_{j=0}^{n-1} e\left(\frac{\xi j}{M}\right) F(j)$$

for $\xi \in [M]$. The DFT modulo M of a distribution $\mathbf{P}$, $\widehat{\mathbf{P}}$, is then the DFT modulo M of its probability mass function (note that one can then equivalently see $\widehat{\mathbf{P}}(\xi)$ as the expectation $\widehat{\mathbf{P}}(\xi) = \mathbb{E}_{X\sim F}[e\left(\frac{\xi X}{M}\right)]$, for $\xi \in [M]$).

The *inverse DFT modulo M* onto the range $[m, m+M-1]$ of $\widehat{F}\colon [M] \to \mathbb{C}$, is the function $F\colon [m, m+M-1] \cap \mathbb{Z} \to \mathbb{C}$ defined by

$$F(j) = \frac{1}{M} \sum_{\xi=0}^{M-1} e\left(-\frac{\xi j}{M}\right) \widehat{F}(\xi),$$

for $j \in [m, m + M - 1] \cap \mathbb{Z}$.

Note that the DFT (modulo M) is a linear operator; moreover, we recall the standard fact relating the norms of a function and of its Fourier transform, that we will use extensively:

Theorem 2.5 (Plancherel's Theorem)**.** *For $M \geq 1$ and $F, G\colon [n] \to \mathbb{C}$, we have (i) $\sum_{j=0}^{n-1} F(j)\overline{G(j)} = \frac{1}{M}\sum_{\xi=0}^{M-1} \widehat{F}(\xi)\overline{\widehat{G}(\xi)}$; and (ii) $\|F\|_2 = \frac{1}{\sqrt{M}}\|\widehat{F}\|_2$, where $\widehat{F}, \widehat{G}$ are the DFT modulo M of F, G, respectively.*

(The latter equality is sometimes referred to as Parseval's theorem.) We also note that, for our PMD testing, we shall need the appropriate generalization of the Fourier transform to the multivariate setting. We leave this generalization to the corresponding section, Section 6.

Tools from Probability We finally recall a classical inequality for sums of independent random variables, due to Bennett [BLM13, Chapter 2]:

Theorem 2.6 (Bennett's inequality)**.** *Let $X = \sum_{i=1}^{n} X_i$, where $X_1, \ldots, X_n$ are independent random variables such that (i) $\mathbb{E}[X_i] = 0$ and (ii) $|X_i| \leq \alpha$ almost surely for all $1 \leq i \leq n$. Letting $\sigma^2 = \mathrm{Var}[X]$, we have, for every $t \geq 0$,*

$$\Pr[X > t] \leq \exp\left(-\frac{\mathrm{Var}[X]}{\alpha^2}\vartheta\left(\frac{\alpha t}{\mathrm{Var}[X]}\right)\right)$$

where $\vartheta(x) = (1+x)\ln(1+x) - x$.

Structural Results on SIIRVs To establish the completeness of our algorithms, we will rely on this lemma from [DKS16b]:

Lemma 2.7 ([DKS16b, Lemma 2.3])**.** *Let $\mathbf{P} \in \mathcal{SIIRV}_{n,k}$ with $\sqrt{\mathrm{Var}_{X\sim\mathbf{P}}[X]} = s$, $1/2 > \delta > 0$, and $M \in \mathbb{Z}_+$ with $M > s$. Let $\widehat{\mathbf{P}}$ be the discrete Fourier transform of $\mathbf{P}$ modulo M. Then, we have*

(i) Let $\mathcal{L} = \mathcal{L}(\delta, M, s) \stackrel{\text{def}}{=} \left\{\xi \in [M-1] \mid \exists a, b \in \mathbb{Z}, 0 \leq a \leq b < k \text{ such that } |\xi/M - a/b| < \frac{\sqrt{\ln(1/\delta)}}{2s}\right\}$. Then, $|\widehat{\mathbf{P}}(\xi)| \leq \delta$ for all $\xi \in [M-1] \setminus \mathcal{L}$. That is, $|\widehat{\mathbf{P}}(\xi)| > \delta$ for at most $|\mathcal{L}| \leq Mk^2 s^{-1}\sqrt{\log(1/\delta)}$ values of ξ.

(ii) At most $4Mks^{-1}\sqrt{\log(1/\delta)}$ many integers $0 \leq \xi \leq M-1$ have $|\widehat{\mathbf{P}}(\xi)| > \delta$.

We also provide a simple structural lemma, bounding the L_2 norm of any (n,k)-SIIRV as a function of k and its variance only:

Lemma 2.8 (Any (n,k)-SIIRV modulo M has small L_2 norm)**.** *If $\mathbf{P} \in \mathcal{S}_{n,k}$ has variance s^2, then the distribution $\mathbf{P}'$ defined as $\mathbf{P}' \stackrel{\text{def}}{=} \mathbf{P} \bmod M$ satisfies $\|\mathbf{P}'\|_2 \leq \sqrt{\frac{8k}{s}}$.*

The proof of this lemma is deferred to Appendix A.

3 Testing Effective Fourier Support

In this section, we prove the following theorem, which will be invoked as a crucial ingredient of our testing algorithms. Broadly speaking, the theorem ensures one can efficiently test whether an unknown distribution $\mathbf{Q}$ has its Fourier transform concentrated on some (small) effective support S (and if this is the case, learn the vector $\widehat{\mathbf{Q}}\mathbf{1}_S$, the restriction of this Fourier transform to S, in L_2 distance).

Theorem 3.1. *Given parameters $M \geq 1$, $\epsilon, b \in (0,1]$, as well as a subset $S \subseteq [M]$ and sample access to a distribution $\mathbf{Q}$ over $[M]$, Algorithm 1 outputs either* reject *or a collection of Fourier coefficients $\widehat{\mathbf{H}'} = (\widehat{\mathbf{H}'}(\xi))_{\xi\in S}$ such that with probability at least $7/10$, all the following statements hold simultaneously.*

1. *if $\|\mathbf{Q}\|_2^2 > 2b$, then it outputs* reject*;*
2. *if $\|\mathbf{Q}\|_2^2 \leq 2b$ and every function $\mathbf{Q}^*\colon [M] \to \mathbb{R}$ with $\widehat{\mathbf{Q}^*}$ supported entirely on S is such that $\|\mathbf{Q} - \mathbf{Q}^*\|_2 > \epsilon$, then it outputs* reject*;*
3. *if $\|\mathbf{Q}\|_2^2 \leq b$ and there exists a function $\mathbf{Q}^*\colon [M] \to \mathbb{R}$ with $\widehat{\mathbf{Q}^*}$ supported entirely on S such that $\|\mathbf{Q} - \mathbf{Q}^*\|_2 \leq \frac{\epsilon}{2}$, then it does not output* reject*;*

4. *if it does not output* reject*, then* $\|\widehat{\mathbf{Q}}\mathbf{1}_S - \widehat{\mathbf{H}'}\|_2 \leq \frac{\epsilon\sqrt{M}}{10}$ *and the inverse Fourier transform (modulo* M*)* $\mathbf{H}'$ *of the Fourier coefficients* $\widehat{\mathbf{H}'}$ *it outputs satisfies* $\|\mathbf{Q} - \mathbf{H}'\|_2 \leq \frac{6\epsilon}{5}$.

Moreover, the algorithm takes $m = O\left(\frac{\sqrt{b}}{\epsilon^2} + \frac{|S|}{M\epsilon^2} + \sqrt{M}\right)$ *samples from* $\mathbf{Q}$*, and runs in time* $O(m\,|S|)$.

Note that the rejection condition in Item 2 is equivalent to $\|\widehat{\mathbf{Q}}\mathbf{1}_{\bar{S}}\|_2 > \epsilon\sqrt{M}$, that is to having Fourier mass more than ϵ^2 outside of S; this is because for any $\mathbf{Q}^*$ supported on S,

$$M\|\mathbf{Q} - \mathbf{Q}^*\|_2^2 = \|\widehat{\mathbf{Q}} - \widehat{\mathbf{Q}^*}\|_2^2 = \|\widehat{\mathbf{Q}}\mathbf{1}_S - \widehat{\mathbf{Q}^*}\mathbf{1}_S\|_2^2 + \|\widehat{\mathbf{Q}}\mathbf{1}_{\bar{S}} - \widehat{\mathbf{Q}^*}\mathbf{1}_{\bar{S}}\|_2^2 \geq \|\widehat{\mathbf{Q}}\mathbf{1}_{\bar{S}} - \widehat{\mathbf{Q}^*}\mathbf{1}_{\bar{S}}\|_2^2 = \|\widehat{\mathbf{Q}}\mathbf{1}_{\bar{S}}\|_2^2$$

and the inequality is tight for $\mathbf{Q}^*$ being the inverse Fourier transform (modulo M) of $\widehat{\mathbf{Q}}\mathbf{1}_S$.

High-level idea. Let $\mathbf{Q}$ be an unknown distribution supported on M consecutive integers (we will later apply this to $\mathbf{Q} \stackrel{\text{def}}{=} \mathbf{P} \bmod M$), and $S \subseteq [M]$ be a set of Fourier coefficients (symmetric with regard to M: $\xi \in S$ implies $-\xi \bmod M \in S$) such that $0 \in S$. We can further assume that we know $b \geq 0$ such that $\|\mathbf{Q}\|_2^2 \leq b$.

Given $\mathbf{Q}$, we can consider its "truncated Fourier expansion" (with respect to S) $\widehat{\mathbf{H}} = \hat{\mathbf{Q}}\mathbf{1}_S$ defined as

$$\widehat{\mathbf{H}}(\xi) \stackrel{\text{def}}{=} \begin{cases} \hat{\mathbf{Q}}(\xi) & \text{if } \xi \in S \\ 0 & \text{otherwise} \end{cases}$$

for $\xi \in [M]$; and let $\mathbf{H}$ be the inverse Fourier transform (modulo M) of $\widehat{\mathbf{H}}$. Note that $\mathbf{H}$ is no longer in general a probability distribution.

To obtain the guarantees of Theorem 3.1, a natural idea is to take some number m of samples from $\mathbf{Q}$, and consider the empirical distribution $\mathbf{Q}'$ they induce over $[M]$. By computing the Fourier coefficients (restricted to S) of this $\mathbf{Q}'$, as well as the Fourier mass "missed" when doing so (i.e., the Fourier mass $\|\widehat{\mathbf{Q}'}\mathbf{1}_{\bar{S}}\|_2^2$ that $\mathbf{Q}'$ puts outside of S) to sufficient accuracy, one may hope to prove Theorem 3.1 with a reasonable bound on m.

The issue is that analyzing *separately* the behavior of $\|\widehat{\mathbf{Q}'}\mathbf{1}_{\bar{S}}\|_2^2$ and $\|\widehat{\mathbf{Q}'}\mathbf{1}_S - \widehat{\mathbf{Q}'}\mathbf{1}_S\|_2^2$ to show that they are both estimated sufficiently accurately, and both small enough, is not immediate. Instead, we will get a bound on both at the same time, by arguing concentration in a different manner – namely, by analyzing a different tester for tolerant identity testing in L_2 norm.

In more detail, letting $\mathbf{H}$ be as above, we have by Plancherel that

$$\sum_{i\in[M]} (\mathbf{Q}'(i) - \mathbf{H}(i))^2 = \|\mathbf{Q}' - \mathbf{H}\|_2^2 = \frac{1}{M}\|\widehat{\mathbf{Q}'} - \widehat{\mathbf{H}}\|_2^2 = \frac{1}{M}\sum_{\xi=0}^{M-1} |\widehat{\mathbf{Q}'}(\xi) - \widehat{\mathbf{H}}(\xi)|^2$$

and, expanding the definition of $\widehat{\mathbf{H}}$ and using Plancherel again, this can be rewritten as

$$M\sum_{i\in[M]} (\mathbf{Q}'(i) - \mathbf{H}(i))^2 = \|\widehat{\mathbf{Q}}\mathbf{1}_S - \widehat{\mathbf{Q}'}\mathbf{1}_S\|_2^2 + \|\mathbf{Q}'\|_2^2 - \|\widehat{\mathbf{Q}'}\mathbf{1}_S\|_2^2.$$

(The full derivation will be given in the proof.) The left-hand side has two non-negative compound terms: the first, $\|\widehat{\mathbf{P}}\mathbf{1}_S - \widehat{\mathbf{Q}'}\mathbf{1}_S\|_2^2$, corresponds to the L_2 error obtained when learning the Fourier coefficients of $\mathbf{Q}$ on S. The second, $\|\mathbf{Q}'\|_2^2 - \|\widehat{\mathbf{Q}'}\mathbf{1}_S\|_2^2 = \|\widehat{\mathbf{Q}'}\mathbf{1}_{\bar{S}}\|_2^2$, is the Fourier mass that our empirical $\mathbf{Q}'$ puts "outside of S."

So if the LHS is small (say, order ϵ^2), then in particular both terms of the RHS will be small as well, effectively giving us bounds on our two quantities in one shot. But this very same LHS is very reminiscent of a known statistic [CDVV14] for testing identity of distributions in L_2. So, one can analyze the number of samples required by analyzing such an L_2 tester instead. This is what we will do in Proposition 3.2.

Algorithm 1 Testing the Fourier Transform Effective Support

Require: parameters $M \geq 1$, $b, \epsilon \in (0,1]$; set $S \subseteq [M]$; sample access to distribution $\mathbf{Q}$ over $[M]$

1: Set $m \leftarrow \left\lceil C(\frac{\sqrt{b}}{\epsilon^2} + \frac{|S|}{M\epsilon^2} + \sqrt{M}) \right\rceil$ ▷ $C > 0$ is an absolute constant
2: Draw $m' \leftarrow \mathrm{Poi}(m)$; if $m' > 2m$, **return** reject
3: Draw m' samples from $\mathbf{Q}$, and let $\mathbf{Q}'$ be the corresponding empirical distribution over $[M]$
4: Compute $\|\mathbf{Q}'\|_2^2$, $\widehat{\mathbf{Q}'}(\xi)$ for every $\xi \in S$, and $\|\widehat{\mathbf{Q}'}\mathbf{1}_S\|_2^2$ ▷ Takes time $O(m\,|S|)$
5: **if** $m'^2\|\mathbf{Q}'\|_2^2 - m' > \frac{3}{2}bm^2$ **then return** reject
6: **else if** $\|\mathbf{Q}'\|_2^2 - \frac{1}{M}\|\widehat{\mathbf{Q}'}\mathbf{1}_S\|_2^2 \geq 3\epsilon^2\left(\frac{m'}{m}\right)^2 + \frac{1}{m'}$ **then return** reject
7: **else**
8: **return** $\widehat{\mathbf{H}'} = (\widehat{\mathbf{Q}'}(\xi))_{\xi \in S}$
9: **end if**

Proof of Theorem 3.1. Given $m' \sim \mathrm{Poi}(m)$ samples from $\mathbf{Q}$, let $\mathbf{Q}'$ be the empirical distribution they define. We first observe that with probability $2^{-\Omega(\epsilon^2 m/b)} < \frac{1}{100}$, we have $m' \in [1 \pm \frac{\epsilon}{100\sqrt{b}}]m$ and thus the algorithm does not output reject in Step 1 (this follows from standard concentration bounds on Poisson random variables). We will afterwards assume this holds. By Plancherel, we have

$$\sum_{i\in[M]} (\mathbf{Q}'(i) - \mathbf{H}(i))^2 = \|\mathbf{Q}' - \mathbf{H}\|_2^2 = \frac{1}{M}\|\widehat{\mathbf{Q}'} - \widehat{\mathbf{H}}\|_2^2 = \frac{1}{M}\sum_{\xi=0}^{M-1} |\widehat{\mathbf{Q}'}(\xi) - \widehat{\mathbf{H}}(\xi)|^2$$

and, expanding the definition of $\widehat{\mathbf{H}}$, this yields

$$\begin{aligned}
\sum_{i\in[M]} (\mathbf{Q}'(i) - \mathbf{H}(i))^2 &= \frac{1}{M}\sum_{\xi\in S} |\widehat{\mathbf{Q}'}(\xi) - \widehat{\mathbf{H}}(\xi)|^2 + \frac{1}{M}\sum_{\xi\notin S} |\widehat{\mathbf{Q}'}(\xi)|^2 \\
&= \frac{1}{M}\sum_{\xi\in S} |\widehat{\mathbf{Q}'}(\xi) - \widehat{\mathbf{Q}}(\xi)|^2 + \frac{1}{M}\sum_{\xi=0}^{M-1} |\widehat{\mathbf{Q}'}(\xi)|^2 - \frac{1}{M}\sum_{\xi\in S} |\widehat{\mathbf{Q}'}(\xi)|^2 \\
&= \frac{1}{M}\left(\|\widehat{\mathbf{Q}}\mathbf{1}_S - \widehat{\mathbf{Q}'}\mathbf{1}_S\|_2^2 + \|\widehat{\mathbf{Q}'}\|_2^2 - \|\widehat{\mathbf{Q}'}\mathbf{1}_S\|_2^2\right) \\
&= \frac{1}{M}\|\widehat{\mathbf{Q}}\mathbf{1}_S - \widehat{\mathbf{Q}'}\mathbf{1}_S\|_2^2 + \|\mathbf{Q}'\|_2^2 - \frac{1}{M}\|\widehat{\mathbf{Q}'}\mathbf{1}_S\|_2^2 \qquad (1)
\end{aligned}$$

where in the last step we invoked Plancherel again to argue that $\frac{1}{M}\|\widehat{\mathbf{Q}'}\|_2^2 = \|\mathbf{Q}'\|_2^2$.

To analyze the correctness of the algorithm (specifically, the completeness), we will adopt the point of view suggested by (1) and analyze instead the statistic $\sum_{i\in[M]}(\mathbf{Q}'(i) - \mathbf{H}(i))^2$, when $\mathbf{H}$ is an explicit (pseudo) distribution on $[M]$ assumed known, and $\mathbf{Q}'$ is the empirical distribution obtained by drawing $\mathrm{Poi}(m)$ samples from some unknown distribution $\mathbf{Q}$. (Namely, we want to see this as a tolerant L_2 identity tester between $\mathbf{Q}$ and $\mathbf{H}$.)

- We first show that, given that $m' = \Omega\left(\frac{|S|}{M\epsilon^2}\right)$, with probability at least $\frac{99}{100}$ we have $\|\widehat{\mathbf{Q}}\mathbf{1}_S - \widehat{\mathbf{H}'}\|_2 \leq \frac{\sqrt{M}\epsilon}{10}$. We note that $m'\widehat{\mathbf{Q}'}(\xi)$ is an sum of m' i.i.d. numbers each of absolute value 1 and mean $\widehat{\mathbf{Q}}(\xi)$ (which has absolute value less than 1). If X is one of these numbers, $|X - \widehat{\mathbf{Q}}(\xi)| \leq 2$ with probability 1 and so the variance of the real and imaginary parts of X is at most 4. Thus the variance of the real and imaginary parts of $m'\widehat{\mathbf{Q}'}(\xi)$ is at most $4m'$. Then we have $\mathbb{E}[|\widehat{\mathbf{Q}}(\xi) - \widehat{\mathbf{Q}'}(\xi)|^2] = \mathbb{E}[(\Re(\widehat{\mathbf{Q}}(\xi) - \widehat{\mathbf{Q}'}(\xi)))^2 + (\Im(\widehat{\mathbf{Q}}(\xi) - \widehat{\mathbf{Q}'}(\xi)))^2] \leq 8/m'$. Summing over S, using that $\mathbf{Q}'$ and $\mathbf{H}'$ have the same Fourier coefficients there, yields

$$\mathbb{E}\left[\sum_{\xi\in S}\left|\widehat{\mathbf{Q}}(\xi) - \widehat{\mathbf{H}'}(\xi)\right|^2\right] \leq \frac{8|S|}{m'} \leq \frac{M\epsilon^2}{10000}$$

and by Markov's inequality we get $\Pr\left[\|\widehat{\mathbf{Q}}\mathbf{1}_S - \widehat{\mathbf{H}'}\|_2^2 \leq \frac{M\epsilon^2}{100}\right] = \Pr\left[\sum_{\xi\in S}\left|\widehat{\mathbf{Q}}(\xi) - \widehat{\mathbf{H}'}(\xi)\right|^2 \leq \frac{M\epsilon^2}{100}\right] \geq \frac{1}{100}$, concluding the proof.

- Then, let us consider Item 1: assume $\|\mathbf{Q}\|_2^2 > 2b$, and set $X \stackrel{\text{def}}{=} m'^2\|\mathbf{Q}'\|_2^2 - m'$. Then,

$$\mathbb{E}[X] = \sum_{i=1}^{M}\mathbb{E}[m'^2\mathbf{Q}'(i)^2] - \sum_{i=1}^{M}\mathbb{E}[m'\mathbf{Q}'(i)] = \sum_{i=1}^{M}(m\mathbf{Q}(i) + m^2\mathbf{Q}(i)^2) - \sum_{i=1}^{M}m\mathbf{Q}(i) = m^2\|\mathbf{Q}\|_2^2$$

since the $m'\mathbf{Q}'(i)$ are distributed as $\mathrm{Poi}(m\mathbf{Q}(i))$. As all $m'\mathbf{Q}'(i)$'s are independent by Poissonization, we also have

$$\mathrm{Var}[X] = \sum_{i=1}^{M}\mathrm{Var}[m'^2\mathbf{Q}'(i)^2 - m'\mathbf{Q}'(i)] = \sum_{i=1}^{M}(2m^2\mathbf{Q}(i)^2 + 4m^3\mathbf{Q}(i)^3) = 2m^2\|\mathbf{Q}\|_2^2 + 4m^3\|\mathbf{Q}\|_3^3$$

and by Chebyshev,

$$\Pr[X \leq \frac{3}{2}m^2b] \leq \Pr\left[|X - \mathbb{E}[X]| > \frac{1}{4}\mathbb{E}[X]\right] \leq 16\frac{\mathrm{Var}[X]}{\mathbb{E}[X]^2} \leq \frac{32}{m^2\|\mathbf{Q}\|_2^2} + \frac{64\|\mathbf{Q}\|_3^3}{m\|\mathbf{Q}\|_2^4}$$

Since $\mathbf{Q}$ is supported on $[M]$, $\|\mathbf{Q}\|_2^2 \geq \frac{1}{M}$ and the first term is at most $\frac{32M}{m^2}$. The second term, by monotonicity of ℓ_p-norms, is at most $\frac{64\|\mathbf{Q}\|_2^3}{m\|\mathbf{Q}\|_2^4} = \frac{48}{m\|\mathbf{Q}\|_2} \leq \frac{48\sqrt{M}}{m}$. The RHS is then at most $\frac{1}{100}$ for a large enough choice of $C > 0$ in the definition of m. Thus, with probability at least $1 - \frac{1}{100}$ we have $m'^2\|\mathbf{Q}'\|_2^2 - m' > \frac{3}{2}b$, and the algorithm outputs reject in Step 5.

Moreover, if $\|\mathbf{Q}\|_2^2 \leq b$, then the same analysis shows that

$$\Pr[X > \frac{3}{2}m^2b] \leq \Pr\left[|X - \mathbb{E}[X]| > \frac{1}{2}\mathbb{E}[X]\right] \leq 4\frac{\mathrm{Var}[X]}{\mathbb{E}[X]^2} \leq \frac{1}{100}$$

and with probability at least $1 - \frac{1}{100}$ the algorithm does not output reject in Step 4.

- Turning now to Items 2 to 4: we assume that the algorithm does not output reject in Step 4 (which by the above happens with probability $99/100$ if $\|\mathbf{Q}\|_2^2 \leq b$; and can be assumed without loss of

generality otherwise, since we then want to argue that the algorithm *does* reject at some point in that case).

By the remark following the statement of the theorem, it is sufficient to show that the algorithm outputs reject (with high probability) if $\|\widehat{\mathbf{Q}}\mathbf{1}_{\bar{S}}\|_2^2 > \epsilon^2 M$, and that if both $\|\mathbf{Q}\|_2^2 \leq b$ and $\|\widehat{\mathbf{Q}}\mathbf{1}_{\bar{S}}\|_2^2 \leq \frac{\epsilon^2}{4}M$ then it does not output reject; and that whenever the algorithm does not output reject, then $\|\widehat{\mathbf{Q}} - \widehat{\mathbf{H}}\|_2 \leq \epsilon^2 M$.

Observe that calling Algorithm 2 with our $m' = \mathrm{Poi}(m)$ samples from $\mathbf{Q}$ (distribution over $[M]$), parameters $\frac{\epsilon}{2}$ and $2b$, and the explicit description of the pseudo distribution $\mathbf{P}^* \stackrel{\text{def}}{=} \frac{m'}{m}\mathbf{H}$ (which one would obtain for $\mathbf{H}$ being the inverse Fourier transform of $\widehat{\mathbf{Q}}\mathbf{1}_S$) would result by Proposition 3.2 (since $m \geq c\frac{\sqrt{2b}}{(\epsilon/2)^2} = 244\sqrt{2}\frac{\sqrt{b}}{\epsilon^2}$, where c is as in Proposition 3.2) in having the following guarantees on $\frac{\sqrt{Z}}{m}$, where Z is the statistic defined in Algorithm 2

- if $\|\mathbf{Q} - \mathbf{P}^*\|_2 \leq \frac{\epsilon}{2}$, then $\frac{\sqrt{Z}}{m} \leq \sqrt{2.9}\epsilon$ with probability at least $3/4$;
- if $\|\mathbf{Q} - \mathbf{P}^*\|_2 \geq \epsilon$, then $\frac{\sqrt{Z}}{m} \geq \sqrt{3.1}\epsilon$ with probability at least $3/4$;

as $\|\mathbf{Q}\|_2^2 \leq 2b$ (note that then $\|\mathbf{H}\|_2^2 \leq b$ as well). Since $\sqrt{M}\|\mathbf{Q} - \mathbf{P}^*\|_2 = \|\widehat{\mathbf{Q}} - \widehat{\mathbf{P}^*}\|_2 = \|\widehat{\mathbf{Q}} - \frac{m}{m'}\widehat{\mathbf{Q}}\mathbf{1}_S\|_2$ and

$$\frac{Z}{m'^2} = \sum_{i=1}^{M}\left((\mathbf{Q}'(i) - \frac{m}{m'}\mathbf{P}^*(i))^2 - \frac{\mathbf{Q}'(i)}{m'}\right) = \sum_{i=1}^{M}(\mathbf{Q}'(i) - \mathbf{H}(i))^2 - \frac{1}{m'}$$

which is equal to $\frac{1}{M}\|\widehat{\mathbf{Q}}\mathbf{1}_S - \widehat{\mathbf{Q}'}\mathbf{1}_S\|_2^2 + \|\mathbf{Q}'\|_2^2 - \frac{1}{M}\|\widehat{\mathbf{Q}'}\mathbf{1}_S\|_2^2 - \frac{1}{m'}$ by Eq. (1), we thus get the following.

- if $\|\widehat{\mathbf{Q}}\mathbf{1}_{\bar{S}}\|_2^2 \leq \frac{\epsilon^2 M}{9}$, then $\|\widehat{\mathbf{Q}} - \widehat{\mathbf{Q}}\mathbf{1}_S\|_2 \leq \frac{\epsilon}{3}\sqrt{M}$, and

  $$\sqrt{M}\|\mathbf{P}^* - \mathbf{Q}\|_2 = \|\widehat{\mathbf{P}^*} - \widehat{\mathbf{Q}}\|_2 \leq \|\widehat{\mathbf{P}^*} - \widehat{\mathbf{Q}}\mathbf{1}_S\|_2 + \|\widehat{\mathbf{Q}}\mathbf{1}_S - \widehat{\mathbf{Q}}\|_2 = \left|\frac{m}{m'} - 1\right|\|\widehat{\mathbf{Q}}\mathbf{1}_S\|_2 + \|\widehat{\mathbf{Q}} - \widehat{\mathbf{Q}}\mathbf{1}_S\|_2$$

  Since we have $m' \in [1 \pm \frac{\epsilon}{100\sqrt{b}}]m$ by the above discussion and $\|\widehat{\mathbf{Q}}\mathbf{1}_S\|_2 \leq \sqrt{2b}\sqrt{M}$, the RHS is upper bounded by $\frac{\epsilon}{6}\sqrt{M} + \frac{\epsilon}{3}\sqrt{M} = \frac{\epsilon}{2}\sqrt{M}$, and $\|\mathbf{P}^* - \mathbf{Q}\|_2 \leq \frac{\epsilon}{2}$. Then $\frac{1}{M}\|\widehat{\mathbf{Q}}\mathbf{1}_S - \widehat{\mathbf{Q}'}\mathbf{1}_S\|_2^2 + \|\mathbf{Q}'\|_2^2 - \frac{1}{M}\|\widehat{\mathbf{Q}'}\mathbf{1}_S\|_2^2 = \frac{Z}{m'^2} + \frac{1}{m'} \leq 2.9\epsilon^2\left(\frac{m'}{m}\right)^2 + \frac{1}{m'}$ with probability at least $3/4$, and in particular $\|\mathbf{Q}'\|_2^2 - \frac{1}{M}\|\widehat{\mathbf{Q}'}\mathbf{1}_S\|_2^2 \leq 2.9\epsilon^2\left(\frac{m'}{m}\right)^2 + \frac{1}{m'} < 3\epsilon^2\left(\frac{m'}{m}\right)^2 + \frac{1}{m'}$;
- if $\|\widehat{\mathbf{Q}}\mathbf{1}_{\bar{S}}\|_2^2 > \epsilon^2 M$, then $\frac{1}{M}\|\widehat{\mathbf{Q}}\mathbf{1}_S - \widehat{\mathbf{Q}'}\mathbf{1}_S\|_2^2 + \|\mathbf{Q}'\|_2^2 - \frac{1}{M}\|\widehat{\mathbf{Q}'}\mathbf{1}_S\|_2^2 = \frac{Z}{m'^2} + \frac{1}{m'} > 3.1\epsilon^2\left(\frac{m'}{m}\right)^2 + \frac{1}{m'}$ with probability at least $3/4$; since by the first part we established we have $\|\widehat{\mathbf{Q}}\mathbf{1}_S - \widehat{\mathbf{Q}'}\mathbf{1}_S\|_2^2 \leq \frac{\epsilon^2 M}{100}$, this implies $\|\mathbf{Q}'\|_2^2 - \frac{1}{M}\|\widehat{\mathbf{Q}'}\mathbf{1}_S\|_2^2 > 3.1\epsilon^2\left(\frac{m'}{m}\right)^2 + \frac{1}{m'} - \frac{\epsilon^2}{100} > 3\epsilon^2\left(\frac{m'}{m}\right)^2 + \frac{1}{m'}$.

This immediately takes care of Items 2 and 3; moreover, this implies that whenever Algorithm 1 does *not* output reject, then the inverse Fourier transform $\mathbf{H}'$ of the collection of Fourier coefficients it

returns (which are supported on S) satisfies

$$\begin{aligned}\|\mathbf{Q}-\mathbf{H}'\|_2^2 = \frac{1}{M}\|\widehat{\mathbf{Q}}-\widehat{\mathbf{H}'}\|_2^2 &= \frac{1}{M}\|\widehat{\mathbf{Q}}\mathbf{1}_S-\widehat{\mathbf{H}'}\|_2^2 + \frac{1}{M}\|\widehat{\mathbf{Q}}\mathbf{1}_{\bar{S}}\|_2^2\\ &\le \frac{\epsilon^2}{100} + \frac{1}{M}\|\widehat{\mathbf{Q}}\mathbf{1}_{\bar{S}}\|_2^2\\ &\le \frac{\epsilon^2}{100} + \epsilon^2 = \frac{101}{100}\epsilon^2\end{aligned}$$

and thus $\|\mathbf{Q}-\mathbf{H}'\|_2 \le \sqrt{\frac{101}{100}}\epsilon < \frac{6}{5}\epsilon$ which establishes Item 4. Finally, by a union bound, all the above holds except with probability $\frac{1}{100}+\frac{1}{100}+\frac{1}{100}+\frac{1}{4} < \frac{3}{10}$. This concludes the proof.

□

3.1 A Tolerant L_2 Tester for Identity to a Pseudodistribution As previously mentioned, one building block in the proof of Theorem 3.1 (and a result that may be of independent interest) is an optimal L_2 identity testing algorithm. Our tester and its analysis are very similar to the tolerant L_2 closeness testing algorithm of Chan et al. [CDVV14], with the obvious simplifications pertaining to identity (instead of closeness). The main difference is that we emphasize here the fact that $\mathbf{P}^*$ need not be an actual distribution: any $\mathbf{P}^*\colon [r] \to \mathbb{R}$ would do, even taking negative values. This will turn out to be crucial for our applications.

Algorithm 2 Tolerant L_2 identity tester

Require: $\epsilon \in (0,1)$, $\mathrm{Poi}(m)$ samples from distributions $\mathbf{P}$ over $[r]$, with X_i denoting the number of occurrences of the i-th domain elements in the samples from $\mathbf{P}$, and $\mathbf{P}^*$ being a fixed, known pseudo distribution over $[r]$.

Ensure: Returns accept if $\|\mathbf{P}-\mathbf{P}^*\|_2 \le \epsilon$ and reject if $\|\mathbf{P}-\mathbf{P}^*\|_2 \ge 2\epsilon$.

Define $Z = \sum_{i=1}^{r}(X_i - m\mathbf{P}^*(i))^2 - X_i$. ▷ Can actually be computed in $O(m)$ time

Return reject if $\frac{\sqrt{Z}}{m} > \sqrt{3}\epsilon$, accept otherwise.

Proposition 3.2. *There exists an absolute constant $c > 0$ such that the above algorithm (Algorithm 2), when given $\mathrm{Poi}(m)$ samples drawn from a distribution $\mathbf{P}$ and an explicit function $\mathbf{P}^*\colon [r] \to \mathbb{R}$ will, with probability at least $3/4$, distinguishes between (a) $\|\mathbf{P}-\mathbf{P}^*\|_2 \le \epsilon$ and (b) $\|\mathbf{P}-\mathbf{P}^*\|_2 \ge 2\epsilon$ provided that $m \ge c\frac{\sqrt{b}}{\epsilon^2}$, where b is an upper bound on $\|\mathbf{P}\|_2^2, \|\mathbf{P}^*\|_2^2$. (Moreover, one can take $c = 61$.)*

Moreover, we have the following stronger statement: in case (a), the statistic Z computed in the algorithm satisfies $\frac{\sqrt{Z}}{m} \le \sqrt{2.9}\epsilon$ with probability at least $3/4$, while in case (b) we have $\frac{\sqrt{Z}}{m} \ge \sqrt{3.1}\epsilon$ with probability at least $3/4$.

4 The SIIRV Tester

We are now ready to describe the algorithm behind Theorem 1.1, and establish the theorem.

4.1 Analyzing the Subroutines We start with a simple fact, that we will use to bound the running time of our algorithm and which follows immediately from [DKS16b, Claim 2.4]:

Fact 4.1. *For S as defined in Step 13, we have*

$$|S| \le Mk^2\frac{C'}{2\widetilde{\sigma}}\sqrt{\ln\frac{1}{\delta}} \le 100C'k^2\sqrt{\ln\frac{4}{\epsilon}}\sqrt{\ln\frac{k}{\epsilon}+\log\log\frac{k}{\epsilon}+\frac{1}{2}\ln(16C'')} \le C''k^2\log^2\frac{k}{\epsilon}$$

Algorithm 3 Algorithm `Test-SIIRV`

Require: sample access to a distribution $\mathbf{P} \in \Delta(\mathbb{N})$, parameters $n, k \geq 1$ and $\epsilon \in (0, 1]$

1: ▷ Let C, C', C'' be sufficiently large universal constants

2: Draw $O(k)$ samples from $\mathbf{P}$ and compute as in Claim 4.2: (a) $\widetilde{\sigma}^2$, a tentative factor-2 approximation to $\mathrm{Var}_{X\sim\mathbf{P}}[X] + 1$, and (b) $\widetilde{\mu}$, a tentative approximation to $\mathbb{E}_{X\sim\mathbf{P}}[X]$ to within one standard deviation.

3: **if** If $\widetilde{\sigma} > 2k\sqrt{n}$ **then**

4: **return** reject ▷ Blatant violation of (n, k)-SIIRV-iness

5: **end if**

6: **if** $\widetilde{\sigma} \leq 2k\sqrt{\ln \frac{10}{\epsilon}}$ **then**

7: Set $M \leftarrow 1 + 2\left\lceil 15k \ln \frac{10}{\epsilon} \right\rceil$, and let $I \leftarrow [\lfloor \widetilde{\mu} \rfloor - \frac{M-1}{2}, \lfloor \widetilde{\mu} \rfloor + \frac{M-1}{2}]$; and $S \leftarrow [M]$

8: Draw $O(1/\epsilon)$ samples from $\mathbf{P}$, to distinguish between $\mathbf{P}(I) \leq 1 - \frac{\epsilon}{4}$ and $\mathbf{P}(I) > 1 - \frac{\epsilon}{5}$. If the former is detected, **return** reject

9: Take $N = C\left(\frac{|S|}{\epsilon^2}\right) = O\left(\frac{k}{\epsilon^2} \log \frac{1}{\epsilon}\right)$ samples from $\mathbf{P}$ to get an empirical distribution $\mathbf{H}$

10: **else**

11: Set $M \leftarrow 1 + 2\left\lceil 4\widetilde{\sigma}\sqrt{\ln(4/\epsilon)} \right\rceil$, and let $I \leftarrow [\lfloor \widetilde{\mu} \rfloor - \frac{M-1}{2}, \lfloor \widetilde{\mu} \rfloor + \frac{M-1}{2}]$

12: Draw $O(1/\epsilon)$ samples from $\mathbf{P}$, to distinguish between $\mathbf{P}(I) \leq 1 - \frac{\epsilon}{4}$ and $\mathbf{P}(I) > 1 - \frac{\epsilon}{5}$. If the former is detected, **return** reject

13: Let $\delta \leftarrow \frac{\epsilon}{C''\sqrt{k \log \frac{k}{\epsilon}}}$, and

$$S \leftarrow \left\{ \xi \in [M-1] : \exists a, b \in \mathbb{Z}, 0 \leq a \leq b < k \text{ s.t. } |\xi/M - a/b| \leq C' \frac{\sqrt{\ln(1/\delta)}}{4\widetilde{\sigma}} \right\}.$$

14: Simulating sample access to $\mathbf{P}' \stackrel{\text{def}}{=} \mathbf{P} \bmod M$, call Algorithm 1 on $\mathbf{P}'$ with parameters M, $\frac{\epsilon}{5\sqrt{M}}$, $b = \frac{16k}{\widetilde{\sigma}}$, and S. If it outputs reject, then **return** reject; otherwise, let $\widehat{\mathbf{H}} = (\widehat{\mathbf{H}}(\xi))_{\xi\in S}$ denote the collection of Fourier coefficients it outputs, and $\mathbf{H}$ their inverse Fourier transform (modulo M) ▷ Do not actually compute $\mathbf{H}$

15: **end if**

16: Projection Step: Check whether $d_{\mathrm{TV}}(\mathbf{H}, \mathcal{SIIRV}_{n,k}) \leq \frac{\epsilon}{2}$ (as in Section 4.3), and **return** accept if it is the case. If not, **return** reject. ▷ Mostly computational step

for a suitably large choice of the constant $C'' > 0$; from which we get $\delta \leq \frac{1}{4\sqrt{|S|}}$.

We then argue that with high probability, the estimates obtained in Step 2 will be accurate enough for our purposes. (The somewhat odd statement below, stating two distinct guarantees where the second implies the first, is due to the following: Eq. (2) will be the guarantee that (the completeness analysis of) our algorithm relies on, while the second, slightly stronger one, will only be used in the particular implementation of the "projection step" (Step 16) from Section 4.3.)

Claim 4.2 (Estimating the first two moments (if $\mathbf{P}$ is a SIIRV))**.** *With probability at least $19/20$ over the $O(k)$ draws from $\mathbf{P}$ in Step 2, the following holds. If $\mathbf{P} \in \mathcal{SIIRV}_{n,k}$, the estimates $\widetilde{\sigma}, \widetilde{\mu}$ defined as the empirical mean and (unbiased) empirical variance meet the guarantees stated in Step 2 of the algorithm, namely*

$$\frac{1}{2} \leq \frac{\widetilde{\sigma}^2}{\mathrm{Var}_{X\sim\mathbf{P}}[X] + 1} \leq 2, \qquad |\widetilde{\mu} - \mathbb{E}_{X\sim\mathbf{P}}[X]| \leq \sqrt{\mathrm{Var}_{X\sim\mathbf{P}}[X]} \tag{2}$$

We even have a quantitatively slightly stronger guarantee: $\frac{2}{3} \le \frac{\widetilde{\sigma}^2}{\mathrm{Var}_{X\sim\mathbf{P}}[X]+1} \le \frac{3}{2}$*, and* $|\widetilde{\mu} - \mathbb{E}_{X\sim\mathbf{P}}[X]| \le \frac{1}{2}\sqrt{\mathrm{Var}_{X\sim\mathbf{P}}[X]}$.

Proof. We handle the estimation of the mean and variance separately.

Estimating the mean. $\widetilde{\mu}$ will be the usual empirical estimator, namely $\widetilde{\mu} \stackrel{\text{def}}{=} \frac{1}{m}\sum_{i=1}^m X_i$ for $X_1, \dots, X_m$ independently drawn from $\mathbf{P}$. Since $\mathbb{E}[\widetilde{\mu}] = \mathbb{E}_{X\sim\mathbf{P}}[X]$ and $\mathrm{Var}[\widetilde{\mu}] = \frac{1}{m}\mathrm{Var}_{X\sim\mathbf{P}}[X]$, Chebyshev's inequality guarantees that

$$\Pr[|\widetilde{\mu} - \mathbb{E}_{X\sim\mathbf{P}}[X]| > \frac{1}{2}\sqrt{\mathrm{Var}_{X\sim\mathbf{P}}[X]}] \le \frac{4}{m}$$

which can be made at most $1/200$ by choosing $m \ge 800$.

Estimating the variance. The variance estimation is exactly the same as in [DDS15, Lemma 6], observing that their argument only requires that $\mathbf{P}$ be the distribution of a sum of independent random variables (not necessarily a Poisson Binomial distribution). Namely, they establish that,[3] letting $\widetilde{\sigma}^2 \stackrel{\text{def}}{=} \frac{1}{m-1}\sum_{i=1}^m (X_i - \frac{1}{m}\sum_{j=1}^m X_j)^2$ be the (unbiased) sample variances, and $s^2 \stackrel{\text{def}}{=} \mathrm{Var}_{X\sim\mathbf{P}}[X]$,

$$\Pr[\left|\widetilde{\sigma}^2 - s^2\right| > \alpha(1+s^2)] \le \frac{4s^4 + k^2 s^2}{\alpha^2(1+s^2)^2}\frac{1}{m} \le \frac{4s^4 + s^2}{\alpha^2(1+s^2)^2}\cdot\frac{k^2}{m} \le \frac{4k^2}{\alpha^2 m}$$

which for $\alpha = 1/3$ is at most $9/200$ by choosing $m \ge 800k$.

A union bound completes the proof, giving a probability of error at most $\frac{1}{200} + \frac{9}{200} = \frac{1}{20}$. □

Claim 4.3 (Checking the effective support)**.** *With probability at least* $19/20$ *over the draws from* $\mathbf{P}$ *in Step 12, the following holds.*

- *if* $\mathbf{P} \in \mathcal{SIIRV}_{n,k}$ *and* (2) *holds, then* $\mathbf{P}(I) \ge 1 - \frac{\epsilon}{5}$ *and the algorithm does not output* reject *in Step 8 nor 12;*
- *if* $\mathbf{P}$ *puts probability mass more than* $\frac{\epsilon}{4}$ *outside of* I*, then the algorithm outputs* reject *in Step 8 or 12.*

Proof. Suppose first $\mathbf{P} \in \mathcal{SIIRV}_{n,k}$ and (2) holds, and set $s \stackrel{\text{def}}{=} \sqrt{\mathrm{Var}_{X\sim\mathbf{P}}[X]}$ and $\mu \stackrel{\text{def}}{=} \mathbb{E}_{X\sim\mathbf{P}}[X]$ as before. By Bennett's inequality applied to X, we have

$$\Pr[X > \mu + t] \le \exp\left(-\frac{s^2}{k^2}\vartheta\left(\frac{kt}{s^2}\right)\right) \tag{3}$$

for any $t > 0$, where $\vartheta\colon \mathbb{R}_+^* \to \mathbb{R}$ is defined by $\vartheta(x) = (1+x)\ln(1+x) - x$.

- If the algorithm reaches Step 8, then $s \le 4k\sqrt{\ln\frac{10}{\epsilon}}$. Setting $t = \alpha\cdot k\ln\frac{10}{\epsilon}$ in Eq. (3) (for $\alpha > 0$ to be determined shortly), and $u = \frac{kt}{s^2} = \alpha\frac{k^2}{s^2}\ln\frac{10}{\epsilon} \ge \frac{\alpha}{16}$,

$$\frac{s^2}{k^2}\vartheta\left(\frac{kt}{s^2}\right) = \alpha\ln\frac{10}{\epsilon}\cdot\frac{\vartheta(u)}{u} \ge \left(16\vartheta\left(\frac{\alpha}{16}\right)\right)\ln\frac{10}{\epsilon} \ge \ln\frac{10}{\epsilon}$$

since $\frac{\vartheta(x)}{x} \ge \frac{\vartheta(\alpha/16)}{\alpha/16}$ for all $x \ge \frac{\alpha}{16}$; the last inequality for $\alpha \ge \alpha^* \simeq 2.08$ chosen to be the solution to $16\vartheta\left(\frac{\alpha^*}{16}\right) = 1$. Thus, $\Pr[X > \mu + t] \le \frac{\epsilon}{10}$. Similarly, we have $\Pr[X < \mu - t] \le \frac{\epsilon}{10}$. As $\mu - 2t \le \mu - s \le \widetilde{\mu} \le \mu + s \le \mu + 2t$, we get $\Pr[X \in I] \ge 1 - \frac{\epsilon}{5}$ as claimed.

[3] [DDS15, Lemma 6] actually only deals with the case $k = 2$; but the bound we state follows immediately from their proof and the simple observation that the excess kurtosis κ of an (n,k)-SIIRV with variance s^2 is at most k^2/s^2.

- If the algorithm reaches Step 12, then $s \geq k\sqrt{\ln \frac{10}{\epsilon}}$ and $M = 1+2\left\lceil 6\widetilde{\sigma}\sqrt{\ln \frac{10}{\epsilon}})\right\rceil \geq 1+2\left\lceil 3s\sqrt{\ln \frac{10}{\epsilon}})\right\rceil$. Setting $t = \beta s\sqrt{\ln \frac{10}{\epsilon}}$ in Eq. (3) (for $\beta > 0$ to be determined shortly), and $u = \frac{kt}{s^2} = \beta\frac{k}{s}\sqrt{\ln \frac{10}{\epsilon}} \leq \beta$,

$$\frac{s^2}{k^2}\vartheta\left(\frac{kt}{s^2}\right) = \frac{t^2}{s^2}\cdot\frac{\vartheta(u)}{u^2} = \beta^2 \ln\frac{10}{\epsilon}\cdot\frac{\vartheta(u)}{u^2} \geq \ln\frac{10}{\epsilon}$$

since $\frac{\vartheta(x)}{x^2} \geq \frac{\vartheta(\beta)}{\beta^2}$ for all $x \in (0,\beta]$; the last inequality for $\beta = e-1 \simeq 1.72$ chosen to be the solution to $\vartheta(\beta) = 1$. Thus, $\Pr[X > \mu + t] \leq \frac{\epsilon}{10}$. Similarly, it holds $\Pr[X < \mu - t] \leq \frac{\epsilon}{10}$. Now note that $\lfloor\widetilde{\mu}\rfloor + (M-1)/2 \geq (\mu - s) + \lceil 2s\sqrt{\ln\frac{10}{\epsilon}})\rceil \geq \mu + t$ and $\lfloor\widetilde{\mu}\rfloor - (M-1)/2 \leq \mu - t$, implying that X is in $[\lfloor\widetilde{\mu}\rfloor - (M-1)/2, \lfloor\widetilde{\mu}\rfloor + (M-1)/2]$ with probability at least $1 - \frac{\epsilon}{5}$ as desired.

To conclude and establish the conclusion of the first item, as well as the second item, recall that distinguishing with probability $19/20$ between the cases $\mathbf{P}(\bar{I}) \leq \frac{\epsilon}{5}$ and $\mathbf{P}(\bar{I}) > \frac{\epsilon}{4}$ can be done with $O(1/\epsilon)$ samples. □

Claim 4.4 (Learning when the effective support is small)**.** *If* $\mathbf{P}$ *satisfies* $\mathbf{P}(I) \geq 1-\frac{\epsilon}{4}$*, and the "If" statement at Step 6 holds, then with probability at least* $19/20$ *the empirical distribution* $\mathbf{H}$ *obtained in Step 9 satisfies (i)* $d_{\mathrm{TV}}(\mathbf{P},\mathbf{H}) \leq \frac{\epsilon}{2}$ *and (ii)* $\|\widehat{\mathbf{P}} - \widehat{\mathbf{H}}\|_2 \leq \frac{\epsilon^2}{100}$.

Proof. The first item, (i), follows from standard bounds on the rate of convergence of the empirical distribution (namely, that $O(r/\epsilon^2)$ samples suffice for it to approximate an arbitrary distribution over support of size r up to total variation distance ϵ). Recalling that in this branch of the algorithm, $S = [M]$ with $M = O(k\log(1/\epsilon))$, the second item, (ii), is proven by the same argument as in (the first bullet in) the proof of Theorem 3.1. □

Claim 4.5 (Any (n,k)-SIIRV puts near all its Fourier mass in S)**.** *If* $\mathbf{P} \in \mathcal{SIIRV}_{n,k}$ *and* (2) *holds, then* $\|\widehat{\mathbf{P}}\mathbf{1}_{\bar{S}}\|_2^2 = \sum_{\xi\notin S}|\widehat{\mathbf{P}}(\xi)|^2 \leq \frac{\epsilon^2}{100}$.

Proof. Since $\mathbf{P} \in \mathcal{SIIRV}_{n,k}$, our assumptions imply that (with the notations of Lemma 2.7) the set of large Fourier coefficients satisfies $\left\{\ \xi \in [M-1]:\ \left|\widehat{\mathbf{P}}(\xi)\right| > \delta\ \right\} \subseteq \mathcal{L}(\delta,M,s) \subseteq S$. Therefore, $\xi \notin S$ implies $|\widehat{\mathbf{P}}(\xi)| \leq \delta$. We then can conclude as follows: applying Lemma 2.7 (ii) with parameter $\delta 2^{-r-1}$ for each $r \geq 0$, this is at most

$$\begin{aligned}\sum_{r\geq 0}(\delta 2^{-r})^2\left|\left\{\ \xi:\ |\widehat{\mathbf{P}}(\xi)| > \delta 2^{-r-1}\ \right\}\right| &\leq \frac{4Mk\delta^2}{s}\sum_{r\geq 0}4^{-r}\sqrt{\log(2^{r+2}/\delta)}\\ &\leq \frac{4Mk\delta^2}{s}\sqrt{\log\frac{1}{\delta}}\sum_{r\geq 0}4^{-r}\sqrt{\log(2^{r+1})}\\ &\leq \frac{12Mk\delta^2}{s}\sqrt{\log\frac{1}{\delta}} = O\left(\epsilon^2\right) \qquad (4)\end{aligned}$$

again at most $\frac{\epsilon^2}{100}$ for big enough C'' in the definition of δ. □

4.2 Putting It Together In what follows, we implicitly assume that I (as defined in Step 11 of Algorithm 3) is equal to $[M]$. This can be done without loss of generality, as this is just a shifting of the interval and all our Fourier arguments are made modulo M.

Lemma 4.6 (Putting it together: completeness)**.** *If $\mathbf{P} \in \mathcal{SIIRV}_{n,k}$, then the algorithm outputs* accept *with probability at least $3/5$.*

Proof. Assume $\mathbf{P} \in \mathcal{SIIRV}_{n,k}$. We condition on the estimates obtained in Step 2 to meet their accuracy guarantees, which by Claim 4.2 holds with probability at least $19/20$: that is, we hereafter assume Eq. (2) holds. Since the variance of any (n,k)-SIIRV is at most $s^2 \le nk^2$, we consequently have $\widetilde{\sigma} \le 2k\sqrt{n}$ and the algorithm does not output reject in Step 3.

- **Case 1:** the branch in Step 6 is taken. In this case, by Claim 4.3 the algorithm does not output reject in Step 8 with probability $19/20$. Since $\mathbf{P}(I) \ge 1 - \frac{\epsilon}{4}$, by Claim 4.4 we get that with probability at least $19/20$ it is the case that $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{H}) \le \frac{\epsilon}{2}$, and therefore the computational check in Step 16 will succeed, and return accept. Overall, by a union bound the algorithm is successful with probability at least $1 - 3/20 > 3/5$.

- **Case 2:** the branch in Step 10 is taken. In this case, by Claim 4.3 the algorithm does not output reject in Step 12 with probability $19/20$. From Lemma 2.8, we know that $\mathbf{P}'$ as defined in Step 14 satisfies $\|\mathbf{P}'\|_2^2 \le \frac{8k}{s} \le \frac{16k}{\widetilde{\sigma}} = b$. Moreover, Claim 4.5 guarantees that $\|\widehat{\mathbf{P}'}\mathbf{1}_{\bar{S}}\|_2 \le \frac{\epsilon}{10\sqrt{M}} = \frac{\epsilon'}{2}$ (for $\epsilon' = \frac{\epsilon}{5\sqrt{M}}$). Since Step 14 calls Algorithm 1 with parameters M, ϵ', b, and S, Item 3 of Theorem 3.1 ensures that (with probability at least $7/10$) the algorithm will not output reject in Step 14, but instead return the S-sparse Fourier transform of some $\mathbf{H}$ supported on $[M]$ with $\|\mathbf{P}' - \mathbf{H}\|_2 \le \frac{6}{5}\epsilon' = \frac{6\epsilon}{25\sqrt{M}}$.

  By Cauchy–Schwarz, we then have $\|\mathbf{P}' - \mathbf{H}\|_1 \le \sqrt{M}\|\mathbf{P}' - \mathbf{H}\|_2 \le \frac{6\epsilon}{25}$, i.e. $d_{\mathrm{TV}}(\mathbf{P}', \mathbf{H}) \le \frac{3\epsilon}{25}$. But since $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{P}') \le \frac{\epsilon}{4}$, we get $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{H}) \le \frac{\epsilon}{4} + \frac{3\epsilon}{25} < \frac{\epsilon}{2}$, and the computational check in Step 16 will succeed, and return accept. Overall, by a union bound the algorithm accepts with probability at least $1 - (1/20 + 1/20 + 3/10) = 3/5$.

□

Lemma 4.7 (Putting it together: soundness)**.** *If $d_{\mathrm{TV}}(\mathbf{P}, \mathcal{SIIRV}_{n,k}) > \epsilon$, then the algorithm outputs* reject *with probability at least $3/5$.*

Proof. We will proceed by contrapositive, and show that if the algorithm returns accept with probability at least $3/5$ then $d_{\mathrm{TV}}(\mathbf{P}, \mathcal{SIIRV}_{n,k}) \le \epsilon$. Depending on the branch of the algorithm followed, we assume the samples taken either in

- Steps 2, 8, 9, meet the guarantees of Claims 4.2 to 4.4 (by a union bound, this happens with probability at least $1 - 3/20 > 2/3$); or

- Steps 2, 12, 14 meet the guarantees of Claims 4.2 and 4.3 and Theorem 3.1 (by a union bound, this happens with probability at least $1 - (1/20 + 1/20 + 3/10) = 3/5$).

In particular, we hereafter assume that $\widetilde{\sigma} \le 2k\sqrt{n}$.

- **Case 1:** the branch in Step 6 is taken.

  By the above discussion, we have $\mathbf{P}(I) \geq 1 - \frac{\epsilon}{4}$ by Claim 4.3 so Claim 4.4 and our conditioning ensure that the empirical distribution $\mathbf{H}$ is such that $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{H}) \leq \frac{\epsilon}{2}$. Since the algorithm did not reject in Step 16, there exists a (n,k)-SIIRV $\mathbf{P}^*$ such that $d_{\mathrm{TV}}(\mathbf{H}, \mathbf{P}^*) \leq \frac{\epsilon}{2}$: by the triangle inequality, $d_{\mathrm{TV}}(\mathbf{P}, \mathcal{SIIRV}_{n,k}) \leq d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}^*) \leq \epsilon$.

- **Case 2:** the branch in Step 10 is taken.

  In this case, we have $\mathbf{P}(I) \geq 1 - \frac{\epsilon}{4}$ by Claim 4.3. Furthermore, as the algorithm did not output reject on Step 14, by Theorem 3.1 we know that the inverse Fourier transform (modulo M) $\mathbf{H}$ of the S-sparse collection of Fourier coefficients $\widehat{\mathbf{H}}$ returned satisfies $\|\mathbf{H} - \mathbf{P}'\|_2 \leq \frac{6\epsilon}{25\sqrt{M}}$ which by Cauchy–Schwarz implies, as both $\mathbf{H}$ and $\mathbf{P}'$ are supported on $[M]$, that $\|\mathbf{H} - \mathbf{P}'\|_1 \leq \frac{6\epsilon}{25}$, or equivalently $d_{\mathrm{TV}}(\mathbf{H}, \mathbf{P}') \leq \frac{3\epsilon}{25}$.

  Finally, since the algorithm outputted accept in Step 16, there exists $\mathbf{P}^* \in \mathcal{SIIRV}_{n,k}$ (supported on $[M]$) such that $d_{\mathrm{TV}}(\mathbf{H}, \mathbf{P}^*) \leq \frac{\epsilon}{2}$, and by the triangle inequality

$$d_{\mathrm{TV}}(\mathbf{P}, \mathbf{P}^*) \leq d_{\mathrm{TV}}(\mathbf{P}, \mathbf{P}') + d_{\mathrm{TV}}(\mathbf{H}, \mathbf{P}') + d_{\mathrm{TV}}(\mathbf{H}, \mathbf{P}^*) \leq \frac{\epsilon}{4} + \frac{3\epsilon}{25} + \frac{\epsilon}{2} \leq \epsilon$$

  and thus $d_{\mathrm{TV}}(\mathbf{P}, \mathcal{SIIRV}_{n,k}) \leq d_{\mathrm{TV}}(\mathbf{P}, \mathbf{P}^*) \leq \epsilon$.

□

Lemma 4.8 (Sample complexity)**.** *The algorithm has sample complexity* $O\left(\frac{kn^{1/4}}{\epsilon^2}\log^{1/4}\frac{1}{\epsilon} + \frac{k^2}{\epsilon^2}\log^2\frac{k}{\epsilon}\right)$.

Proof. Algorithm 3 takes samples in Steps 2, 8, 12, and 14. The sample complexity is dominated by Steps 9 and 14, which take respectively N and

$$\begin{aligned} O\left(\frac{\sqrt{b}}{(\epsilon/\sqrt{M})^2} + \frac{|S|}{M(\epsilon/\sqrt{M})^2} + \sqrt{M}\right) &= O\left(\frac{\sqrt{k\widetilde{\sigma}}}{\epsilon^2}\sqrt[4]{\log\frac{1}{\epsilon}} + \frac{|S|}{\epsilon^2} + \sqrt{\widetilde{\sigma}}\sqrt[4]{\log\frac{1}{\epsilon}}\right) \\ &= O\left(\frac{kn^{1/4}}{\epsilon^2}\log^{1/4}\frac{1}{\epsilon} + \frac{k^2}{\epsilon^2}\log^2\frac{k}{\epsilon}\right) \end{aligned}$$

samples; recalling that Step 3 ensured that $\widetilde{\sigma} \leq 2k\sqrt{n}$ and that $|S| = O\left(k^2\log^2\frac{k}{\epsilon}\right)$ by Fact 4.1. □

Lemma 4.9 (Time complexity)**.** *The algorithm runs in time* $O\left(\frac{k^4n^{1/4}}{\epsilon^2}\log^4\frac{k}{\epsilon}\right) + T(n,k,\epsilon)$*, where* $T(n,k,\epsilon) = n(k/\epsilon)^{O(k\log(k/\epsilon))}$ *is the running time of the projection subroutine of Step 16.*

Proof. The running time, depending on the branch taken, is either $O(N) + T(n,k,\epsilon)$ for the first or $O\left(|S|\left(\frac{kn^{1/4}}{\epsilon^2}\log^{1/4}\frac{1}{\epsilon} + \frac{k^2}{\epsilon^2}\log^2\frac{k}{\epsilon}\right)\right) + T(n,k,\epsilon)$ for the second (the latter from the running time of Algorithm 1). Recalling that $|S| = O\left(k^2\log^2\frac{k}{\epsilon}\right)$ by Fact 4.1 yields the claimed running time. □

4.3 The Projection Subroutines

Algorithm 4 Algorithm `Project-k-SIIRV`

Require: Parameters n,ϵ; the approximate Fourier coefficients $(\widehat{\mathbf{H}}(\xi))_{\xi\in S}$ modulo M, of a distribution $\mathbf{P}$ known to be effectively supported on I and to have a Fourier transform effectively supported on S of the form given in Step 13 of Algorithm 3, with $\widetilde{\sigma}^2$ and $\widetilde{\mu}$, an approximation to $\mathbb{E}_{X\sim\mathbf{P}}[X]$ to within half a standard deviation.

1: Compute $\mathcal{C}$, an $\frac{\epsilon}{5\sqrt{|S|}}$-cover in total variation distance of all (n,k)-SIIRVs.
2: **for** each $\mathbf{Q}\in\mathcal{C}$ **do**
3: **if** the mean $\mu_{\mathbf{Q}}$ and variance $\sigma_{\mathbf{Q}}$ of $\mathbf{Q}$ satisfy $|\widetilde{\mu}-\mu_{\mathbf{Q}}|\le\widetilde{\sigma}$ and $2(\sigma_{\mathbf{Q}}+1)\ge\widetilde{\sigma}+1\ge(\sigma_{\mathbf{Q}}+1)/2$ **then**
4: Compute $\widehat{\mathbf{Q}}(\xi)$ for $\xi\in S$.
5: **if** $\sum_{\xi\in S}|\widehat{\mathbf{H}}-\widehat{\mathbf{Q}}|^2\le\frac{\epsilon^2}{5}$ **then return** accept
6: **end if**
7: **end if**
8: **end for**
9: **return** reject ▷ we did not return accept for any $\mathbf{Q}\in\mathcal{C}$

4.3.1 The Projection Step for (n,k)-SIIRVs We can use the proper ϵ-cover given in [DKS16b] to find a (n,k)-SIIRV near $\mathbf{P}$ by looking at $\widehat{\mathbf{H}}$.

Lemma 4.10. *If Algorithm `Project-k-SIIRV` is given inputs that satisfy its assumptions and we have that $\sum_{\xi\in S}|\widehat{\mathbf{H}}-\widehat{\mathbf{P}}|^2\le(3\epsilon/25)^2$, $d_{\mathrm{TV}}(\mathbf{P},\mathbf{H})\le 6\epsilon/25$, and that if $\mathbf{P}\in\mathcal{SIIRV}_{n,k}$ then $\widetilde{\sigma}^2$ is a factor-1.5 approximation to $\mathrm{Var}_{X\sim\mathbf{P}}[X]+1$, then it distinguishes between (i) $\mathbf{P}\in\mathcal{SIIRV}_{n,k}$ and (ii) $d_{\mathrm{TV}}(\mathbf{P},\mathcal{SIIRV}_{n,k})>\epsilon$. The algorithm runs in time $n\,(k/\epsilon)^{O(k\log(k/\epsilon))}$.*

4.3.2 The Case $k=2$ For the important case of Poisson Binomial distributions, that is $(n,2)$-SIIRVs, we can dispense with using a cover at all. [DKS16c] gives an algorithm that can properly learn Poisson binomial distributions in time $(1/\epsilon)^{O(\log\log 1/\epsilon)}$. The algorithm works by first learning the Fourier coefficients in S, which we have already computed here, and checks if one of many systems of polynomial inequalities has a solution: if the Fourier coefficients are close to those of a $(n,2)$-SIIRV, then there will be a solution to one of these systems. This allows us to test whether or not we are close to a $(n,2)$-SIIRV.

More precisely, we can handle this in two cases: the first, when the variance s^2 of $\mathbf{P}$ is relatively small, corresponding to $\widetilde{\sigma}\le\alpha/\epsilon^2$ (for some absolute constant $\alpha>0$). In this case, we use the following lemma:

Lemma 4.11. *Let $\mathbf{P}$ be a distribution with variance $O(1/\epsilon^2)$. Let $\widetilde{\mu}$ and $\widetilde{\sigma}^2$ be approximations to the mean μ and variance s^2 of $\mathbf{P}$ with $|\widetilde{\mu}-\mu|\le\widetilde{\sigma}$ and $2(\sigma+1)\ge\widetilde{\sigma}+1\ge(\sigma+1)/2$. Suppose that $\mathbf{P}$ is effectively supported on an interval I and that its DFT modulo M is effectively supported on S, the set of integers $\xi\le\ell\stackrel{\mathrm{def}}{=}O(\log(1/\epsilon))$. Let $\widehat{\mathbf{H}}(\xi)$ be approximations to $\widehat{\mathbf{P}}(\xi)$ for all $\xi\in S$ with $\sum_{\xi\in S}|\widehat{\mathbf{H}}(\xi)-\widehat{\mathbf{P}}(\xi)|^2\le\frac{\epsilon^2}{16}$. There is an algorithm that, given n,ϵ,$\widetilde{\mu}$, $\widetilde{\sigma}$ and $\widehat{\mathbf{H}}(\xi)$, distinguishes between (i) $\mathbf{P}\in\mathcal{PBD}_n$ and (ii) $d_{\mathrm{TV}}(\mathbf{P},\mathcal{PBD}_n)>\epsilon$, in time at most $(1/\epsilon)^{O(\log\log 1/\epsilon)}$.*

If $\widetilde{\sigma}\ge\alpha/\epsilon^2$ (corresponding to a "big variance" $s^2=\Omega(1/\epsilon^2)$), then we take an additional $O(|S|/\epsilon^2)$ samples from $\mathbf{P}$ and use them to learn a shifted binomial using algorithms `Learn-Poisson` and `Locate-Binomial` from [DDS15] that is within $O(\epsilon/\sqrt{|S|})$ total variation distance from $\mathbf{P}$. If these succeed, we can check if its Fourier coefficients are close using the method in Algorithm 4 (`Project-k-SIIRV`). As we can compute the Fourier coefficients of a shifted binomial easily, this overall takes time $\mathrm{poly}(1/\epsilon)$.

5 The General Tester

In this section, we abstract the ideas underlying the (n, k)-SIIRV from Section 4, to provide a general testing framework. In more detail, our theorem (Theorem 5.1) has the following flavor: if $\mathcal{P}$ is a property of distributions such that every $\mathbf{P} \in \mathcal{P}$ has both (i) small effective support and (ii) sparse effective Fourier support, then one can test membership to $\mathcal{P}$ with $O(\sqrt{sM}/\epsilon^2 + s/\epsilon^2)$ samples (where M and s are the bounds on the effective support and effective Fourier support, respectively). As a caveat, we do require that the sparse effective Fourier support S be independent of $\mathbf{P} \in \mathcal{P}$, i.e., is a characteristic of the class $\mathcal{P}$ itself.

The high-level idea is then quite simple: the algorithm proceeds in three stages, namely the *effective support test*, the *Fourier effective support test*, and the *projection step*. In the first, it takes some samples from $\mathbf{P}$ to identify what should be the effective support I of $\mathbf{P}$, if $\mathbf{P}$ did have the property: and then checks that indeed $|I| \leq M$ (as it should) and that $\mathbf{P}$ puts probability mass $1 - O(\epsilon)$ on I.

In the second stage, it invokes the Fourier testing algorithm of Section 3 to verify that $\widehat{\mathbf{P}}$ indeed puts very little Fourier mass outside of S; and, having verified this, learns very accurately the set of Fourier coefficients of $\mathbf{P}$ on this set S, in L_2 distance.

At this point, either the algorithm has detected that $\mathbf{P}$ violates some required characteristic of the distributions in $\mathcal{P}$, in which case it has rejected already; or is guaranteed to have *learned* a good approximation $\mathbf{H}$ of $\mathbf{P}$, by the Fourier learning performed in the second stage. It only remains to perform the third stage, which "projects" this good approximation $\mathbf{H}$ of $\mathbf{P}$ onto $\mathcal{P}$ to verify that $\mathbf{H}$ is close to some distribution $\mathbf{P}^* \in \mathcal{P}$ (as it should if indeed $\mathbf{P} \in \mathcal{P}$).

Algorithm 5 Algorithm `Test-Fourier-Sparse-Class`

Require: sample access to a distribution $\mathbf{P} \in \Delta(\mathbb{N})$, parameter $\epsilon \in (0, 1]$, $b \in (0, 1]$, functions $S\colon (0, 1] \to 2^{\mathbb{N}}$, $M\colon (0, 1] \to \mathbb{N}$, $q_I\colon (0, 1] \to \mathbb{N}$, and procedure PROJECT$_{\mathcal{P}}$ as in Theorem 5.1

Effective Support

Take $q_I(\epsilon)$ samples from $\mathbf{P}$ to identify a "candidate set" I. ▷ Guaranteed to work w.p. $19/20$ if $\mathbf{P} \in \mathcal{P}$.

Draw $O(1/\epsilon)$ samples from $\mathbf{P}$, to distinguish between $\mathbf{P}(I) \geq 1 - \frac{\epsilon}{5}$ and $\mathbf{P}(I) < 1 - \frac{\epsilon}{4}$. ▷ Correct w.p. $19/20$.

if $|I| > M(\epsilon)$ or we detected that $\mathbf{P}(I) > \frac{\epsilon}{4}$ **then**

return reject

end if

Fourier Effective Support

Simulating sample access to $\mathbf{P}' \stackrel{\text{def}}{=} \mathbf{P} \bmod M(\epsilon)$, call Algorithm 1 on $\mathbf{P}'$ with parameters $M(\epsilon)$, $\frac{\epsilon}{5\sqrt{M(\epsilon)}}$, b, and $S(\epsilon)$.

if Algorithm 1 returned reject **then**

return reject

end if

Let $\widehat{\mathbf{H}} = (\widehat{\mathbf{H}}(\xi))_{\xi \in S(\epsilon)}$ denote the collection of Fourier coefficients it outputs, and $\mathbf{H}$ their inverse Fourier transform (modulo $M(\epsilon)$) ▷ Do not actually compute $\mathbf{H}$ here.

Projection Step

Call PROJECT$_{\mathcal{P}}$ on parameters ϵ and $\mathbf{H}$, and **return** accept if it does, reject otherwise.

Theorem 5.1 (General Testing Statement)**.** *Assume $\mathcal{P} \subseteq \Delta(\mathbb{N})$ is a property of distributions satisfying the following. There exist $S\colon (0,1] \to 2^{\mathbb{N}}$, $M\colon (0,1] \to \mathbb{N}$, and $q_I\colon (0,1] \to \mathbb{N}$ such that, for every $\epsilon \in (0,1]$,*

1. *Fourier sparsity: for all $\mathbf{P} \in \mathcal{P}$, the Fourier transform (modulo $M(\epsilon)$) of $\mathbf{P}$ is concentrated on $S(\epsilon)$: namely, $\|\widehat{\mathbf{P}}\mathbf{1}_{\overline{S(\epsilon)}}\|_2^2 \leq \frac{\epsilon^2}{100}$.*
2. *Support sparsity: for all $\mathbf{P} \in \mathcal{P}$, there exists an interval $I(\mathbf{P}) \subseteq \mathbb{N}$ with $|I(\mathbf{P})| \leq M(\epsilon)$ such that (i) $\mathbf{P}$ is concentrated on $I(\mathbf{P})$: namely, $\mathbf{P}(I(\mathbf{P})) \geq 1 - \frac{\epsilon}{5}$ and (ii) $I(\mathbf{P})$ can be identified with probability at least $19/20$ from $q_I(\epsilon)$ samples from $\mathbf{P}$.*
3. *Projection: there exists a procedure $\textsc{Project}_{\mathcal{P}}$ which, on input $\epsilon \in (0,1]$ and the explicit description of a distribution $\mathbf{H} \in \Delta(\mathbb{N})$, runs in time $T(\epsilon)$; and outputs accept if $d_{\mathrm{TV}}(\mathbf{H}, \mathcal{P}) \leq \frac{2\epsilon}{5}$, and reject if $d_{\mathrm{TV}}(\mathbf{H}, \mathcal{P}) > \frac{\epsilon}{2}$ (and can answer either otherwise).*
4. *(Optional) L_2-norm bound: there exists $b \in (0,1]$ such that, for all $\mathbf{P} \in \mathcal{P}$, $\|\mathbf{P}\|_2^2 \leq b$.*

Then, there exists a testing algorithm for $\mathcal{P}$, in the usual standard sense: it outputs either accept or reject, and satisfies the following.

1. *if $\mathbf{P} \in \mathcal{P}$, then it outputs accept with probability at least $3/5$;*
2. *if $d_{\mathrm{TV}}(\mathbf{P}, \mathcal{P}) > \epsilon$, then it outputs reject with probability at least $3/5$.*

The algorithm takes

$$O\left(\frac{\sqrt{|S(\epsilon)|\, M(\epsilon)}}{\epsilon^2} + \frac{|S(\epsilon)|}{\epsilon^2} + q_I(\epsilon)\right)$$

samples from $\mathbf{P}$ (if Item 4 holds, one can replace the above bound by $O\left(\frac{\sqrt{b}M(\epsilon)}{\epsilon^2} + \frac{|S(\epsilon)|}{\epsilon^2} + q_I(\epsilon)\right)$); and runs in time $O\left(m\,|S| + T(\epsilon)\right)$, where m is the sample complexity.

Moreover, whenever the algorithm outputs accept, it also learns $\mathbf{P}$*; that is, it provides a hypothesis $\mathbf{H}$ such that $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{H}) \leq \epsilon$ with probability at least $3/5$.*

We remark that the statement of Theorem 5.1 can be made slightly more general; specifically, one can allow the procedure $\textsc{Project}_{\mathcal{P}}$ to have sample access to $\mathbf{P}$ and err with small probability, and further provide it with the Fourier coefficients $\widehat{\mathbf{H}}$ learnt in the previous step.

Proof of Theorem 5.1. For convenience, we hereafter write S and M instead of $S(\epsilon)$ and $M(\epsilon)$, respectively. Before establishing the theorem, which will be a generalization of (the second branch of) Algorithm 3, we note that it is sufficient to prove the version including Item 4. This is because, if no bound b is provided, one can fall back to setting $b \stackrel{\text{def}}{=} \frac{|S|+1}{M}$: indeed, for any $\mathbf{P} \in \mathcal{P}$,

$$\|\mathbf{P}\|_2^2 = \|\widehat{\mathbf{P}}\|_2^2 = \|\widehat{\mathbf{P}}\mathbf{1}_S\|_2^2 + \|\widehat{\mathbf{P}}\mathbf{1}_{\bar{S}}\|_2^2 = \frac{1}{M}\sum_{\xi \in S}|\widehat{\mathbf{P}}(\xi)|^2 + \|\widehat{\mathbf{P}}\mathbf{1}_{\bar{S}}\|_2^2 \leq \frac{|S|}{M} + \frac{\epsilon^2}{100M} = \frac{|S| + \frac{\epsilon^2}{100}}{M} \tag{5}$$

from Item 1 and the fact that $|\widehat{\mathbf{P}}(\xi)| \leq 1$ for any $\xi \in [M]$. Then, we have $\sqrt{b}M \leq \sqrt{2\frac{|S|}{M}}M = \sqrt{2\,|S|\,M}$, concluding the remark.

The algorithm is given in Algorithm 5. Its sample complexity and running time are immediate from the assumptions on the input parameters, and its description; we thus focus on establishing its correctness.

- Completeness: suppose $\mathbf{P} \in \mathcal{P}$. Then, by definition of q_I and M (Item 2 of the theorem), we have that with probability at least $19/20$ the interval I identified in Step 2 satisfies $\mathbf{P}(I) \geq 1 - \frac{\epsilon}{5}$ and $|I| \leq M$. In this case, also with probability at least $19/20$ the check in Step 3 succeeds, and the algorithm does not output reject there.

  The call to Algorithm 1 in Step 9 then, with probability at least 7/10, does not output reject, but instead Fourier coefficients $\widehat{H}$ (supported on S) of some $\mathbf{H}$ such that $\mathbf{H}' = \mathbf{H} \bmod M$ satisfies $\|\mathbf{H}' - \mathbf{P}'\|_2 \leq \frac{6}{5} \cdot \frac{\epsilon}{5\sqrt{M}} = \frac{6\epsilon}{25\sqrt{M}}$ (this is because of the definition of b and Item 1, which ensure the assumptions of Theorem 3.1 are met). Thus $\|\mathbf{H}' - \mathbf{P}'\|_1 \leq \sqrt{M}\|\mathbf{H}' - \mathbf{P}'\|_2 \leq \frac{6\epsilon}{25}$. Since $\|\mathbf{P} - \mathbf{P}'\|_2 \leq 2 \cdot \frac{\epsilon}{4}$ (as $\mathbf{P}(I) \geq 1 - \frac{\epsilon}{4}$ and $\mathbf{P}' = \mathbf{P} \bmod M$), by the triangle inequality

$$d_{\mathrm{TV}}(\mathbf{P}, \mathbf{H}') = \frac{1}{2}\|\mathbf{H}' - \mathbf{P}'\|_1 \leq \frac{3\epsilon}{25} + \frac{\epsilon}{4} < \frac{2\epsilon}{5}$$

  and the algorithm returns accept in Step 16 (as promised by Item 3).

  Overall, by a union bound the algorithm is correct with probability at least $1 - (\frac{1}{20} + \frac{1}{20} + \frac{3}{10}) \geq \frac{3}{5}$.

- Soundness: we proceed by contrapositive, and show that if the algorithm returns accept with probability at least $3/5$ then $d_{\mathrm{TV}}(\mathbf{P}, \mathcal{P}) \leq \epsilon$. We hereafter assume the guarantees of Steps 2, 3, and 9 hold, which by a union bound is the case with probability at least $1 - (\frac{1}{20} + \frac{1}{20} + \frac{3}{10}) \geq \frac{3}{5}$.

  Since the algorithm passed Step 5, we have $\mathbf{P}(I) \geq 1 - \frac{\epsilon}{4}$ and $|I| \leq M$. Furthermore, as the algorithm did not output reject on Step 9, by Theorem 3.1 we know that the inverse Fourier transform (modulo M) $\mathbf{H}$ of the S-sparse collection of Fourier coefficients $\widehat{\mathbf{H}}$ returned satisfies, for $\mathbf{H}' \stackrel{\text{def}}{=} \mathbf{H} \bmod M$,

$$\|\mathbf{H}' - \mathbf{P}'\|_2 \leq \frac{6\epsilon}{25\sqrt{M}}$$

  which by Cauchy–Schwarz implies that $\|\mathbf{H} - \mathbf{P}'\|_1 \leq \frac{6\epsilon}{25}$, or equivalently $d_{\mathrm{TV}}(\mathbf{H}, \mathbf{P}') \leq \frac{3\epsilon}{25}$.

  Finally, since the algorithm outputted accept in Step 16, there exists $\mathbf{P}^* \in \mathcal{P}$ (supported on $[M]$) such that $d_{\mathrm{TV}}(\mathbf{H}, \mathbf{P}^*) \leq \frac{\epsilon}{2}$, and by the triangle inequality

$$d_{\mathrm{TV}}(\mathbf{P}, \mathbf{P}^*) \leq d_{\mathrm{TV}}(\mathbf{P}, \mathbf{P}') + d_{\mathrm{TV}}(\mathbf{H}, \mathbf{P}') + d_{\mathrm{TV}}(\mathbf{H}, \mathbf{P}^*) \leq \frac{\epsilon}{4} + \frac{3\epsilon}{25} + \frac{\epsilon}{2} \leq \epsilon$$

  and thus $d_{\mathrm{TV}}(\mathbf{P}, \mathcal{P}) \leq d_{\mathrm{TV}}(\mathbf{P}, \mathbf{P}^*) \leq \epsilon$.

□

6 The PMD Tester

In this section, we generalize our Fourier testing approach to higher dimensions, and leverage it to design a testing algorithm for the class of Poisson Multinomial distributions – thus establishing Theorem 1.3 (restated below).

Theorem 6.1 (Testing PMDs)**.** *Given parameters $k, n \in \mathbb{N}$, $\epsilon \in (0, 1]$, and sample access to a distribution $\mathbf{P}$ over $\mathbb{N}$, there exists an algorithm (Algorithm 7) which outputs either accept or reject, and satisfies the following.*

1. *if $\mathbf{P} \in \mathcal{PMD}_{n,k}$, then it outputs accept with probability at least $3/5$;*

2. *if* $d_{\mathrm{TV}}(\mathbf{P}, \mathcal{PMD}_{n,k}) > \epsilon$*, then it outputs* reject *with probability at least* $3/5$.

Moreover, the algorithm takes $O\left(\frac{n^{(k-1)/4}k^{2k}\log(k/\epsilon)^k}{\epsilon^2}\right)$ *samples from* $\mathbf{P}$*, and runs in time* $n^{O(k^3)}\cdot(1/\epsilon)^{O(k^3\frac{\log(k/\epsilon)}{\log\log(k/\epsilon)})^{k-1}}$ *or alternatively in time* $n^{O(k)}\cdot 2^{O(k^{5k}\log(1/\epsilon)^{k+2}}$.

The reason for the two different running times is that, for the projection step, one can use either the cover given by [DKS16c] or that given by [DDKT16], which yield the two statements. In contrast to Section 4 and Section 5, for PMDs we will have to use a *multidimensional* Fourier transform, which is a little more complicated – and we define next.

Let $M \in \mathbb{Z}^{k\times k}$ be an integer $k \times k$ matrix. We consider the integer lattice $L = L(M) = M\mathbb{Z}^k \stackrel{\text{def}}{=} \{p \in \mathbb{Z}^k \mid p = Mq, q \in \mathbb{Z}^k\}$, and its dual lattice $L^* = L^*(M) \stackrel{\text{def}}{=} \left\{\, \xi \in \mathbb{R}^k : \xi \cdot x \in \mathbb{Z} \text{ for all } x \in L \,\right\}$. Note that $L^* = (M^T)^{-1}\mathbb{Z}^k$, and that L^* is not necessarily integral. The quotient $\mathbb{Z}^k/L$ is the set of equivalence classes of points in $\mathbb{Z}^k$ such that two points $x, y \in \mathbb{Z}^k$ are in the same equivalence class if, and only if, $x - y \in L$. Similarly, the quotient $L^*/\mathbb{Z}^k$ is the set of equivalence classes of points in L^* such that any two points $x, y \in L^*$ are in the same equivalence class if, and only if, $x - y \in \mathbb{Z}^k$.

The *Discrete Fourier Transform (DFT) modulo* M, $M \in \mathbb{Z}^{k\times k}$, of a function $F\colon \mathbb{Z}^k \to \mathbb{C}$ is the function $\widehat{F}_M\colon L^*/\mathbb{Z}^k \to \mathbb{C}$ defined as $\widehat{F}_M(\xi) \stackrel{\text{def}}{=} \sum_{x\in\mathbb{Z}^k} e(\xi \cdot x)F(x)$. (We will omit the subscript M when it is clear from the context.) Similarly, for the case that F is a probability mass function, we can equivalently write $\widehat{F}(\xi) = \mathbb{E}_{X\sim F}\left[e(\xi \cdot X)\right]$. The *inverse DFT* of a function $\widehat{G}\colon L^*/\mathbb{Z}^k \to \mathbb{C}$ is the function $G\colon A \to \mathbb{C}$ defined on a *fundamental domain* A of $L(M)$ as follows: $G(x) = \frac{1}{|\det(M)|}\sum_{\xi\in L^*/\mathbb{Z}^k}\widehat{G}(x)e(-\xi \cdot x)$. Note that these operations are inverse of each other, namely for any function $F\colon A \to \mathbb{C}$, the inverse DFT of $\widehat{F}$ is identified with F.

With this in hand, Algorithm 1 easily generalizes to high dimension:

Algorithm 6 Testing the Fourier Transform Effective Support in high dimension

Require: parameters, a $k \times k$ matrix M, $b, \epsilon \in (0, 1]$; a fundamental domain A of $L(M)$; sample access to distribution $\mathbf{Q}$ over A

1: Set $m \leftarrow \left\lceil C(\frac{\sqrt{b}}{\epsilon^2} + \sqrt{\det(M)})\right\rceil$ ▷ $C > 0$ is an absolute constant; $C = 2000$ works.
2: Draw $m' \leftarrow \mathrm{Poi}(m)$; if $m' > 2m$, **return** reject
3: Draw m' samples from $\mathbf{Q}$, and let $\mathbf{Q}'$ be the corresponding empirical distribution over $[M]$
4: Compute $\|\mathbf{Q}'\|_2^2$, $\widehat{\mathbf{Q}'}(\xi)$ for every $\xi \in S$, and $\left\|\widehat{\mathbf{Q}'}\mathbf{1}_S\right\|_2^2$ ▷ Takes time $O\left(m\,|S|\right)$
5: **if** $m'^2\|\mathbf{Q}'\|_2^2 - m' > \frac{3}{2}bm^2$ **then return** reject
6: **else if** $\|\mathbf{Q}'\|_2^2 - \left\|\widehat{\mathbf{Q}'}\mathbf{1}_S\right\|_2^2 \geq 3\epsilon^2 + \frac{1}{m'}$ **then return** reject
7: **else**
8: **return** $(\widehat{\mathbf{Q}'}(\xi))_{\xi\in S}$
9: **end if**

Crucially, we observe that the proof of Theorem 3.1 nowhere requires that $[M]$ be a set of M consecutive integers, but only that it is a fundamental domain of the lattice used in the DFT. Consequently, Theorem 3.1 also applies in this high dimensional setting, with appropriate notation. Note that the size of any fundamental domain is $\det(M)$ which appears in place of M in the sample complexity.

The proof of correctness of Algorithm 7 is very similar to that of Algorithm 3, except that we need results from the proof of correctness of the PMD Fourier learning algorithm of [DKS16d]; we will only sketch these ingredients here. That I is an effective support of a PMD whose mean and covariance matrix we

Algorithm 7 Algorithm `Test-PMD`

Require: sample access to a distribution $\mathbf{P} \in \Delta(\mathbb{N}^k)$, parameters $n, k \geq 1$ and $\epsilon \in (0, 1]$

▷ Let C, C', C'' be sufficiently large universal constants

Draw $m_0 = O(k^4)$ samples from X, and let $\widehat{\mu}$ be the sample mean and $\widehat{\Sigma}$ the sample covariance matrix.

Compute an approximate spectral decomposition of $\widehat{\Sigma}$, i.e., an orthonormal eigenbasis v_i with corresponding eigenvalues λ_i, $i \in [k]$.

Set $M \in \mathbb{Z}^{k \times k}$ to be the matrix whose i^{th} column is the closest integer point to the vector $C\left(\sqrt{k \log(k/\epsilon)\lambda_i + k^2 \log^2(k/\epsilon)}\right) v_i$.

Set $I \leftarrow \mathbb{Z}^k \cap (\widehat{\mu} + M \cdot (-1/2, 1/2]^k)$

Draw $O(1/\epsilon)$ samples from $\mathbf{P}$, and **return** reject if any falls outside of I

Let $S \subseteq (\mathbb{R}/\mathbb{Z})^k$ to be the set of points $\xi = (\xi_1, \ldots, \xi_k)$ of the form $\xi = (M^T)^{-1} \cdot v + \mathbb{Z}^k$, for some $v \in \mathbb{Z}^k$ with $\|v\|_2 \leq C^2 k^2 \log(k/\epsilon)$.

Define $\mathbf{P} \bmod M$ to be the distribution obtained by sampling X from $\mathbf{P}$ and if it lies outside in I, returning X, else returning $X + Mb$ for the uniwue $b \in \mathbb{Z}^k$ such that $X + Mb \in I$.

Simulating sample access to $\mathbf{P}' \stackrel{\text{def}}{=} \mathbf{P} \bmod M$, call Algorithm 6 on $\mathbf{P}'$ with parameters M, $\frac{\epsilon}{5\sqrt{\det(M)}}$, $b = \frac{|S|+1}{\det(M)}$, and S. If it outputs reject, then **return** reject; otherwise, let $\widehat{\mathbf{H}} = (\widehat{\mathbf{H}}(\xi))_{\xi \in S}$ denote the collection of Fourier coefficients it outputs, and $\mathbf{H}$ their inverse Fourier transform (modulo M) onto I.
▷ Do not actually compute $\mathbf{H}$

Compute a proper $\epsilon/6\sqrt{|S|}$-cover $\mathcal{C}$ of all PMDs using the algorithm from [DKS16d].

for each $\mathbf{Q} \in \mathcal{C}$ **do**

  if the mean $\mu_{\mathbf{Q}}$ and covariance matrix $\Sigma_{\mathbf{Q}}$ satisfy $(\widehat{\mu} - \mu_{\mathbf{Q}})^T (\Sigma + I)^{-1} (\widehat{\mu} - \mu_{\mathbf{Q}}) \leq 1$ and $2(\Sigma_{\mathbf{Q}} + I) \geq \widehat{\Sigma} + I \geq (\Sigma_{\mathbf{Q}} + I)/2$. **then**

    Compute $\widehat{\mathbf{Q}}(\xi)$ for $\xi \in S$.

    if $\sum_{\xi \in S} |\widehat{\mathbf{H}} - \widehat{\mathbf{Q}}|^2 \leq \epsilon^2/16$ **then return** accept

    end if

  end if

end for

return reject if we do not accept for any $\mathbf{Q} \in \mathcal{C}$.

have estimated to within appropriate error with high probability follows from Lemmas 3.3–3.6 of [DKS16d], the last of which gives that the probability mass outside of I is at most $\epsilon/10$, smaller than that claimed for I in the (n, k)-SIIRV algorithm. Lemma 3.3 gives, if $\mathbf{P}$ is a PMD, that the mean and covariance satisfy $(\widehat{\mu} - \mu)^T (\Sigma + I)^{-1} (\widehat{\mu} - \mu) = O(1)$ and $2(\Sigma_{\mathbf{Q}} + I) \geq \widehat{\Sigma} + I \geq (\Sigma_q + I)/2$. Again, with more samples, we can strengthen this to $(\widehat{\mu} - \mu)^T (\Sigma + I)^{-1} (\widehat{\mu} - \mu) = \frac{1}{2}$ and $(3/2)(\Sigma + I) \geq \widehat{\Sigma} + I \geq (\Sigma + I)/(3/2)$ with $O(k^4)$ samples.

The effective support of the Fourier transform of a PMD is given by the following proposition:

Proposition 6.2 (Proposition 2.4 of [DKS16d])**.** *Let S be as in the algorithm. With probability at least $99/100$, the Fourier coefficients of $\mathbf{P}$ outside S satisfy $\sum_{\xi \in (L^*/\mathbb{Z}^k) \setminus S} |\widehat{\mathbf{P}}(\xi)| < \epsilon/10$.*

This holds not just for $\mathbf{P}$, but any (n, k)-PMD $\mathbf{Q}$ whose mean $\mu_{\mathbf{Q}}$ and covariance matrix $\Sigma_{\mathbf{Q}}$ satisfy $(\widehat{\mu} - \mu_{\mathbf{Q}})^T (\Sigma + I)^{-1} (\widehat{\mu} - \mu) = O(1)$ and $2(\Sigma_{\mathbf{Q}} + I) \geq \widehat{\Sigma} + I \geq (\Sigma_{\mathbf{Q}} + I)/2$.

We need to show that this L_1 bound is stronger than the L_2 bound we need. Since every individual

$\xi \notin S$ has $|\widehat{\mathbf{P}}(\xi)| < \epsilon/10$, we have

$$\sum_{\xi \in (L^*/\mathbb{Z}^k)\setminus S} |\widehat{\mathbf{P}}(\xi)|^2 \leq \sum_{\xi \in (L^*/\mathbb{Z}^k)\setminus S} \epsilon |\widehat{\mathbf{P}}(\xi)|/10 \leq \epsilon^2/100$$

and so S is an effective support of the DFT modulo M.

To show that the value of b is indeed a bound on $\|\mathbf{P}\|_2^2$, we can use (5), yielding that $\|\mathbf{P}\|_2^2 \leq (|S| + 1)/\det(M) = b$, where $\det(M)$ here is indeed the size of I.

The proof of correctness of the algorithm and the projection step is now very similar to the (n,k)-SIIRV case. We need to get bounds on the sample and time complexity. We can bound the size of S using

$$\begin{aligned} |S| &\leq \left| \left\{ \, v \in \mathbb{Z}^k \, : \, \|v\|_2 \leq C^2 k^2 \log(k/\epsilon) \, \right\} \right| \leq \left| \left\{ \, v \in \mathbb{Z}^k \, : \, \|v\|_\infty \leq C^2 k^2 \log(k/\epsilon) \, \right\} \right| \\ &= \left(1 + 2\lfloor C^2 k^2 \log(k/\epsilon) \rfloor\right)^k = O(k^2 \log(k/\epsilon))^k \end{aligned}$$

We can bound $\det(M)$ in terms of the L_2 norms of its columns using Hadamard's inequality

$$\det(M) \leq \prod_{i=1}^k \|M_i\|_2 \leq \prod_{i=1}^k \left(C \left(\sqrt{k \log(k/\epsilon) \lambda_i + k^2 \log^2(k/\epsilon)} \right) + \sqrt{k} \right)$$

recalling that λ_i are the eigenvalues of $\widehat{\Sigma}$ which satisfies $2(\Sigma_{\mathbf{Q}} + I) \geq \widehat{\Sigma} + I$. We need a bound on $\|\Sigma\|_2$. Each individual summand k-CRV (categorical random variable) is supported on unit vectors, the distance between any two of which is $\sqrt{2}$. Therefore we have that $\|\Sigma\|_2 \leq 2n$. Then $\lambda_i \leq 4n+1$ for every $1 \leq i \leq k$; moreover, since the k coordinates must sum to n, $\widehat{\Sigma}$ has rank at most $k-1$ and so at least one of the λ_i's is zero. Combining these observations, we obtain

$$\det(M) \leq \sqrt{k^2 \log^2 \frac{k}{\epsilon}} \cdot \left(C^2 k (4n+2) \log \frac{k}{\epsilon} + k^2 \log^2 \frac{k}{\epsilon} \right)^{\frac{k-1}{2}} = k \log \frac{k}{\epsilon} \cdot O\left(n k^2 \log \frac{k}{\epsilon} \right)^{\frac{k-1}{2}}.$$

With high constant probability, the number of samples we need is then

$$\begin{aligned} &O\left(\frac{\sqrt{|S| \det M}}{\epsilon^2} + \frac{|S|}{\epsilon^2} + q_I(\epsilon) \right) = \frac{1}{\epsilon^2} \sqrt{k \log \frac{k}{\epsilon}} \cdot O\left(n k^2 \log \frac{k}{\epsilon} \right)^{\frac{k-1}{4}} + \frac{O(k^2 \log(k/\epsilon))^k}{\epsilon^2} + O(k^4) \\ &= O(n^{(k-1)/4} k^{2k} \log(k/\epsilon)^k / \epsilon^2) \end{aligned}$$

The time complexity of the algorithm is dominated by the projection step. By Proposition 4.9 and Corollary 4.12 of [DKS16d], we can produce a proper ϵ-cover of $\mathcal{PMD}_{n,k}$ of size $n^{O(k^3)} \cdot (1/\epsilon)^{O(k^3 \frac{\log(k/\epsilon)}{\log\log(k/\epsilon)})^{k-1}}$ in time also $n^{O(k^3)} \cdot (1/\epsilon)^{O(k^3 \frac{\log(k/\epsilon)}{\log\log(k/\epsilon)})^{k-1}}$. Note that producing an $(\epsilon/6\sqrt{|S|})$-cover, as $= \epsilon/O(k^2 \log(k/\epsilon))^{k/2}$, takes time $n^{O(k^3)} \cdot (1/\epsilon)^{O(k^3 \frac{\log(k/\epsilon)}{\log\log(k/\epsilon)})^{k-1}}$ (which is also the size of the resulting cover). Hence the running time of the algorithm is at most $n^{O(k^3)} \cdot (1/\epsilon)^{O(k^3 \frac{\log(k/\epsilon)}{\log\log(k/\epsilon)})^{k-1}}$.

Alternatively, [DDKT16] gives an ϵ-cover of size $n^{O(k)} \cdot \min 2^{\mathrm{poly}(k/\epsilon)}, 2^{O(k^{5k} \log(1/\epsilon)^{k+2}}$ that can also be constructed in polynomial time. By using this result, one needs to take time $n|S|\mathrm{poly}(\log(1/\epsilon))$ to compute the Fourier coefficients. Applying this to get an $\epsilon/O(k^2 \log(k/\epsilon))^{k/2}$-cover means that unfortunately we are always doubly exponential in k. In this case, the running time of the algorithm is $n^{O(k)} \cdot 2^{O(k^{5k} \log(1/\epsilon)^{k+2}}$.

7 The Discrete Log-Concavity Tester

Theorem 7.1 (Testing Log-Concavity)**.** *Given parameters $n \in \mathbb{N}$, $\epsilon \in (0,1]$, and sample access to a distribution $\mathbf{P}$ over $\mathbb{Z}$, there exists an algorithm which outputs either* accept *or* reject*, and satisfies the following.*

1. *if $\mathbf{P} \in \mathcal{LCV}_n$, then it outputs* accept *with probability at least $3/5$;*
2. *if $d_{\mathrm{TV}}(\mathbf{P}, \mathcal{LCV}_n) > \epsilon$, then it outputs* reject *with probability at least $3/5$.*

where $\mathcal{LCV}_n$ denotes the class of (discrete) log-concave distributions over $\{0, \ldots, n-1\}$. Moreover, the algorithm takes $O(\sqrt{n}/\epsilon^2 + \log(1/\epsilon)/\epsilon^{5/2})$ samples from $\mathbf{P}$; and runs in time $O(\sqrt{n} \cdot \mathrm{poly}(1/\epsilon))$.

We will sketch the proof and algorithm here. We first remark that the Maximum Likelihood Estimator (MLE) for log-concave distributions can be formulated as a convex program [DR11], which can be solved in sample polynomial time. One advantage of the MLE for log-concave distributions is that it properly learns log-concave distributions (over support size M) to within Hellinger distance ϵ using $O(\log(M/\epsilon)/\epsilon^{5/2})$ samples[4]. Note that the squared Hellinger distance satisfies:

$$\mathrm{d_H}\left(\mathbf{P}, \mathbf{Q}\right)^2 = \sum_x (\sqrt{\mathbf{P}(x)} - \sqrt{\mathbf{Q}(x)})^2 = \sum_x \frac{(\mathbf{P}(x) - \mathbf{Q}(x))^2}{(\sqrt{\mathbf{P}} + \sqrt{\mathbf{Q}})^2} \geq \frac{\|\mathbf{P} - \mathbf{Q}\|_2}{2\max\{\mathbf{P}(x), \mathbf{Q}(x)\}} \ .$$

Further, it is known that a log-concave distribution with variance σ^2 is effectively supported in an interval of length $M = O(\log(1/\epsilon)\sigma)$ centered at the mean, and that its maximum probability is $O(1/\sigma)$ (See Fact 7.6). Thus, by learning a log-concave distribution properly to within $\epsilon/\log(1/\epsilon)$ Hellinger distance, one also learns it to within $\frac{\epsilon}{\sqrt{M}}$ L_2-distance.

A log-concave distribution $\mathbf{P}$ has L_2 norm bounded by $\|\mathbf{P}\|_2^2 \leq \max_x \mathbf{P}(x) \leq O(1/\sigma)$. It is easy to show using concentration bounds(Fact 7.6) that $\mathbf{P} \bmod M$ also has L_2 norm $O(1/\sqrt{\sigma})$. We will prove in Proposition 7.2 that its DFT modulo M is effectively supported on a known set S of size $|S| = O(\log(1/\epsilon)^2/\epsilon^2)$.

Thus our algorithm (Algorithm 8) will work as follows: First we estimate the mean and variance under the assumption of log-concavity. We construct an interval I of length $M = O(\log(1/\epsilon)\sigma)$ which would be containing the effective support if we were log-concave; and reject if it is not the case, i.e., too much probability mass falls outside I. Then we properly learn $\mathbf{P}$ to within $\epsilon/\log(1/\epsilon)$ Hellinger distance using the MLE of $O(\log(M/\epsilon)/\epsilon^{5/2})$ samples,[5] giving a hypothesis $\mathbf{H}$. At this point, we reject if our estimates for the variance is far from that of $\mathbf{H}$. Then we run an L_2 identity tester between $\mathbf{P}$ and $\mathbf{H}$, i.e., test whether the empirical distribution $\mathbf{Q}$ of $O(M/\sigma\epsilon^2)$ samples is far in L_2 from $\mathbf{H}$ using Proposition 3.2 . To do this efficiently, we compute $\|\mathbf{Q}'\|_2^2 - \|\widehat{\mathbf{Q}}\mathbf{1}_S\|_2^2/M + \|\widehat{\mathbf{Q}'}\mathbf{1}_S - \widehat{\mathbf{H}}\mathbf{1}_S\|_2^2/M$ which is close to $\|\mathbf{Q}' - \mathbf{H}\|_2^2$ since $\widehat{\mathbf{H}}$ is effectively supported on S as it is a log-concave distribution whose stamdard deviation is at least half of our estimate.

[4] We note that a similar, slightly stronger result is already known for *continuous* log-concave distributions, which can be learned to Hellinger distance ϵ from only $O(\epsilon^{-5/2})$ samples [KS16]. The proof of this result, however, does not seem to generalize to discrete log-concave distributions, which is our focus here; thus, we establish in Appendix B the learning result we require, namely an upper bound on the sample complexity of the MLE estimator for learning the class of log-concave distributions over $\{0, \ldots, M-1\}$ in Hellinger distance (Theorem B.1).

[5] Note that we here invoke the MLE estimator not on the full domain, but on the effective support, which contains at least $1 - O(\epsilon^2)$ probability mass. This conditioning overall does not affect the sample complexity nor the distances, as it can only cause $O(\epsilon^2)$ error in total variation (and thus $O(\epsilon)$ in Hellinger distance).

Algorithm 8 Algorithm `Test-log-concave`

Require: sample access to a distribution $\mathbf{P} \in \Delta(\mathbb{N})$, parameter $\epsilon \in (0, 1]$

▷ Let C, C', C'' be sufficiently large universal constants

Draw $O(1)$ samples from $\mathbf{P}$ and compute their mean $\widetilde{\mu}$ and let $\widetilde{\sigma}$ be 1 plus their standard deviation.

Set $M \leftarrow 1 + 2\lceil C\widetilde{\sigma}\ln(1/\epsilon)\rceil$, and let $I \leftarrow [\lfloor\widetilde{\mu}\rfloor - \frac{M-1}{2}, \lfloor\widetilde{\mu}\rfloor + \frac{M-1}{2}]$

Draw $O(1/\epsilon^2)$ samples from $\mathbf{P}$, to distinguish between $\mathbf{P}(I) \le 1 - \frac{\epsilon^2}{4}$ and $\mathbf{P}(I) > 1 - \frac{\epsilon^2}{5}$. If the former is detected, **return** reject

Draw $O(\log(M/\epsilon)/\epsilon^{5/2})$ samples from $\mathbf{P}$ and let T be the subset of these samples in I. Compute the MLE H over all discrete log-concave distributions for T using a convex program.

Compute the standard deviation $\sigma_{\mathbf{H}}$ of $\mathbf{H}$. If $1 + \sigma_{\mathbf{H}} \le \widetilde{\sigma}/2$ or $\sigma_{\mathbf{H}} \ge 2\widetilde{\sigma}$, then **return** reject.

Set $S \leftarrow \left\{ \xi \in [M-1] : |\xi| \le C' \log(1/\epsilon)^2/\epsilon^2 \right\}$

Let $m = C''/(\epsilon^2\sqrt{\widetilde{\sigma}})$ and draw m' from $\mathrm{Poi}(m)$. Take m' samples from $\mathbf{P}$ and let $\mathbf{Q}'$ be their empirical distibution.

Compute $\widehat{\mathbf{Q}'}(\xi)$ and $\widehat{\mathbf{H}}(\xi)$ for every $\xi \in S$.

if $\|\mathbf{Q}'\|_2^2 - \|\widehat{\mathbf{Q}'}\mathbf{1}_S\|_2^2/M + \|\widehat{\mathbf{Q}'}\mathbf{1}_S - \widehat{\mathbf{H}}\mathbf{1}_S\|_2^2/M > 3m^2\epsilon^2$ **then**

    return reject

else

    return accept

end if

To do this in time $O(\sqrt{n} \cdot \mathrm{poly}(1/\epsilon))$, we need to compute the Fourier coefficients efficiently. The MLE for log-concave distributions is a piecewise exponential distribution with a number of pieces at most the number of samples [DR11], which is $O(\log(M/\epsilon)/\epsilon^{5/2})$ in this case. Using the expression for the sum of a geometric series gives a simple closed-form expression for $\widehat{\mathbf{H}}(\xi)$ that we can compute in time $O(\log(M/\epsilon)/\epsilon^{5/2})$.

Proposition 7.2. *Let $\mathbf{P}$ be a discrete log-concave distribution with variance σ^2 and $M = O(\log(1/\epsilon)\sigma)$ be the size of its effective support. Then its Discrete Fourier transform is effectively supported on a known set S of size $|S| = O(\log(1/\epsilon)^2/\epsilon^2)$.*

Proof. First we show that for any unimodal distribution, we can relate the maximum probability to the size of the effective support.

Lemma 7.3. *Let $\mathbf{P}$ be a unimodal distribution supported on $\mathbb{Z}$ such that the probability of the mode is $\mathbf{P}_{\max}$. Then the DFT modulo M of $\mathbf{P}$ at $\xi \in [-M/2, M/2)$ has $\widehat{\mathbf{P}}(\xi) = O(\mathbf{P}_{\max}M/|\xi|)$.*

Proof. Let m be the mode of $\mathbf{P}$. Then we have

$$\widehat{\mathbf{P}}(\xi) = \sum_{j=-\infty}^{m-1} \mathbf{P}(j)\exp\left(-2\pi i\frac{\xi j}{M}\right) + \sum_{j=m}^{\infty} \mathbf{P}(j)\exp\left(-2\pi i\frac{\xi j}{M}\right).$$

We will apply summation by parts to these two series. Let $g(x) = \sum_{j=m+1}^{x} \exp(-2\pi i\xi j/M)$ and $g(m) = 0$. By a standard result on geometric series, we have $g(x) = -\frac{\exp(-2\pi i\xi(x+1)/M) - \exp(-2\pi i\xi(m+1)/M)}{1-\exp(-2\pi i\xi/M)}$.

Claim 7.4. *$|g(x)| = O(M/\xi)$ for all integers $x \ge m$.*

Proof. The modulus of the numerator $|\exp(-2\pi i\xi(x+1)/M) - \exp(-2\pi i\xi(m+1)/M)|$ is at most 2. We thus only need to find a lower bound for $|1 - \exp(-2\pi i\xi/M|$.

$$|1 - \exp(-2\pi i\xi/M)|^2 = (1 - \cos(2\pi\xi/M))^2 + \sin(2\pi\xi/M)^2 = 2 - 2\cos(2\pi\xi/M) = \Omega((\xi/M)^2) \ ,$$

and so $|g(x)| \leq 2/\sqrt{\Omega((\xi/M)^2)} = O(M/|\xi|)$. □

Now consider the following, for any $n > m$:

$$\sum_{j=m+1}^{n} \mathbf{P}(j)(g(j) - g(j-1)) + \sum_{j=m+1}^{n} g(j)(\mathbf{P}(j+1) - \mathbf{P}(j)) = \mathbf{P}(n+1)g(n) - \mathbf{P}(m+1)g(m) \ .$$

Now $g(m) = 0$ and $\mathbf{P}(n+1) \to 0$ as $n \to \infty$ while $g(n+1)$ is bounded for all n. Hence, the RHS tends to 0 as $n \to \infty$ and we have:

$$\begin{aligned} |\sum_{j=m+1}^{\infty} \mathbf{P}(j)\exp(-2\pi i\xi j/M)| = |\sum_{j=m+1}^{\infty} \mathbf{P}(j)(g(j) - g(j-1))| &= |\sum_{j=m+1}^{\infty} g(j)(\mathbf{P}(j+1) - \mathbf{P}(j))| \\ &\leq O(M/\xi) \cdot \sum_{j=m+1}^{\infty} (\mathbf{P}(j) - \mathbf{P}(j+1)) = O(\mathbf{P}_{\max}M/\xi) \ . \end{aligned}$$

Similarly, we can show that $\sum_{j=-\infty}^{m-1} \mathbf{P}(j)\exp(-2\pi i\xi j/M) = O(\mathbf{P}_{\max}M/\xi)$ since $\mathbf{P}$ is monotone there as well. □

Then we can get a bound on the size of the effective support:

Lemma 7.5. *Let $\mathbf{P}$ be a unimodal distribution supported on $\mathbb{Z}$ such that the probability of the mode is $\mathbf{P}_{\max}$ and let $\epsilon \leq 1/M$. Then the DFT modulo M of $\mathbf{P}$ has $\sum_{|\xi|>\ell} |\widehat{\mathbf{P}}|^2 \leq \epsilon^2/100$, where $\ell = \Theta(\mathbf{P}_{\max}^2 M^2/\epsilon^2)$.*

Proof.

$$\sum_{|\xi|>\ell} |\widehat{P}|^2 \leq 2\sum_{\xi=\ell+1}^{M/2} O(\mathbf{P}_{\max}M/\xi)^2 \leq O(\mathbf{P}_{\max}M)^2 \sum_{\xi=\ell+1}^{\infty} 1/\xi^2 \leq O(\mathbf{P}_{\max}^2 M^2/\ell) \leq \frac{\epsilon^2}{100} \ .$$

□

For log-concave distributions, we can relate $\mathbf{P}_{\max}$ and M as follows,

Fact 7.6. *Let $\mathbf{P}$ be a discrete log-concave distribution with mean μ and variance σ^2. Then*

- *$\mathbf{P}$ is unimodal;*
- *its probability mass function satisfies $\mathbf{P}(x) = \exp(-O((x-\mu)/\sigma))/\sigma$; and*
- $\Pr[|X - \mu| \geq \Omega(\sigma\log(1/\epsilon))] \leq \epsilon$.

Since $\mathbf{P}_{\max} = O(1/\sigma)$, we can take $M = O(\sigma\log(1/\epsilon))) = O(\log(1/\epsilon)/\mathbf{P}_{\max})$. Substituting this into Lemma 7.5 completes the proof of the proposition. □

8 Lower Bound for PMD Testing

In this section, we obtain a lower bound to complement our upper bound for testing Poisson Multinomial Distributions. Namely, we prove the following:

Theorem 8.1. *There exists an absolute constant $c \in (0,1)$ such that the following holds. For any $k \leq n^c$, any testing algorithm for the class of $\mathcal{PMD}_{n,k}$ must have sample complexity $\Omega\left(\left(\frac{4\pi}{k}\right)^{k/4} \frac{n^{(k-1)/4}}{\epsilon^2}\right)$.*

The proof will rely on the lower bound framework of [CDGR17], reducing testing $\mathcal{PMD}_{n,k}$ to testing identity to some suitable hard distribution $\mathbf{P}^* \in \mathcal{PMD}_{n,k}$. To do so, we need to (a) choose a convenient $\mathbf{P}^* \in \mathcal{PMD}_{n,k}$; (b) prove that testing identity to $\mathbf{P}^*$ requires that many samples (we shall do so by invoking the [VV14] instance-by-instance lower bound method); (c) provide an agnostic learning algorithm for $\mathcal{PMD}_{n,k}$ with small enough sample complexity, for the reduction to go through. Invoking [CDGR17, Theorem 18] with these ingredients will then conclude the argument.

Proof of Theorem 8.1. In what follows, we choose our "hard instance" $\mathbf{P}^* \in \mathcal{PMD}_{n,k}$ to be the PMD obtained by summing n i.i.d. random variables, all uniformly distributed on $\{e_1, \dots, e_k\}$. This takes care of point (a) above.

To show (b), we will rely on a result of Valiant and Valiant, which showed in [VV14] that testing identity to any discrete distribution $\mathbf{P}$ required $\Omega\left(\|\mathbf{P}^{-\max}_{-\epsilon}\|_{2/3}/\epsilon^2\right)$ samples, where $\mathbf{P}^{-\max}_{-\epsilon}$ is the vector obtained by zeroing out the largest entry of $\mathbf{P}$, as well as a cumulative ϵ mass of the smallest entries. Since $\|\mathbf{P}^{-\max}_{-\epsilon}\|_{2/3}$ is rather cumbersome to analyze, we shall instead use a slightly looser bound, considering $\|\mathbf{P}\|_2$ as a proxy.

Fact 8.2. *For any discrete distribution $\mathbf{P}$, we have $\|\mathbf{P}\|_{2/3} \geq \frac{1}{\|\mathbf{P}\|_2}$. More generally, for any vector x we have $\|x\|_{2/3} \geq \frac{\|x\|_1^2}{\|x\|_2}$.*

Proof. It is sufficient to prove the second statement, which implies the first. This is in turn a straightforward application of Hölder's inequality, with parameters $(4, \frac{4}{3})$: $\|x\|_1 = \sum_i |x|_i^{1/2} |x|_i^{1/2} \leq \left(\sum_i |x|_i^2\right)^{1/4} \left(|x|_i^{2/3}\right)^{3/4}$. Squaring both sides yields the claim. □

Fact 8.3. *For our distribution $\mathbf{P}^*$, we have $\|\mathbf{P}^*\|_2 = \Theta\left(\frac{k^{k/4}}{(4\pi n)^{(k-1)/4}}\right)$.*

Proof. It is not hard to see that, from any $\mathbf{n} = (n_1, \dots, n_k) \in \mathbb{N}^k$ such that $\sum_{i=1}^k n_i = n$, $\mathbf{P}^*(\mathbf{n}) = \frac{1}{k^n}\binom{n}{n_1,\dots,n_k}$ (where $\binom{n}{n_1,\dots,n_k}$ denotes the multinomial coefficient). From there, we have

$$\|\mathbf{P}^*\|_2^2 = \frac{1}{k^{2n}} \sum_{n_1+\dots+n_k=n} \binom{n}{n_1,\dots,n_k}^2 \underset{n\to\infty}{\sim} \frac{1}{k^{2n}} \cdot k^{2n} \frac{k^{k/2}}{(4\pi n)^{(k-1)/2}}$$

where the equivalent is due to Richmond and Shallit [RS08]. □

However, from Fact 8.2 we want to get a hold on $\|\mathbf{P}^{*-\max}_{-\epsilon}\|_2$, not $\|\mathbf{P}^*\|_2$ (since $\|\mathbf{P}^{*-\max}_{-\epsilon}\|_1^2 \geq 1-\Omega(\epsilon)$, we then will have our lower bound on $\|\mathbf{P}^{*-\max}_{-\epsilon}\|_{2/3}$). Fortunately, the two are related: namely, $\|\mathbf{P}^{*-\max}_{-\epsilon}\|_2 \leq \|\mathbf{P}^*\|_2$, so $\frac{1}{\|\mathbf{P}^{*-\max}_{-\epsilon}\|_2} \geq \frac{1}{\|\mathbf{P}^*\|_2}$ which is the direction we need.

Combining the three facts above establishes (b), providing a lower bound of $q_{\text{hard}}(n,k,\epsilon) = \Omega\left(\frac{(4\pi n)^{(k-1)/4}}{k^{k/4}\epsilon^2}\right)$ for testing identity to $\mathbf{P}^*$. It only remains to establish (c):

Lemma 8.4. *There exists a (not necessarily efficient) agnostic learner for $\mathcal{PMD}_{n,k}$, with sample complexity* $q_{\mathrm{agn}}(n,k,\epsilon) = \frac{1}{\epsilon^2}\left(O(k^2\log n) + O\left(\frac{k\log(k/\epsilon)}{\log\log(k/\epsilon)}\right)^k\right)$.

Proof. This is implied by a result of [DKS16d], which establishes the existence of a (proper) ϵ-cover $\mathcal{M}_{n,k,\epsilon}$ of $\mathcal{PMD}_{n,k}$ such that $|\mathcal{M}_{n,k,\epsilon}| \leq n^{O(k^2)} \cdot (1/\epsilon)^{O\left(\frac{k\log(k/\epsilon)}{\log\log(k/\epsilon)}\right)^{k-1}}$. By standard arguments, this yields information-theoretically an agnostic learner with sample complexity $O\left(\frac{\log|\mathcal{M}_{n,k,\epsilon}|}{\epsilon^2}\right)$. □

Having (a), (b), and (c), an application of [CDGR17, Theorem 18] yields that, as long as $q_{\mathrm{agn}}(n,k,\epsilon) = o(q_{\mathrm{hard}}(n,k,\epsilon))$ then testing membership to $\mathcal{PMD}_{n,k}$ requires $\Omega\left(q_{\mathrm{hard}}(n,k,\epsilon)\right)$ samples as well. This in particular holds for $k = o(n^c)$ (where e.g. $c < 1/9$) and $\epsilon = 1/2^{O(n)}$.

□

References

[AD15] J. Acharya and C. Daskalakis. Testing Poisson Binomial Distributions. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA*, pages 1829–1840, 2015.

[ADK15] J. Acharya, C. Daskalakis, and G. Kamath. Optimal testing for properties of distributions. In *Proceedings of NIPS'15*, 2015.

[An95] M. Y. An. Log-concave probability distributions: Theory and statistical testing. Technical Report Economics Working Paper Archive at WUSTL, Washington University at St. Louis, 1995.

[Bar88] A. D. Barbour. Stein's Method and Poisson Process Convergence. *Journal of Applied Probability*, 25:pp. 175–184, 1988.

[BCI+08] C. Borgs, J. T. Chayes, N. Immorlica, A. T. Kalai, V. S. Mirrokni, and C. H. Papadimitriou. The myth of the folk theorem. In *STOC*, pages 365–372, 2008.

[BDS12] A. Bhaskara, D. Desai, and S. Srinivasan. Optimal hitting sets for combinatorial shapes. In *15th International Workshop, APPROX 2012, and 16th International Workshop, RANDOM 2012*, pages 423–434, 2012.

[Ben03] V. Bentkus. On the dependence of the Berry-Esseen bound on dimension. *Journal of Statistical Planning and Inference*, 113:385–402, 2003.

[BFR+00] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing that distributions are close. In *IEEE Symposium on Foundations of Computer Science*, pages 259–269, 2000.

[BHJ92] A.D. Barbour, L. Holst, and S. Janson. *Poisson Approximation*. Oxford University Press, New York, NY, 1992.

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013.

[Can15] C. L. Canonne. A survey on distribution testing: Your data is big. but is it blue? *Electronic Colloquium on Computational Complexity (ECCC)*, 22:63, 2015.

[CDGR17] C. L. Canonne, I. Diakonikolas, T. Gouleakis, and R. Rubinfeld. Testing shape restrictions of discrete distributions. *Theory of Computing Systems*, 2017.

[CDS17] Y. Cheng, I. Diakonikolas, and A. Stewart. Playing anonymous games using simple strategies. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, Proceedings of SODA '17, pages 616–631, Philadelphia, PA, USA, 2017. Society for Industrial and Applied Mathematics.

[CDVV14] S. Chan, I. Diakonikolas, P. Valiant, and G. Valiant. Optimal algorithms for testing closeness of discrete distributions. In *SODA*, pages 1193–1203, 2014.

[CGS11] L. Chen, L. Goldstein, and Q.-M. Shao. *Normal Approximation by Stein's Method*. Springer, 2011.

[Che52] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statist.*, 23:493–507, 1952.

[CL97] S.X. Chen and J.S. Liu. Statistical applications of the Poisson-Binomial and Conditional Bernoulli Distributions. *Statistica Sinica*, 7:875–892, 1997.

[CL10] L. H. Y. Chen and Y. K. Leong. From zero-bias to discretized normal approximation. 2010.

[DDKT16] C. Daskalakis, A. De, G. Kamath, and C. Tzamos. A size-free CLT for poisson multinomials and its applications. In *Proceedings of STOC'16*, 2016.

[DDO+13] C. Daskalakis, I. Diakonikolas, R. O'Donnell, R.A. Servedio, and L. Tan. Learning Sums of Independent Integer Random Variables. In *FOCS*, pages 217–226, 2013.

[DDS12] C. Daskalakis, I. Diakonikolas, and R.A. Servedio. Learning Poisson Binomial Distributions. In *STOC*, pages 709–728, 2012.

[DDS15] C. Daskalakis, I. Diakonikolas, and R. A. Servedio. Learning Poisson Binomial Distributions. *Algorithmica*, 72(1):316–357, 2015.

[De15] A. De. Beyond the central limit theorem: asymptotic expansions and pseudorandomness for combinatorial sums. In *FOCS*, 2015.

[DK16] I. Diakonikolas and D. M. Kane. A new approach for testing properties of discrete distributions. In *FOCS*, pages 685–694, 2016. Full version available at abs/1601.05557.

[DKN15] I. Diakonikolas, D. M. Kane, and V. Nikishkin. Testing Identity of Structured Distributions. In *Proceedings of SODA'15*, 2015.

[DKS16a] I. Diakonikolas, D. M. Kane, and A. Stewart. Efficient robust proper learning of log-concave distributions. *CoRR*, abs/1606.03077, 2016.

[DKS16b] I. Diakonikolas, D. M. Kane, and A. Stewart. Optimal Learning via the Fourier Transform for Sums of Independent Integer Random Variables. In *COLT*, volume 49 of *JMLR Workshop and Conference Proceedings*, pages 831–849. JMLR.org, 2016. Full version available at arXiv:1505.00662.

[DKS16c] I. Diakonikolas, D. M. Kane, and A. Stewart. Properly learning Poisson binomial distributions in almost polynomial time. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016*, pages 850–878, 2016. Full version available at arXiv:1511.04066.

[DKS16d] I. Diakonikolas, D. M. Kane, and A. Stewart. The Fourier Transform of Poisson Multinomial Distributions and its Algorithmic Applications. In *Proceedings of STOC'16*, 2016. Full version available at arXiv:1511.03592.

[DKT15] C. Daskalakis, G. Kamath, and C. Tzamos. On the structure, covering, and learning of poisson multinomial distributions. In *FOCS*, 2015.

[DP07] C. Daskalakis and C. H. Papadimitriou. Computing equilibria in anonymous games. In *FOCS*, pages 83–93, 2007.

[DP08] C. Daskalakis and C. H. Papadimitriou. Discretized multinomial distributions and nash equilibria in anonymous games. In *FOCS*, pages 25–34, 2008.

[DP09a] C. Daskalakis and C. Papadimitriou. On Oblivious PTAS's for Nash Equilibrium. In *STOC*, pages 75–84, 2009.

[DP09b] D. Dubhashi and A. Panconesi. *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press, Cambridge, 2009.

[DP14] C. Daskalakis and C. H. Papadimitriou. Approximate Nash equilibria in anonymous games. *Journal of Economic Theory*, 2014.

[DR11] L. Dumbgen and K. Rufibach. logcondens: Computations related to univariate log-concave density estimation. *J. Statist. Software*, 39(6), 2011.

[GKM15] P. Gopalan, D. M. Kane, and R. Meka. Pseudorandomness via the discrete fourier transform. In *FOCS*, 2015.

[GMRZ11] P. Gopalan, R. Meka, O. Reingold, and D. Zuckerman. Pseudorandom generators for combinatorial shapes. In *STOC*, pages 253–262, 2011.

[GT14] P. W. Goldberg and S. Turchetta. Query complexity of approximate equilibria in anonymous games. *CoRR*, abs/1412.6455, 2014.

[Hoe63] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[Kru86] J. Kruopis. Precision of approximation of the generalized binomial distribution by convolutions of poisson measures. *Lithuanian Mathematical Journal*, 26(1):37–49, 1986.

[KS16] A. K. H. Kim and R. J. Samworth. Global rates of convergence in log-concave density estimation. *Ann. Statist.*, 44(6):2756–2779, 12 2016. Available at http://arxiv.org/abs/1404.2298.

[Loh92] W. Loh. Stein's method and multinomial approximation. *Ann. Appl. Probab.*, 2(3):536–554, 08 1992.

[LR05] E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, 2005.

[LV07] L. Lovász and S. Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures and Algorithms*, 30(3):307–358, 2007.

[NP33] J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.

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

[Poi37] S.D. Poisson. *Recherches sur la Probabilitè des jugements en matié criminelle et en matiére civile*. Bachelier, Paris, 1837.

[Pre83] E. L. Presman. Approximation of binomial distributions by infinitely divisible ones. *Theory Probab. Appl.*, 28:393–403, 1983.

[Roo99] B. Roos. On the Rate of Multivariate Poisson Convergence. *Journal of Multivariate Analysis*, 69(1):120 – 134, 1999.

[Roo10] B. Roos. Closeness of convolutions of probability measures. *Bernoulli*, 16(1):23–50, 2010.

[RS08] L. B. Richmond and J. Shallit. Counting Abelian Squares. *ArXiv e-prints*, July 2008.

[Rub12] R. Rubinfeld. Taming big probability distributions. *XRDS*, 19(1):24–28, 2012.

[Sta89] Richard P. Stanley. Log-concave and unimodal sequences in algebra, combinatorics, and geometry. *Annals of the New York Academy of Sciences*, 576(1):500–535, 1989.

[SW14] A. Saumard and J. A. Wellner. Log-concavity and strong log-concavity: A review. *Statist. Surv.*, 8:45–114, 2014.

[Val08] P. Valiant. Testing symmetric properties of distributions. In *STOC*, pages 383–392, 2008.

[vdG00] S. A. van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge University Press, 2000.

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

[VV11] G. Valiant and P. Valiant. Estimating the unseen: an $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs. In *STOC*, pages 685–694, 2011.

[VV14] G. Valiant and P. Valiant. An automatic inequality prover and instance optimal identity testing. In *FOCS*, 2014.

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

[WS95] W. H. Wong and X. Shen. Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist.*, 23(2):339–362, 1995.

A Omitted Proofs

In this appendix, we provide the proofs of the lemmas and technical results omitted in the main body.

A.1 From Section 2

Proof of Lemma 2.8. By Plancherel, we have $\|\mathbf{P}'\|_2^2 = \frac{1}{M}\sum_{\xi=0}^{M-1}|\widehat{\mathbf{P}'}(\xi)|^2 = \frac{1}{M}\sum_{\xi=0}^{M-1}|\widehat{\mathbf{P}}(\xi)|^2$, the second equality due to the definition of $\widehat{\mathbf{P}'}$. Indeed, for any $\xi \in [M]$,

$$\widehat{\mathbf{P}'}(\xi) = \sum_{j=0}^{M-1} e^{-2i\pi\frac{j\xi}{M}}\mathbf{P}'(j) = \sum_{j=0}^{M-1} e^{-2i\pi\frac{j\xi}{M}} \sum_{\substack{j'\in\mathbb{N}\\ j'=j \bmod M}} \mathbf{P}(j') = \sum_{j=0}^{M-1}\sum_{\substack{j'\in\mathbb{N}\\ j'=j \bmod M}} e^{-2i\pi\frac{j'\xi}{M}}\mathbf{P}(j')$$
$$= \sum_{j\in\mathbb{N}} e^{-2i\pi\frac{j'\xi}{M}}\mathbf{P}(j') = \widehat{\mathbf{P}}(\xi)$$

as $u \mapsto e^{-2i\pi u}$ is 1-periodic. Since $\left|\widehat{\mathbf{P}}(\xi)\right| \le 1$ for every $\xi \in [M]$ (as $\widehat{\mathbf{P}}(\xi) = \mathbb{E}_{j\sim\mathbf{P}}[e^{-2i\pi\frac{j\xi}{M}}]$), we can upper bound the RHS as

$$\frac{1}{M}\sum_{\xi=0}^{M-1}|\widehat{\mathbf{P}}(\xi)|^2 \le \frac{1}{M}\sum_{r\ge 0}\sum_{\xi:\frac{1}{2^{r+1}}<\left|\widehat{\mathbf{P}}(\xi)\right|\le\frac{1}{2^r}}\left|\widehat{\mathbf{P}}(\xi)\right|^2 \le \frac{1}{M}\sum_{r\ge 0}\frac{1}{2^{2r}}\left|\left\{\,\xi\in[M]:\ \frac{1}{2^{r+1}}<\left|\widehat{\mathbf{P}}(\xi)\right|\,\right\}\right|.$$

Invoking Lemma 2.7(ii) with parameter δ set to $\frac{1}{2^{r+1}}$, we get that $\left|\left\{\,\xi\in[M]:\ \frac{1}{2^{r+1}}<\left|\widehat{\mathbf{P}}(\xi)\right|\,\right\}\right| \le 4Mks^{-1}\sqrt{r+1}$, from which

$$\|\mathbf{P}'\|_2^2 \le \frac{4k}{s}\sum_{r\ge 0}\frac{\sqrt{r+1}}{2^{2r}} \le \frac{8k}{s}$$

as desired. □

A.2 From Section 3

Proof of Proposition 3.2. Letting X_i denote the number of occurrences of the i-th domain element in the samples from $\mathbf{P}$, define $Z_i = (X_i - m\mathbf{P}^*(i))^2 - X_i$. Since X_i is distributed as $\mathrm{Poi}(m\cdot p_i)$, $\mathbb{E}[Z_i] = m^2(\mathbf{P}(i)-\mathbf{P}^*(i))^2$; thus, Z is an unbiased estimator for $m^2\|\mathbf{P}-\mathbf{P}^*\|_2^2$. (Note that this holds even when $\mathbf{P}^*$ is allowed to take negative values.)

We compute the variance of Z_i via a straightforward calculation involving standard expressions for the moments of a Poisson distribution: getting

$$\mathrm{Var}[Z] = \sum_{i=1}^{r}\mathrm{Var}[Z_i] = \sum_{i=1}^{r}\left(4m^3(\mathbf{P}(i)-\mathbf{P}^*(i))^2\mathbf{P}(i) + 2m^2\mathbf{P}(i)^2\right).$$

By Cauchy–Schwarz, and since $\sum_{i=1}^{r} \mathbf{P}(i)^2 \le b$ by assumption, we have

$$\begin{aligned}\sum_{i=1}^{r}(\mathbf{P}(i)-\mathbf{P}^*(i))^2\mathbf{P}(i) &= \sum_{i=1}^{r}(\mathbf{P}(i)-\mathbf{P}^*(i))\cdot(\mathbf{P}(i)-\mathbf{P}^*(i))\mathbf{P}(i)\\ &\le \sqrt{\sum_{i=1}^{r}(\mathbf{P}(i)-\mathbf{P}^*(i))^2\sum_{i=1}^{r}\mathbf{P}(i)^2(\mathbf{P}(i)-\mathbf{P}^*(i))^2}\\ &\le \sqrt{\sum_{i=1}^{r}(\mathbf{P}(i)-\mathbf{P}^*(i))^2 b\sum_{i=1}^{r}(\mathbf{P}(i)-\mathbf{P}^*(i))^2} = \sqrt{b}\|\mathbf{P}-\mathbf{P}^*\|_2^2\end{aligned}$$

and so

$$\mathrm{Var}[Z] \le 4m^3\sqrt{b}\|\mathbf{P}-\mathbf{P}^*\|_2^2 + 2m^2 b.$$

For convenience, let $\eta \stackrel{\text{def}}{=} \frac{1}{10}$, and write $\rho \stackrel{\text{def}}{=} \frac{\|\mathbf{P}-\mathbf{P}^*\|_2}{\epsilon}$ – so that we need to distinguish $\rho \le 1$ from $\rho \ge 2$. If $\rho \le 1$, i.e. $\mathbb{E}[Z] \le m^2\epsilon^2$, then

$$\Pr[Z > (3-\eta)m^2\epsilon^2] = \Pr[|Z-\mathbb{E}[Z]| > m^2\epsilon^2(((3-\eta)-\gamma)-\rho^2)]$$

while if $\rho \ge 2$, i.e. $\mathbb{E}[Z] \ge 4m^2\epsilon^2$, then

$$\Pr[Z < (3+\eta)m^2\epsilon^2] = \Pr[\mathbb{E}[Z]-Z > m^2(\|p-q\|_2^2-(3+\eta)\epsilon^2)] \le \Pr[|Z-\mathbb{E}[Z]| > m^2\epsilon^2(\rho^2-(3+\eta))].$$

In both cases, by Chebyshev's inequality, the test will be correct with probability at least $3/4$ provided $m \ge c\sqrt{b}/\epsilon^2$ for some suitable choice of $c > 0$, since (where

$$\begin{aligned}\Pr[|Z-\mathbb{E}[Z]| > m^2\epsilon^2|\rho^2-(3\pm\eta)|] &\le \frac{\mathrm{Var}[Z]}{m^4\epsilon^4(\rho^2-(3\pm\eta))^2}\\ &\le \frac{4m^3\sqrt{b}\rho^2\epsilon^2+2m^2b}{m^4\epsilon^4(\rho^2-(3\pm\eta))^2} = \frac{\rho^2}{(\rho^2-(3\pm\eta))^2}\cdot\frac{4\sqrt{b}}{m\epsilon^2}+\frac{1}{(\rho^2-(3\pm\eta))^2}\cdot\frac{2b}{m^2\epsilon^4}\\ &\le \frac{20\sqrt{b}}{m\epsilon^2}+\frac{5b}{2m^2\epsilon^4} \le \frac{20}{c}+\frac{5}{2c^2} \le \frac{1}{3}\end{aligned}$$

as $\max_{\rho\in[0,1]}\frac{\rho^2}{(\rho^2-(3\pm\eta))^2} \le 5$ and $\max_{\rho\in[0,1]}\frac{1}{(\rho^2-(3\pm\eta))^2} \le \frac{5}{4}$ and the last inequality holds for $c \ge 61$. □

A.3 From Section 4.3

Proof of Lemma 4.10. By Theorem 3.7 of [DKS16b], there is an algorithm that can compute an ϵ-cover of all (n,k)-SIIRVs of size $n\,(k/\epsilon)^{O(k\log(1/\epsilon))}$ that runs in time $n\,(k/\epsilon)^{O(k\log(1/\epsilon))}$. Note the way the cover is given, allows us to compute the Fourier coefficients $\widehat{\mathbf{Q}}(\xi)$ for any ξ for each $\mathbf{Q} \in \mathcal{C}$ in time $\mathrm{poly}(k/\epsilon)$.

Since $\epsilon/\sqrt{|S|} = 1/\mathrm{poly}(k/\epsilon)$, Step 1 takes time $n\,(k/\epsilon)^{O(k\log(k/\epsilon))}$ and outputs a cover of size $n\,(k/\epsilon)^{O(k\log(k/\epsilon))}$. As each iteration takes time $|S|$, the whole algorithm takes $n\,(k/\epsilon)^{O(k\log(k/\epsilon))}$ time.

Note that each $\mathbf{Q}$ that passes Step 3 is effectively supported on I by (3) and has Fourier transform supported on S by Claim 4.5.

- Suppose that $\mathbf{P} \in \mathcal{SIIRV}_{n,k}$. Then there is a (n,k)-SIIRV $\mathbf{Q} \in \mathcal{C}$ with $d_{\mathrm{TV}}(\mathbf{P},\mathbf{Q}) \le \epsilon/5\sqrt{|S|}$. We need to show that if the algorithm considers $\mathbf{Q}$, it accepts. From standard concentration bounds,

one gets that the expectations of $\mathbf{P}$ and $\mathbf{Q}$ are within $O(\epsilon\sqrt{\log(1/\epsilon)})$ standard deviations of $\mathbf{P}$ and the variances of $\mathbf{P}$ and $\mathbf{Q}$ are within $O(\epsilon \log(1/\epsilon))$ multiplicative error. Thus $\mathbf{Q}$ passes the condition of Step 3. Since $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) \leq \epsilon/(5\sqrt{|S|})$, we have that $|\widehat{\mathbf{P}}(\xi) - \widehat{\mathbf{Q}}(\xi)| \leq \epsilon/(5\sqrt{|S|})$ for all ξ. In particular, we have $\sum_{\xi \in S} |\widehat{\mathbf{H}} - \widehat{\mathbf{Q}}|^2 \leq \epsilon^2/25$. Thus by the triangle inequality for L_2 norm, we have $\sum_{\xi \in S} |\widehat{\mathbf{H}} - \widehat{\mathbf{Q}}|^2 \leq (\epsilon/5 + 3\epsilon/25)^2 \leq (\epsilon/\sqrt{5})^2$. Thus the algorithm accepts.

- Now suppose that the algorithm accepts. We need to show that $\mathbf{P}$ has total variation distance at most ϵ from some (n,k)-SIIRV. We will show that $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) \leq \epsilon$ for the $\mathbf{Q}$ which causes the algorithm to accept. Since the algorithm accepts, $\sum_{\xi \in S} |\widehat{\mathbf{H}} - \widehat{\mathbf{Q}}|^2 \leq \epsilon^2/25$. For $x \notin S$, $\widehat{\mathbf{H}}(\xi) = 0$ and so $\sum_{\xi \notin S} |\widehat{\mathbf{H}} - \widehat{\mathbf{Q}}|^2 = \sum_{\xi \notin S} |\widehat{\mathbf{Q}}|^2 \leq \epsilon^2/100$ by Claim 4.5. By Plancherel, the distributions $\mathbf{Q}' \stackrel{\text{def}}{=} \mathbf{Q} \bmod M$, $\mathbf{H}' \stackrel{\text{def}}{=} \mathbf{H} \bmod M$ satisfy

$$\|\mathbf{Q}' - \mathbf{H}'\|_2^2 = \frac{1}{M} \sum_{\xi=0}^{M-1} |\widehat{\mathbf{H}} - \widehat{\mathbf{Q}}|^2 \leq \frac{\epsilon^2}{20M}.$$

Thus $d_{\mathrm{TV}}(\mathbf{Q}', \mathbf{H}') \leq \frac{\epsilon}{4}$. By definition $\mathbf{H}$ has probability 0 outside I and by (3), $\mathbf{Q}$ has at most $\frac{\epsilon}{5}$ probability outside I, Thus $d_{\mathrm{TV}}(\mathbf{Q}, \mathbf{H}) \leq \frac{\epsilon}{4} + \frac{\epsilon}{5} \leq \frac{\epsilon}{2}$ and by the triangle inequality $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) \leq d_{\mathrm{TV}}(\mathbf{Q}, \mathbf{H}) + d_{\mathrm{TV}}(\mathbf{P}, \mathbf{H}) \leq \epsilon/2 + 6\epsilon/25 \leq \epsilon$ as required.

□

Proof of Lemma 4.11. We use Steps 4 and 5 of Algorithm `Proper-Learn-PBD` in [DKS16c]. Step 5 checks if one of a system of polynomials has a solution. If such a solution is found, it corresponds to an $(n,2)$-SIIRV $\mathbf{Q}$ that has $\sum_{|\xi| \leq \ell} |\widehat{\mathbf{H}}(\xi) - \widehat{\mathbf{Q}}(\xi)|^2 \leq \epsilon^2/4$ and so we accept. If no systems have a solution, then there is no such $(n,2)$-SIIRV and so we reject. The conditions of this lemma are enough to satisfy the conditions of Theorem 11 of [DKS16c], though we need that the constant C' used to define $|S|$ is sufficiently large to cover the $\ell = O(\log(1/\epsilon)$ from that paper. This theorem means that if $\mathbf{P}$ is a $(n,2)$-SIIRV, then we accept.

We need to show that if the algorithm finds a solution, then it is within ϵ of a Poisson Binomial distribution. The system of equations ensures that $\sum_{|\xi| \leq \ell} |\widehat{\mathbf{H}}(\xi) - \widehat{\mathbf{Q}}(\xi)|^2 \leq \epsilon^2/4$. Now the argument is similar to that for (n,k)-SIIRVs. For $x \notin S$, $\widehat{\mathbf{H}}(\xi) = 0$ and so $\sum_{\xi \notin S} |\widehat{\mathbf{H}} - \widehat{\mathbf{Q}}|^2 = \sum_{\xi \notin S} |\widehat{\mathbf{Q}}|^2 \leq \epsilon^2/100$ by Claim 4.5. By Plancherel, the distributions $\mathbf{Q}' \stackrel{\text{def}}{=} \mathbf{Q} \bmod M$, $\mathbf{H}' \stackrel{\text{def}}{=} \mathbf{H} \bmod M$ satisfy

$$\|\mathbf{Q}' - \mathbf{H}'\|_2^2 = \frac{1}{M} \sum_{\xi=0}^{M-1} |\widehat{\mathbf{H}} - \widehat{\mathbf{Q}}|^2 \leq \frac{\epsilon^2}{20M}.$$

Thus $d_{\mathrm{TV}}(\mathbf{Q}', \mathbf{H}') \leq \frac{\epsilon}{4}$. By definition $\mathbf{H}$ has probability 0 outside I and by (3), $\mathbf{Q}$ has at most $\frac{\epsilon}{5}$ probability outside I, Thus $d_{\mathrm{TV}}(\mathbf{Q}, \mathbf{H}) \leq \frac{\epsilon}{4} + \frac{\epsilon}{5} \leq \frac{\epsilon}{2}$ and by the triangle inequality $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) \leq d_{\mathrm{TV}}(\mathbf{Q}, \mathbf{H}) + d_{\mathrm{TV}}(\mathbf{P}, \mathbf{H}) \leq \epsilon/2 + 6\epsilon/25 \leq \epsilon$ as required. □

B Learning Discrete Log-Concave Distributions in Hellinger Distance

Recall that the Hellinger distance between two probability distributions over a domain $\mathbb{D}$ is defined as

$$\mathrm{d_H}(p, q) \stackrel{\text{def}}{=} \frac{1}{\sqrt{2}} \|\sqrt{p} - \sqrt{q}\|_2$$

where the 2- norm is to be interpreted as either the ℓ_2 distance or L^2 distance between the pmf or pdf's of p, q, depending on whether $\mathbb{D}$ is $\mathbb{Z}$ or $\mathbb{R}$. In particular, one can extend this metric to the set of *pseudo*-distributions over $\mathbb{D}$, relaxing the requirement that the measures sum to one. We let $\mathcal{F}_{\mathbb{D}}$ denote the set of pseudo-distributions over $\mathbb{D}$. The *bracketing entropy* of a family of functions $\mathcal{G} \subseteq \mathbb{R}^{\mathbb{D}}$ with respect to the Hellinger distance (for parameter ϵ) if then the minimum cardinality of a collection $\mathcal{C}$ of pairs $(g_L, g_U) \in \mathcal{F}_{\mathbb{D}}^2$ such that every $f \in \mathcal{G}$ is "bracketed" between the elements of some pair in $\mathcal{C}$:

$$\mathcal{N}_{[]}(\epsilon, \mathcal{G}, \mathrm{d_H}) \stackrel{\text{def}}{=} \min\left\{ N \in \mathbb{N} : \ \exists \mathcal{C} \subseteq \mathcal{F}_{\mathbb{D}}^2,\ |\mathcal{C}| = N,\ \forall f \in \mathcal{G}, \exists (g_L, g_U) \in \mathcal{C} \text{ s.t. } g_L \leq f \leq g_U \text{ and } \mathrm{d_H}\left(g_L, g_H\right) \leq \epsilon \right\}$$

Theorem B.1. *Let $\hat{p}_m$ denote the maximum likelihood estimator (MLE) for discrete log-concave distributions on a sample of size m. Then, the minimax supremum risk satisfies*

$$\sup_{p \in \mathcal{LCV}_n} \mathbb{E}_p[\mathrm{d_H}\left(\hat{p}_m, p\right)^2] = O\left(\frac{\log^{4/5}(mn)}{m^{4/5}}\right).$$

Note that it is known that for *continuous* log-concave distributions over $\mathbb{R}$, the rate of the MLE is $O(m^{-4/5})$ [KS16]; this result, however, does not generalize to discrete log-concavity, as it crucially relies on a scaling argument which does not work in the discrete case. On the other hand, one can derive a rate of convergence to learn discrete log-concave distributions in *total variation distance* (using another estimator than the MLE), getting again $O(m^{-4/5})$ in that case [DKS16a]. However, due to the loose upper bound relating total variation and Hellinger distance, this latter result only implies an $O(m^{-2/5})$ convergence rate in Hellinger distance, which is quadratically worse than what we would hope for.

Thus, the result above, while involving a logarithmic dependence on the support size, has the advantage of getting the "right" rate of convergence. (While this additional dependence does not matter for our purposes, we believe a modification of our techniques would allow one to get rid of it, obtaining a rate of $\tilde{O}\left(m^{-4/5}\right)$ instead.) We however conjecture that the tight rate of convergence should be $O(m^{-4/5})$, as in the continuous case (i.e., without the dependence on the domain size n nor the extra logarithmic factors in m).

In order to prove Theorem B.1, we obtain along the way several interesting results on discrete (and continuous) log-concave distributions, namely a bound on their bracketing entropy (Theorem B.2) and an approximation result (Theorem B.3), which we believe are of independent interest.

In what follows, $\mathbb{D}$ will denote either $\mathbb{R}$ or $\mathbb{Z}$; we let $\mathcal{LCV}(\mathbb{D})$ denote the set of log-concave distributions over $\mathbb{D}$, and $\mathcal{LCV}_n \subseteq \mathcal{LCV}(\mathbb{Z})$ be the subset of log-concave distributions supported on $\{0, \ldots, n-1\}$.

Theorem B.2. *For every $\epsilon \in (0, 1)$,*

$$\mathcal{N}_{[]}(\epsilon, \mathcal{LCV}_n, \mathrm{d_H}) \leq \left(\frac{n}{\epsilon}\right)^{O(1/\sqrt{\epsilon})}$$

A crucial element in to establish Theorem B.2 will be the following theorem, which shows that log-concave distributions are well-approximated (in Hellinger distance) by piecewise-constant pseudo-distributions with few pieces:

Theorem B.3. *Let $\mathbb{D}$ be either $\mathbb{R}$ or $\mathbb{Z}$. For every $p \in \mathcal{LCV}(\mathbb{D})$ and $\epsilon \in (0, 1)$, there exists a pseudo-distribution g such that (i) g is piecewise-linear with $O\left(1/\sqrt{\epsilon}\right)$ pieces; (ii) g is supported on an interval $[a, b]$ with $p(\mathbb{D} \setminus [a, b]) = O(\epsilon^2)$; and (iii) $\mathrm{d_H}\left(p, g\right) \leq \epsilon$. (Moreover, one can choose to enforce $g \leq p$, or $p \leq g$, on $[a, b]$).*

The proof of Theorem B.3 will be very similar to that of [DKS16a, Theorem 12]; specifically, we will use the following (reformulation of a) lemma due to Diakonikolas, Kane, and Stewart:

Lemma B.4 ([DKS16a, Lemma 14], rephrased)**.** *Let $\mathbb{D}$ be either $\mathbb{R}$ or $\mathbb{Z}$. Let f be a log-concave function defined on an interval $I \subseteq \mathbb{D}$, and suppose that $f(I) \subseteq [a, 2a]$ for some constant $a > 0$. Furthermore, suppose that the logarithmic derivative of f (or, if $\mathbb{D} = \mathbb{Z}$, the log-finite difference of f) varies by at most $1/|I|$ on I; then, for any $\epsilon \in (0, 1)$ there exists two piecewise linear functions $g^\ell, g^u \colon I \mapsto \mathbb{R}$ with $O(1/\sqrt{\epsilon})$ pieces such that*

$$\left|f(x) - g^j(x)\right| = O(\epsilon) f(x), \qquad j \in \{\ell, u\} \tag{6}$$

for all $x \in I$, and with $g^\ell \leq f \leq g^u$.

Proof. Observe that it suffices to establish Eq. (6) for a piecewise linear function $g \colon I \mapsto \mathbb{R}$ with $O(1/\sqrt{\epsilon})$ pieces; indeed, then in order to obtain g^ℓ, g^u from g, it will be sufficient to scale it by respectively $(1+\alpha\epsilon)^{-1}$ and $(1 + \alpha\epsilon)$ (for a suitably big absolute constant $\alpha > 0$), thus ensuring both Eq. (6) and $g^\ell \leq f \leq g^u$. We therefore focus hereafter on obtaining such a pseudo-distribution g.

For ease of notation, we write h for the logarithmic derivative (or log-finite difference) of f (e.g., in the continuous case, $h = (\ln f)'$). By rescaling f, we may assume without loss of generality that $a = 1$. Note that h is then bounded on I, i.e. $|h| \leq c/|I|$ for some absolute constant $c > 0$. We now partition I into subintervals $J_1, J_2, \ldots, J_\ell$ so that (i) each J_i has length at most $\epsilon^{1/2}|I|$, and (ii) h varies by at most $\epsilon^{1/2}/|I|$ on each J_i. This can be achieved with $\ell = O(1/\sqrt{\epsilon})$ by placing an interval boundary every $\epsilon^{1/2}|I|$ distance as well as every time h passes a multiple of $\epsilon^{1/2}/|I|$.

We now claim that on each interval J_i there exists a linear function g_i so that $|g_i(x) - f(x)| = O(\epsilon)f(x)$ for all $x \in J_i$. Letting g be g_i on J_i will complete the proof. Fix any i, and write $J_i = [s_i, t_i]$. Letting $\alpha_0 \in h(J_i)$ be an arbitrary value in the range spanned by h on J_i, observe that for any $x \in J_i$ there exists $\alpha_x \in h(J_i)$ such that

$$f(x) = f(s_i)e^{\alpha_x(x-s_i)}$$

from which we have

$$\begin{aligned}
f(x) &= f(s_i)e^{\alpha_0(x-s_i)+(\alpha_x-\alpha_0)(x-s_i)} = f(s_i)e^{\alpha_0(x-s_i)}e^{(\alpha_x-\alpha_0)(x-s_i)} \\
&= f(s_i)\left(1 + \alpha_0(x-s_i) + O(\epsilon)\right)\left(1 + O((\alpha_x-\alpha_0)(x-s_i))\right) \\
&= f(s_i)\left(1 + \alpha_0(x-s_i) + O(\epsilon)\right)\left(1 + O(\epsilon)\right) \\
&= f(s_i) + \alpha_0 f(s_i)(x-s_i) + O(\epsilon)
\end{aligned}$$

recalling that $|\alpha_0|, |\alpha_x| = O(1/|I|)$, $|x - s_i| \leq \epsilon^{1/2}|I|$, and $|\alpha_x - \alpha_0| \leq \epsilon^{1/2}/|I|$, so that $|\alpha_0(x-s_i)| = O(\epsilon^{1/2})$ and $|(\alpha_x-\alpha_0)(x-s_i)| = O(\epsilon)$. This motivates defining the affine function g_i as

$$g_i(x) \stackrel{\text{def}}{=} f(s_i) + \alpha_0 f(s_i)(x-s_i), \qquad x \in J_i$$

from which

$$\begin{aligned}
\left|\frac{f(x)-g_i(x)}{f(x)}\right| &= \left|1 - \frac{f(s_i)+\alpha_0 f(s_i)(x-s_i)}{f(s_i)e^{\alpha_x(x-s_i)}}\right| = \left|1 - \frac{1+\alpha_0(x-s_i)}{e^{\alpha_x(x-s_i)}}\right| \\
&= \left|1 - \frac{1+\alpha_0(x-s_i)}{1+\alpha_x(x-s_i)+O(\epsilon)}\right| = \left|1 - \left(1+\alpha_0(x-s_i)\right)\left(1-\alpha_x(x-s_i)+O(\epsilon)\right)\right| \\
&= \left|(\alpha_x-\alpha_0)(x-s_i) + O(\epsilon)\right| = O(\epsilon)
\end{aligned}$$

as claimed. This concludes the proof. □

We will also rely on the following proposition, from the same paper:

Proposition B.5 ([DKS16a, Proposition 15])**.** *Let f be a log-concave distribution on $\mathbb{D}$ (as before, either $\mathbb{R}$ or $\mathbb{Z}$). Then there exists a partition of $\mathbb{D}$ into disjoint intervals $I_1, I_2, \ldots$ and a constant $C > 0$ such that*

- *f satisfies the hypotheses of Lemma B.4 on each I_i.*
- *For each m, there are most Cm values of i so that $f(I_i) > 2^{-m}$.*

(Moreover, f is monotone on each I_i.)

We are now ready to prove Theorem B.3:

Proof of Theorem B.3. Fix any $\epsilon \in (0,1)$, and $p \in \mathcal{LCV}(\mathbb{D})$. We divide $\mathbb{D}$ into intervals as described in Proposition B.5. Call these intervals $I_1, I_2, \ldots$ sorted so that $p(I_i)$ is decreasing in i. Therefore, we have that $p(I_m) \leq 2^{-m/C}$.

For $1 \leq m \leq M \stackrel{\text{def}}{=} 2C\log(1/\epsilon)$, let $\epsilon_m \stackrel{\text{def}}{=} \epsilon 2^{m/(3C)}$; we use Lemma B.4 to approximate p in I_m by two piecewise linear functions g^ℓ_m, g^u_m so that (i) g^j_m has at most $O(1/\sqrt{\epsilon_m})$ pieces and (ii) p and g^j_m are, on I_m, within a multiplicative $(1 \pm O(\epsilon_m))$ factor with $g^\ell_m \leq p \leq g^u_m$. For $j \in \{\ell, u\}$, let g^j be the piecewise linear function that is g^j_m on I_m for $1 \leq m \leq M$, and 0 elsewhere. g^j is then piecewise linear on

$$\sum_{m=1}^{M} O(\epsilon_m^{-1/2}) = \sum_{m=1}^{M} O\left(\epsilon^{-1/2} 2^{-\frac{m}{6C}}\right) = O(\epsilon^{-1/2})$$

intervals.

Let I be defined as the smallest interval such that $\bigcup_{m=1}^{M} I_m \subseteq I$. By definition, g is 0 outside of I, and moreover the total mass of p there is

$$\sum_{m=M+1}^{\infty} p(I_m) \leq \sum_{m=M+1}^{\infty} \frac{1}{2^{m/C}} = O\left(2^{-M/C}\right) = O\left(\epsilon^2\right)$$

By replacing g^j by $\max(g^j, 0)$, we may ensure that it is non-negative (while at most doubling the number of pieces without increasing the distance from p). This establishes the first two items of the theorem; we now turn to the third.

The Hellinger distance between p and g^j satisfies, letting $J \stackrel{\text{def}}{=} \bigcup_{m=1}^{M} I_m$,

$$\begin{aligned}
2\mathrm{d_H}\left(p, g^j\right)^2 &= \|\sqrt{p} - \sqrt{g^j}\|_2^2 = \int_{\mathbb{D}} \left(\sqrt{p(x)} - \sqrt{g^j(x)}\right)^2 \mu(dx) \\
&= \int_{\mathbb{D}\setminus J} \left(\sqrt{p(x)} - \sqrt{g^j(x)}\right)^2 \mu(dx) + \int_{J} \left(\sqrt{p(x)} - \sqrt{g^j(x)}\right)^2 \mu(dx) \\
&= \int_{\mathbb{D}\setminus J} p(x)\mu(dx) + \sum_{m=1}^{M} \int_{I_m} p(x) \left(1 - \sqrt{1 \pm O(\epsilon_m)}\right)^2 \mu(dx) \\
&\leq O(\epsilon^2) + \sum_{m=1}^{M} \int_{I_m} p(x) \left(1 - \sqrt{1 \pm O(\epsilon_m)}\right)^2 \mu(dx) \\
&= O(\epsilon^2) + \sum_{m=1}^{M} \int_{I_m} p(x) O(\epsilon_m^2) \mu(dx) = O(\epsilon^2) + \sum_{m=1}^{M} O\left(\epsilon_m^2 p(I_m)\right) \\
&= O(\epsilon^2) + \sum_{m=1}^{M} O\left(\epsilon^2 2^{\frac{2m}{3C}} 2^{\frac{-m}{C}}\right) = O(\epsilon^2) + \sum_{m=1}^{M} O\left(\epsilon^2 2^{\frac{-m}{3C}}\right) \\
&= O(\epsilon^2) + O(\epsilon^2) = O(\epsilon^2)
\end{aligned}$$

establishing the third item. (By dividing ϵ by a sufficiently big absolute constant before applying the above, one gets (i), (ii), and (iii) with $\mathrm{d_H}\left(p, g^j\right) \leq \epsilon$ as desired.) For technical reasons (that we will need in the proof of Theorem B.2), instead of defining $[a, b]$ to be our interval I, we choose $[a, b]$ to be I augmented with up to two of the remaining I_m's (those directly on the left and right of I, defining g_m^ℓ, g_m^u on these two additional pieces as before by Lemma B.4). This does not change the fact that the piecewise linear function obtained on $[a, b]$ has $O(\epsilon^{-1/2})$ pieces (we only added $o(\epsilon^{-1/2})$ pieces), and $p(\mathbb{D}\setminus[a, b]) \leq p(\mathbb{D}\setminus I) = O(\epsilon^2)$. Finally, it is easy to see that this only changes, as per the computation above, the Hellinger distance by $O(\epsilon^2)$ as well. (The advantage of this technicality is that now, the two end intervals in the union constituting $[a, b]$ have each total probability mass $O(\epsilon^2)$ under p, which will come in handy later.) It then only remains to choose g to be either g^ℓ or g^u, depending on whether one wants a lower- or upperbound on f (on $[a, b]$). $\square$

We can finally prove Theorem B.2:

Proof of Theorem B.2. We can slightly strengthen the proof of Theorem B.3 for the case of $\mathcal{LCV}_n$, by imposing some restriction on the form of the 'approximating distributions" g. Namely, for any $\epsilon \in (0, 1)$, fix any $p \in \mathcal{LCV}_n$ and consider the construction of g^ℓ, g^u as in the proof of Theorem B.3. Clearly, we can assume $[a, b] \subseteq \{0, \ldots, n-1\}$.

Now, we modify g^j as follows (for $j \in \{\ell, u\}$): for $1 \leq m \leq M$, consider the interval $I_m = [a_m, b_m]$, and the corresponding "piece" g_m^j of g on I_m. We let $\tilde{g}_m^j$ be the pseudo-distribution defined from g_m^j as follows: it is affine on I_m, with

$$\tilde{g}_m^u(a_m) \stackrel{\text{def}}{=} \left\lceil g^u(a_m) \frac{M\,|I_m|}{2\epsilon^2} \right\rceil \frac{2\epsilon^2}{M\,|I_m|}, \qquad \tilde{g}_m^u(a_m) \stackrel{\text{def}}{=} \left\lceil g^u(b_m) \frac{M\,|I_m|}{2\epsilon^2} \right\rceil \frac{2\epsilon^2}{M\,|I_m|}$$

and

$$\tilde{g}_m^\ell(a_m) \stackrel{\text{def}}{=} \left\lfloor g^\ell(a_m) \frac{M\,|I_m|}{2\epsilon^2} \right\rfloor \frac{2\epsilon^2}{M\,|I_m|}, \qquad \tilde{g}_m^\ell(a_m) \stackrel{\text{def}}{=} \left\lfloor g^\ell(b_m) \frac{M\,|I_m|}{2\epsilon^2} \right\rfloor \frac{2\epsilon^2}{M\,|I_m|}$$

i.e. g^j_m is g "rounded up" (resp. down) to the near multiple of $\frac{\epsilon^2}{M|I_m|}$ on the endpoints. We then let $\tilde{g}^j$ be the correspond piecewise-affine pseudo-distribution defined by piecing together the $\tilde{g}^j_m$'s. Clearly, by construction $\tilde{g}^\ell$ and $\tilde{g}^u$ still satisfies (i) and (ii) of Theorem B.3, and $\tilde{g}^\ell \le p \le \tilde{g}^u$. As for (iii), observe that at all $1 \le m \le M$ and $k \in I_m$ we have $\left|\tilde{g}^j(k) - g^j(k)\right| \le \frac{2\epsilon^2}{M|I_m|}$, from which

$$\mathrm{d_H}\left(p, \tilde{g}^j\right) \le \mathrm{d_H}\left(p, g^j\right) + \mathrm{d_H}\left(g, \tilde{g}^j\right) \le \epsilon + \sqrt{d_{\mathrm{TV}}\left(g^j, \tilde{g}^j\right)} \le \epsilon + \sqrt{\frac{1}{2}\sum_{m=1}^{M} |I_m| \cdot \frac{2\epsilon^2}{M\,|I_m|}} = 2\epsilon$$

showing that we get (up to a constant factor loss in the distance) (iii) as well. Given this, we get that specifying $(\tilde{g}^\ell, \tilde{g}^u)$ can be done by the list of the $O(1/\sqrt{\epsilon})$ endpoints along with the value of each $\tilde{g}^j$ for all of these endpoints. Now, given the two endpoints, one gets the size of the corresponding interval I_m (which is at most n), and the two values to specify are a multiple of $\epsilon^2/(M\,|I_m|)$ in $[0,1]$. (If we were to stop here, we would get the existence of an ϵ-cover $\mathcal{C}'_\epsilon$ of $\mathcal{LCV}_n$ in Hellinger distance of size $(n/\epsilon)^{O(1/\sqrt{\epsilon})}$.)

One Last Step: Outside $[a,b]$. To get the bracketing bound we seek, we need to do one last modification to our pair $(\tilde{g}^\ell, \tilde{g}^u)$. Specifically, in the above we have one issue when approximating p: namely, that outside of their common support $\{a, \dots, b\}$, both $\tilde{g}^j$'s are 0. While this is fine for the lower bound $\tilde{g}^\ell$, this is not for $\tilde{g}^u$, as it needs to dominate p outside of $\{a, \dots, b\}$ as well, where p may have $O(\epsilon^2)$ probability mass. Thus, we need to adapt the construction above, as follows (we treat the setting of $\tilde{g}^u$ on $\{b+1, \dots, n\}$, the definition on $\{0, \dots, a-1\}$ is similar).

First, observe if $p(b+1) = 0$, we are done, as then by monotonicity we must have $(k) = 0$ for all $k \ge b+1$, and so setting $\tilde{g}^u = 0$ on $\{b+1, \dots, n\}$ suffices. Thus, we hereafter assume $p(b+1) > 0$; and, for $b+1 \le k \le n$, set

$$\tilde{g}^u(k) \stackrel{\mathrm{def}}{=} \alpha e^{\beta(k-(b+1))}$$

where $\alpha \stackrel{\mathrm{def}}{=} \left\lceil p(b+1)\frac{n}{2\epsilon^2}\right\rceil \frac{2\epsilon^2}{n}$ and $\beta \stackrel{\mathrm{def}}{=} \left\lceil \frac{n}{\epsilon} \ln \frac{p(b+2)}{p(b+1)}\right\rceil \frac{\epsilon}{n}$ (so that $\beta \le 0$). Then $\tilde{g}^u(b+1) \ge p(b+1)$, and for $b+1 < k \le n$

$$\frac{\tilde{g}^u(k)}{\tilde{g}^u(k-1)} = e^\beta \ge \frac{p(b+2)}{p(b+1)} \ge \frac{p(k)}{p(k-1)}$$

(the last inequality due to the log-concavity of p). This implies $\tilde{g}^u \ge p$ on $\{b+1, \dots, n\}$ as desired; and, thanks to the rounding, there are only $O(n/\epsilon^2)$ different possibilities for the tail of $\tilde{g}^u$. In view of bounding the Hellinger distance between p and $\tilde{g}^u$ added by this modification, which is upper bounded by the (square root) of the total variation distance this added, recall that $p(\{b+1, \dots, n\}) = O(\epsilon^2)$ by construction, and that

$$\tilde{g}^u(\{b+1, \dots, n\}) = \sum_{k=b+1}^{n} \alpha e^{\beta(k-(b+1))} = \frac{\alpha}{1-e^\beta}.$$

Thus, the Hellinger distance incurred on $\{b+1, \dots, n\}$ is at most $\sqrt{O(\epsilon^2) + \frac{\alpha}{1-e^\beta}}$; and to conclude, it only remains to show that $\frac{\alpha}{1-e^\beta} = O(\epsilon^2)$.

To show this last point, let $I_m = [c,b]$ be the rightmost interval in the decomposition from Proposition B.5. Recall that we are guaranteed that p is non-increasing on I_m; further, by inspection of the proof of [DKS16a, Proposition 15], we also have that I_m is *maximal*, in the sense that b is the rightmost point k such that $[c,k]$ satisfies the assumptions of Lemma B.4. Using first the monotoncity, we have

$$p(b+1) \le p(b) \le \frac{p(I_m)}{b-c} \le \frac{O(\epsilon^2)}{b-c}$$

that last inequality by construction (from the technicality we enforced in the end of the proof of Theorem B.3); and therefore $\alpha \leq \frac{O(\epsilon^2)}{b-c} + \frac{\epsilon^2}{n} = \frac{O(\epsilon^2)}{b-c}$.
In order to obtain an upper bound on β, we rely on the maximality of I_m, leading to two cases to consider:

- The first is that $p(b+1) < \frac{p(c)}{2}$; in which case $p(b+2) \leq p(b+1) < \frac{p(c)}{2}$; which implies that

$$\frac{1}{2} > \frac{p(b+2)}{p(c)} = \frac{p(b+2)}{p(b+1)} \cdot \frac{p(b+1)}{p(b)} \cdots \frac{p(c+1)}{p(c)} \geq \left(\frac{p(b+2)}{p(b+1)}\right)^{b-c+2}$$

  the last inequality by log-concavity. In turn, we get

$$\beta \leq \ln \frac{p(b+2)}{p(b+1)} + \frac{\epsilon}{n} \leq -\frac{\ln 2}{b-c+2} + \frac{\epsilon}{n}.$$

- The second is that $\ln \frac{p(c+1)}{p(c)} - \ln \frac{p(b+1)}{p(b)} > \frac{1}{b-c+1}$. In this case,

$$\ln \frac{p(b+2)}{p(b+1)} \leq \ln \frac{p(b+1)}{p(b)} < \ln \frac{p(c+1)}{p(c)} - \frac{1}{b-c+1} \leq -\frac{1}{b-c+1} < -\frac{\ln 2}{b-c+2}$$

  (the last inequality as $b-c \geq 0$) and therefore $\beta \leq -\frac{\ln 2}{b-c+2} + \frac{\epsilon}{n}$ as in the first case.

Combining these two bounds, we obtain

$$\frac{\alpha}{1-e^{\beta}} \leq \frac{O(\epsilon^2)}{b-c} \cdot \frac{1}{1-e^{\frac{\epsilon}{n}} e^{-\frac{\ln 2}{b-c+2}}} = O(\epsilon^2)$$

the last inequality for $\epsilon < \frac{\ln 2}{2}$ (using the fact that $1 \leq b-c \leq n$). This concludes the proof: as discussed, we then have that our setting of $\bar{g}^u$ outside of $[a,b]$ only causes an addition Hellinger distance of $\sqrt{O(\epsilon^2) + \frac{\alpha}{1-e^{\beta}}} = \sqrt{O(\epsilon^2)} = O(\epsilon)$.

□

We are, at last, ready to prove our main theorem:

Proof of Theorem B.1. Recall the following theorem, due to Wong and Shen [WS95] (see also [vdG00, Theorem 7.4], [KS16, Theorem 17]):

Theorem B.6 ([WS95, Theorem 2])**.** *There exist positive constants $\tau_1, \tau_2, \tau_3, \tau_4 > 0$ such that, for all $\epsilon \in (0,1)$, if*

$$\int_{\epsilon^2/2^8}^{\sqrt{2}\epsilon} \sqrt{\mathcal{N}_{[]}(u/\tau_1, \mathcal{G}, \mathrm{d_H})}\, du \leq \tau_2 m^{1/2} \epsilon^2 \tag{7}$$

and $\tilde{p}_n$ is an estimator that approximates $\hat{p}_m$ within error η (i.e., solves the maximization problem within additive error η) with $\eta \leq \tau_3 \epsilon^2$, then

$$\Pr\left[\, \mathrm{d_H}\left(\tilde{p}_m, p\right) \geq \epsilon \,\right] \leq 5 \exp(-\tau_4 m \epsilon^2).$$

To apply this theorem, define the function $J_n\colon (0,1) \to \mathbb{R}$ by $J(x) \stackrel{\text{def}}{=} \int_{x^2}^{x} \sqrt{\ln \frac{n}{u}} u^{-1/4}\, du$. By (tedious) computations, one can verify that $J_n(x) \sim_{x\to 0} \frac{4}{3} x^{3/4} \sqrt{\ln \frac{n}{x}}$; this, combined with the bound of Theorem B.2, yields that for any $\epsilon \in (0,1)$

$$\int_{\epsilon^2/2^8}^{\sqrt{2}\epsilon} \sqrt{\mathcal{N}_{[]}(u/\tau_1, \mathcal{LCV}_n, \mathrm{d_H})}\, du = O\left(\epsilon^{3/4} \sqrt{\ln \frac{n}{\epsilon}}\right).$$

Thus, setting, for $m \geq 1$, $\epsilon_m \stackrel{\text{def}}{=} C m^{-2/5} (\ln(mn))^{2/5}$ for a sufficiently big absolute constant $C > 0$ ensures that ϵ_m satisfies (7). Let $\rho_m \stackrel{\text{def}}{=} 1/\epsilon_m$. It follows that any estimator which, on a sample of size m, approximates the log-concave MLE to within an additive $\eta_m \stackrel{\text{def}}{=} \tau_3 \epsilon_m^2$ has minimax error

$$\begin{aligned}
\rho_m^2 \sup_{p\in\mathcal{LCV}_n} \mathbb{E}_p[\mathrm{d_H}(\tilde{p}_m, p)^2] &= \sup_{p\in\mathcal{LCV}_n} \int_0^\infty \Pr\left[\rho_m^2 \mathrm{d_H}(\tilde{p}_n, p)^2 \geq t\right] dt \\
&= \sup_{p\in\mathcal{LCV}_n} \int_0^\infty \Pr\left[\mathrm{d_H}(\tilde{p}_n, p) \geq \sqrt{t}\rho_m^{-1}\right] dt \\
&\leq 1 + \sup_{p\in\mathcal{LCV}_n} \int_1^\infty \Pr\left[\mathrm{d_H}(\tilde{p}_n, p) \geq \sqrt{t}\rho_m^{-1}\right] dt \\
&= 1 + \sup_{p\in\mathcal{LCV}_n} \int_1^\infty \Pr\left[\mathrm{d_H}(\tilde{p}_n, p) \geq \sqrt{t}\epsilon_m\right] dt \\
&\leq 1 + 5 \sup_{p\in\mathcal{LCV}_n} \int_1^\infty \exp(-\tau_4 m t \epsilon_m^2)\, dt \\
&= 1 + 5 \sup_{p\in\mathcal{LCV}_n} \int_1^\infty \exp(-\tau_4 C m^{1/2} \ln(mn) t)\, dt \\
&= O(1)
\end{aligned}$$

where we used the fact that if $\epsilon_t > \epsilon_m$, then ϵ_t satisfies (7) as well (and applied it to $\epsilon_t = \sqrt{t}\epsilon_m$). This concludes the proof. □

ECCC ISSN 1433-8092
https://eccc.weizmann.ac.il