Nearly Optimal Learning and Sparse Covers for Sums of Independent Integer Random Variables

Ilias Diakonikolas*
University of Edinburgh
ilias.d@ed.ac.uk.

Daniel M. Kane†
University of California, San Diego
dakane@cs.ucsd.edu.

Alistair Stewart‡
University of Edinburgh
stewart.al@gmail.com.

May 4, 2015

Abstract

For $k \in \mathbb{Z}_+$, a *k-SIIRV of order $n \in \mathbb{Z}_+$* is the discrete probability distribution of the sum of n mutually independent random variables each supported on $\{0, 1, \ldots, k-1\}$. We denote by $\mathcal{S}_{n,k}$ the set of all k-SIIRV's of order n. In this paper we prove two main results:

- We give a near-sample optimal and computationally efficient algorithm for learning k-SIIRVs from independent samples under the total variation distance (L_1 distance). Our algorithm uses $\widetilde{O}(k/\epsilon^2)$ samples and runs in $\widetilde{O}(k^3/\epsilon^2)$ time. The sample size of our algorithm is optimal up to logarithmic factors, as $\Omega(k/\epsilon^2)$ samples are information-theoretically necessary to learn a single random variable supported on $\{0, 1, \ldots, k-1\}$.
- We prove nearly tight bounds on the size of ϵ-covers for $\mathcal{S}_{n,k}$ under the total variation distance. In particular, we show that for all $k, n \in \mathbb{Z}_+$ and $\epsilon \leq 1/k$, $\mathcal{S}_{n,k}$ admits an ϵ-cover of size $n \cdot (1/\epsilon)^{O(k \cdot \log(1/\epsilon))}$ that can be constructed in polynomial time. We also prove a nearly matching lower bound: For $k \in \mathbb{Z}_+$ and $n = \Omega(\log(1/\epsilon))$ any ϵ-cover for $\mathcal{S}_{n,k}$ has size at least $n \cdot (1/\epsilon)^{\Omega(k \cdot \log(1/\epsilon))}$. Using the structural understanding obtained from our construction, we prove that the sample complexity of learning 2-SIIRVs is $\Omega((1/\epsilon^2)\sqrt{\log(1/\epsilon)})$.

The unifying idea of our upper bounds is an analysis of the structure of the Fourier Transform of k-SIIRVs. Our learning algorithm relies on a structural property of the Fourier transform of k-SIIRVs, namely that it has small effective support. Our lower bounds employ a combination of geometric and analytic arguments.

*Supported by EPSRC grant EP/L021749/1 and a Marie Curie Career Integration grant.
†Some of this work was performed while visiting the University of Edinburgh.
‡Supported by EPSRC grant EP/L021749/1.

1 Introduction

This paper is concerned with sums of independent integer random variables:

Definition. For $k \in \mathbb{Z}_+$, a *k-IRV* is a random variable supported on $\{0, 1, \ldots, k-1\}$. A *k-SIIRV of order n* is any random variable $X = \sum_{i=1}^{n} X_i$ where the X_i's are independent k-IRVs. We will denote by $\mathcal{S}_{n,k}$ the set of probability distributions of all k-SIIRVs of order n.

For convenience, throughout this paper, we will often blur the distinction between a random variable and its distribution. In particular, we will use the term k-SIIRV for a random variable or its corresponding distribution, and the distinction will be clear from the context.

Sums of independent integer random variables comprise a rich class of discrete distributions that arise in many settings. The special case of $k = 2$, $\mathcal{S}_{n,2}$, was first considered by Poisson [Poi37] as a non-trivial extension of the Binomial distribution and are known as Poisson Binomial distributions (PBDs). In application domains, PBDs 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. We remark that k-SIIRVs are of fundamental interest and have been extensively studied in probability and statistics. For example, tail bounds on k-SIIRVs form an important special case of Chernoff/Hoeffding bounds [Che52, Hoe63, DP09b]. Moreover, there is a long line of research on approximate limit theorems for sums of independent integer random variables, dating back several decades (see e.g., [Pre83, Kru86, BHJ92]) and [CL10, CGS11] for some recent results.

In this paper, we prove two main results: (1) We give a near-sample optimal efficient algorithm for learning k-SIIRVs from independent samples, resolving the main open questions of [DDS12b, DDO$^+$13]. (2) We prove nearly-tight upper and lower bounds on the size of ϵ-covers for k-SIIRVs. Our cover upper bound is constructive, i.e., we also give an efficient algorithm to construct a near-minimum size cover. As we explain below, our cover bounds have interesting implications in learning and computational game theory.

In the following, we state our results in detail and elaborate on their context and the connections between them.

Learning k-SIIRVs. The main motivation of this work has been the problem of learning an unknown k-SIIRV given access to independent samples drawn from it. More specifically, given access to independent samples drawn from an unknown random variable $X \in \mathcal{S}_{n,k}$, we must output a hypothesis random variable H such that with probability at least $2/3$ the total variation distance $d_{\mathrm{TV}}(X, H)$ between X and H is at most ϵ.

We note that this is the standard definition of *density estimation* [DG85, DL01], i.e., PAC learning an unknown probability distribution from samples [KMR$^+$94], which is a natural analogue of Valiant's well-known PAC model for learning Boolean functions [Val84] to the unsupervised setting. Density estimation is a classical topic in statistics and machine learning with a rich history and extensive literature (see e.g., [BBBB72, DG85, Sil86, Sco92, DL01]). The reader is referred to [Ize91] for a survey of statistical techniques in this context. In recent years, a large body of work in theoretical computer science has been studying these questions from a computational perspective; see e.g., [KMR$^+$94, FM99, AK01, CGG02, VW02, FOS05, BS10, KMV10, MV10, DDS12a, DDS12b, DDO$^+$13, CDSS14].

The ultimate goal in density estimation is to design a learning algorithm that is both computationally and information-theoretically efficient. The "gold standard" in this setting is an algorithm with information-theoretically optimal sample size and running time polynomial (or, ideally, linear) in its sample size. As our main learning result, we give a near-sample optimal and computationally efficient algorithm for learning k-SIIRVs:

Theorem 1 (Main Learning Result)**.** *There is a learning algorithm for k-SIIRVs with the following properties: Let X be any k-SIIRV of order n. The algorithm uses $\widetilde{O}(k/\epsilon^2)$ samples from X, runs in time $\widetilde{O}(k^3/\epsilon^2)$, and with probability at least $2/3$ outputs a (succinct description of a) hypothesis H such that $d_{\mathrm{TV}}(H, X) \leq \epsilon$.*

We remark that even learning a single k-IRV to total variation distance ϵ information-theoretically requires $\Omega(k/\epsilon^2)$ samples, Hence, the sample complexity of our algorithm is provably optimal up to logarithmic factors.

Comparison to Previous Work. The very special case of $k = 2$ (i.e., the problem of learning 2-SIIRVs/PBDs) was previously studied by [DDS12b]. [DDS12b] gave two algorithms for learning 2-SIIRVs: one that uses $\widetilde{O}(1/\epsilon^3)$ samples and runs in sample near-linear time, and one that uses $\widetilde{O}(1/\epsilon^2)$ samples and runs in time $(1/\epsilon)^{\mathrm{polylog}(1/\epsilon)}$. Obtaining a near-sample optimal and *computationally efficient* learning algorithm for PBDs was posed as one of the main open problems in [DDS12b]. As a corollary of Theorem 1, we obtain a near-sample optimal and *near-linear time* algorithm for learning PBDs.

More recently, [DDO+13] studied the problem of learning k-SIIRVs for general $k \geq 2$. They obtained an algorithm for this problem that uses $\mathrm{poly}(k/\epsilon)$ samples and runs in $\mathrm{poly}(k/\epsilon)$ time. We remark that the degree of these polynomials is quite high: the sample complexity (and, hence, running time) of the [DDO+13] algorithm is $\Omega(k^9/\epsilon^6)$. Even understanding the sample complexity of learning k-SIIRVs for $k > 2$ (i.e., without any computational considerations) has been an open problem. Theorem 1 gives a tight upper bound on the sample complexity of this learning problem up to logarithmic factors, and does so with a computationally efficient algorithm.

Given our $\widetilde{O}(k/\epsilon^2)$ sample upper bound, it would be tempting to conjecture that $\Theta(k/\epsilon^2)$ is in fact the optimal sample complexity of learning k-SIIRVs, i.e., that there exists an $O(k/\epsilon^2)$ sample algorithm. If true, this would imply that learning a k-SIIRV is as easy as learning a k-IRV. For example, in the case $k = 2$, such a statement would mean that learning an arbitrary PBD up to total variation distance ϵ is as easy (up to a constant factor) as distinguishing a fair coin from an ϵ-biased coin. Perhaps surprisingly, we show that this is not the case:

Theorem 2 (Sample Lower Bound)**.** *Any algorithm that learns an arbitrary 2-SIIRV (PBD) to total variation distance ϵ with probability at least $1/10$ must use $\Omega((1/\epsilon^2)\sqrt{\log(1/\epsilon)})$ samples.*

Theorem 2 provides a separation between learning PBDs and learning Binomial distributions. We conjecture that the optimal sample complexity for learning k-SIIRVs is $\Theta((k/\epsilon^2)\sqrt{\log(1/\epsilon)})$.

Our learning algorithm and our information-theoretic lower bound both rely on a novel structural understanding of the space of k-SIIRVs. We elaborate on our techniques in Section 1.2 below.

Sparse Covers for k-SIIRVs. Let $(\mathcal{X}, d)$ be a metric space. Given $\delta > 0$, a subset $\mathcal{Y} \subseteq \mathcal{X}$ is said to be a *δ-cover of $\mathcal{X}$* with respect to the metric $d : \mathcal{X}^2 \to \mathbb{R}_+$ if for every $\mathbf{x} \in \mathcal{X}$ there exists some $\mathbf{y} \in \mathcal{Y}$ such that $d(\mathbf{x}, \mathbf{y}) \leq \delta$. There may exist many δ-covers of $\mathcal{X}$, but one is typically interested in one with the minimum cardinality. The *δ-covering number* of $(\mathcal{X}, d)$ is the minimum cardinality of any δ-cover of $\mathcal{X}$. Intuitively, the covering number of a metric space captures the "size" of the space. A sparse cover provides useful structural information about the underlying space that can be exploited in a variety of applications.

Covering numbers (and their logarithms, known as *metric entropy* numbers) were first defined by A. N. Kolmogorov in the 1950's and have since played a central role in a number of areas, including approximation theory, geometric functional analysis (see, e.g., [Dud74, Mak86, BGL07] and the books [KT59, Lor66, CS90, ET96]), geometric approximation algorithms [Hp11], information theory, statistics, and machine learning (see, e.g., [Yat85, Bir86, HI90, HO97, YB99, GS13] and the books [vdVW96, DL01, Tsy08]).

As our second main result, we prove nearly tight upper and lower bounds on the covering numbers of k-SIIRVs under the total variation distance metric. Our upper bound is constructive in the sense that it implies an efficient algorithm to computer a near-optimal size cover. We prove:

Theorem 3 (Sparse Cover Bound)**.** *For $\epsilon \leq 1/k$, there exists an ϵ-cover $\mathcal{S}_{n,k,\epsilon} \subseteq \mathcal{S}_{n,k}$ of $\mathcal{S}_{n,k}$ under the total variation distance of size $|\mathcal{S}_{n,k,\epsilon}| \leq n \cdot (1/\epsilon)^{O(k \log(1/\epsilon))}$ that can be constructed in polynomial time. Moreover, any ϵ-cover for $\mathcal{S}_{n,k}$ has size at least $n \cdot (1/\epsilon)^{\Omega(k \log(1/\epsilon))}$.*

Comparison to Previous Work. The best previous upper bound on the cover size of PBDs ($k = 2$) is $n^2 + n \cdot (1/\epsilon)^{O(\log^2(1/\epsilon))}$ due to Daskalakis and Papadimitriou [DP09a, DP14a]. For $k > 2$, the main theorem of [DDO+13] implies a (non-proper) cover of size $n \cdot 2^{\mathrm{poly}(k/\epsilon)}$. Before our work, no non-trivial lower bound on the cover size was known. We view the inherent quasi-polynomial dependence on $1/\epsilon$ in the cover size as a rather surprising result.

Beyond its independent interest as a structural result, Theorem 3 has consequences in learning theory and computational game theory. In particular, our proper cover upper bound combined with our non-proper learning algorithm implies that there exists a *proper* learning algorithm for k-SIIRVs, i.e., an algorithm that outputs a k-SIIRV as its hypothesis, that draws $\widetilde{O}(k/\epsilon^2)$ samples and runs in time $(k/\epsilon)^{O(k \log(k/\epsilon))}$. This cover-based approach needs to perform some sort of enumeration over a cover, hence it cannot run in $\mathrm{poly}(k/\epsilon)$ time because of our lower bound. Obtaining a $\mathrm{poly}(k/\epsilon)$ time proper algorithm requires new ideas and is an interesting open problem for future work.

We now point out a connection between covers for k-SIIRVs and approximate Nash equilibrium computation. In a sequence of papers, Daskalakis and Papadimitriou [DP07, DP09a, DP14a, DP14b] showed that a constructive upper bound on the cover size of generalized multinomial distributions (i.e., sums of independent random variables with support $\{\mathbf{e}_i\}_{i=1}^k$, where $\mathbf{e}_i$ is the standard unit vector along dimension i in $\mathbb{R}^k$) implies an additive PTAS for the problem of computing ϵ-Nash equilibria in anonymous games [Mil96, Blo99, Blo05] with k strategies per player.

Note that for $k = 2$, generalized multinomial distributions correspond exactly to PBDs. Hence, our improved constructive upper bound on the cover size of PBDs implies an improved $\mathrm{poly}(n) \cdot (1/\epsilon)^{O(\log(1/\epsilon))}$ time algorithm for computing ϵ-Nash equilibria in anonymous games with 2 strategies per player. Our matching lower bound on the cover size implies that the "cover based approach" introduced by Daskalakis and Papadimitriou cannot lead to an FPTAS for this computational problem. We believe that this implication is interesting in light of the recent PPAD-hardness [CDO15] of computing exact Nash equilibria in anonymous games.

1.1 Preliminaries For a distribution $\mathbf{P}$ supported on $[m] = \{0, 1, \ldots, m\}$, $m \in \mathbb{Z}_+$, we write $\mathbf{P}(i)$ to denote the value $\Pr_{X \sim \mathbf{P}}[X = i]$ of the probability density function (pdf) at point i, and $\mathbf{P}(\leq i)$ to denote the value $\Pr_{X \sim \mathbf{P}}[X \leq i]$ of the cumulative density function (cdf) at point i. For $S \subseteq [n]$, we write $\mathbf{P}(S)$ to denote $\sum_{i \in S} \mathbf{P}(i)$.

The *total variation distance* between two distributions $\mathbf{P}$ and $\mathbf{Q}$ supported on a finite domain A is $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) \stackrel{\mathrm{def}}{=} \max_{S \subseteq A} |\mathbf{P}(S) - \mathbf{Q}(S)| = (1/2) \cdot \|\mathbf{P} - \mathbf{Q}\|_1$. Similarly, if X and Y are random variables, their total variation distance $d_{\mathrm{TV}}(X, Y)$ is defined as the total variation distance between their distributions. Another useful notion of distance between distributions/random variables is the *Kolmogorov distance*, defined as $d_{\mathrm{K}}(\mathbf{P}, \mathbf{Q}) \stackrel{\mathrm{def}}{=} \sup_{x \in \mathbb{R}} |\mathbf{P}(\leq x) - \mathbf{Q}(\leq x)|$. Note that for any pair of distributions $\mathbf{P}$ and $\mathbf{Q}$ supported on a finite subset of $\mathbb{R}$ we have that $d_{\mathrm{K}}(\mathbf{P}, \mathbf{Q}) \leq d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q})$.

Let $\mathcal{F}$ be a family of probability distributions. Given $\delta > 0$, a subset $\mathcal{G} \subseteq \mathcal{F}$ is said to be a (proper) *δ-cover of $\mathcal{F}$* with respect to the metric $d(\cdot, \cdot)$ if for every distribution $\mathbf{P} \in \mathcal{F}$ there exists some $\mathbf{Q} \in \mathcal{G}$ such that $d(\mathbf{P}, \mathbf{Q}) \leq \delta$. The δ-covering number for $(\mathcal{F}, d)$ is the minimum cardinality of a δ-cover. The δ-packing number for $(\mathcal{F}, d)$ is the maximum number of points (distributions) in $\mathcal{F}$ at pairwise distance at least δ from each other.

1.2 Prior Work and our Techniques In this section we give an overview of our techniques and compare them to previous approaches.

Prior Work. A crucial ingredient for the PBD cover upper bound of [DP09a, DP14a] is the following moment matching theorem that they establish: If two PBDs agree on their first $\Omega(\log(1/\epsilon))$ moments, then their total variation distance is at most ϵ. We show that this structural result is in fact tight: In Proposition 29 (Appendix A), we give an explicit example of two PBDs over $k+1$ variables that agree exactly on the first k moments and have total variation distance $2^{-\Omega(k)}$. Unfortunately, such a moment matching statement cannot be true for k-SIIRVs, even for $k = 3$. Intuitively, this is because knowledge about moments fails to account for potential periodic structure of the probability mass function that comes into play for $k > 2$. For example, $\Omega(n)$ moments do not suffice to distinguish between the cases that a 3-SIIRV of order n is supported on the even versus the odd integers. More specifically, in Proposition 30 (Appendix A), we give an explicit example of two 3-SIIRV of order $n/2$ that agree exactly on the first $n-1$ moments and have disjoint supports.

Previous works on learning PBDs [DDS12b] and k-SIIRVs [DDO$^+$13] rely on a certain limit theorem that they establish about the structure of these distributions. In both cases, this limit theorem has the form of a "regularity" lemma: Any k-SIIRV is either ϵ-close in total variation distance to being $L = \Theta(k^9/\epsilon^4)$- "sparse", i.e., it is supported on a set of at most L consecutive integers, or ϵ-close to being "structured" (see Theorem 8 for a precise statement). In the former case, the distribution can be learned by "brute force" (since the support is "small"), and in the latter case one can exploit the structure to learn with a small number of samples as well. The overall algorithm proceeds by running a subroutine for each case followed by a hypothesis testing to select the correct hypothesis. Unfortunately, the sparse case is a bottleneck for the sample complexity of this approach, as any algorithm to learn an arbitrary distribution over support L requires $\Omega(L/\epsilon^2)$ samples. Hence, one needs to exploit the structure of k-SIIRVs beyond the aforementioned.

Our Techniques. The unifying idea of our upper bounds is an analysis of the structure of the *Fourier Transform* of k-SIIRVs. The Fourier transform is a natural tool to consider given that previous moment-based techniques fail to detect periodic structure. On the other hand, this type of structure is easily detectable by considering the Fourier transform. Recall that the Fourier transform of a sum of independent random variables is the product of the Fourier transforms of the individual variables. Moreover, if two random variables have similar Fourier transforms, they also have similar distributions. These two basic facts are the starting point of our analysis.

Our learning algorithm makes essential use of a new structural property exhibited by the Fourier transform of k-SIIRVs, namely that it has *small effective support* (see Lemma 6). Assuming that the effective support is explicitly known, our new structural result suggests an extremely simple learning algorithm: Use samples from the distribution to estimate its Fourier transform at the points of the effective support, set the Fourier transform to 0 everywhere else, and compute the inverse transform. By exploiting the sparsity of the Fourier transform, it can be shown that this algorithm achieves total variation distance ϵ after $\widetilde{O}(k/\epsilon^2)$ samples. Our actual learning algorithm (Section 2) is only slightly more complicated than the above two line description, because the effective support of the Fourier transform is not precisely known in advance.

Our cover upper bound hinges on showing that the Fourier transform of a k-SIIRV is necessarily of low complexity, i.e., it can be succinctly described up to small error. In particular, since the Fourier transform is smooth, we show (Lemma 13), roughly, that its logarithm can be well approximated by a low degree Taylor polynomial on intervals of length $O(1/k)$. (Our actual statement is somewhat more complicated as it needs to account for roots of the Fourier transform close to the unit circle.) Therefore, providing approximations to the low-degree Taylor coefficients of the logarithm of the Fourier transform provides a concise approximate description of the distribution.

Our lower bounds take a geometric view of the problem. At a high-level, we consider the function that maps the set of $n(k-1)$ parameters defining a k-SIIRV to the corresponding probability mass function. We show that there exists a region of the space of distributions where this function is invertible. For $k = 2$, we in fact show that the distribution of any PBD with distinct parameters lies in the interior of this region. This structural understanding allows us to use certain appropriately defined expectations to extract the effect of individual parameters on the distribution. We show, roughly, that when $n = \Theta(\log(1/\epsilon))$, the effects of changing each of the $n(k-1)$ parameters in this region are sufficiently distinct, so that we can isolate the effect of a single parameter. In other words, $\Omega(k \log(1/\epsilon))$ parameters are effectively independent, which intuitively implies the $(1/\epsilon)^{\Omega(k \log(1/\epsilon))}$ lower bound on the cover size. To prove our sample lower bound, at a high-level, we combine the aforementioned construction with Assouad's lemma [Ass83].

Remark. Recently, [CDSS14] introduced a general approach for learning structured distribution families based on piecewise polynomial approximations. While the [CDSS14] is very general, it provably does not apply to k-SIIRVs, $k > 2$, in the sense that it cannot lead to a learning algorithm for this class with sample complexity independent of n. The reason is that *any* piecewise polynomial approximation for k-SIIRVs of order n necessarily incurs at least a polynomial dependence on n.

Structure of the Paper. In Section 2 we describe and analyze our learning algorithm for k-SIIRVs. Section 3 contains our cover upper bound construction. Our lower bounds are given in Sections 4 and 5 respectively.

2 Learning Sums of Independent Integer Random Variables

In this section we prove Theorem 1, which we state below in more detail for the sake of completeness.

Theorem 4. *There is an algorithm* `Learn-SIIRV` *that for any $\mathbf{P} \in \mathcal{S}_{n,k}$ and $\epsilon > 0$, takes $O(k \log^2(k/\epsilon)/\epsilon^2)$ samples from $\mathbf{P}$, runs in time $\widetilde{O}(k^3/\epsilon^2)$ and returns a (succinct description of a) hypothesis $\mathbf{H}$ so that with probability at least $2/3$ we have that $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{H}) < \epsilon$.*

Our algorithm uses the Discrete Fourier Transform, which we now define.

Definition 5. For $x \in \mathbb{R}$ we will denote $e(x) \stackrel{\text{def}}{=} \exp(2\pi i x)$. The *Discrete Fourier Transform (DFT) modulo M* of a function $F : [n] \to \mathbb{C}$ is the function $\widehat{F} : [M-1] \to \mathbb{C}$ defined as $\widehat{F}(\xi) = \sum_{j=0}^{n} e(-\xi j/M) F(j)$, for integers $\xi \in [M-1]$. The DFT modulo M of a distribution $\mathbf{P}$, $\widehat{\mathbf{P}}$ is the DFT modulo M of its probability mass function. The *inverse DFT modulo M* onto the range $[m, m+M-1]$ of $\widehat{F} : [M-1] \to \mathbb{C}$, is the function $F : [m, m+M-1] \cap \mathbb{Z} \to \mathbb{C}$ defined by $F(j) = \frac{1}{M} \sum_{\xi=0}^{M-1} e(\xi j/M) \widehat{F}(\xi)$, for $j \in [m, m+M-1] \cap \mathbb{Z}$. The L_2 norm of the DFT is defined as $\|\widehat{F}\|_2 = \sqrt{\frac{1}{M} \sum_{\xi=0}^{M-1} |\widehat{F}(\xi)|^2}$.

We start by giving an intuitive explanation of our approach. The Fourier transform $\widehat{\mathbf{Q}}$ of the empirical distribution $\mathbf{Q}$ provides an approximation to the Fourier transform $\widehat{\mathbf{P}}$ of $\mathbf{P}$. In particular, if we take N samples from $\mathbf{P}$, we expect that the empirical Fourier transform $\widehat{\mathbf{Q}}$ has error $O(N^{-1/2})$ at each point. This implies that the expected L_2 error $\|\widehat{\mathbf{Q}} - \widehat{\mathbf{P}}\|_2$ is $O(N^{-1/2})$, and thus by applying the inverse Fourier transform, would yield a distribution with L_2 error of $O(N^{-1/2})$ from $\mathbf{P}$. This guarantee may sound good, but unfortunately, the distribution $\mathbf{P}$ has effective support of size approximately $s\sqrt{\log(1/\epsilon)}$, where $s = \sqrt{\mathrm{Var}_{X \sim \mathbf{P}}[X]}$, and thus the resulting distribution will likely have L_1 error of $O(N^{-1/2} s^{1/2} \log^{1/4}(1/\epsilon))$ from $\mathbf{P}$. This bound is prohibitively large, especially when the standard deviation of $\mathbf{P}$ is large.

This obstacle can be circumvented by relying on a new structural result that we believe may be of independent interest. *We show that for any k-SIIRV with large variance, its Fourier Transform*

will have small effective support. In particular, for any k-SIIRV with standard deviation s and $\epsilon > 0$ we consider its Discrete Fourier transform modulo M, and show the set of points in $[M-1]$ whose Fourier transform is bigger than ϵ in magnitude has size at most $O(Mks^{-1}\sqrt{\log(1/\epsilon)})$. By choosing M to be approximately $s\sqrt{\log(1/\epsilon)}$, i.e., of the same order as the effective support of $\mathbf{P}$, we conclude that the effective support of $\widehat{\mathbf{P}}$ (modulo M) is $O(k\log(1/\epsilon))$.

If the effective support for $\widehat{\mathbf{P}}$ was explicitly known, we could truncate our empirical Discrete Fourier transform $\widehat{\mathbf{Q}}$ (modulo M) outside this set and reduce the L_2 error $\|\widehat{\mathbf{Q}} - \widehat{\mathbf{P}}\|_2$ to $N^{-1/2}k^{1/2}s^{-1/2}\log^{1/4}(1/\epsilon)$. This in turn would correspond to an L_1 error of $O(N^{-1/2}k^{1/2}\sqrt{\log(1/\epsilon)})$. Unfortunately, we do not know exactly where the support of the Fourier transform is, so we will need to approximate it by calculating the empirical DFT where the support might be and then simply truncating this empirical DFT whenever it is sufficiently small. Fortunately, we do have some idea of where the support is and it is not hard to show that we can truncate at all of the appropriate points with high probability.

Algorithm `Learn-SIIRV`
Input: sample access to a k-SIIRV $\mathbf{P}$ and $\epsilon > 0$.

Let C be a sufficiently large universal constant.

1. Draw $O(1)$ samples from $\mathbf{P}$ and with confidence probability 19/20 compute: (a) $\widetilde{\sigma}^2$, a factor 2 approximation to $\mathrm{Var}_{X\sim\mathbf{P}}[X]+1$, and (b) $\widetilde{\mu}$, an approximation to $\mathbb{E}_{X\sim\mathbf{P}}[X]$ to within one standard deviation.
2. Take $N = C^3k/\epsilon^2\ln^2(k/\epsilon)$ samples from $\mathbf{P}$ to get an empirical distribution $\mathbf{Q}$.
3. If $\widetilde{\sigma} \le 4k\ln(4/\epsilon)$, then output $\mathbf{Q}$. Otherwise, proceed to next step.
4. Set $M \stackrel{\text{def}}{=} 1 + 2\lceil 6\widetilde{\sigma}\sqrt{\ln(4/\epsilon)})\rceil$. Let

   $$S \stackrel{\text{def}}{=} \{\xi \in [M-1] \mid \exists a,b \in \mathbb{Z}, 0 \le a \le b < k \text{ such that } |\xi/M - a/b| \le O(\log(k/\epsilon)/M)\} \ .$$

   For each $\xi \in S$, compute the DFT modulo M of $\mathbf{Q}$ at ξ, $\widehat{\mathbf{Q}}(\xi)$.
5. Compute $\widehat{\mathbf{H}}$ which is defined as $\widehat{\mathbf{H}}(\xi) = \widehat{\mathbf{Q}}(\xi)$ if $\xi \in S$ and $|\widehat{\mathbf{Q}}(\xi)| \ge R := 2C^{-1}\epsilon/\sqrt{k\ln(k/\epsilon)}$, and $\widehat{\mathbf{H}}(\xi) = 0$ otherwise.
6. Output $\widehat{\mathbf{H}}$ which is a succinct representation of $\mathbf{H}$, the inverse DFT of $\widehat{\mathbf{H}}$ modulo M onto the range $[\lfloor\widetilde{\mu}\rfloor - (M-1)/2, \lfloor\widetilde{\mu}\rfloor + (M-1)/2]$.

The bulk of our analysis will depend on showing that the Fourier transform of $\mathbf{P}$ has appropriately small effective support. To do this we need the following lemma:

Lemma 6. *Let $\mathbf{P} \in \mathcal{S}_{n,k}$ with $\sqrt{\mathrm{Var}_{X\sim\mathbf{P}}[X]} = s$, $\delta > 0$, and $M \in \mathbb{Z}_+$. Let $\widehat{\mathbf{P}}$ be the discrete Fourier transform of $\mathbf{P}$ modulo M. Then, we have*

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

(ii) At most $2Mks^{-1}\sqrt{\log(1/\delta)}$ many integers $0 \le \xi \le M-1$ have $|\widehat{\mathbf{P}}(\xi)| > \delta$.

Before we proceed with the proof of the lemma some comments are in order. Statement (i) of the lemma exhibits an explicit set $\mathcal{L}$ of cardinality $O(Mk^2s^{-1}\sqrt{\log(1/\delta)})$ that contains all the points $\xi \in [M-1]$ such that $|\widehat{\mathbf{P}}(\xi)| > \delta$. Note that the set $\mathcal{L}$ can be efficiently computed from M, δ, s, and does not otherwise depend on the particular k-SIIRV $\mathbf{P}$. Statement (ii) of the lemma shows that the effective support $\mathcal{L}' = \mathcal{L}'(\delta) = \{\xi \in [M-1] \mid |\widehat{\mathbf{P}}(\xi)| > \delta\}$ is in fact significantly smaller than $\mathcal{L}$, namely $|\mathcal{L}'| = O(Mks^{-1}\sqrt{\log(1/\delta)})$. This part of the lemma is non-constructive in the sense that it does not provide an explicit description for $\mathcal{L}'$ (beyond the fact that $\mathcal{L}' \subseteq \mathcal{L}$). We remark that the upper bound on the size of the effective support will be crucial for the analysis of our algorithm.

Proof of Lemma 6. Since $\mathbf{P} \in \mathcal{S}_{n,k}$, for $X \sim \mathbf{P}$, we have $X = \sum_{i=1}^{n} X_i$ where each $X_i \sim \mathbf{P}_i$ for a k-IRV $\mathbf{P}_i$. Let $Y_i = X_i - X_i'$ be the difference of two independent copies of X_i. Let $p_{ij} = \Pr[|Y_i| = j]$. Note that Y_i is a symmetric random variable. Consider its DFT modulo M which we will write as $\widehat{Y_i}$. We have the following sequence of (in)equalities:

$$\begin{aligned}
|\widehat{\mathbf{P}_i}(\xi)|^2 = \widehat{\mathbf{P}_i}(\xi)\widehat{\mathbf{P}_i}(-\xi) &= \widehat{Y_i}(\xi) \\
&= \sum_{j=0}^{k-1} p_{ij} \cos\left(\frac{2\pi\xi j}{M}\right) = 1 - \sum_{j=1}^{k-1} p_{ij}\left(1 - \cos\left(\frac{2\pi\xi j}{M}\right)\right) \\
&\leq 1 - 8\sum_{j=1}^{k-1} p_{ij}[\xi j/M]^2 \leq \exp\left(-8\sum_{j=1}^{k-1} p_{ij}[\xi j/M]^2\right),
\end{aligned}$$

where $[x]$, $x \in \mathbb{R}$, denotes the distance between x and its nearest integer. For the last two inequalities, we used that $\cos 2\pi x \leq 1 - 8x^2$ when $|x| \leq 1/2$, and $e^{-x} \geq 1 - x$ when $x \geq 0$.

Therefore, we have that $|\widehat{\mathbf{P}}(\xi)|^2 = \prod_{i=1}^{n} |\widehat{\mathbf{P}_i}(\xi)|^2 \leq \exp(-8\sum_{i=1}^{n}\sum_{j=1}^{k-1} p_{ij}[\xi j/M]^2)$. Taking square roots, we obtain

$$|\widehat{\mathbf{P}}(\xi)| \leq \exp\Big(-4\sum_{i=1}^{n}\sum_{j=1}^{k-1} p_{ij}[\xi j/M]^2\Big). \tag{1}$$

Note that we can relate the variance of $\mathbf{P}$ to the p_{ij}'s as follows:

$$s^2 = \mathrm{Var}[X] = \sum_{i=1}^{n} \mathrm{Var}[X_i] = \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}[Y_i^2] = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{k-1} p_{ij} j^2 .$$

Using (1), we get

$$|\widehat{\mathbf{P}}(\xi)| \leq \exp\left(-8s^2\left(\min_j \big(\frac{[\xi j/M]}{j}\big)^2\right)\right).$$

To complete the proof of (i), we will need a simple counting argument given in the following claim:

Claim 7. *For $a \in \mathbb{R}_+$ $j \in \mathbb{Z}_+$, there are at most $2Maj$ integers $0 \leq \xi \leq M-1$ with the following property: there exists $c \in \mathbb{Z}$ with $0 \leq c \leq j$ such that $|\xi/M - c/j| < a$. Therefore, there are at most $2Ma$ integers $0 \leq \xi \leq M-1$ with $[\xi j/M] < a$.*

Proof. For each c satisfying $1 \leq c \leq j-1$ there are either $\lfloor 2Ma \rfloor$ or $\lfloor 2Ma \rfloor - 1$ integers $0 \leq \xi \leq M-1$ with $|\frac{\xi}{M} - \frac{c}{j}| < a$. For $c = 0$ and $c = j$ there are either $\lfloor Ma \rfloor$ or $\lfloor Ma \rfloor - 1$ integers with $|\frac{\xi}{M} - \frac{c}{j}| < a$. Finally, note that $|\frac{\xi}{M} - \frac{c}{j}| < a$ for some $1 \leq c \leq j-1$ if and only if $[j\xi/M] < aj$. □

An application of the above claim for $a = (1/2)\sqrt{\ln(1/\delta)}$ implies that there are at most $\sum_{j=1}^{k-1} 2Mjs^{-1}\sqrt{\ln(1/\delta)}/2 \leq Mk^2s^{-1}\sqrt{\ln(1/\delta)}$ integers $0 \leq \xi \leq M-1$ with $\min_j \left(\frac{[\xi j/M]}{j}\right)^2 < \ln(1/\delta)/(4s^2)$. For all other integers we have $|\widehat{\mathbf{P}}(\xi)| \leq \delta$, which completes the proof of (i).

To prove (ii) we proceed by a probabilistic argument as follows: Consider evaluating the RHS of (1) with ξ being an integer random variable uniformly distributed in $[M-1]$. For $1 \leq j \leq k-1$, let N_j be the indicator random variable for the event that $[\xi j/M] < ks^{-1}\sqrt{\ln(1/\delta)}/2$. Note that by Claim 7 it follows that $\mathbb{E}[N_j] \leq ks^{-1}\sqrt{\ln(1/\delta)}$.

Note that $[\xi j/M] \geq \sqrt{1-N_j} \cdot ks^{-1}\sqrt{\ln(1/\delta)}/2$. Plugging this into (1) gives

$$|\widehat{\mathbf{P}}(\xi)| \leq \exp\left(-\frac{k^2}{s^2}\ln(1/\delta)\sum_{i=1}^{n}\sum_{j=1}^{k-1} p_{ij}(1-N_j)\right).$$

Since $s^2 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{k-1} p_{ij}j^2 \leq \frac{k^2}{2}\sum_{i=1}^{n}\sum_{j=1}^{k-1} p_{ij}$, it follows that $\theta := \sum_{i=1}^{n}\sum_{j=1}^{k-1} p_{ij} \geq 2s^2/k^2$. Therefore,

$$\mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{k-1} p_{ij}N_j\right] \leq \theta \cdot ks^{-1}\sqrt{\ln(1/\delta)}.$$

By Markov's inequality, except with probability $2ks^{-1}\sqrt{\ln(1/\delta)}$, we have that $\sum_{i=1}^{n}\sum_{j=1}^{k-1} p_{ij}N_j \leq \frac{\theta}{2}$. In this event, we have $\sum_{i=1}^{n}\sum_{j=1}^{k-1} p_{ij}(1-N_j) \geq \frac{\theta}{2}$ and hence

$$|\widehat{\mathbf{P}}(\xi)| \leq \exp\left(-\frac{k^2}{s^2}\ln(1/\delta)\sum_{i=1}^{n}\sum_{j=1}^{k-1} p_{ij}(1-N_j)\right) \leq \exp\left(-\frac{k^2}{s^2}\ln(1/\delta)\frac{\theta}{2}\right) \leq \delta.$$

Since ξ is uniformly distributed on $[M-1]$, it follows that $|\widehat{\mathbf{P}}(\xi)| > \delta$ for at most $2Mks^{-1}\sqrt{\ln(1/\delta)}$ integers ξ in $[M-1]$. This completes the proof of (ii). □

We are now ready to prove Theorem 4.

Proof of Theorem 4. Note that it is straightforward to verify the sample complexity bound. The running time of the algorithm is dominated by computing the DFT $\widehat{\mathbf{Q}}$. Since the support of $\mathbf{Q}$ is at most N, for each $\xi \in S$, we sum at most N terms to calculate $\widehat{\mathbf{Q}}(\xi)$. Therefore, the overall running time is $O(N \cdot |S|) = O(k\log^2(k/\epsilon)/\epsilon^2 \cdot k^2\log(k/\epsilon)) = O(k^3\log^3(k/\epsilon)/\epsilon^2)$ as claimed.

To show correctness, we will prove that the expected squared L_2 norm between $\widehat{\mathbf{H}}$ and $\widehat{\mathbf{P}}$ is small, i.e., that $\|\widehat{\mathbf{H}} - \widehat{\mathbf{P}}\|_2^2 = (1/M)\cdot\sum_{\xi=0}^{M-1}|\widehat{\mathbf{H}}(\xi) - \widehat{\mathbf{P}}(\xi)|^2$ has small expected value.

It is easy to see that, after drawing a constant number of samples, the quantities $\widetilde{\mu}$ and $\widetilde{\sigma}$ can be estimated to satisfy the required conditions with probability at least 19/20. (This follows for example by Lemma 6 of [DDS12b] with $\epsilon = 1/2$). We will henceforth condition on this event.

If $\widetilde{\sigma} \leq 4k\ln(4/\epsilon)$, then $s \leq 2k\ln(4/\epsilon)+1$, and Bernstein's inequality implies that $X \sim \mathbf{P}$ is within $O(k\log(1/\epsilon))$ of the mean with probability $1-\epsilon/2$. In this case, $O(k\log(1/\epsilon)/\epsilon^2) \leq N$ samples are sufficient to give that $d_{\mathrm{TV}}(\mathbf{P},\mathbf{Q}) \leq \epsilon$ with probability 2/3. (This follows from the fact that any distribution over support of size L can be learned with $O(L/\epsilon^2)$ samples to total variation distance ϵ.) We henceforth assume that we have $|\mu - \widetilde{\mu}| \leq s$, $s \geq \widetilde{\sigma}/2 \geq 2k\ln(4/\epsilon)$ and $\widetilde{\sigma} \leq 2s$.

Since $M = 1 + 2\lceil 6\widetilde{\sigma}\sqrt{\ln(4/\epsilon)})\rceil$, a random variable $X \sim \mathbf{P}$ lies in $[\lfloor\widetilde{\mu}\rfloor - (M-1)/2, \lfloor\widetilde{\mu}\rfloor + (M-1)/2]$ with probability at least $1 - \frac{\epsilon}{2}$. Indeed, an application of Bernstein's inequality for X yields that

$$\Pr(X > \mu + t) \leq \exp\left(-\frac{t^2}{2s^2 + \frac{2}{3}kt}\right)$$

where μ is the mean of $\mathbf{P}$, for any $t > 0$. For $t = 2s\sqrt{\ln(4/\epsilon)}$, we have $t^2 = (\ln(4/\epsilon))4s^2$ and $2s^2 + \frac{2}{3}kt = 2s^2 + \frac{4}{3}ks\sqrt{\ln(4/\epsilon)} \leq \frac{8}{3}s^2 \leq 4s^2$. Thus $\Pr(X > \mu + t) \leq \epsilon/4$. Similarly, it holds $\Pr(X < \mu - t) \leq \epsilon/4$. Now note that $\lfloor\widetilde{\mu}\rfloor + (M-1)/2 \geq (\mu - s) + \lceil 3s\sqrt{\ln(4/\epsilon)})\rceil \geq \mu + t$ and $\lfloor\widetilde{\mu}\rfloor - (M-1)/2 \leq \mu - t$. Hence, X is in $[\lfloor\widetilde{\mu}\rfloor - (M-1)/2, \lfloor\widetilde{\mu}\rfloor + (M-1)/2]$ with probability at least $1 - \epsilon/2$ as desired.

Fix $T = R/2 = C^{-1}\epsilon/(\sqrt{k\ln(k/\epsilon)})$. We analyze separately the contribution to the squared L_2 norm coming from ξ with $|\widehat{\mathbf{P}}(\xi)| > T$ and with $|\widehat{\mathbf{P}}(\xi)| \leq T$. Let us denote $\mathcal{L}'(T) = \{\xi \in [M-1] \mid |\widehat{\mathbf{P}}(\xi)| > T\}$. First consider

$$(1/M) \cdot \sum_{\xi \in \overline{\mathcal{L}'}(T)} |\widehat{\mathbf{H}}(\xi) - \widehat{\mathbf{P}}(\xi)|^2.$$

We first claim that with high probability $\widehat{\mathbf{H}}(\xi) = 0$ for all $\xi \in \overline{\mathcal{L}'}(T)$. This happens automatically when $\xi \notin S$, where the S is defined in the algorithm description. Note that $|S| = O(k^2 \log(k/\epsilon))$. For $\xi \in S \setminus \mathcal{L}'(T)$, we note that $\widehat{\mathbf{Q}}(\xi)$ is an average of N i.i.d. numbers each of absolute value 1 and mean $\widehat{\mathbf{P}}(\xi)$ (which has absolute value less than 1). Note that if $|\widehat{\mathbf{Q}}(\xi) - \widehat{\mathbf{P}}(\xi)| \geq R - T$, then either the real or the imaginary part is at least $(R-T)/\sqrt{2}$. By a Chernoff bound, the probability that for a given $\xi \in S \setminus \mathcal{L}'(T)$, $\Re(\widehat{\mathbf{Q}}(\xi) - \widehat{\mathbf{P}}(\xi)) \geq (R-T)/\sqrt{2}$ is at most $2\exp(-N(R-T)^2/4)$. The same is true of the imaginary part so by a union bound the probability that $|\widehat{\mathbf{Q}}(\xi) - \widehat{\mathbf{P}}(\xi)| \geq R - T$ is at most $4\exp(-N(R-T)^2/4)$. Again by a union bound we get that the probability that any $\xi \in S \setminus \mathcal{L}'(T)$ has $|\widehat{\mathbf{Q}}(\xi) - \widehat{\mathbf{P}}(\xi)| \geq R - T$ is at most $O(k^2 \log(k/\epsilon)\exp(-N(R-T)^2/4)) = O(k^2\log(k/\epsilon)\exp(-C\ln(k/\epsilon))) = O(\epsilon^{C-1})$. So, except with probability $O(\epsilon^{C-1})$, for all ξ in $S \setminus \mathcal{L}'(T)$ we have $|\widehat{\mathbf{Q}}(\xi) - \widehat{\mathbf{P}}(\xi)| < R - T$ and so $|\widehat{\mathbf{Q}}(\xi)| \leq R$. In fact, the total expected contribution to the squared L_2 norm coming from cases when $\widehat{\mathbf{H}}(\xi)$ is not identically 0 on all such ξ is also $O(\epsilon^{C-1})$. Therefore, up to negligible error, the squared L_2 error coming from this range is at most

$$\sum_{r \geq 0} (T2^{-r})^2 \left(\frac{\#\{\xi : |\widehat{\mathbf{P}}(\xi)| > T2^{-r-1}\}}{M} \right).$$

Applying Lemma 6 (ii) with $\delta := T2^{-r-1}$ for each $r \geq 0$, this is at most

$$\begin{aligned}\sum_{r \geq 0} (T2^{-r})^2 \left(\frac{\#\{\xi : |\widehat{\mathbf{P}}(\xi)| > T2^{-r-1}\}}{M} \right) &\leq \sum_{r \geq 0} T^2 4^{-r} 2ks^{-1}\sqrt{\log(2^r/T)} \\ &\leq 4T^2 ks^{-1}\sqrt{\log(1/T)}.\end{aligned}$$

Next we consider the remaining contribution

$$(1/M) \cdot \sum_{\xi \in \mathcal{L}'(T)} |\widehat{\mathbf{H}}(\xi) - \widehat{\mathbf{P}}(\xi)|^2.$$

We note by Lemma 6 (i) applied with $\delta := T$, $\mathcal{L}'(T) \subseteq \mathcal{L}(T, M, s)$. Since $\sqrt{\ln(1/T)}/2s = O(\log(k/\epsilon)/M)$, we can choose the constant in the definition of S so that $\mathcal{L}(T, M, s) \subseteq S$. So, for $\xi \in \mathcal{L}'(T)$, we do compute $\widehat{\mathbf{Q}}(\xi)$ and then either $\widehat{\mathbf{H}}(\xi) = \widehat{\mathbf{Q}}(\xi)$ or $|\widehat{\mathbf{Q}}(\xi)| < R$ and $\widehat{\mathbf{H}}(\xi) = 0$. In either case, we have that $|\widehat{\mathbf{H}}(\xi) - \widehat{\mathbf{Q}}(\xi)| < R$. Recall that the expected size of $|\widehat{\mathbf{Q}}(\xi) - \widehat{\mathbf{P}}(\xi)|^2$ is $1/N$ for any $\xi \in [M-1]$. So, for $\xi \in \mathcal{L}'(T)$, the expected squared error at ξ satisfies $\mathbb{E}[|\widehat{\mathbf{H}}(\xi) - \widehat{\mathbf{P}}(\xi)|^2] \leq 2(R^2 + N^{-1})$.

We note that by Lemma 6 (ii) applied with $\delta := T$, we have $|\mathcal{L}'(T)| \leq 2ks^{-1}\sqrt{\ln(1/T)}$. So the expected size of the L_2^2 error on $\mathcal{L}'(T)$ has

$$\mathbb{E}[(1/M) \cdot \sum_{\xi \in \mathcal{L}'(T)} |\widehat{\mathbf{H}}(\xi) - \widehat{\mathbf{P}}(\xi)|^2] \leq 2(R^2 + N^{-1})(2ks^{-1}\sqrt{\ln(1/T)})$$

Combining the above results, we find that the expected L_2^2 error between $\widehat{\mathbf{H}}$ and $\widehat{\mathbf{P}}$ is at most

$$2(R^2 + N^{-1} + T^2)(2ks^{-1}\sqrt{\log(1/T)}) = O(C^{-1}s^{-1}\epsilon^2/\sqrt{\log(k/\epsilon)}).$$

Therefore, if C is sufficiently large, the Markov inequality yields that, with probability $\frac{2}{3}$, we have $\|\widehat{\mathbf{H}} - \widehat{\mathbf{P}}\|_2^2 < \epsilon^2/M$.

At this point, we would like to use Plancherel's theorem followed by Cauchy-Schwartz to complete the proof. Formally, since $\mathbf{P}$ may be supported outside $[\lfloor\widetilde{\mu}\rfloor - (M-1)/2, \lfloor\widetilde{\mu}\rfloor + (M-1)/2]$, we cannot use Plancherel's theorem directly to show that $\|\mathbf{H} - \mathbf{P}\|_2 = \|\widehat{\mathbf{H}} - \widehat{\mathbf{P}}\|_2$. Instead, consider the function $\mathbf{P}' : [\lfloor\widetilde{\mu}\rfloor - (M-1)/2, \lfloor\widetilde{\mu}\rfloor + (M-1)/2] \cap \mathbb{Z} \to [0,1]$ defined as $\mathbf{P}'(i) = \sum_{j \equiv i \pmod M} \mathbf{P}(j)$ for $\lfloor\widetilde{\mu}\rfloor - (M-1)/2 \leq i \leq \lfloor\widetilde{\mu}\rfloor + (M-1)/2$. Note that $\widehat{\mathbf{P}'} = \widehat{\mathbf{P}}$ by the definition of the DFT modulo M, since $e(\xi j/M) = e(\xi i/M)$ when $j \equiv i \pmod M$ for all $\xi \in [M-1]$ and $i, j \in [n]$. Thus, $\|\widehat{\mathbf{H}} - \widehat{\mathbf{P}'}\|_2^2 < \epsilon^2/M$ and Plancherel's theorem gives $\|\mathbf{H} - \mathbf{P}'\|_2 = \|\widehat{\mathbf{H}} - \widehat{\mathbf{P}'}\|_2 < \epsilon/\sqrt{M}$. Since $\mathbf{P}'$ has support at most M, an application of Cauchy-Schwartz gives $\|\mathbf{H} - \mathbf{P}'\|_1 \leq \|\mathbf{H} - \mathbf{P}'\|_2\sqrt{M} < \epsilon$.

Since $X \sim \mathbf{P}$ is in $[\lfloor\widetilde{\mu}\rfloor - (M-1)/2, \lfloor\widetilde{\mu}\rfloor + (M-1)/2]$ with probability at least $1 - \epsilon/2$, we have $\|\mathbf{P} - \mathbf{P}'\|_1 \leq \epsilon$ and so $\|\mathbf{P} - \mathbf{H}\|_1 \leq \|\mathbf{P} - \mathbf{P}'\|_1 + \|\mathbf{H} - \mathbf{P}'\|_1 \leq 2\epsilon$. Since $\widehat{\mathbf{H}}(0) = \widehat{\mathbf{Q}}(0) = 1$, it follows that $\sum_{i=0}^{n} \mathbf{H}(i) = 1$. Also, by symmetry, all the $\mathbf{H}(i)$'s are real. This completes the proof of Theorem 4. □

3 Cover Size Upper Bound and Efficient Construction

We start by establishing an upper bound on the cover size and then proceed to describe our efficient algorithm for the construction of a proper cover with near minimum size. To prove the desired upper bound on the size of the cover, we proceed as follows: We start (Section 3.1) by reducing the cover size problem to the case that the order n of the k-SIIRV is at most $\text{poly}(k/\epsilon)$. In the second and main step (Section 3.2), we prove the desired upper bound for the polynomially sparse case. Our efficient algorithm for the cover construction (Section 3.3) is based on dynamic programming and follows a similar case analysis.

3.1 Reduction to Sparse Case.

Our starting point is the following theorem:

Theorem 8. *[[DDO+13], Theorem I.2] Let $\mathbf{P} \in \mathcal{S}_{n,k}$ be a k-SIIRV of order n. Then, for any $\epsilon > 0$, $\mathbf{P}$ is either*

1. *a distribution with variance at most $\text{poly}(k/\epsilon)$; or*
2. *ϵ-close to a distribution $\mathbf{P}'$ such that for a random variable $X \sim \mathbf{P}'$, we have $X = cZ + Y$ for some $1 \leq c \leq k-1$, where Y, Z are independent random variables such that: (i) Y is distributed as a c-IRV, and (ii) Z is a discretized normal random variable with parameters $\frac{\mu}{c}, \frac{\sigma^2}{c^2}$ where $\mu = \mathbb{E}[X]$ and $\sigma^2 = \text{Var}[X]$.*

The above theorem allows us to reduce the problem of constructing an $O(\epsilon)$-cover for $\mathcal{S}_{n,k}$ to the problem of constructing an ϵ-cover for $\mathcal{S}_{n',k}$, where $n' = \text{poly}(k/\epsilon)$. Indeed, given an arbitrary k-SIIRV $\mathbf{P} \in \mathcal{S}_{n,k}$ we proceed as follows: If $\mathbf{P}$ belongs to Case 1 of the above theorem, then we show (Lemma 9) that there exists a translation of a k-SIIRV with $n' = \text{poly}(k/\epsilon)$ variables that is ϵ-close to $\mathbf{P}$. We show in the following subsection (Proposition 10) that $\mathcal{S}_{n',k}$ admits an ϵ-cover of size $(1/\epsilon)^{O(k\log(1/\epsilon))}$. Since there are $O(kn)$ possible translations, this gives a 2ϵ-cover of size $n(1/\epsilon)^{O(k\log(1/\epsilon))}$ for k-SIIRVs in Case 1.

Moreover, it is not difficult to show that there exists an ϵ-cover for distributions in Case 2 with at most $n \cdot (k/\epsilon)^{O(k)}$ points. In particular, we claim that for distributions in sub-case 2(i) there exists an ϵ-cover of size $(1/\epsilon)^{O(k)}$, and for distributions in sub-case 2(ii) there exists an ϵ-cover of

size $O(n)$. Assuming these claims, the sub-additivity of total variation distance (Proposition 41) implies that distributions in Case 2 have a 2ϵ-cover of size $n \cdot (1/\epsilon)^{O(k)}$ as desired.

Note that the random variable Y in Case 2(i) is distributed as a k-IRV, i.e., it has support k. It is well-known and easy to show that the set of all distributions over a domain of size k has an ϵ-cover of size $(1/\epsilon)^{O(k)}$. It remains to show that we can ϵ-cover the set of discretized normal distributions of Case2(ii) with $O(nk/\epsilon)$ points. To do this, we exploit the fact that the variance of such distributions is large. Let $\sigma_{\min} = \Omega(k^9/\epsilon^3)$ be the minimum variance of a k-SIIRV X in Case 2. Note that the discrete Gaussian in Case 2 has a variance of $\mathrm{Var}[X]/c^2$. Hence, we want to ϵ-cover the set of discrete Gaussians with standard deviation σ in the interval $[\sigma_{\min}, \sigma_{\max}]$, where $\sigma_{\max} = O(\sqrt{n}k)$, and mean value μ in the interval $[0, n(k-1)]$. Consider the following discretization of the space (σ^2, μ): We first define a geometric grid on σ^2 with ratio $(1+\epsilon)$, i.e., $\sigma_i^2 = \sigma_{\min}^2(1+\epsilon)^i$, where where $0 \leq i \leq i_{\max}$ and $i_{\max} = O((1/\epsilon)\cdot\log(n))$. For every fixed i, we define an additive grid on the means, so that $|\mu_{j+1} - \mu_j| \leq \epsilon \cdot \sigma_i$. A combination of Propositions 40 and 42 implies that this grid defines an ϵ-cover. Note that the total size of the described grid on (σ^2, μ) is

$$\sum_{i=0}^{i_{\max}} \frac{n(k-1)}{\epsilon \cdot \sigma_i} = \sum_{i=0}^{i_{\max}} \frac{n(k-1)}{\epsilon \cdot \sigma_{\min}(1+\epsilon)^{i/2}} = O(n),$$

where the last inequality follows from the lower bound on $\sigma_{\min}$ and the elementary inequality $\sum_i (1+\epsilon)^{-i/2} = O(1/\epsilon)$.

The following lemma completes our reduction to the $n = \mathrm{poly}(k/\epsilon)$ case:

Lemma 9. *Let $\mathbf{P} \in \mathcal{S}_{n,k}$ be a k-SIIRV with $\mathrm{Var}_{X\sim\mathbf{P}}[X] = V$. For any $0 < \delta < 1/4$, there exists $\mathbf{Q} \in \mathcal{S}_{n,k}$ with $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) = O(\delta V)$ such that all but $O(k + V/\delta)$ of the k-IRV's defining $\mathbf{Q}$ are constant.*

The proof of Lemma 9 is deferred to Appendix B.1. Note that an application of the lemma for $\delta = \epsilon/V$ completes the proof.

3.2 Cover Upper Bound for Sparse Support.

In this subsection we prove the desired upper bound on the cover size for the sparse case:

Proposition 10. *Fix arbitrary constants $c, C > 0$. Consider n, k, ϵ satisfying $\epsilon \leq k^{-c}$ and $n \leq (k/\epsilon)^C$. Then there exists an ϵ-cover of $\mathcal{S}_{n,k}$ under d_{TV} of size $(1/\epsilon)^{O_{c,C}(k\log(1/\epsilon))}$.*

Our proof proceeds by analyzing the Fourier transform of the probability density functions of k-SIIRVs. We will need the following definitions.

Basic Definitions. For $\xi \in \mathbb{R}$, recall that we use the notation $e(\xi) \stackrel{\text{def}}{=} \exp(2\pi i\xi)$. For a probability density function (pdf) $\mathbf{P}$ over $\mathbb{R}$, its Fourier Transform is the function $\widehat{\mathbf{P}} : [0,1) \to \mathbb{C}$ defined by $\widehat{\mathbf{P}}(\xi) = \mathbb{E}_{y\sim\mathbf{P}}[\exp(-2\pi i y\xi)] = \mathbb{E}_{y\sim\mathbf{P}}[e(-y\xi)]$. Note that Parseval's identity states that for two pdf's $\mathbf{P}$ and $\mathbf{Q}$ we have $\|\mathbf{P} - \mathbf{Q}\|_2 = \|\widehat{\mathbf{P}} - \widehat{\mathbf{Q}}\|_2$. In our context, $\mathbf{P}$ and $\mathbf{Q}$ are going to be supported on a discrete set A, in which case we have $\|\mathbf{P} - \mathbf{Q}\|_2 = \left(\sum_{a\in A}(\mathbf{P}(a) - \mathbf{Q}(a))^2\right)^{1/2}$. On the other hand, $\widehat{\mathbf{P}}$ and $\widehat{\mathbf{Q}}$ are Lebesgue measuable and we have $\|\widehat{\mathbf{P}} - \widehat{\mathbf{Q}}\|_2 = \left(\int_0^1 |\widehat{\mathbf{P}}(\xi) - \widehat{\mathbf{Q}}(\xi)|^2 d\xi\right)^{1/2}$.

An equivalent way to view the Fourier transform is as a function defined on the unit circle in the complex plane. For our purposes, we will need to analyze the corresponding polynomial defined over the entire complex plane. Namely, we will consider the probability generating function $\widetilde{\mathbf{P}} : \mathbb{C} \to \mathbb{C}$ of $\mathbf{P}$ defined as $\widetilde{\mathbf{P}}(z) = \mathbb{E}_{y\sim\mathbf{P}}[z^y]$. Note that when $|z| = 1$, this function agrees with the Fourier transform, i.e., $\widehat{\mathbf{P}}(\xi) = \widetilde{\mathbf{P}}(e(-\xi))$.

At a high-level, our proof is conceptually simple: For a k-SIIRV $\mathbf{P}$, we would like to show that the logarithm of its Fourier transform $\log \widehat{\mathbf{P}}(\xi)$ is determined up to an additive ϵ by its degree $O(\log(1/\epsilon))$ Taylor polynomial. Assuming this holds, it is relatively straightforward to prove the desired upper bound on the cover size. Unfortunately, such a statement cannot be true in general for the following reason: the function $\widetilde{\mathbf{P}}(z)$ may have roots near (or on) the unit circle, in which case the logarithm of the Fourier transform is either very big or infinite at certain points. Intuitively, we would like to show that the magnitude of $\widetilde{\mathbf{P}}(z)$ close to a root is small. Unfortunately, this is not necessarily true.

We circumvent this problem as follows: We partition the unit circle into $O(k)$ arcs each of length $O(1/k)$. We perform a case analysis based on the number of roots that are close to an arc. If there are at least $\Omega(\log(1/\epsilon))$ roots of $\widetilde{\mathbf{P}}(z)$ close to a particular arc, then we show (Lemma 12(i)) that the magnitude of $\widetilde{\mathbf{P}}(z)$ within the arc is going to be negligibly small. Otherwise, we consider the polynomial $q(z)$ obtained by $\widetilde{\mathbf{P}}(z)$ after dividing by the corresponding roots, and show that $\log q(z)$ is determined up to an additive ϵ by its degree $O(\log(1/\epsilon))$ Taylor polynomial within the arc (see Lemma 13). Using the aforementioned structural understanding, to prove the cover upper bound, we define a "succinct" description of the Fourier Transform based on the logarithm of $q(z)$ and appropriate discretization of $O(\log(1/\epsilon))$ nearby roots.

Note that we take advantage of the fact that our distributions are supported over a domain of size $\ell = \mathrm{poly}(k/\epsilon)$, in order to relate their total variation distance to the L_∞ distance between their Fourier transforms. In particular, we have the following simple fact:

Fact 11. *For any pair of pdfs $\mathbf{P}, \mathbf{Q}$ over $[\ell]$, we have $\|\mathbf{P} - \mathbf{Q}\|_1 \le \sqrt{\ell+1}\|\widehat{\mathbf{P}} - \widehat{\mathbf{Q}}\|_\infty$.*

Indeed, note that $\|\mathbf{P} - \mathbf{Q}\|_1 \le \sqrt{\ell+1}\|\mathbf{P} - \mathbf{Q}\|_2 = \sqrt{\ell+1}\|\widehat{\mathbf{P}} - \widehat{\mathbf{Q}}\|_2 \le \sqrt{\ell+1}\|\widehat{\mathbf{P}} - \widehat{\mathbf{Q}}\|_\infty$, where the equality is Parseval's identity.

For the rest of this section we fix an arbitrary $\mathbf{P} \in \mathcal{S}_{n,k}$ and analyze the polynomial $\widetilde{\mathbf{P}}(x)$. We start with the following important lemma whose proof is deferred to Appendix B.2:

Lemma 12. *Fix $x \in \mathbb{C}$ with $|x| = 1$. Suppose that $\rho_1, \ldots, \rho_m$ are roots of $\widetilde{\mathbf{P}}(x)$ (listed with appropriate multiplicity) which have $|\rho_i - x| \le \frac{1}{2k}$. Then, we have the following:*

(i) $|\widetilde{\mathbf{P}}(x)| \le 2^{-m}$.

(ii) For the polynomial $q(x) = \widetilde{\mathbf{P}}(x)/\prod_{i=1}^m (x - \rho_i)$, we have that $|q(x)| \le k^m$.

Our main lemma for this section shows that we can ϵ-approximate the Taylor series of $q(x)$ by only considering the first $O(\log(1/\epsilon))$ terms:

Lemma 13. *Fix $w \in \mathbb{C}$ with $|w| = 1$. Suppose that $\rho_1, \ldots, \rho_m$ are all the roots of $\widetilde{\mathbf{P}}(x)$ (listed with appropriate multiplicity) which have $|\rho_i - w| \le \frac{1}{3k}$. Let $q(x) = \frac{\widetilde{\mathbf{P}}(x)}{\prod_{i=1}^m (x-\rho_i)}$ and let the Taylor series of $\ln(q(x))$ at w be $\ln q(x) = \sum_{j=0}^\infty c_j (x-w)^j$. Then, we have that $|c_j| \le nk(3k)^j$, for all $j \ge 1$, and the real part of c_0 is at most $m \ln k$.*

Fix $0 < \epsilon \le 1/(12mk)$ and an integer ℓ satisfying $\ell \ge \log(9nk)$. For ρ'_j with $|\rho'_j - \rho_j| \le \epsilon$ for $j \in \{1, \ldots, m\}$, and c'_j with $|c'_j - c_j| \le \epsilon$ for $j \in \{1, \ldots, \ell\}$ we have: For all $x \in \mathbb{C}$ with $|x| = 1$ and $|x - w| \le \frac{1}{6k}$

$$\left| \widetilde{\mathbf{P}}(x) - \Big(\prod_{j=1}^m (x - \rho'_j) \Big) \exp\Big(\sum_{j=0}^\ell c'_j (x-w)^j \Big) \right| \le O\left(\epsilon m k + n k 2^{-\ell} \right). \tag{2}$$

Proof. We start by noting that, by the triangle inequality, Lemma 12 applies to all points $x \in \mathbb{C}$ with $|x| = 1$ and $|x - w| \leq \frac{1}{6k}$. Observe that $c_0 = \ln[q(w)]$ and by Lemma 12(ii) $|q(w)| \leq k^m$. This gives the claim on the real part of c_0.

Note that $\ln(q(x))$ can be expressed as a sum of the form

$$\ln(q(x)) = c_0 + \sum_{h=1}^{R} \ln(1 - (x-w)/(r_h - w)) \ ,$$

where $c_0 = \ln[q(w)]$, r_j are the roots of $q(x)$, and $R \leq n(k-1)$ is the degree of $q(x)$. By the definition of q, it follows that $|r_h - w| > \frac{1}{3k}$ for all $1 \leq h \leq R$.

Inserting the standard Taylor series $\ln(1+y) = \sum_{j=0}^{\infty} \frac{y^j}{j}$ gives

$$\ln(q(x)) = c_0 + \sum_{h=1}^{R} \sum_{j=0}^{\infty} \frac{(-1)^j (x-w)^j}{j \cdot (r_h - w)^j}.$$

Considering the $(x-w)^j$ term above gives $c_j = \frac{(-1)^j}{j} \sum_{j=1}^{R} (r_j - w)^{-j}$. Therefore,

$$|c_j| \leq R(3k)^j \leq nk(3k)^j \ .$$

This gives the desired bound on $|c_j|$, $j \geq 1$.

We now proceed to prove (2). We start by considering the difference

$$\sum_{j=0}^{\ell} c'_j (x-w)^j - \ln(q(x)) \ ,$$

for x in the appropriate range. Since $|x - w| \leq \frac{1}{6k} \leq 1/2$ and $|c'_j - c_j| \leq \epsilon$, we have

$$\left| \sum_{j=0}^{\ell} c'_j (x-w)^j - \sum_{j=0}^{\ell} c_j (x-w)^j \right| \leq \epsilon \cdot \sum_{j=0}^{\ell} 2^{-j} \leq 2\epsilon \ .$$

So, we need to consider the error introduced by truncating the Taylor series after the first ℓ terms. We have

$$\begin{aligned} \left| \sum_{j=0}^{\ell} c_j (x-w)^j - \ln(q(x)) \right| &= \left| \sum_{j>\ell} c_j (x-w)^j \right| \\ &\leq \sum_{j>\ell} nk(3k)^j (6k)^{-j} \\ &= nk2^{-\ell} \end{aligned}$$

Therefore, by the triangle inequality,

$$\left| \sum_{j=0}^{\ell} c'_j (x-w)^j - \log(q(x)) \right| \leq 2\epsilon + nk2^{-\ell}.$$

Thus, the multiplicative error in this approximation, i.e.,

$$\frac{1}{q(x)} \exp\left(\sum_{j=0}^{\ell} c'_j (x-w)^j \right) = \frac{1}{\widetilde{\mathbf{P}}(x)} \left(\prod_{j=1}^{m} (x - \rho_j) \right) \exp\left(\sum_{j=0}^{\ell} c'_j (x-w)^j \right)$$

is between $\exp(-(2\epsilon + nk2^{-\ell}))$ and $\exp(2\epsilon + nk2^{-\ell})$. Since $|\widetilde{\mathbf{P}}(x)| \leq 1$ and by our assumptions on ℓ, $2\epsilon + nk2^{-\ell} \leq 1$, we have that

$$\left| \widetilde{\mathbf{P}}(x) - \left(\prod_{j=1}^{m} (x - \rho_j) \right) \exp\left(\sum_{j=0}^{\ell} c'_j (x-w)^j \right) \right| \leq e \cdot (2\epsilon + nk2^{-\ell}).$$

We next replace each ρ_j by the corresponding ρ'_j one at a time. By a simple induction, we will show that for all $1 \leq h \leq m$

$$\left| \widetilde{\mathbf{P}}(x) - \left(\prod_{j=1}^{h} (x - \rho'_j) \right) \left(\prod_{j=h+1}^{m} (x - \rho_j) \right) \exp\left(\sum_{j=0}^{\ell} c'_j (x-w)^j \right) \right| \leq e \cdot (2\epsilon + nk2^{-\ell}) + 4hk\epsilon. \quad (3)$$

We have just shown this for $h = 0$. So, we assume (3) for $0 \leq h \leq m-1$ and seek to prove it for $h+1$. For simplicity, we rewrite (3) as

$$\left| \widetilde{\mathbf{P}}(x) - (x - \rho_h) f_h(x) \right| \leq e \cdot (2\epsilon + nk2^{-\ell}) + 4hk\epsilon$$

where $f_h(x) = \left(\prod_{j=1}^{h-1} (x - \rho'_j) \right) \left(\prod_{j=h+1}^{m} (x - \rho_j) \right) \exp\left(\sum_{j=0}^{\ell} c'_j (x-w)^j \right)$.

Note that the RHS of (3) satisfies

$$e \cdot (2\epsilon + nk2^{-\ell}) + 4hk\epsilon \leq e \cdot (2\epsilon + nk2^{-\ell}) + 4mk\epsilon \leq 1 \ ,$$

by our assumptions on ϵ and ℓ. Since $|\widetilde{\mathbf{P}}(x)| \leq 2^{-m} \leq 1$, we have $|(x - \rho_h) f_h(x)| \leq 2$ or $|f_h(x)| \leq \frac{2}{|x - \rho_h|} \leq 4k$. Now if we replace $(x - \rho_h) f_h(x)$ with $(x - \rho'_h) f_h(x)$, we introduce an error of $|(x - \rho_h) f_h(x) - (x - \rho'_h) f_h(x)| = |\rho'_h - \rho_h| |f_h(x)| \leq \epsilon \cdot 4k$. Hence,

$$\left| \widetilde{\mathbf{P}}(x) - (x - \rho'_h) f_h(x) \right| \leq e \cdot (\ell\epsilon + nk2^{-\ell}) + 4(h+1)k\epsilon$$

But this is just (3) for $h+1$, completing the induction.

Taking $h = m$ in (3) gives:

$$\left| \widetilde{\mathbf{P}}(x) - \left(\prod_{j=1}^{m} (x - \rho'_j) \right) \exp\left(\sum_{j=0}^{\ell} c'_j (x-w)^j \right) \right| \leq e \cdot (2\epsilon + nk2^{-\ell}) + 4mk\epsilon$$

as required. □

We are now prepared to prove Proposition 10.

Proof of Proposition 10. By replacing ϵ by a power of itself, we may assume that $\epsilon \leq k^{-1}$ and that $n \leq \epsilon^{-1}$. We may additionally assume that ϵ is sufficiently small.

It suffices to find a subset T of $\mathcal{S}_{n,k}$ of appropriate size so that for any $\mathbf{P} \in \mathcal{S}_{n,k}$ there is some $\mathbf{Q} \in T$ so that $|\widetilde{\mathbf{P}}(z) - \widetilde{\mathbf{Q}}(z)| \leq \epsilon^2$ for all $|z| = 1$, as Fact 11 would then imply that $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) \leq \epsilon$.

We begin by defining some parameters. Let m be an integer larger than $3\log(1/\epsilon)$. Let ℓ be an integer larger than $\log(nk/\epsilon^3)$ and $\delta > 0$ a real number smaller than $\epsilon^3/(mk + \ell)$. Additionally, we divide the unit circle of $\mathbb{C}$ into $O(k)$ arcs each of length at most $1/(3k)$.

To each $\mathbf{P} \in \mathcal{S}_{n,k}$ we associate the following data:

- For each arc in our partition with midpoint w_I, define $q(z)$ as in Lemma 13. Then we define $\mathbf{P}_I$ as follows:
  - If $\widetilde{\mathbf{P}}(z)$ has at least m roots within distance $1/(3k)$ of w_I or if $|q(w_I)| < \epsilon^3 \exp(-nk)$, we let $\mathbf{P}_I = \mathbf{Small}$.
  - Otherwise, we let $\mathbf{P}_I$ consist of the following data:
    * Roundings of the roots of $\widetilde{\mathbf{P}}(z)$ that are within $1/(3k)$ of w_I to the nearest complex numbers whose real and imaginary parts are multiples of $\delta/2$.
    * Roundings of the first ℓ Taylor coefficients of $\log(q)$ about w_I to the nearest complex numbers whose real and imaginary parts are multiples of $\delta/2$.

We then let $D(\mathbf{P})$ be the sequence $\{\mathbf{P}_I\}_{I \text{ an arc in the partition}}$. For each value V that can be obtained as $D(\mathbf{P})$ for some $\mathbf{P} \in \mathcal{S}_{n,k}$, we pick one such $\mathbf{P}$ called $\mathbf{Q}_V$. We define our cover T to be the set of all such $\mathbf{Q}_V$. In order to show that this is an appropriate cover, we need to show two claims:

1. The number of possible values of $D(\mathbf{P})$ is at most $(1/\epsilon)^{O(k \log(1/\epsilon))}$. This implies that $|T|$ is appropriately small.
2. If $\mathbf{P}, \mathbf{Q} \in \mathcal{S}_{n,k}$ have $D(\mathbf{P}) = D(\mathbf{Q})$, then $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) \leq \epsilon$. This will imply that T is a cover, since given any $\mathbf{P} \in \mathcal{S}_{n,k}$, we may take $\mathbf{Q} = \mathbf{Q}_{D(\mathbf{P})} \in T$.

The first claim is relatively straightforward. For each of $O(k)$ arcs, I, we have that $\mathbf{P}_I$ is either **Small** or a sequence of $O(\log(1/\epsilon))$ complex numbers, each of which can take only $\mathrm{poly}(1/\delta)$ many possible values. Thus, the number of possible values for $\mathbf{P}_I$ is at most $\delta^{-O(\log(1/\epsilon))} = (1/\epsilon)^{O(\log(1/\epsilon))}$. The number of possible values for $D(\mathbf{P})$ is at most this raised to the number of arcs, which is $(1/\epsilon)^{O(k \log(1/\epsilon))}$.

The second claim is slightly more involved. We note that it is sufficient to show that if $D(\mathbf{P}) = D(\mathbf{Q})$, then $|\tilde{\mathbf{P}}(z) - \tilde{\mathbf{Q}}(z)| \leq \epsilon^2$ for all unit norm z. In particular, we show the stronger claim that for any of our arcs I if $\mathbf{P}_I = \mathbf{Q}_I$, then $|\widetilde{\mathbf{P}}(z) - \widetilde{\mathbf{Q}}(z)| = O(\epsilon^3)$ for all $z \in I$.

If $\mathbf{P}_I = \mathbf{Q}_I = \mathbf{Small}$, we claim that $|\tilde{\mathbf{P}}(z)|, |\tilde{\mathbf{Q}}(z)| = O(\epsilon^3)$ for all $z \in I$. It suffices to show this merely for $\mathbf{P}$. On the one hand, if $\widetilde{\mathbf{P}}(z)$ has more than m roots near w_I, this follows from the first part of Lemma 12. On the other hand, if $|q(w_I)| \leq \epsilon^3 \exp(-nk)$, then for any other $z \in I$ we have that

$$q(z) = q(w_I) \exp\left(\sum_{i=1}^{\infty} c_i (z - w_I)^i \right),$$

where by Lemma 13, $|c_i| \leq nk(3k)^i$. Therefore, for $z \in I$, since $|z - w_I| \leq 1/(6k)$, we have by Lemma 12 that

$$|\widetilde{\mathbf{P}}(z)| \leq |q(z)| \leq |q(w_I)| \exp(nk) \leq \epsilon^3.$$

If $\mathbf{P}_I = \mathbf{Q}_I \neq \mathbf{Small}$, we note by Lemma 13 that for $z \in I$ that both of $\widetilde{\mathbf{P}}(z)$ and $\widetilde{\mathbf{Q}}(z)$ are within $O(mk\delta + \ell\delta + nk2^{-\ell}) = O(\epsilon^3)$ of $\prod_{j=1}^{M}(z - \rho'_j) \exp\left(\sum_{j=0}^{\ell} c'_j (z - w_I)^j\right)$, where the ρ'_j are the roundings of nearby roots and c'_j the roundings of the Taylor coefficients given by the data $p_I = q_I$. Thus, again in this case, $|\widetilde{\mathbf{P}}(z) - \widetilde{\mathbf{Q}}(z)| \leq O(\epsilon^3)$ for all $z \in I$.

This completes the proof of Proposition 10.

□

3.3 Efficient Cover Construction. In this section, we give an algorithm to construct a near-minimum size cover in output polynomial time:

Theorem 14. *Let n, k be positive integers and $\epsilon > 0$. There exists an algorithm that runs in time $n\,(k/\epsilon)^{O(k\log(1/\epsilon))}$ and returns a proper ϵ-cover for $\mathcal{S}_{n,k}$, i.e., a cover consisting of $n\,(k/\epsilon)^{O(k\log(1/\epsilon))}$ k-SIIRVs each given as an explicit sum of k-IRVs.*

Our algorithm builds on the existential upper bound established in the previous subsections. We first construct an ϵ-cover for k-SIIRVs in Case 2 of Theorem 8, i.e., k-SIIRVs whose variance is more than a sufficiently large polynomial in k/ϵ. By Theorem 8 each such k-SIIRV is ϵ-close to a random variable of the form $cZ+Y$, where $1 \le c \le k-1$ is an integer, Z is a discrete Gaussian and Y is a c-IRV. In Section 3.1 we exploited this structural fact to construct a non-proper cover for k-SIIRVs in this case. We remark that this non-proper cover may contain "spurious" points, i.e., points not close to a large variance k-SIIRV. Efficiently constructing a proper cover without spurious points for the high variance case requires careful arguments and is deferred to Appendix B.3.

We now focus our attention to Case 1. By Lemma 9, we have that all such k-SIIRVs can be approximated by a constant plus a sum of $\mathrm{poly}(k/\epsilon)$ k-IRVs. Since there are only nk possibilities for this constant, and all such possibilities are easily obtainable, it suffices to find an explicit ϵ-cover for $\mathcal{S}_{n,k}$ when $n = \mathrm{poly}(k/\epsilon)$.

A simple but useful observation is that we can round each coordinate probability for each of our k-IRVs to a multiple of $\epsilon/(nk)$ and introduce an error of $O(\epsilon)$ in total variation distance. Therefore, it suffices to find a cover of $\mathcal{S}'_{n,k}$, a sum of $n = \mathrm{poly}(k/\epsilon)$ independent k-IRVs, where each of their coordinate probabilities is a multiple of $\frac{1}{N}$ for some integer $N = \mathrm{poly}(k/\epsilon)$. We will henceforth call such a k-IRV *N-discrete k-IRV*.

Our main workhorse here will once again be Lemma 13. The cover we construct will be much the same as in Proposition 10, but we will now explicitly produce SIIRVs that obtain every possible value of D. Fortunately, the Taylor series of the log of the Fourier transform is additive in the composite k-IRVs, and so there exists an appropriate dynamic program to solve this problem.

Let $\delta > 0$ be given by a sufficiently small polynomial in ϵ/k, and let m be an integer at least a sufficiently large multiple of $\log(1/\epsilon)$. We divide the unit circle into arcs I with midpoints w_I as described in the proof of Proposition 10. For any N-discrete k-IRV, $\mathbf{P}$, we associate the following data. For each interval I, let $\rho_{1,I}, \ldots, \rho_{r_I,I}$ be the roots of $\widetilde{\mathbf{P}}$ that are within distance $1/(3k)$ of w_I, and let $q(z) = \frac{\widetilde{\mathbf{P}}(z)}{\prod(z-\rho_{i,I})}$. For $1 \le j \le r_I$, let $\rho'_{j,I}$ be a rounding of $\rho_{j,I}$ with $\rho'_j, I = (a+bi)\delta$ for some $a, b \in \mathbb{Z}$ and $|\rho'_{j,I} - \rho_{j,I}| \le \delta$. For $1 \le j \le m$, let $c'_{j,I}$ be a rounding of $c_{j,I}$ with $c'_{j,I} = (a+bi)\delta$ for some $a, b \in \mathbb{Z}$ and $|c'_{i,I} - c_{i,I}| \le \delta$, where the $c_{k,I}$ are the coefficients of first $m+1$ terms of the Taylor series $\ln q(z) = \sum_{j=0}^{\infty} c_j (z - w_I)^j$. Let $\mathbf{P}_I$ be the data consisting of the list $(\rho'_{1,I}, \ldots, \rho'_{r_I,I})$ and the vector $(c'_{0,I}, c'_{1,I}, \ldots, c'_{m,I})$. We let $D(\mathbf{P})$ be the sequence of $\mathbf{P}_I$ over all intervals I.

Given a sequence $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_h$ of k-IRVs, we let $D(\mathbf{P}_1, \ldots, \mathbf{P}_k)$ be given by the following data for each I:

- The first m elements of the concatenation of the lists of approximate roots of $\prod_{i=1}^{h} \widetilde{\mathbf{P}_i}(z)$ near w_I.
- The list of elements $\sum_{i=1}^{h} c'_{j,I}(\mathbf{P}_i)$ for $0 \le j \le m$, with the exception that the $j = 0$ term is replaced by $-\infty$ if for any $h' < h$ we have that the real part of $\sum_{i=1}^{h'} c'_{0,I}(\mathbf{P}_i)$ is less than $-nk - m - m\ln k$.

Our algorithm will follow from three important claims:

Claim 15. *We have the following:*

(i) $D(\mathbf{P}_1, \ldots, \mathbf{P}_h)$ *can be computed in* $\text{poly}(k/\epsilon)$ *time from* $D(\mathbf{P}_1, \ldots, \mathbf{P}_{h-1})$ *and* $D(\mathbf{P}_h)$.

(ii) There are only $(k/\epsilon)^{O(k \log(1/\epsilon))}$ *possible values for* $D(\mathbf{P}_1, \ldots, \mathbf{P}_h)$ *for any* $h \leq n$.

(iii) If $D(\mathbf{P}_1, \ldots, \mathbf{P}_n) = D(\mathbf{Q}_1, \ldots, \mathbf{Q}_n)$ *and* $\mathbf{P}$,$\mathbf{Q}$ *are the distributions of* $\sum_{i=1}^n X_i$ *and* $\sum_{i=1}^n Y_i$ *for* $X_i \sim \mathbf{P}_i$ *and* $Y_i \sim \mathbf{Q}_i$ *then* $d_{\text{TV}}(\mathbf{P}, \mathbf{Q}) \leq \epsilon$.

Proof. The first statement follows from the fact that the lists of roots in $D(\mathbf{P}_1, \ldots, \mathbf{P}_h)$ are obtained by concatenating those in $D(\mathbf{P}_1, \ldots, \mathbf{P}_{h-1})$ with those in $D(\mathbf{P}_h)$, and truncating if necessary. And moreover that $\sum_{i=1}^h c'_{j,I}(\mathbf{P}_i)$ is obtained by adding $c'_{j,I}(\mathbf{P}_h)$ to $\sum_{i=1}^{h-1} c'_{j,I}(\mathbf{P}_i)$ (with the term remaining $-\infty$ if it was in $D(\mathbf{P}_1, \ldots, \mathbf{P}_{h-1})$).

For the second statement note that for each of the $O(I)$ intervals, we store $O(\log(1/\epsilon))$ complex numbers whose real and imaginary parts are each multiples of δ. As each of these numbers (with the exception of a $-\infty$ term) have size at most $\text{poly}(k/\epsilon)$ and $\delta = \text{poly}(\epsilon/k)$, there are only $\text{poly}(k/\epsilon)^{O(k \log(1/\epsilon))}$ many possible values for $D(\mathbf{P}_1, \ldots, \mathbf{P}_h)$.

The third statement is true for essentially the same reasons as in the proof of Proposition 10. Once again, we simply need to show that for each interval I it holds $|\widetilde{\mathbf{P}}(z) - \widetilde{\mathbf{Q}}(z)| \leq (\epsilon/k)^c$ for all $z \in I$ and c a sufficiently large constant. Note that the listed roots are simply δ-approximations of the (first m) roots of $\widetilde{\mathbf{P}}$ and $\widetilde{q}$ within distance $1/(3k)$ of w_I, and the $\sum_{i=1}^n c'_{j,I}(\mathbf{P}_i)$ are within distance $n\delta$ of the coefficients of the Taylor expansion of the logarithm of $q(z)$ about w_I. If we have m nearby roots, both $\widetilde{\mathbf{P}}$ and $\widetilde{\mathbf{Q}}$ are small for all z in this range. Otherwise, unless there is a $-\infty$ in $D(\mathbf{P}) = D(\mathbf{Q})$, they are close by Lemma 13. If we do have a $-\infty$ then

$$\Re\left(\sum_{i=1}^{h'} c'_{0,I}(\mathbf{P}_i)\right) < -nk - m - m \ln k$$

for some $h' \leq h$. Since the later $c_{0,I}(\mathbf{P}_i)$ and $c_{0,I}(\mathbf{Q}_i)$ have $\Re c_{0,I}(\mathbf{P}_i) \leq m_i \ln k$ and $\Re c_{0,I}(\mathbf{P}_i) \leq m_i \ln k$ by Lemma 13, this means that $|q(w_I)| < e^{-m} e^{-nk}$, and as in Proposition 10, this implies that both $\widetilde{\mathbf{P}}$ and $\widetilde{\mathbf{Q}}$ are sufficiently small. □

We can now present the algorithm for producing our cover. The basic idea is to use a dynamic program to come up with one representative collection of $\mathbf{P}_1, \ldots, \mathbf{P}_h$ to obtain each achievable value of D. The algorithm is as follows:

Algorithm `Cover-SIIRV`
Input: $k, \epsilon > 0$ and $n, N = \text{poly}(k/\epsilon)$.

1. Define δ and m as above.
2. Let $L_0 = \{(D(\emptyset), \emptyset)\}$.
3. For $h = 1$ to n
4. Let L_h be the set of terms of the form $(D(\mathbf{P}_1, \ldots, \mathbf{P}_h), (\mathbf{P}_1, \ldots, \mathbf{P}_h))$ where $(D(\mathbf{P}_1, \ldots, \mathbf{P}_{h-1}), (\mathbf{P}_1, \ldots, \mathbf{P}_{h-1})) \in L_{h-1}$ and $\mathbf{P}_h$ is an N-discrete k-IRV.
5. Use a hash table to remove from L_h all but one term with each possible value of $D(\mathbf{P}_1, \ldots, \mathbf{P}_h)$
6. End for
7. Return the list of distributions $\sum_{i=1}^{n} X_i$ with $X_i \sim \mathbf{P}_i$ for each $(D(\mathbf{P}_1, \ldots, \mathbf{P}_n), (\mathbf{P}_1, \ldots, \mathbf{P}_n)) \in L_n$.

To prove that this produces a cover, we claim by induction on h that L_h contains an element that achieves each possible value of $D(\mathbf{P}_1, \ldots, \mathbf{P}_h)$. This is clearly true for $h = 0$. Given that it holds for $h-1$, Claim 15(i) implies that the non-deduped version of L_h also satisfies this property, and deduping clearly does not destroy it. Therefore L_n contains (exactly one) element for each possible value of $D(\mathbf{P}_1, \ldots, \mathbf{P}_n)$. Therefore, by Claims 15(ii) and (iii), the algorithm will return a cover of the appropriate size. For the runtime, we note that the initial size of L_h before deduping is the product of the size of L_{h-1} and the number of N-discrete k-IRVs, which by Claim 15(ii) is $\text{poly}(k/\epsilon)^{k \log(1/\epsilon)}$. Each of these elements are generated in $\text{poly}(k/\epsilon)$ time, and the deduping process takes only polynomial time per element. Therefore, the final runtime is $\text{poly}(k/\epsilon)^{k \log(1/\epsilon)}$. This completes the proof of Theorem 14.

4 Cover Size Lower Bound

In this section we prove our lower bound on the cover size of k-SIIRVs. In Section 4.1 we show the desired lower bound for the case of PBDs ($k = 2$). In Section 4.2 we generalize this construction for general k-SIIRVs.

4.1 Cover Size Lower Bound for PBDs.

We start by providing an explicit lower bound on the cover size of PBDs. In particular, we show the following:

Theorem 16. *For all $0 < \epsilon \leq e^{-42}$ and $n \in \mathbb{Z}$ such that $7 \leq n \leq \frac{1}{6}\ln(1/\epsilon)$, there is an ϵ-packing of $\mathcal{S}_{n,2}$ under d_{TV} with cardinality $(1/\epsilon)^{\Omega(n)}$.*

We begin with the following useful lemma:

Lemma 17. *Let $\mathbf{P}$ and $\mathbf{Q}$ be PBDs given by parameters p_i and q_i for $1 \leq i \leq n$, for some $n \geq 7$. Suppose that for all i, $1 \leq i \leq n$, it holds $|p_i - i/(n+1)| \leq 1/4(n+1)$ and $|q_i - i/(n+1)| \leq 1/4(n+1)$. Then,*

$$d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) \geq \max_i |p_i - q_i| \cdot e^{-3n}.$$

Proof. Let $\epsilon = |p_i - q_i| e^{-3n}$. For a distribution $\mathbf{P}$ supported on $[n]$, define $r_{\mathbf{P}}(p)$ to be the polynomial

$$r_{\mathbf{P}}(p) = \mathbb{E}_{X \sim \mathbf{P}} \left[(p-1)^X \cdot p^{n-X} \right] = \sum_{i=0}^{n} \mathbf{P}(i)(p-1)^i p^{n-i}.$$

For a PBD $\mathbf{P} \in \mathcal{S}_{n,2}$ and $X \sim \mathbf{P}$ with $X = \sum_{i=1}^{n} X_i$ for $X_i \sim \mathrm{Ber}(p_i)$, we have that

$$\begin{aligned}
r_{\mathbf{P}}(p) &= \mathbb{E}\left[(p-1)^X p^{n-X}\right] \\
&= \mathbb{E}\left[(p-1)^{\sum_{i=1}^n X_i} \cdot p^{\sum_{i=1}^n (1-X_i)}\right] \\
&= \mathbb{E}\left[\prod_{i=1}^{n} (p-1)^{X_i} p^{1-X_i}\right] \\
&= \prod_{i=1}^{n} \mathbb{E}\left[(p-1)^{X_i} p^{1-X_i}\right] \\
&= \prod_{i=1}^{n} (p_i(p-1) + (1-p_i)p) \\
&= \prod_{i=1}^{n} (p - p_i)\ .
\end{aligned}$$

Hence, the roots of the polynomial $r_{\mathbf{P}}$ are exactly the parameters p_i of the PBD $\mathbf{P} \in \mathcal{S}_{n,2}$. We have the following simple claim:

Claim 18. *Let $\mathbf{P}, \mathbf{Q} \in \mathcal{S}_{n,2}$ such that $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) < \epsilon$. Then for any $p \in [0,1]$, we have that*

$$|r_{\mathbf{P}}(p) - r_{\mathbf{Q}}(p)| < 2\epsilon.$$

Proof. We have the following sequence of (in)equalities:

$$\begin{aligned}
|r_{\mathbf{P}}(p) - r_{\mathbf{Q}}(p)| &= \left| \sum_{i=0}^{n} (\mathbf{P}(i) - \mathbf{Q}(i))(p-1)^i p^{n-i} \right| \\
&\leq \sum_{i=0}^{n} |(\mathbf{P}(i) - \mathbf{Q}(i))| \cdot \left|(p-1)^i p^{n-i}\right| \\
&\leq \sum_{i=0}^{n} |\mathbf{P}(i) - \mathbf{Q}(i)| = 2 d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) \\
&< 2\epsilon\ ,
\end{aligned}$$

where the second line is the triangle inequality and the third line uses the fact that $|(p-1)^i p^{n-i}| \leq 1$ for all $i \in [n]$ and $p \in [0,1]$. □

Hence, to prove the lemma, it suffices to show that for some $p \in [0,1]$ that

$$|r_{\mathbf{P}}(p) - r_{\mathbf{Q}}(p)| \geq 2\epsilon.$$

In particular, we show this for $p = p_i$. Noting that $r_{\mathbf{P}}(p_i) = 0$, it suffices to show that $|r_{\mathbf{Q}}(p_i)| \geq 2\epsilon$. We now proceed to prove this fact. If $j \neq i$ we have that,

$$|p_i - q_j| \geq \frac{|i-j|}{n+1} - \left| p_i - \frac{i}{n+1} \right| - \left| q_j - \frac{j}{n+1} \right| \geq \frac{1}{2(n+1)}.$$

Therefore, we have that

$$|r_{\mathbf{Q}}(p_i)| = \prod_{j=1}^{n} |p_i - q_j| \geq |p_i - q_i| \cdot \prod_{j \neq i} \frac{|i-j|}{2(n+1)}.$$

We note that

$$\begin{aligned} \prod_{j \neq i} \frac{|i-j|}{(n+1)} &= (i-1)!(n-i)! \geq \frac{n!}{\binom{n-1}{i-1}} \\ &\geq \frac{(n/e)^n}{2^{n-1}}, \end{aligned} \tag{4}$$

where we use the elementary inequalities $n! \geq (n/e)^n$ and $\binom{n-1}{i^*-1} \leq 2^{n-1}$. Applying this to the above, we find that

$$|r_{\mathbf{Q}}(p_i)| = \frac{|p_i - q_i|}{e(n+1)(4e)^n} \geq \frac{2|p_i - q_i|}{e^{3n}} \geq 2\epsilon.$$

□

Proof of Theorem 16. Given $\epsilon > 0$ and $n \in \mathbb{Z}$ satisfying the condition of the theorem, we define an explicit ϵ-packing for $\mathcal{S}_{n,2}$ as follows: Let $s = \lfloor \epsilon^{-1/2} \rfloor$. For a vector $\mathbf{a} = (a_1, \ldots, a_n) \in [s]^n$, let

$$p_i^{\mathbf{a}} = \frac{i}{n+1} + \frac{a_i \sqrt{\epsilon}}{4n}, \quad i \in \{1, \ldots, n\}$$

be the parameters of a PBD $\mathbf{P_a} \in \mathcal{S}_{n,2}$. We claim that the set of PBDs $\left\{\mathbf{P_a}\right\}_{\mathbf{a} \in [s]^n}$ satisfies the conditions of the theorem, i.e., for all $\mathbf{a}, \mathbf{b} \in [s]^n$, $\mathbf{a} \neq \mathbf{b}$ implies $d_{\mathrm{TV}}(\mathbf{P_a}, \mathbf{P_b}) \geq \epsilon$.

In particular, if $\mathbf{a} \neq \mathbf{b}$, then there must be some i so that $a_i \neq b_i$. Then, by Lemma 17, we have that

$$d_{\mathrm{TV}}(\mathbf{P_a}, \mathbf{P_b}) \geq |p_i^{\mathbf{a}} - p_i^{\mathbf{b}}| e^{-3n} \geq \frac{\sqrt{\epsilon}}{4n} e^{-3n} \geq \frac{\epsilon^{3/4}}{4n} \geq \epsilon.$$

□

As a simple corollary we obtain the desired lower bound:

Corollary 19. *For all $0 < \epsilon < 1$ and $n = \Omega(\log(1/\epsilon))$, any ϵ-cover of $\mathcal{S}_{n,2}$ under d_{TV} must be of size $n \cdot (1/\epsilon)^{\Omega(\log 1/\epsilon)}$.*

Proof. We will assume without loss of generality that ϵ is smaller than an appropriately small positive constant. First note that if there exists a 3ϵ-packing for $\mathcal{S}_{n,2}$ of cardinality M, then any ϵ-cover for $\mathcal{S}_{n,2}$ must be of cardinality at least M. Indeed, for every $\mathbf{Q}_i$, $i = 1, \ldots, M$, in the 3ϵ-packing, consider the (non-empty) set $N_\epsilon(\mathbf{Q}_i)$ of points $\mathbf{P}$ in the ϵ-cover with $d_{\mathrm{TV}}(\mathbf{Q}_i, \mathbf{P}) \leq \epsilon$. If $\mathbf{P} \in N_\epsilon(\mathbf{Q}_i)$ and $j \neq i$, we have $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}_j) \geq d_{\mathrm{TV}}(\mathbf{Q}_j, \mathbf{Q}_i) - d_{\mathrm{TV}}(\mathbf{Q}_i, \mathbf{P}) \geq 2\epsilon$. That is, the sets $N_\epsilon(\mathbf{Q}_i)$ are each non-empty and mutually disjoint, which implies that the size of any ϵ-cover is at least M.

By Theorem 16, for any $0 < \epsilon \leq e^{-42}/3$, if we fix $n_0 = \lfloor \frac{1}{6} \ln(1/3\epsilon) \rfloor$, there is a 3ϵ-packing for $\mathcal{S}_{n_0,2}$ of size $(1/\epsilon)^{\Omega(\log(1/\epsilon))}$. From the argument of the previous paragraph, any ϵ-cover for $\mathcal{S}_{n_0,2}$ is of size $(1/\epsilon)^{\Omega(\log(1/\epsilon))}$.

To prove the desired lower bound of $n \cdot (1/\epsilon)^{\Omega(\log(1/\epsilon))}$ we construct appropriate "shifts" of the set $\mathcal{S}_{n_0,2}$ as follows: Consider the set $\mathcal{S}_{n,2}$ where $n \geq r(n_0+1)$ for some $r \in \mathbb{Z}_+$. For $0 \leq i < r$, let $\mathcal{S}^i_{n,2}$ be the subset of $\mathcal{S}_{n,2}$ where $i(n_0+1)$ of the parameters p_j are equal to 1, and at most n_0 other p_j's are non-zero. Note that for $i \neq j$ any elements of $\mathcal{S}^i_{n,2}$ and $\mathcal{S}^j_{n,2}$ have disjoint supports. Therefore, any ϵ-cover of $\mathcal{S}_{n,2}$ must contain disjoint ϵ-covers for $\mathcal{S}^i_{n,2}$ for each i. Note also that $\mathcal{S}^i_{n,2}$ is isomorphic to $\mathcal{S}_{n_0,2}$ for each i, and thus has minimal ϵ-cover size at least $(1/\epsilon)^{\Omega(\log(1/\epsilon))}$. Therefore, any ϵ-cover of $\mathcal{S}_n$ must have size at least $\lfloor n/n_0 \rfloor \cdot (1/\epsilon)^{\Omega(\log(1/\epsilon))} = n(1/\epsilon)^{\Omega(\log(1/\epsilon))}$.

□

4.2 Cover Size Lower Bound for k-SIIRVs. In this section, we prove our cover lower bound for k-SIIRVs:

Theorem 20. *For $0 < \epsilon \leq e^{-12}(2k)^{-9}$ and $n \leq \lfloor \frac{1}{12}\log(1/\epsilon) \rfloor$, there is an ϵ-packing of $\mathcal{S}_{n,k}$ under d_{TV} with cardinality $(1/\epsilon)^{\Omega(nk)}$.*

Proof. We consider k-SIIRVs close to the $(k-1)$ multiple of the PBD $\mathbf{P}_0$ with parameters $p_i = \frac{i}{n+1}$ we used for the explicit lower bound in 4.1. Let $m \in \mathbb{Z}_+$ and $0 < \delta < 1$ be parameters that will be fixed later. Given an $\mathbf{a} \in [m]^{n(k-2)}$, which will index by a_{ij}, for $i \in \{1,\ldots,n\}$ and $j \in \{1,\ldots,k-2\}$, we define a k-SIIRV $\mathbf{P_a}$ as follows. For each i, we take a k-IRV Y_i with pdf defined as follows:

$$\begin{aligned}
\Pr[Y_i = 0] &= (1-p_i)\left(1 - \delta \cdot \sum_j a_{ij}\right), \\
\Pr[Y_i = j] &= \delta \cdot a_{ij}, \quad 1 \leq j \leq k-2, \\
\Pr[Y_i = k-1] &= p_i\left(1 - \delta \sum_j a_{ij}\right).
\end{aligned}$$

For convenience, we will denote $\gamma_{\mathbf{a},i} = \left(1 - \delta \cdot \sum_j a_{ij}\right)$. We claim that the set of distributions $\mathbf{P_a}$, $\mathbf{a} \in [m]^{n(k-2)}$, is an ϵ-packing. To prove this statement we proceed similarly to the proof of Theorem 16. For a distribution $\mathbf{P}$, we will consider the expectations

$$r_{\mathbf{P},ij} = \sum_{l=0}^{n} p_i^{n-l}(p_i-1)^l \mathbf{P}(l(k-1)+j)$$

for $i \in \{1,\ldots,n\}$ and $j \in \{1,\ldots,k-2\}$. Similarly to Claim 18, we have the following

Claim 21. *Let $\mathbf{P}, \mathbf{Q} \in \mathcal{S}_{n,k}$ such that $d_{\mathrm{TV}}(\mathbf{P},\mathbf{Q}) < \epsilon$. Then for any $i \in \{1,\ldots,n\}$ and $j \in \{1,\ldots,k-2\}$, we have that*

$$|r_{\mathbf{P},ij} - r_{\mathbf{Q},ij}| < 2\epsilon.$$

Proof. We have the following sequence of (in)equalities:

$$\begin{aligned}
|r_{\mathbf{P},ij} - r_{\mathbf{Q},ij}| &= \left|\sum_{l=0}^{n} (\mathbf{P}(l(k-1)+j) - \mathbf{Q}(l(k-1)+j)) p_i^{n-l}(p_i-1)^l\right| \\
&\leq \sum_{i=0}^{n} |(\mathbf{P}(l(k-1)+j) - \mathbf{Q}(l(k-1)+j)| \cdot \left|p_i^{n-l}(p_i-1)^l\right| \\
&\leq \sum_{i=0}^{n} |\mathbf{P}(l(k-1)+j) - \mathbf{Q}(l(k-1)+j)| \leq 2d_{\mathrm{TV}}(\mathbf{P},\mathbf{Q}) \\
&< 2\epsilon,
\end{aligned}$$

where the second line is the triangle inequality and the third line uses the fact that $|p_i^{n-l}(p_i-1)^l| \leq 1$ for all $l \in [n]$ and $i \in \{1,\ldots,n\}$. □

By the above claim, to complete the proof, it suffices to show that $|r_{\mathbf{P_a},ij} - r_{\mathbf{P_b},ij}| \geq 2\epsilon$ whenever $a_{ij} \neq b_{ij}$. To prove this statement, we exploit the fact that these k-SIIRVs are close to a multiple of $\mathbf{P}_0$, by ignoring terms in the expectations that are $O(\delta^2)$.

Let $Y = \sum_{i=1}^{n} Y_i$ with $Y \sim \mathbf{P_a}$ for a given $\mathbf{a} \in [m]^{n(k-2)}$. We define several events depending on which coordinates Y_i are equal to 0 or $k-1$, and consider their contribution to the expectation $r_{\mathbf{P_a},ij}$ separately.

Firstly, let $A_{\geq 2}$ be the event that more than one Y_i is not 0 or $k-1$. The probability that any fixed Y_i is not 0 or $k-1$ is small, namely

$$\sum_{j=1}^{k-2} \Pr[Y_i = j] = \sum_{j=1}^{k-2} \delta a_{ij} \leq (k-2)m\delta \ .$$

Hence,

$$\Pr[A_{\geq 2}] \leq \binom{n}{2} ((k-2)m\delta)^2 \leq \frac{1}{2} \cdot (n(k-2)m\delta)^2 \ .$$

The contribution of $A_{\geq 2}$ to $r_{\mathbf{P_a},ij}$ is $r_{\mathbf{P_a},ij,A_{\geq 2}} := \sum_{l=0}^{n} p_i^{n-l}(p_i-1)^l \Pr_{Y\sim\mathbf{P_a}}[Y = l(k-1)+j \cap A_{\geq 2}]$, and therefore

$$|r_{\mathbf{P_a},ij,A_{\geq 2}}| \leq \frac{1}{2}(n(k-2)m\delta)^2 \ ,$$

since $|p_i^{n-l}(p_i-1)^l| \leq 1$.

Secondly, let A_0 be the event that all Y_i's are 0 or $k-1$. If A_0 occurs then Y is a multiple of $k-1$. Thus, for $l \in [n]$ and $j \in \{1,\ldots,k-2\}$, we have $\Pr_{Y\sim\mathbf{P_a}}[Y = l(k-1)+j \cap A_0] = 0$. The contribution of A_0 to $r_{\mathbf{P_a},ij}$ is

$$r_{\mathbf{P_a},ij,A_0} := \sum_{l=0}^{n} p_i^{n-l}(p_i-1)^l \Pr_{Y\sim\mathbf{P_a}}[Y = l(k-1)+j \cap A_0] = 0 \ .$$

Finally, for $i \in \{1,\ldots,n\}$, let B_i be the event that Y_i is the only k-IRV that takes a value between 1 and $k-2$. The probability of all other Y_h, with $h \neq i$, being either 0 or $k-1$ is $\prod_{h\neq i} \gamma_{\mathbf{a},h}$. We consider the RVs $X_{-i} = \sum_{h\neq i} X_h$, where $X_h \sim \mathrm{Ber}(p_h)$. That is, $X_{-i} \sim \mathbf{P}_{-i} \in \mathcal{S}_{n-1,2}$, i.e., it is a PBD with parameters p_h for $h \neq i$. Then, the conditional probability $\Pr\left[\sum_{h\neq i} Y_h = l(k-1) | (B_i \cup A_0)\right]$ is equal to $\Pr[X_{-i} = l] = \mathbf{P}_{-i}(l)$ for all $l \in [n]$. So, for all $l \in [n]$ and $j \in \{1,\ldots,k-2\}$ we have

$$\begin{aligned}
\Pr[Y = l(k-1)+j \cap B_i] &= \Pr\left[\sum_{h\neq i} Y_h = l(k-1) \cap (B_i \cup A_0)\right] \Pr[Y_i = j] \\
&= \left(\prod_{h\neq i} \gamma_{\mathbf{a},h}\right) \mathbf{P}_{-i}(l)\delta a_{ij}
\end{aligned}$$

Then, the contribution of B_i to $r_{\mathbf{P_a},gj}$ is

$$\begin{aligned} r_{\mathbf{P_a},gj,B_i} &:= \sum_{l=0}^{n} p_g^{n-l}(p_g-1)^l \mathrm{Pr}_{Y\sim\mathbf{P_a}}\left[Y = l(k-1)+j \cap B_i\right] \\ &= \left(\prod_{h\neq i}\gamma_{\mathbf{a},h}\right)\cdot \delta a_{ij}\cdot \sum_{l=0}^{n} p_g^{n-l}(p_g-1)^l \mathbf{P}_{-i}(l) \\ &= \left(\prod_{h\neq i}\gamma_{\mathbf{a},h}\right)\cdot \delta a_{ij}\cdot r_{\mathbf{P}_{-i}}(p_g) \\ &= \left(\prod_{h\neq i}\gamma_{\mathbf{a},h}\right)\cdot \delta a_{ij}\cdot \prod_{h\neq i}(p_h-p_g)\ , \end{aligned}$$

where $r_{\mathbf{P}_{-i}}$ above is as defined in the previous section, and when $g \neq i$, the second product includes the term $p_g - p_g = 0$, so $r_{\mathbf{P_a},gj,B_i} = 0$. Summing these contributions to the expectation $r_{\mathbf{P_a},ij}$ gives:

$$\begin{aligned} r_{\mathbf{P_a},ij} &= r_{\mathbf{P_a},ij,A_{\geq 2}} + r_{\mathbf{P_a},ij,A_0} + \sum_{g=1}^{n} r_{\mathbf{P_a},ij,B_g} \\ &= r_{\mathbf{P_a},ij,A_{\geq 2}} + r_{\mathbf{P_a},ij,B_i} \\ &= r_{\mathbf{P_a},ij,A_{\geq 2}} + \prod_{h\neq i}\gamma_{\mathbf{a},h}\cdot\delta a_{ij}\cdot\prod_{h\neq i}(p_h-p_i) \end{aligned}$$

Now consider $\mathbf{a}$ and $\mathbf{b}$ which for some $i \in \{1,2,...,n\}$ and $j \in \{1,2,...,k-2\}$ have $a_{ij} \neq b_{ij}$. We have that $\prod_{h\neq i}|p_h - p_i| \geq e^{-3n}$ by Equation (4), and thus,

$$\prod_{h\neq i}\gamma_{\mathbf{a},h} = \prod_{h\neq i}\left(1-\delta\sum_j a_{hj}\right) \geq (1-(k-2)m\delta)^{n-1} \geq (1-(n-1)(k-2)m\delta),$$

$|a_{ij}-b_{ij}| \geq 1$, and $|r_{\mathbf{P_a},ij,A_{\geq 2}}| \leq \frac{1}{2}(n(k-2)m\delta)^2$.

We obtain the following sequence of inequalities:

$$\begin{aligned} & |r_{\mathbf{P_a},ij} - r_{\mathbf{P_b},ij}| = |r_{\mathbf{P_a},ij,B_i} - r_{\mathbf{P_b},ij,B_i} + r_{\mathbf{P_a},ij,A_{\geq 2}} - r_{\mathbf{P_b},ij,A_{\geq 2}}| \\ \geq\ & \Big|\prod_{h\neq i}(p_h-p_g)\Big(\prod_{h\neq i}\gamma_{\mathbf{a},h}\delta a_{ij} - \prod_{h\neq i}\gamma_{\mathbf{b},h}\delta b_{ij}\Big)\Big| - (n(k-2)m\delta)^2 \\ \geq\ & e^{-3n}\Big|\prod_{h\neq i}\gamma_{\mathbf{a},h}\delta\Big|\cdot|a_{ij}-b_{ij}| - e^{-3n}\delta b_{ij}\Big|\prod_{h\neq i}\gamma_{\mathbf{a},h} - \prod_{h\neq i}\gamma_{\mathbf{b},h}\Big| - (n(k-2)m\delta)^2 \\ \geq\ & e^{-3n}(1-(n-1)(k-2)m\delta)\delta - \delta m\Big(\Big|1-\prod_{h\neq i}\gamma_{\mathbf{a},h}\Big| + \Big|1-\prod_{h\neq i}\gamma_{\mathbf{b},h}\Big|\Big) - (n(k-2)m\delta)^2 \\ \geq\ & e^{-3n}\delta - 2\delta mn(k-2)m\delta - 2(n(k-2)m\delta)^2 \\ \geq\ & e^{-3n}\delta - 3(n(k-2)m\delta)^2\ . \end{aligned}$$

Recall that by assumption $\epsilon \leq e^{-12}(2k)^{-9}$. We set $n = \lfloor \frac{1}{12}\log(1/\epsilon)\rfloor$, $\delta = 3\epsilon^{3/4}$, and $m = \lfloor \frac{\epsilon^{-1/4}}{2n^2(k-2)^2}\rfloor$. Then, $e^{-3n}\delta \geq 3\epsilon$ and $3(n(k-2)m\delta)^2 \leq \epsilon$. So, we have that $|r_{\mathbf{P_a},ij} - r_{\mathbf{P_b},ij}| \geq 2\epsilon$ as required. Also, $\gamma_{\mathbf{a}} \geq 1-\sqrt{\epsilon} \geq 0$, so the k-IRVs are indeed well-defined.

Therefore, we have exhibited a set of $m^{n(k-2)}$ k-SIIRVs that have pairwise total variation distance at least ϵ. The proof follows by observing that $m^{n(k-2)} = (1/\epsilon)^{\Omega(k\log 1/\epsilon)}$. □

5 Sample Complexity Lower Bound

5.1 A Useful Structural Result. In this section, we prove a novel structural result for the space of PBDs (Lemma 22). This allows us to obtain a simple non-constructive lower bound on the cover size of PBDs under the Kolmogorov distance metric. More importantly, this lemma is crucial for the sample complexity lower bound of the following subsection.

Before we state our lemma, we provide some basic intuition. The set of all distributions supported on $[n]$ is n-dimensional (viewed as a metric space). Note that each $\mathbf{P} \in \mathcal{S}_{n,2}$ is defined by n parameters. It turns out that $\mathcal{S}_{n,2}$ is also n-dimensional in a precise sense. This intuition is formalized in the following lemma:

Lemma 22. *(i) Given any $\mathbf{P} \in \mathcal{S}_{n,2}$ with distinct parameters in $(0,1)$, there is a radius $\delta = \delta(\mathbf{P})$ such that any distribution $\mathbf{Q}$ with support $[n]$ that satisfies $d_{\mathrm{K}}(\mathbf{P},\mathbf{Q}) \leq \delta$ can also be expressed as a PBD, i.e., $\mathbf{Q} \in \mathcal{S}_{n,2}$.*

(ii) Let $\mathbf{P}_0 \in \mathcal{S}_{n,2}$ be the PBD with parameters $p_i = \frac{i}{n+1}$, $1 \leq i \leq n$. Then any distribution $\mathbf{Q}$ with support $[n]$ that satisfies $d_{\mathrm{K}}(\mathbf{P}_0,\mathbf{Q}) \leq 2^{-9n}$ is itself a PBD with parameters q_i such that $|q_i - p_i| \leq \frac{1}{4(n+1)}$.

Proof. We consider the space of cumulative distribution functions (CDF's) of all distributions of support $[n]$. Let T_n be the set of sequences $0 \leq x_1 \leq x_2 \leq \ldots \leq x_n \leq 1$. Consider the map $\mathcal{P}_n : T_n \to T_n$ defined as follows: For $\mathbf{p} = (p_1,\ldots,p_n) \in T_n$ (i.e., with ordered parameters $0 \leq p_1 \leq \ldots \leq p_n \leq 1$), let $\mathbf{P}$ be the corresponding PBD in $\mathcal{S}_{n,2}$. For $i \in \{1,\ldots,n\}$, let $(\mathcal{P}_n(\mathbf{p}))_i = \mathbf{P}(< i)$. Namely, $\mathcal{P}_n$ maps a sequence of probabilities to the sequence of probabilities defining the CDF of the corresponding PBD.

The basic idea of the proof is that the mapping $\mathcal{P}_n$ is invertible in a neighborhood of a point $\mathbf{p}$ with distinct coordinates. This allows us to uniquely obtain the distinct parameters of a PBD $\mathbf{P} \in \mathcal{S}_{n,2}$ from its CDF. We will make essential use of the inverse function theorem for $\mathcal{P}_n$, which we now recall:

Theorem 23 (Inverse function theorem [Rud76])**.** *Let $F : S \to \mathbb{R}^n$, $S \subseteq \mathbb{R}^n$, be a continuously differentiable function and $\mathbf{x}$ be a point in the interior of S such that the Jacobian matrix of F, $\mathrm{Jac}(F)(\mathbf{x})$, is non-singular. Then there exists an inverse function, F^{-1}, of F in a neighborhood of $F(\mathbf{x})$. Furthermore the inverse function F^{-1} is continuously differentiable and its Jacobian matrix satisfies $\mathrm{Jac}(F^{-1})(F(\mathbf{x})) = (\mathrm{Jac}(F)(\mathbf{x}))^{-1}$.*

We will apply the inverse function theorem for $\mathcal{P}_n$ at the point $\mathbf{p}$ defining the distinct parameters of the PBD $\mathbf{P}$ in the statement of the theorem. It is easy to see that $\mathcal{P}_n$ is continuously differentiable. The main part of the argument involves proving that the Jacobian matrix of $\mathcal{P}_n$ at $\mathbf{p}$, $\mathrm{Jac}(\mathcal{P}_n)(\mathbf{p})$, is non-singular.

Recall that $\mathrm{Jac}(\mathcal{P}_n)(\mathbf{p})$ is the $n \times n$ matrix whose (i,j) entry is the partial derivatives of $(\mathcal{P}_n)_i$ in direction j, i.e., $(\mathrm{Jac}(\mathcal{P}_n)(\mathbf{p}))_{ij} = \frac{\partial(\mathcal{P}_n(\mathbf{p}))_i}{\partial p_j}$. We start by showing the following lemma:

Lemma 24. *For a PBD $\mathbf{P} \in \mathcal{S}_{n,2}$ with parameters $\mathbf{p}$, we have*

$$M(\mathbf{p}) \cdot \mathrm{Jac}(\mathcal{P}_n)(\mathbf{p}) = -\mathrm{diag}\left(\prod_{j \neq i}(p_i - p_j)\right) \tag{5}$$

where $M(\mathbf{p})$ is the $n \times n$ matrix with entries $(M(\mathbf{p}))_{ij} = (1-p_i)^{j-1} p_i^{n-j}$, $1 \leq i,j \leq n$. Here, for $x \in \mathbb{R}^n$, we denote by $\mathrm{diag}(x)$ the diagonal matrix with entries $(\mathrm{diag}(x))_{ii} = x_i$.

Proof. To calculate the partial derivative $\frac{\partial(\mathcal{P}_n(\mathbf{p}))_i}{\partial p_j}$, we isolate the effect of the parameter p_j from the other variables. In particular, for $X \sim \mathbf{P}$, i.e., $X = \sum_{i=1}^n X_i$, with $X_i \sim \mathrm{Ber}(p_i)$, we can write $X = X_{-j} + X_j$, where $X_{-j} = \sum_{i \neq j} X_i$. Note that $X_j \sim \mathbf{P}_{-j} \in \mathcal{S}_{n-1,2}$, i.e., it is the $(n-1)$ parameter PBD with parameters p_i for $i \neq j$. Now, for $1 \leq i \leq n$, we can write

$$(\mathcal{P}_n(\mathbf{p}))_i = \mathbf{P}(< i) = \mathbf{P}_{-j}(< (i-1)) + (1-p_j)\mathbf{P}_{-j}(i-1).$$

The derivative of this quantity with respect to p_j equals

$$\frac{\partial(\mathcal{P}_n(\mathbf{p}))_i}{\partial p_j} = -\mathbf{P}_{-j}(i-1).$$

Therefore, the j-th column of $\mathrm{Jac}(\mathcal{P}_n)(\mathbf{p})$ equals -1 times the pdf of the distribution $\mathbf{P}_{-j}$. This allows us to consider multiplying on the right by $\mathrm{Jac}(\mathcal{P}_n)(\mathbf{p})$ as taking the expectations of certain distributions. In particular, for $y \in \mathbb{R}^n$ and any $1 \leq j \leq n$, we have that

$$(y^T \mathrm{Jac}(\mathcal{P}_n)(\mathbf{p}))_j = -\sum_{i=1}^n y_i \mathbf{P}_{-j}(i-1) = -\mathbb{E}\left[y_{X_{-j}+1}\right] \ .$$

Therefore, for $1 \leq i, j \leq n$, we can write

$$\begin{aligned}
(M(\mathbf{p}) \cdot \mathrm{Jac}(\mathcal{P}_n)(\mathbf{p}))_{ij} &= -\sum_{k=1}^n (p_i-1)^{k-1} p_i^{n-k} \mathbf{P}_{-j}(k-1) \\
&= -\mathbb{E}\left[(p_i-1)^{X_{-j}} p_i^{n-X_{-j}-1}\right] \\
&= -\mathbb{E}\left[\prod_{k \neq j} (p_i-1)^{X_k} p_i^{1-X_k}\right] \\
&= -\prod_{k \neq j} \mathbb{E}\left[(p_i-1)^{X_k} p_i^{1-X_k}\right] \\
&= -\prod_{k \neq j} \left[(p_i-1)p_k + p_i(1-p_k)\right] \\
&= -\prod_{k \neq j} (p_i - p_k) \ .
\end{aligned}$$

Note that for $i \neq j$, the above product contains the term $(p_i - p_i)$ and so is equal to 0. When $i = j$, we have $(M(\mathbf{p}) \cdot \mathrm{Jac}(\mathcal{P}_n)(\mathbf{p}))_{ii} = -\prod_{k \neq i}(p_i - p_k)$ completing the proof of the lemma. □

We are now ready to prove part (i) of Lemma 22. To this end, consider a PBD $\mathbf{P}$ with distinct parameters $\mathbf{p}$, i.e., $p_i \neq p_j$ for $i \neq j$, such that $p_i \in (0,1)$ for all i. Note that $\mathbf{p}$ lies in the interior of T_n. Moreover, for all i, we have $\prod_{j \neq i}(p_i - p_j) \neq 0$ and therefore the matrix $\mathrm{diag}(\prod_{j \neq i}(p_i - p_j))$ appearing in (5) is non-singular. It follows from Lemma 24 that both matrices on the LHS of (5) are non-singular. In particular, $\mathrm{Jac}(\mathcal{P}_n)(\mathbf{p})$ is non-singular, hence we can apply the inverse function theorem. As a corollary, there exists an inverse mapping $\mathcal{P}_n^{-1}$ in some neighborhood of $\mathcal{P}_n(\mathbf{p})$. Specifically, there is some $\delta > 0$ such that $\mathcal{P}_n^{-1}$ is defined at every $\mathbf{x} \in T_n$ with $\|\mathbf{x} - \mathcal{P}_n(\mathbf{p})\|_\infty \leq \delta$.

Let $\mathbf{Q}$ be a distribution over $[n]$ satisfying $d_{\mathrm{K}}(\mathbf{P}, \mathbf{Q}) \leq \delta$. Equivalently, if $\mathbf{y} = (\mathbf{Q}(< i))_{i=1}^n \in T_n$ is the CDF of $\mathbf{Q}$, then $\|\mathcal{P}_n(\mathbf{p}) - \mathbf{y}\|_\infty \leq \delta$. Thus $\mathcal{P}_n^{-1}$ is defined at $\mathbf{y}$ and $\mathbf{q} = \mathcal{P}_n^{-1}(\mathbf{y}) \in T_n$ are the parameters of a PBD with distribution $\mathbf{Q}$. Thus, $\mathbf{Q}$ is a PBD with parameters $\mathbf{q}$, which completes

the proof of (i). Note that the proof also implies that $\mathbf{Q}$ in this neighborhood can be taken to be $\mathcal{P}_n(\mathbf{q}')$ for $\mathbf{q}'$ in some small neighborhood of $\mathbf{p}$.

To prove part (ii) of Lemma 22, we use a geometric argument. Recall that the parameters of $\mathbf{P}_0$ are $\mathbf{p}_0 = \left(\frac{1}{n+1}, \ldots, \frac{n}{n+1}\right)$. Let $S \subseteq T_n$ be the set of vectors $\mathbf{p}$ with $\|\mathbf{p} - \mathbf{p}_0\|_\infty \leq \frac{1}{4(n+1)}$. By Lemma 17 we have that any $\mathbf{Q}$ in $\mathcal{P}_n(\partial S)$ satisfies $d_{\mathrm{TV}}(\mathbf{P}_0, \mathbf{Q}) \geq \frac{e^{-3n}}{4(n+1)}$, and therefore $d_{\mathrm{K}}(\mathbf{P}_0, \mathbf{Q}) \geq \frac{e^{-3n}}{8(n+1)^2} \geq 2^{-9n}$.

Let B be the set of distributions $\mathbf{Q}$ on $[n]$ so that $d_{\mathrm{K}}(\mathbf{P}_0, \mathbf{Q}) \leq 2^{-9n}$. We claim that $\mathcal{P}_n(S) \cap B = B$. To begin, note that S is compact, and therefore this intersection is closed. On the other hand, since $\mathcal{P}_n(\partial S)$ is disjoint from B, this intersection is $\mathcal{P}_n(\mathrm{int}(S)) \cap B$. On the other hand, since $\mathcal{P}_n$ has non-singular Jacobian on $\mathrm{int}(S)$, the open mapping theorem implies that $\mathcal{P}_n(\mathrm{int}(S)) \cap B$ is an open subset of B. Therefore, $\mathcal{P}_n(S) \cap B$ is both a closed and open subset of B, and therefore, since B is connected, it must be all of B. This completes the proof of part (ii).

□

As a simple application of our structural lemma, we obtain a non-constructive lower bound on the cover size under the Kolmogorov distance metric:

Theorem 25. *For any $\epsilon > 0$ and $n = \Omega(\log(1/\epsilon))$ any ϵ-cover of $\mathcal{S}_{n,2}$ under d_{K} must have size at least $n \cdot (1/\epsilon)^{\Omega(\log(1/\epsilon))}$.*

Proof. Note that by an argument identical to that of Corollary 19 it suffices to prove a packing lower bound of $(1/\epsilon)^{\Omega(\log(1/\epsilon))}$ for $n = \Theta(\log(1/\epsilon))$.

To that end, fix $n = n_0 = \lfloor \frac{1}{18} \log_2(1/\epsilon) \rfloor$. Then, we have $2^{-9n} \geq \sqrt{\epsilon}$. By Lemma 22(ii), there is a PBD $\mathbf{P}_0 \in \mathcal{S}_{n,2}$, such that any distribution $\mathbf{Q}$ with support $[n]$ and $d_{\mathrm{K}}(\mathbf{P}_0, \mathbf{Q}) \leq \sqrt{\epsilon}$ is in $\mathcal{S}_{n,2}$. We will give an ϵ-packing lower bound for this subset of PBDs.

Let us denote by $\mathbf{z} \in T_n$ the vector defining the CDF of $\mathbf{P}_0$, i.e., $\mathbf{z} = (\mathbf{P}_0(< i))_{i=1}^n$. Let $S \subseteq \mathbb{R}^n$ be the set of points $\mathbf{x} \in \mathbb{R}^n$ with $\|\mathbf{x} - \mathbf{z}\|_\infty \leq \sqrt{\epsilon}$. Note that S is an n-cube with side length $2\sqrt{\epsilon}$.

We claim that every $\mathbf{x} \in S$ is the CDF of a PBD $\mathbf{Q} \in \mathcal{S}_{n,2}$. By Lemma 22, this follows immediately if $\mathbf{x} \in T_n$, i.e., if $\mathbf{x}$ is the CDF of a distribution. So, it suffices to show that $S \subseteq T_n$. Suppose for the sake of contradiction that there is a point $\mathbf{y} \in S \setminus T_n$. Then, there is a point $\mathbf{x} \in S$ such that $\mathbf{x}$ lies on the boundary of T_n. For such a point $\mathbf{x}$, one of the inequalities $0 \leq x_1 \leq x_2 \leq \ldots \leq x_n \leq 1$ is tight. Thus, $\mathbf{x}$ is the CDF of a distribution $\mathbf{Q}$ which has $\mathbf{Q}(i) = 0$ for some i. Since $\mathbf{x} \in S \cap T_n$, $\mathbf{Q}$ is a PBD with parameters given by Lemma 22. In particular $\mathbf{Q}$ does not have any parameters equal to 0 or 1. Thus, we have $\mathbf{Q}(i) > 0$ for all $i \in [n]$, a contradiction.

Therefore, any ϵ-cover of $\mathcal{S}_{n,2}$ in Kolmogorov distance induces an ϵ-cover of the same size in L_∞ distance of the CDFs of distributions in $\mathcal{S}_{n,2}$. If s is the size of such a cover, then we have s n-cubes of side length ϵ whose union contains S. Recall that S is an n-cube of side length $\sqrt{\epsilon}$. The volume of each of these s n-cubes is $(2\epsilon)^n$ and the volume of S is $(2\sqrt{\epsilon})^n$. The volume of the union of s n-cubes is at most $s \cdot (2\epsilon)^n$ and hence $s \cdot (2\epsilon)^n \geq (2\sqrt{\epsilon})^n$ or $s = (1/\epsilon)^{\Omega(n)}$, which competes the proof. □

5.2 Sample complexity lower bound for PBDs. In this subsection we prove our sample complexity lower bound of $\omega(1/\epsilon^2)$ for learning PBDs to total variation distance ϵ. Our proves uses a combination of information-theoretic arguments and the structural lemma of the previous subsection. In particular, we show:

Theorem 26 (Sample Lower Bound)**.** *Let $\mathcal{A}$ be any algorithm which, given as input n, ϵ, and sample access to an unknown $\mathbf{P} \in \mathcal{S}_{n,2}$ outputs a hypothesis distribution $\mathbf{H}$ such that $\mathbb{E}[d_{\mathrm{TV}}(\mathbf{H}, \mathbf{P})] \leq \epsilon$. Then $\mathcal{A}$ must use $\Omega((1/\epsilon^2) \cdot \sqrt{\log(1/\epsilon)})$ samples.*

Our main information-theoretic tool to prove our lower bound is Assouad's Lemma [Ass83]. We recall the statement of the lemma (see, e.g., [DG85]), tailored to discrete distributions below:

Theorem 27. *[Theorem 5, Chapter 4, [DG85]] Let $r \geq 1$ be an integer. For each $\mathbf{b} \in \{-1,1\}^r$, let $\mathbf{P}_\mathbf{b}$ be a probability distribution over a finite set A. For $1 \leq \ell \leq r$ and $\mathbf{b} \in \{-1,1\}^r$, we denote by $\mathbf{b}^{(\ell,+)}$ (resp. $\mathbf{b}^{(\ell,-)}$) the vector with $\mathbf{b}_i^{(\ell,+)} = \mathbf{b}_i$ (resp. $\mathbf{b}_i^{(\ell,-)} = \mathbf{b}_i$) for $i \neq \ell$ and $\mathbf{b}_\ell^{(\ell,+)} = 1$ (resp. $\mathbf{b}_\ell^{(\ell,-)} = -1$). Suppose there exists a partition $A_0, A_1, \ldots, A_r$ of A such that for all $\mathbf{b} \in \{-1,1\}^r$ and all $1 \leq \ell \leq r$, the following inequalities are valid:*

(a) $\sum_{x \in A_\ell} |\mathbf{P}_{\mathbf{b}^{(\ell,+)}}(x) - \mathbf{P}_{\mathbf{b}^{(\ell,-)}}(x)| \geq \alpha$, *and*

(b) $\sum_{x \in A} \sqrt{\mathbf{P}_{\mathbf{b}^{(\ell,+)}}(x)\mathbf{P}_{\mathbf{b}^{(\ell,-)}}(x)} \geq 1 - \gamma > 0$.

Then, for any any algorithm $\mathcal{A}$ that draws s samples from an unknown $\mathbf{P} \in \mathbf{P}_\mathbf{b}$ and outputs a hypothesis distribution $\mathbf{H}$, there is some $\mathbf{b} \in \{-1,1\}^r$ such that if the target distribution $\mathbf{P}$ is $\mathbf{P}_\mathbf{b}$,

$$\mathbb{E}\left[d_{\mathrm{TV}}(\mathbf{P}, \mathbf{H})\right] \geq (r\alpha/4)(1 - \sqrt{2s\gamma}).$$

Recall that PBDs are discrete log-concave distributions. We will use the following basic properties of log-concave distributions:

Lemma 28. *There exists a universal constant $c > 0$ such that the following holds: For any log-concave distribution $\mathbf{P}$ supported on the integers and standard deviation σ, there exist at least $\Omega(\sigma)$ consecutive integers with probability mass under $\mathbf{P}$ at least $c \cdot \frac{1}{1+\sigma}$.*

The simple proof is deferred to Appendix C.1. We are now ready to prove Theorem 26.

Proof of Theorem 26. Ideally, we would like to use the set of PBDs whose parameters are explicitly described in Theorem 16 in our application of Assouad's lemma. Unfortunately, however, this particular set is not in a form that allows a direct application of the theorem. The difficulty lies in the fact that it is not clear how to isolate the changes between distributions in disjoint intervals using explicit parameters.

We therefore proceed with an indirect approach making essential use of Lemma 22(ii). We start from the PBD $\mathbf{P}_0$ in the statement of the lemma and we perturb its pdf appropriately to construct our "hypercube" distributions $\mathbf{P}_\mathbf{b}$. The lemma guarantees that, if the perturbation is small enough, all these distributions are indeed PBDs.

Observe that the variance of $\mathbf{P}_0$ is $\Omega(n)$ since $\Omega(n)$ parameters p_i lie in $[1/4, 3/4]$. By Lemma 28, there exist $r = \Omega(\sqrt{n})$ consecutive integers, an integer m, $0 \leq m \leq n$, and a real value t with $t \geq c \cdot r$, such that for all i, with $m \leq i \leq m + 2r$, we have

$$\mathbf{P}(i) \geq \frac{2}{t}\,.$$

For n sufficiently large, we can assume that $2^{-9n} \leq c$ and therefore $\frac{1}{t} \geq \frac{2^{-9n}}{r}$.

We are now ready to define our "hypercube" of PBDs. For $\mathbf{b} \in \{-1,1\}^r$, consider the distribution $\mathbf{P}_\mathbf{b}$ with

$$\mathbf{P}_\mathbf{b}(i) = \begin{cases} \mathbf{P}_0(i) & \text{if } i < m,\ i > m + 2r, \text{ or } \mathbf{b}_{\lfloor \frac{1}{2}(i-m) \rfloor} = -1 \\ \mathbf{P}_0(i) - \frac{2^{-9n}}{r} & \text{if } \mathbf{b}_{\lfloor \frac{1}{2}(i-m) \rfloor} = 1 \text{ and } i \text{ is even} \\ \mathbf{P}_0(i) + \frac{2^{-9n}}{r} & \text{if } \mathbf{b}_{\lfloor \frac{1}{2}(i-m) \rfloor} = 1 \text{ and } i \text{ is odd} \end{cases}$$

Note that all these distributions are PBDs as follows from Lemma 22(ii) since

$$d_{\mathrm{K}}(\mathbf{P_b}, \mathbf{P}_0) \leq d_{\mathrm{TV}}(\mathbf{P_b}, \mathbf{P}_0) = 2^{-9n} .$$

For $0 \leq i \leq r-1$, the sets $A_{i+1} = \{m+2i, m+2i+1\}$ define the partition of the domain. We can now apply Assouad's lemma to this instance.

For $\mathbf{b} \in \{-1,1\}^r$ we can write

$$\sum_{x \in A_\ell} |\mathbf{P}_{\mathbf{b}^{(\ell,+)}}(x) - \mathbf{P}_{\mathbf{b}^{(\ell,-)}}(x)| = \frac{2 \cdot 2^{-9n}}{r} .$$

Similarly,

$$\begin{aligned}
\sum_{i=0}^{n} \left(\sqrt{\mathbf{P}_{\mathbf{b}^{(\ell,+)}}(i)} - \sqrt{\mathbf{P}_{\mathbf{b}^{(\ell,-)}}(i)} \right)^2 &= \sum_{i=m+2\ell, m+2\ell+1} \left(\frac{\mathbf{P}_{\mathbf{b}^{(\ell,+)}}(i) - \mathbf{P}_{\mathbf{b}^{(\ell,-)}}(i)}{\sqrt{\mathbf{P}_{\mathbf{b}^{(\ell,+)}}(i))} + \sqrt{\mathbf{P}_{\mathbf{b}^{(\ell,-)}}(i)}} \right)^2 \\
&= \sum_{i=m+2\ell, m+2\ell+1} \left(\frac{2^{-9n}/r}{\sqrt{\mathbf{P}_{\mathbf{b}^{(\ell,+)}}(i))} + \sqrt{\mathbf{P}_{\mathbf{b}^{(\ell,-)}}(i)}} \right)^2 \\
&\geq \sum_{i=m+2\ell, m+2\ell+1} \left(\frac{2^{-9n}/r}{2\sqrt{1/t}} \right)^2 \\
&= \frac{2^{-18n} \cdot c}{2r} ,
\end{aligned}$$

where the first inequality uses the fact that

$$\mathbf{P_b}(i) \geq \mathbf{P}_0(i) - \frac{2^{-9n}}{r} \geq \frac{2}{t} - \frac{1}{t} \geq \frac{1}{t},$$

for $m \leq i \leq m+2k$.

Therefore, the parameters in Assouad's Lemma are

$$\alpha := \frac{2 \cdot 2^{-9n}}{r}, \quad \gamma = \frac{2^{-18n} \cdot c}{2r}, \quad \text{and} \quad s = \frac{1}{8\gamma}$$

from which we obtain that that there is a $\mathbf{P_b}$ with

$$\mathbb{E}\left[d_{\mathrm{TV}}(\mathbf{H}, \mathbf{P_b})\right] \geq (r\alpha/4) \cdot (1 - \sqrt{2s\gamma}) = \frac{2^{-9n}}{4}.$$

Hence, for $\epsilon = 2^{-9n-2}$, if the number of samples satisfies

$$s \leq \frac{1}{8\gamma} = \frac{r \cdot 2^{18n}}{4c} = O(2^{18n}\sqrt{n}) = O\left((1/\epsilon^2)\sqrt{\log(1/\epsilon)}\right),$$

then $\mathbb{E}\left[d_{\mathrm{TV}}(\mathbf{H}, \mathbf{P_a})\right] \geq \epsilon$, completing the proof of the theorem. □

References

[AK01] S. Arora and R. Kannan. Learning mixtures of arbitrary Gaussians. In *Proceedings of the 33rd Symposium on Theory of Computing*, pages 247–257, 2001.

[Ass83] P. Assouad. Deux remarques sur l'estimation. *C. R. Acad. Sci. Paris Sér. I*, 296:1021–1024, 1983.

[BBBB72] R.E. Barlow, D.J. Bartholomew, J.M. Bremner, and H.D. Brunk. *Statistical Inference under Order Restrictions*. Wiley, New York, 1972.

[BGL07] R. Blei, F. Gao, and W. V. Li. Metric entropy of high dimensional distributions. *Proceedings of the American Mathematical Society (AMS)*, 135(12):4009 – 4018, 2007.

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

[Bir86] L. Birgé. On estimating a density using Hellinger distance and some other strange facts. *Probab. Theory Relat. Fields*, 71(2):271–291, 1986.

[Blo99] M. Blonski. Anonymous games with binary actions. *Games and Economic Behavior*, 28(2):171 – 180, 1999.

[Blo05] M. Blonski. The women of cairo: Equilibria in large anonymous games. *Journal of Mathematical Economics*, 41(3):253 – 264, 2005.

[BS10] M. Belkin and K. Sinha. Polynomial learning of distribution families. In *FOCS*, pages 103–112, 2010.

[CDO15] X. Chen, D. Durfee, and A. Orfanou. On the complexity of nash equilibria in anonymous games. In *STOC*, 2015.

[CDSS14] S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Efficient density estimation via piecewise polynomial approximation. In *STOC*, pages 604–613, 2014.

[CGG02] M. Cryan, L. Goldberg, and P. Goldberg. Evolutionary trees can be learned in polynomial time in the two state general Markov model. *SIAM Journal on Computing*, 31(2):375–397, 2002.

[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.

[CS90] B. Carl and I. Stephani. *Entropy, compactness and the approximation of operators*, volume 98 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1990.

[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.

[DDS12a] C. Daskalakis, I. Diakonikolas, and R.A. Servedio. Learning k-modal distributions via testing. In *SODA*, pages 1371–1385, 2012.

[DDS12b] C. Daskalakis, I. Diakonikolas, and R.A. Servedio. Learning Poisson Binomial Distributions. In *STOC*, pages 709–728, 2012.

[DG85] L. Devroye and L. Györfi. *Nonparametric Density Estimation: The L_1 View*. John Wiley & Sons, 1985.

[DL01] L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. Springer Series in Statistics, Springer, 2001.

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

[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.

[DP14a] C. Daskalakis and C. Papadimitriou. Sparse covers for sums of indicators. *Probability Theory and Related Fields*, pages 1–27, 2014.

[DP14b] C. Daskalakis and C. H. Papadimitriou. Approximate nash equilibria in anonymous games. *Journal of Economic Theory*, 2014.

[Dud74] R.M Dudley. Metric entropy of some classes of sets with differentiable boundaries. *Journal of Approximation Theory*, 10(3):227 – 236, 1974.

[ET96] D. E. Edmunds and H. Triebel. *Function spaces, entropy numbers, differential operators*, volume 120 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1996.

[FM99] Y. Freund and Y. Mansour. Estimating a mixture of two product distributions. In *Proceedings of the 12th Annual COLT*, pages 183–192, 1999.

[FOS05] J. Feldman, R. O'Donnell, and R. Servedio. Learning mixtures of product distributions over discrete domains. In *Proc. 46th IEEE FOCS*, pages 501–510, 2005.

[GS13] A. Guntuboyina and B. Sen. Covering numbers for convex functions. *Information Theory, IEEE Transactions on*, 59(4):1957–1965, April 2013.

[HI90] R. Hasminskii and I. Ibragimov. On density estimation in the view of kolmogorov's ideas in approximation theory. *Ann. Statist.*, 18(3):999–1010, 1990.

[HO97] D. Haussler and M. Opper. Mutual information, metric entropy and cumulative relative entropy risk. *Ann. Statist.*, 25(6):2451–2492, 1997.

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

[Hp11] S. Har-peled. *Geometric Approximation Algorithms*. American Mathematical Society, Boston, MA, USA, 2011.

[Ize91] A. J. Izenman. Recent developments in nonparametric density estimation. *Journal of the American Statistical Association*, 86(413):205–224, 1991.

[KMR+94] M. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R. Schapire, and L. Sellie. On the learnability of discrete distributions. In *Proc. 26th STOC*, pages 273–282, 1994.

[KMV10] A. T. Kalai, A. Moitra, and G. Valiant. Efficiently learning mixtures of two Gaussians. In *STOC*, pages 553–562, 2010.

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

[KT59] A. N. Kolmogorov and V. M. Tihomirov. ε-entropy and ε-capacity of sets in function spaces. *Uspehi Mat. Nauk*, 14:3–86, 1959.

[Lor66] G. G. Lorentz. Metric entropy and approximation. *Bull. Amer. Math. Soc.*, 72:903–937, 1966.

[Mak86] Y. Makovoz. On the kolmogorov complexity of functions of finite smoothness. *Journal of Complexity*, 2(2):121 – 130, 1986.

[Mil96] I. Milchtaich. Congestion games with player-specific payoff functions. *Games and Economic Behavior*, 13(1):111 – 124, 1996.

[MV10] A. Moitra and G. Valiant. Settling the polynomial learnability of mixtures of Gaussians. In *FOCS*, pages 93–102, 2010.

[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.

[Rud76] W. Rudin. *Principles of mathematical analysis*. McGraw-Hill Book Co., New York, 1976. International Series in Pure and Applied Mathematics.

[Sco92] D.W. Scott. *Multivariate Density Estimation: Theory, Practice and Visualization*. Wiley, New York, 1992.

[Sil86] B. W. Silverman. *Density Estimation*. Chapman and Hall, London, 1986.

[Tsy08] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 2008.

[Val84] L. G. Valiant. A theory of the learnable. In *Proc. 16th Annual ACM Symposium on Theory of Computing (STOC)*, pages 436–445. ACM Press, 1984.

[vdVW96] A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

[VW02] S. Vempala and G. Wang. A spectral algorithm for learning mixtures of distributions. In *Proceedings of the 43rd Annual Symposium on Foundations of Computer Science*, pages 113–122, 2002.

[Yat85] Y. G. Yatracos. Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *Annals of Statistics*, 13:768–774, 1985.

[YB99] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Ann. Statist.*, 27(5):1564–1599, 1999.

Appendix

A Lower Bounds on Matching Moments

We start by giving an explicit example of two PBDs over $k+1$ variables that agree exactly on the first k moments and have total variation distance $2^{-\Omega(k)}$.

Proposition 29. *Let $\mathbf{P}, \mathbf{Q} \in \mathcal{S}_{k+1,2}$ be PBD's with parameters $p_i = (1+\cos(\frac{2\pi i}{k+1}))/2$ and $q_i = (1+\cos(\frac{2\pi i+\pi}{k+1}))/2$ respectively, where $1 \leq i \leq k+1$. Then $\mathbf{P}$ and $\mathbf{Q}$ agree on their first k moments and have $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) \geq 4^{-k}$.*

Proof. Let $X = \sum_{i=1}^{k+1} X_i$, where X_i are independent Bernoulli variables, and suppose that $X \sim \mathbf{P}$. We note that, for $m \leq k$, the random variable X^m can be expressed as a degree m polynomial in the X_i's. Therefore, the m-th moment of $\mathbf{P}$ is a degree m symmetric polynomial of the p_i's. Similarly, the m-th moment of $\mathbf{Q}$ must be the same symmetric polynomial of the q_i. Therefore, to show that the first k moments of $\mathbf{P}$ and $\mathbf{Q}$ agree, it suffices to show that the first k elementary symmetric polynomials in the p_i have the same values as the corresponding polynomials of the q_i's.

Note that the p_i are the roots of $T_{k+1}(2x-1)-1$ and that the q_i are the roots of $T_{k+1}(2x-1)+1$, where T_{k+1} is the $(k+1)$-st Chebychev polynomial. Therefore, for $m \leq k$, the m-th elementary symmetric polynomial in the p_i is $[x^{k+1-m}](-1)^m 2^{-2k-1} T_{k+1}(2x+1)$ and the same holds for the q_i. Thus, the first k moments of $\mathbf{P}$ and $\mathbf{Q}$ agree. To bound the total variation distance from below we observe that

$$\prod_{i=1}^{k+1} p_i = \mathbf{P}(k+1) = [x^0](-1)^{k+1} 2^{-2k-1}(T_{k+1}(2x+1)-1),$$

and

$$\prod_{i=1}^{k+1} q_i = \mathbf{Q}(k+1) = [x^0](-1)^{k+1} 2^{-2k-1}(T_{k+1}(2x+1)+1).$$

Therefore, the probability that $\mathbf{P} = k+1$ and the probability that $\mathbf{Q} = k+1$ differ by 4^{-k}. This implies the appropriate bound in their variational distance and completes the proof. □

We also show that matching moments does not suffice for the case of k-SIIRVs, even for $k = 3$:

Proposition 30. *For n an even integer, there exist $\mathbf{P}, \mathbf{Q} \in \mathcal{S}_{n/2,3}$ with disjoint supports such that their first $n-1$ moments agree.*

Proof. We first show that there exist such $\mathbf{P}$ and $\mathbf{Q}$ with $\mathbf{P}$ supported on even numbers and $\mathbf{Q}$ supported on odd numbers, so that

$$\mathbf{P}(2j) = 2^{-n+1}\binom{n}{2j},$$

and

$$\mathbf{Q}(2j+1) = 2^{-n+1}\binom{n}{2j+1}.$$

We begin by showing that $\mathbf{P} \in \mathcal{S}_{n/2,3}$. Since $\sum_j 2^{-n+1}\binom{n}{2j} = 1$, we will show that the polynomial $\widetilde{\mathbf{P}}(z) = \sum_j 2^{-n+1}\binom{n}{2j}z^{2j}$ factors as a product of $n/2$ quadratic polynomials with non-negative coefficients. To prove this, we note that it suffices to show that all roots of $\widetilde{\mathbf{P}}$ are pure imaginary; then, the natural factorization into quadratics using complex conjugate pairs will complete the argument. For this, we observe that $\widetilde{\mathbf{P}}(z) = 2^{-n}((1+z)^n + (1-z)^n)$. Therefore, z is a root of $\widetilde{\mathbf{P}}$ only when $|1+z| = |1-z|$, or when z is equidistant from 1 and -1, which happens only when the real part of z is 0, i.e., when z is pure imaginary.

Similarly, we show that $\mathbf{Q} \in \mathcal{S}_{n/2,3}$. Once again $\sum_j 2^{-n+1}\binom{n}{2j+1} = 1$, and so we merely need to show that $\widetilde{\mathbf{Q}}(z) = \sum_j 2^{-n+1}\binom{n}{2j+1}z^{2j+1}$ factors into quadratics with non-negative coefficients. Since $\widetilde{\mathbf{Q}}(z) = 2^{-n}((1+z)^n - (1-z)^n)$, it also has only purely imaginary roots.

It remains to show that $\mathbf{P}$ and $\mathbf{Q}$ have identical first $n-1$ moments. For this, it suffices to show that $\widetilde{\mathbf{P}}(z)^{(k)}(1) = \widetilde{\mathbf{Q}}(z)^{(k)}(1)$ for all $0 \le k < n$. Indeed, we have that

$$\widetilde{\mathbf{P}}(z)^{(k)}(1) - \widetilde{\mathbf{Q}}(z)^{(k)}(1) = 2^{1-n}\frac{\partial^k}{\partial z^k}(1-z)^n|_{z=1} = \frac{2^{1-n}(1-z)^{n-k}n!}{(n-k)!}|_{z=1} = 0.$$

This completes the proof. □

B Omitted Proofs from Section 3

B.1 Proof of Lemma 9.

For convenience, we restate Lemma 9:

Lemma 9. *Let $\mathbf{P} \in \mathcal{S}_{n,k}$ be a k-SIIRV with $\mathrm{Var}[X] = V$. For any $0 < \delta < 1/4$, there exists $\mathbf{Q} \in \mathcal{S}_{n,k}$ with $d_{\mathrm{TV}}(\mathbf{P},\mathbf{Q}) = O(\delta V)$ such that all but $O(k + V/\delta)$ of the k-IRV's defining $\mathbf{Q}$ are constant.*

Proof. For a k-IRV A let $m(A)$ be an index i so that $\Pr[A = i]$ is maximized. Let $d(A) = \Pr[A \neq m(A)]$ be the probability A assigns to values in $[k] \setminus \{i\}$. Suppose that $d(A) \le 1/2$. Then we have that

$$d(A)/2 \le (1/2)\cdot \Pr(A \neq A') \le (1/2)\cdot \mathbb{E}[|A - A'|^2] = \mathrm{Var}[A] \le \mathbb{E}[|A - m(A)|^2] \le k^2 \cdot d(A),$$

where A' is an independent copy of A. The leftmost inequality follows from our assumption that $d(A) \le 1/2$. The proof of the lemma will make repeated applications of the following claim:

Claim 31. *Let A, B be independent k-IRV's with $m(A) = m(B)$ and $d(A) + d(B) \le 1/2$. Then there exist independent k-IRV's C and D, where D is a constant, $d(C) = d(A) + d(B)$, and $d_{\mathrm{TV}}(A + B, C + D) = O(d(A)d(B))$.*

Proof. Let $m(A) = m(B) = i$. Let $d(A) = \delta_1, d(B) = \delta_2$. Let A' be the random variable A conditioned on A not equaling i, and B' be the random variable B conditioned on it not equaling i. Note that A is a mixture of i and A' and B a mixture of i and B'. Furthermore $A + B$ equals $2i$ with probability $(1-\delta_1)(1-\delta_2)$, $i + A'$ with probability $\delta_1(1-\delta_2)$, $i + B'$ with probability $(1-\delta_1)\delta_2$ and $A' + B'$ with probability $\delta_1\delta_2$.

Let D be the random variable that is deterministically i and C be the random variable that equals i with probability $1 - \delta_1 - \delta_2$, A' with probability δ_1, and B' with probability δ_2. Then $C + D$ equals $2i$, $i + A'$, $i + B'$ and $A' + B'$ with probabilities $1 - \delta_1 - \delta_2$, δ_1, δ_2, and 0. These probabilities are within an additive $\delta_1\delta_2$ of the corresponding probabilities for $A + B$ and therefore $d_{\mathrm{TV}}(A + B, C + D) = O(\delta_1\delta_2)$. Note that $C = i$ with probability $1 - \delta_1 - \delta_2$, so $d(C) = \delta_1 + \delta_2$, which completes the proof. □

For a random variable $X \sim \mathbf{P}$, we have that $X = \sum_{i=1}^{n} A_i$ where the A_i's are independent k-IRV's. We iteratively modify $\mathbf{P}$ as follows: If two of the non-constant component k-IRV's of $\mathbf{P}$ are A and B, with $m(A) = m(B)$ and $d(A), d(B) < \delta$, then we replace the pair A and B with the pair C and D as described by the above claim. Notice that every step reduces the number of non-constant component variables, and therefore this process terminates, giving a k-SIIRV $\mathbf{Q}$ with for $Y \sim \mathbf{Q}$, $Y = \sum_{i=1}^{n} B_i$.

By construction, for each $1 \leq i \leq k$, $\mathbf{Q}$ has at most one non-constant component variable with $m(B_j) = i$ and $d(B_j) < \delta$. Claim 31 implies the sum of the d's of the component variables does not increase in any iteration, and therefore

$$\sum_{j=1}^{n} d(B_j) \leq \sum_{j=1}^{n} d(A_j) \leq 2 \sum_{j=1}^{n} \mathrm{Var}[A_j] = 2\mathrm{Var}[X] = 2V \ ,$$

where the second inequality uses the aforementioned lower bound on the variance of a k-IRV. Hence, the number of non-constant component variables in $\mathbf{Q}$ is at most $k + 2V\delta^{-1}$.

It remains to show that $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) = O(\delta V)$. Let A, B and C, D be the k-IRV's of Claim 31. Then $d_{\mathrm{TV}}(A + B, C + D) = O(d(A)d(B)) = O([d(C)^2 + d(D)^2] - [d(A)^2 + d(B)^2])$. That is, the total variation distance error introduced by replacing A, B by C, D is at most a constant times the amount that the sum of the squares of the d's of the component variables increases by. Repeated application of this observation combined with the sub-additivity of total variation distance gives $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) = O\left(\sum_{j=1}^{n} d(B_j)^2 - \sum_{j=1}^{n} d(A_j)^2\right)$. On the other hand, note that all of the B_j's that are not also A_j satisfy $d(B_j) \leq 2\delta$. Therefore, we have that $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) \leq O\left(\sum_{j:d(B_j)\leq 2\delta} d(B_j)^2\right) = O\left(\delta \sum_j d(B_j)\right) = O(\delta V)$, which completes the proof. □

B.2 Proof of Lemma 12. For convenience, we restate Lemma 12:

Lemma 12. *Fix $x \in \mathbb{C}$ with $|x| = 1$. Suppose that $\rho_1, \ldots, \rho_m$ are roots of $\widetilde{\mathbf{P}}(x)$ (listed with appropriate multiplicity) which have $|\rho_i - x| \leq \frac{1}{2k}$. Then, we have the following:*

(i) $|\widetilde{\mathbf{P}}(x)| \leq 2^{-m}$.

(ii) For the polynomial $q(x) = \widetilde{\mathbf{P}}(x) / \prod_{i=1}^{m}(x - \rho_i)$, we have that $|q(x)| \leq k^m$.

To prove our lemma, we will make essential use of the following simple lemma:

Lemma 32. *For any polynomial $p(x) \in \mathbb{C}[x]$ of degree d where the sum of the absolute values of the coefficients of p is at most 1, we have the following: Fix $z \in \mathbb{C}$ with $|z| = 1$. Suppose that p has roots $\rho_1, \ldots, \rho_m$ with $|\rho_i - z| \leq \frac{1}{2d}$, for $i \in \{1, \ldots, m\}$. Then, the following hold:*

(i) $|p(z)| \leq 2^{-m}$,

(ii) for the polynomial $q(x) = p(x) / \prod_{i=1}^{m}(x - \rho_i)$ we have that $|q(z)| \leq d^m$.

Proof. The lemma is proved by repeated applications of the following claim:

Claim 33. *Let $p(x) \in \mathbb{C}[x]$ be a degree-d polynomial such that the sum of the absolute values of the coefficients of p is at most 1. Let ρ be a root of $p(x)$ and $q(x)$ be the polynomial $\frac{p(x)}{x-\rho}$. Then, the sum of the absolute values of the coefficients of q is at most d.*

Proof. We write the coefficients of $p(x)$ and $q(x)$ as $p(x) = \sum_{i=0}^{d} p_i x^i$ and $q(x) = \sum_{i=0}^{d-1} q_i x^i$. Since $p(x) = (x-\rho)q(x)$, for $1 \le i \le d-1$, we have

$$p_i = q_{i-1} - \rho q_i \ , \tag{6}$$

and similarly $p_d = q_{d-1}$, $p_0 = -\rho q_0$.

We consider two cases based on the magnitude of ρ. First, suppose that $|\rho| \le 1$. Since $q_{d-1} = p_d$ and, by (6), $q_{i-1} = p_i + \rho q_i$, for $1 \le i \le d-1$, an easy induction gives that $q_i = \sum_{j=i+1}^{d} p_j \rho^{j-i-1}$ for $0 \le i \le d-1$. Summing and taking absolute values gives:

$$\begin{aligned}\sum_{i=0}^{d-1} |q_i| &\le \sum_{i=0}^{d-1}\sum_{j=i+1}^{d} |p_j||\rho|^{j-i-1} = \sum_{i=1}^{d}(|p_i| \sum_{j=0}^{i-1} |\rho|^j) \\ &\le \sum_{i=1}^{d} |p_i| i \le d \sum_{i=1}^{d} |p_i| \le d \ .\end{aligned}$$

Second, suppose $|\rho| > 1$. Then, $\frac{1}{|\rho|} < 1$. We have $q_0 = -\frac{1}{\rho} p_0$ and by (6), for $1 \le i \le d-1$, $q_i = \frac{1}{\rho}(q_{i-1} - p_i)$. By an easy induction, for $0 \le i \le d$, $q_i = -\sum_{j=0}^{i} p_j \frac{1}{\rho^{i-j}}$. Summing and taking absolute values gives:

$$\begin{aligned}\sum_{i=0}^{d-1} |q_i| &\le \sum_{i=0}^{d-1}\sum_{j=0}^{i} |p_j| \frac{1}{|\rho|^{i-j}} = \sum_{i=0}^{d-1}(|p_i| \sum_{j=i}^{d-1} \frac{1}{|\rho|^{d-1-i}}) \\ &\le \sum_{i=0}^{d-1} |p_i|(d-1-i) \le d \sum_{i=0}^{d-1} |p_i| \le d \ .\end{aligned}$$

□

By repeated applications of the claim it follows that the polynomial $q(x)$ has the sum of the absolute values of its coefficients at most d^m. Since $|z| = 1$, it follows that $|q(z)| \le d^m$ which gives (ii). To show (i) we note that

$$|p(z)| = |q(z)| \cdot \prod_{i=1}^{m} |z - \rho_i| \le |q(z)| \cdot (1/2d)^m \le 2^{-m} \ .$$

This completes the proof of Lemma 32. □

Proof of Lemma 12. Note that $\widetilde{\mathbf{P}}(x)$ is the degree $n(k-1)$ polynomial defined by $\widetilde{\mathbf{P}}(x) = \sum_{i=0}^{n(k-1)} \mathbf{P}(i) x^i$. Note that the sum of the absolute values of $\widetilde{\mathbf{P}}$'s coefficients is 1. However, to apply Lemma 32 directly to $\widetilde{\mathbf{P}}$ we would need the roots to be at distance at most $\frac{1}{2n(k-1)}$.

Note that $\widetilde{\mathbf{P}}(x)$ factors as $\prod_{i=1}^{n} p_i(x)$, where $p_i(x) = \mathbb{E}[x^{X_i}]$ is a degree $k-1$ polynomial that is determined by the i-th k-IRV. It is clear that the coefficients of $p_i(x)$ are non-negative and sum to 1, hence we may apply Lemma 32 to $p_i(x)$. Suppose that $p_i(x)$ has m_i roots with $|\rho_i - x| \le \frac{1}{2k}$. Lemma 32(i) implies that $|p_i(x)| \le 2^{-m_i}$. Since $\widetilde{\mathbf{P}}(x) = \prod_{i=1}^{n} p_i(x)$, this yields part (i) of Lemma 12.

Lemma 32(ii) implies that the polynomial $q_i(x) = p_i(x) / \prod_{j \in S_i}(x - \rho_j)$, for $S_i \subseteq \{1, \ldots, m\}$ with $|S_i| = m_i$, satisfies $|q_i(x)| \le k^{m_i}$. Note that $q(x) = \prod_{i=1}^{n} q_i(x)$. Therefore, $|q(x)| \le \prod_i k^{m_i} = k^m$, giving part (ii) of Lemma 12. □

B.3 Proper Cover Construction for the High Variance Case. Exhausting over the $k-1$ possible values of c, we can assume that c is known to the algorithm. Before proceeding further, we will need further structural information about the k-SIIRVs in this case. We start with the following simple lemma giving an upper bound on the total variation distance between two high variance k-SIIRVs:

Lemma 34. *For $\epsilon > 0$, let X, X' be k-SIIRVs with $\mathrm{Var}[X], \mathrm{Var}[X'] \geq \mathrm{poly}(k/\epsilon)$ for a sufficiently large $\mathrm{poly}(k/\epsilon)$ that have $d_{\mathrm{TV}}(X, Y + cZ) \leq \epsilon$ and $d_{\mathrm{TV}}(X', Y' + cZ') \leq \epsilon$ for c-IRVs Y,Y' and discrete Gaussians Z,Z', with $\mathbb{E}[X] = c\mathbb{E}[Z]$, $\mathrm{Var}[X] = c^2\mathrm{Var}[Z]$, $\mathbb{E}[X'] = c\mathbb{E}[Z']$ and $\mathrm{Var}[X'] = c^2\mathrm{Var}[Z']$. Then we have that*

$$d_{\mathrm{TV}}(X, X') \leq 4\epsilon + d_{\mathrm{TV}}\left(X \pmod c, X' \pmod c\right) + \frac{1}{2}\frac{|\mathbb{E}[X] - \mathbb{E}[X']|}{\sqrt{\mathrm{Var}[X]}} + \frac{1}{2}\frac{|\mathrm{Var}[X] - \mathrm{Var}[X']|}{\mathrm{Var}[X]}$$

where $X \pmod c$ is the c-IRV with $\Pr[X \pmod c = i] = \Pr[X \equiv i \pmod c]$ for $i \in [c]$.

Proof. Using Proposition 40, since $d_{\mathrm{TV}}(X, Y + cZ) \leq \epsilon$ with $Y \equiv Y + cZ \pmod c$, we have $d_{\mathrm{TV}}\left(X \pmod c, Y\right) \leq \epsilon$. Similarly, $d_{\mathrm{TV}}\left(X' \pmod c, Y'\right) \leq \epsilon$. By a combination of Propositions 40 and 42, we have that $d_{\mathrm{TV}}(Z, Z') \leq \frac{1}{2}\left(\frac{|\mathbb{E}[Z] - \mathbb{E}[Z']|}{\sqrt{\mathrm{Var}[Z]}} + \frac{|\mathrm{Var}[Z] - \mathrm{Var}[Z']|}{\mathrm{Var}[Z]}\right)$. Since $\mathbb{E}[X] = c\mathbb{E}[Z]$, $\mathrm{Var}[X] = c^2\mathrm{Var}[Z]$, $\mathbb{E}[X'] = c\mathbb{E}[Z']$ and $\mathrm{Var}[X'] = c^2\mathrm{Var}[Z']$ it follows that

$$\frac{|\mathbb{E}[Z] - \mathbb{E}[Z']|}{\sqrt{\mathrm{Var}[Z]}} + \frac{|\mathrm{Var}[Z] - \mathrm{Var}[Z']|}{\mathrm{Var}[Z]} = \frac{|\mathbb{E}[X] - \mathbb{E}[X']|}{\sqrt{\mathrm{Var}[X]}} + \frac{|\mathrm{Var}[X] - \mathrm{Var}[X']|}{\mathrm{Var}[X]}.$$

Therefore,

$$\begin{aligned} d_{\mathrm{TV}}(Y + cZ, Y' + cZ') &\leq d_{\mathrm{TV}}(Y, Y') + d_{\mathrm{TV}}(Z, Z') \\ &\leq 2\epsilon + d_{\mathrm{TV}}\left(X \pmod c, X' \pmod c\right) + \\ &+ \frac{1}{2}\left(\frac{|\mathbb{E}[X] - \mathbb{E}[X']|}{\sqrt{\mathrm{Var}[X]}} + \frac{|\mathrm{Var}[X] - \mathrm{Var}[X']|}{\mathrm{Var}[X]}\right). \end{aligned}$$

By another application of the triangle inequality, we have that $d_{\mathrm{TV}}(X, X') \leq d_{\mathrm{TV}}(X, Y + cZ) + d_{\mathrm{TV}}(Y + cZ, Y' + cZ') + d_{\mathrm{TV}}(Y' + cZ', X') \leq 2\epsilon + d_{\mathrm{TV}}(Y + cZ, Y' + cZ)$, which completes the proof. □

To use the above lemma, we need a way to characterize the constant c in the statement of Theorem 8, namely to show that the theorem applies to both X and X' for *the same value* of c. For a k-IRV A, let $m(A)$ be an index i so that $\Pr[A = i]$ is maximized. The following result is implicit in the proof of Theorem 8 in [DDO+13] (in particular, in Theorem 4.3 of that paper):

Lemma 35 ([DDO+13])**.** *Given a k-SIIRV $X = \sum_{i=1}^{n} X_i$ with $\mathrm{Var}[X] \geq \mathrm{poly}(k/\epsilon)$, let $\mathcal{H}$ be the set of integers b such that $\sum_{i=1}^{n} \Pr[X_i - m(X_i) = c] \geq \Theta(k^7/\epsilon^2)$ and $c = \gcd(\mathcal{H})$. Then there is a c-IRV Y and a discrete Gaussian Z with $d_{\mathrm{TV}}(X, Y + cZ) \leq \epsilon$.*

Let $X \in \mathcal{S}_{n,k}$ be a k-SIIRV with $\mathrm{Var}[X] \geq \mathrm{poly}(k/\epsilon)$ as in Case 2 of Theorem 8. Our main claim is that, up to ϵ error in total variation distance, we can assume that X has a special structure. In particular, we can take all but one of the component IRVs of X to be constant modulo c, with the last one being a c-IRV. More formally, we claim that there is a k-SIIRV X' with $d_{\mathrm{TV}}(X, X') \leq \epsilon$, such that $X' = \sum_{i=1}^{n} X'_i$ with

- For $1 \leq i \leq H$, where $H = \Theta(k^7/\epsilon^2)$, X'_i is either 0 or c each with equal probability.
- For $1 \leq i \leq n-1$, X'_i is constant modulo c.
- X'_n is a c-IRV.

where c is as in Lemma 35.

We can construct such an X' from X as follows. For $1 \leq i \leq H$, we replace X_i with the X'_i above that is 0 or c with equal probability. For $H+1 \leq i \leq n-1$, we replace each X_i by X_i conditioned on the event that $X_i \pmod c = m(X_i) \pmod c$. Finally we take X'_n to be $(X_n - \sum_{i=1}^{n-1} X'_i) \pmod c$ noting that $\sum_{i=1}^{n-1} X'_i \pmod c$ is a constant.

We now show that the above procedure only changes the expectation and variance by $|\mathbb{E}[X] - \mathbb{E}[X']| \leq \text{poly}(k/\epsilon)$ and $|\text{Var}[X] - \text{Var}[X']| \leq \text{poly}(k/\epsilon)$. Note that for two arbitrary k-IRVs, A and B, we have that $|\mathbb{E}[A] - \mathbb{E}[B]| \leq k$ and $|\text{Var}[A] - \text{Var}[B]| \leq k^2$. Thus,

$$|\mathbb{E}[X_n + \sum_{i=1}^{H} X_i] - \mathbb{E}[X'_n + \sum_{i=1}^{H} X'_i]| \leq (H+1)k \leq \text{poly}(k/\epsilon)$$

and

$$|\text{Var}[X_n + \sum_{i=1}^{H} X_i] - \text{Var}[X'_n + \sum_{i=1}^{H} X'_i]| \leq (H+1)k^2 \leq \text{poly}(k/\epsilon).$$

For the remaining variables $H+1 \leq i \leq n-1$, we have $d_{\text{TV}}(X_i, X'_i) \leq \Pr[X_i - m(X_i) \not\equiv 0 \pmod c]$ and so $|\mathbb{E}[X_i] - \mathbb{E}[X'_i]| \leq k \Pr[X_i - m(X_i) \not\equiv 0 \pmod c]$ and $|\text{Var}[X_i] - \text{Var}[X'_i]| \leq k^2 \Pr[X_i - m(X_i) \not\equiv 0 \pmod c]$. For each integer $0 \leq b \leq k-1$ that does not divide c, by Lemma 35, we must have that $b \notin \mathcal{H}$ and hence $\sum_{i=1}^{n} \Pr[X_i - m(X_i) = b] = O(k^7/\epsilon^2)$. Thus, $\sum_{i=1}^{n} \Pr[X_i - m(X_i) \not\equiv 0 \pmod c] = O(k^8/\epsilon^2)$.

If $\text{Var}[X]$ is a sufficiently large $\text{poly}(k/\epsilon)$, then $\text{Var}[X']$ is large enough that we can apply Theorem 8 and Lemma 35 to X'. Note that $\sum_{i=1}^{n} \Pr[|X'_i - m(X'_i)| = c] \geq \sum_{i=1}^{H} \Pr[|X'_i - m(X'_i)| = c] = H/2$. We thus have that either $c \in \mathcal{H}$ or $-c \in \mathcal{H}$. Since for b that does not divide c, we have $\sum_{i=1}^{n} \Pr[X'_i - m(X'_i) = b] = \Pr[X'_n - m(X'_n) = b] \leq 1$ and thus $b \notin (H)$, we have that $\gcd(\mathcal{H}) = c$. Thus, for X with sufficiently large $\text{poly}(k/\epsilon)$ variance, we have that $d_{\text{TV}}(X, Y + cZ) \leq \epsilon/10$ and $d_{\text{TV}}(X', Y' + cZ') \leq \epsilon/10$ for the same $1 \leq c \leq k-1$ and c-IRVs Y, Y' and discrete Gaussians Z, Z'. In conclusion, we can apply Lemma 34 to X and X'. We have that $X' \pmod c = X'_n = X \pmod c$. We have shown that $\mathbb{E}[X] - \mathbb{E}[X'] \leq \text{poly}(k/\epsilon)$ and $\text{Var}[X] - \text{Var}[X'] \leq \text{poly}(1/\epsilon)$. If $\text{Var}[X]$ is a sufficiently large $\text{poly}(k/\epsilon)$ then we can make the contributions of each of these to $d_{\text{TV}}(X, X')$ in Lemma 34 smaller than $\epsilon/10$. Then we have $d_{\text{TV}}(X, X') \leq \epsilon$.

Since every k-SIIRV X in Case 2 is ϵ-close to an X' of the aforementioned form, to compute a proper cover for this case, we can consider only k-SIIRVs of the form stated above. By a similar argument as above, our cover only needs to ensure that the triple of $X \pmod c, \mathbb{E}[X], \text{Var}[X]$ is sufficiently close to any such triple achievable by an element of $\mathcal{S}_{n,k}$ of this form. Obtaining a cover of $X \pmod c$ is easy, as we only need to deal with the single term X_n that is non-constant modulo c, and produce a cover for c-IRVs. Indeed, it is straightforward to produce such a cover of size $O(k/\epsilon)^k$.

As explained in Section 3.1, we have an explicit cover for the discrete Gaussian random variables that can appear in this setting. However, we are left with the difficulty of producing an explicit k-SIIRV approximating one of these c times a discrete Gaussian whenever such an approximation is possible. Fortunately, we note that we only need to be able to approximately match the mean and the variance. Note that as above, the $H = \text{poly}(k/\epsilon)$ components that we are requiring to be

either 0 or c, and the one that is a c-IRV can be assumed to have negligible effect on the final mean and variance if we had a sufficiently large $\text{poly}(k/\epsilon)$ threshold for the variance.

Let C be the largest multiple of c that is at most k. Let $\mathcal{S}_{n,k,c}$ be the set of k-SIIRVs on n components all of which are constant modulo c. For a given $\sigma > \text{poly}(k/\epsilon)$ and μ we need to determine whether or not there is an element of $\mathcal{S}_{n,k,c}$ whose mean and variance match μ and σ to within $\epsilon\sigma$, and if so to produce one. To do this, we first need a couple of observations about which μ, σ are attainable.

Observation 36. *For $\mathbf{P} \in \mathcal{S}_{n,k,c}$, $\text{Var}_{X\sim\mathbf{P}}[X] < nC^2/4$.*

Proof. This is because any k-IRV that is constant modulo c has a distance of at most C between its minimum and maximum values, and thus has variance at most $C^2/4$. □

Observation 37. *For $\mathbf{P} \in \mathcal{S}_{n,k,c}$ and $X \sim \mathbf{P}$, if $\mathbb{E}[X] \leq nC/2$, then $\text{Var}[X] \leq C\mathbb{E}[X] - \mathbb{E}[X]^2/n$.*

Proof. We note that in the range in question the quantity $C\mathbb{E}[X] - \mathbb{E}[X]^2/n$ is increasing in $\mathbb{E}[X]$, and therefore, we may show that for any given achievable variance the minimum possible expectation satisfies this inequality. Note that for the minimum achievable expectation, we may assume that each of the component IRVs is deterministically 0 modulo c, since otherwise we could subtract a constant from it, which would decrease the expectation and leave the variance unchanged. The observation now follows given that for any k-IRV, Y that has $\Pr[Y \pmod c = 0] = 1$ it holds $\text{Var}[Y] = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 \leq C\mathbb{E}[Y] - \mathbb{E}[Y]^2$. □

Observation 38. *For $\mathbf{P} \in \mathcal{S}_{n,k,c}$ and $X \sim \mathbf{P}$, if $\mathbb{E}[X] \geq nk - nC/2$, then $\text{Var}[X] \leq C(nk - \mathbb{E}[X]) - (nk - \mathbb{E}[X])^2/n$.*

Proof. This follows from the previous observation by considering the random variable $nk - X$. □

We now claim that any pair of expectation and variance μ and σ^2 not disallowed by the above observations may be approximated by an explicitly computable element of $\mathcal{S}_{n,k,c}$. Note that, by symmetry, we may assume that $\mu \leq nk/2$. If $\mu \geq 2\sigma^2/C$, we may make $\lfloor 4\sigma^2/C^2 \rfloor \leq n$ of our IRVs either x_i or $x_i + C$ with equal probability for some integers $0 \leq x_i \leq k-1$ and all other X_i with $H+1 \leq i \leq n-1$ constant. By adjusting the x_i's and the constants, we can make the expectation of X satisfy $|\mathbb{E}[X] - \mu| \leq 1$ so long as $\mu \geq 2\sigma^2/C$, and the variance $\text{Var}[X] = C^2\lfloor 4\sigma^2/C^2 \rfloor$ satisfies $|\text{Var}[X] - \sigma^2| \leq 1$.

Otherwise, if $\mu \leq 2\sigma^2/C$, let $\sigma^2 = C\mu \cdot q$ with $1 > q > 1/2$. We then use a sum of k-IRVs that are 0 with probability q and C with probability $1-q$, and some k-IRVs that are deterministically 0. If we have a many IRVs of the first type, then we get a mean and variance of $\mathbb{E}[X] = a(1-q)C$ and $\text{Var}[X] = aq(1-q)C$. Letting a be approximately $\text{Var}[X]/(q(1-q)C)$ completes the argument. We simply need to verify that in this case $a \leq n$ i.e., that $\sigma^2/(q(1-q)C) \leq n$. Indeed, note that

$$\text{Var}[X]/(q(1-q)C) = \frac{\text{Var}[X]}{(\text{Var}[X]/(C\mathbb{E}[X]))(1 - (\text{Var}[X]/(C\mathbb{E}[X])))C} = \frac{C\mathbb{E}[X]^2}{C\mathbb{E}[X] - \text{Var}[X]} \leq n$$

by Observation 37. This shows that given a discrete Gaussian, Z so that cZ approximates some element of $\mathcal{S}_{n,k,c}$, we can efficiently find such an element. In Section 3.1 we gave an appropriately small cover of the set of such Gaussians, which consists of a grid of means and variances of size $O(n)$. It is easy to construct such a grid and by the above, we can construct an X with $|\mathbb{E}[X] - c\mu| \leq \text{poly}(k/\epsilon)$ and $|\text{Var}[X] - c^2\sigma^2| \leq \text{poly}(k/\epsilon)$ for each μ, σ^2 in the grid that is not disallowed by our observations. Thus, we can efficiently find a cover of the elements of $\mathcal{S}_{n,k}$ satisfying Case 2 of Theorem 8.

C Omitted Proofs from Section 4

C.1 Proof of Lemma 28. For completeness, we restate the lemma below.

Lemma 28. *There exists a universal constant $c > 0$ such that the following holds: For any log-concave distribution $\mathbf{P}$ supported on the integers and standard deviation σ, there exist at least $\Omega(\sigma)$ consecutive integers with probability mass under $\mathbf{P}$ at least $c \cdot \frac{1}{1+\sigma}$.*

Proof. Note that if $\sigma \leq 1$, taking the mode trivially satisfies this property.

Without loss of generality we can assume that 0 is the mode of $\mathbf{P}$. We know that $\sum_{x\in\mathbb{Z}} x^2\mathbf{P}(x) = \Theta(\sigma^2)$. Let $\sigma_+^2 = \sum_{x>0} x^2\mathbf{P}(x)$. Let t_+ be the largest integer so that $\mathbf{P}(t_+ + 1)/\mathbf{P}(t_+) \leq e^{1/t_+}$. We note that

$$\sum_{x>0} x^2\mathbf{P}(x) \leq \sum_{x=0}^{\infty} x^2\mathbf{P}(t_+)e^{-(x-t_+)/t_+} = \Theta(t_+^3\mathbf{P}(0)),$$

and

$$\sum_{x>0} x^2\mathbf{P}(x) \geq \mathbf{P}(t_+)\sum_{x=0}^{t_+} x^2 = \Theta(t_+^3\mathbf{P}(0)).$$

Also note that

$$\sum_{x>0} \mathbf{P}(x) \leq \sum_{x=0}^{\infty} \mathbf{P}(t_+)e^{-(x-t_+)/t_+} = \Theta(t_+\mathbf{P}(0)).$$

Similarly, defining σ_- and t_-, we find that $\sigma^2 = \Theta(\sigma_+^2+\sigma_-^2) = \Theta(\mathbf{P}(0)(t_+^3+t_-^3))$. Thus, $\max(t_+, t_-)^3\mathbf{P}(0) = \Theta(\sigma^2)$ and $\max(t_+, t_-)\mathbf{P}(0) = \Omega(1)$. Without loss of generality this maximum is t_+. Note that for all $0 \leq x \leq t_+$ that $\mathbf{P}(x) = \Theta(\mathbf{P}(t_+))$. This implies that $t_+\mathbf{P}(0) = O(1)$, and thus, by the above is $\Theta(1)$. Therefore, it follows by the variance bounds that $t_+^2 = \Omega(\sigma^2)$, so $t_+ = \Theta(\sigma)$. Hence, $x = 0, 1, \ldots, t_+$ are $\Omega(\sigma)$ terms on which the value of $\mathbf{P}$ is $\Omega(1/t_+) = \Omega(1/\sigma)$. This completes the proof. □

D Basic Facts from Probability

Definition 39. Let $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}^{\geq 0}$. We let $Z(\mu, \sigma^2)$ denote the *discretized normal* distribution. The definition of $Z \sim Z(\mu, \sigma^2)$ is that we first draw a normal $G \sim \mathrm{N}(\mu, \sigma^2)$ and then we set $Z = \lfloor G \rceil$; i.e., G rounded to the nearest integer.

We begin by recalling some basic facts concerning total variation distance, starting with the "data processing inequality for total variation distance":

Proposition 40 (Data Processing Inequality for Total Variation Distance)**.** *Let X, X' be two random variables over a domain Ω. Fix any (possibly randomized) function F on Ω (which may be viewed as a distribution over deterministic functions on Ω) and let $F(X)$ be the random variable such that a draw from $F(X)$ is obtained by drawing independently x from X and f from F and then outputting $f(x)$ (likewise for $F(X')$). Then we have $d_{\mathrm{TV}}(F(X), F(X')) \leq d_{\mathrm{TV}}(X, X')$.*

Next we recall the subadditivity of total variation distance for independent random variables:

Proposition 41. *Let A, A', B, B' be integer random variables such that (A, A') is independent of (B, B'). Then $d_{\mathrm{TV}}(A + B, A' + B') \leq d_{\mathrm{TV}}(A, A') + d_{\mathrm{TV}}(B, B')$.*

We will use the following standard result which bounds the variation distance between two normal distributions in terms of their means and variances:

Proposition 42. *Let $\mu_1, \mu_2 \in \mathbb{R}$ and $0 < \sigma_1 \leq \sigma_2$. Then $d_{\mathrm{TV}}(\mathrm{N}(\mu_1, \sigma_1^2), \mathrm{N}(\mu_2, \sigma_2^2)) \leq \frac{1}{2}\left(\frac{|\mu_1-\mu_2|}{\sigma_1} + \frac{\sigma_2^2-\sigma_1^2}{\sigma_1^2}\right)$.*