JMLR: Workshop and Conference Proceedings vol 49:1–19, 2016

Optimal Learning via the Fourier Transform for Sums of Independent Integer Random Variables

Ilias Diakonikolas DIAKONIK@USC.EDU
Dept. of Computer Science, University of Southern California, Los Angeles, CA, USA

Daniel M. Kane DAKANE@CS.UCSD.EDU
Dept. of Computer Science & Engineering, and Mathematics, University of California, San Diego, CA, USA

Alistair Stewart ALISTAIS@USC.EDU
Dept. of Computer Science, University of Southern California, Los Angeles, CA, USA

Abstract

We study the structure and learnability of sums of independent integer random variables (SIIRVs). For $k \in \mathbb{Z}_+$, a *k-SIIRV of order $n \in \mathbb{Z}_+$* is the 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-SIIRVs of order n.

How many samples are required to learn an arbitrary distribution in $\mathcal{S}_{n,k}$? In this paper, we tightly characterize the sample and computational complexity of this problem. More precisely, we design a computationally efficient algorithm that uses $\widetilde{O}(k/\epsilon^2)$ samples, and learns an arbitrary k-SIIRV within error ϵ, in total variation distance. Moreover, we show that the *optimal* sample complexity of this learning problem is $\Theta((k/\epsilon^2)\sqrt{\log(1/\epsilon)})$, i.e., we prove an upper bound and a matching information-theoretic lower bound. Our algorithm proceeds by learning the Fourier transform of the target k-SIIRV in its effective support. Its correctness relies on the *approximate sparsity* of the Fourier transform of k-SIIRVs – a structural property that we establish, roughly stating that the Fourier transform of k-SIIRVs has small magnitude outside a small set.

Along the way we prove several new structural results about k-SIIRVs. As one of our main structural contributions, we give an efficient algorithm to construct a sparse *proper* ϵ-cover for $\mathcal{S}_{n,k}$, in total variation distance. We also obtain a novel geometric characterization of the space of k-SIIRVs. Our characterization allows us to prove a tight lower bound on the size of ϵ-covers for $\mathcal{S}_{n,k}$ – establishing that our cover upper bound is optimal – and is the key ingredient in our tight sample complexity lower bound.

Our approach of exploiting the sparsity of the Fourier transform in distribution learning is general, and has recently found additional applications. In a subsequent work Diakonikolas et al. (2015c), we use a generalization of this idea to obtain the first computationally efficient learning algorithm for Poisson multinomial distributions. In Diakonikolas et al. (2015b), we build on our Fourier-based approach to obtain the fastest known proper learning algorithm for Poisson binomial distributions (2-SIIRVs).

Keywords: density estimation, distribution learning, sums of independent random variables, Fourier transform, metric entropy

[1]

1. Extended abstract. Full version appears as [CoRR, abs/1505.00662, v2] Diakonikolas et al. (2015a).

1. Introduction

1.1. Motivation and Background

We study sums of independent integer random variables:

Definition. For $k \in \mathbb{Z}_+$, a *k-IRV* is any 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 the random variable or its corresponding distribution, and the distinction will be clear from the context.

Sums of independent integer random variables (SIIRVs) comprise a rich class of distributions that arise in many settings. The special case of $k = 2$, $\mathcal{S}_{n,2}$, was first considered by Poisson Poisson (1837) as a non-trivial extension of the Binomial distribution, and is known as Poisson binomial distribution (PBD). In application domains, SIIRVs have many uses in research areas such as survey sampling, case-control studies, and survival analysis, see e.g., Chen and Liu (1997) for a survey of the many practical uses of these distributions. We remark that these distributions are of fundamental interest and have been extensively studied in probability and statistics. For example, tail bounds on SIIRVs form an important special case of Chernoff/Hoeffding bounds Chernoff (1952); Hoeffding (1963); Dubhashi and Panconesi (2009). Moreover, there is a long line of research on approximate limit theorems for SIIRVs, dating back several decades (see e.g., Presman (1983); Kruopis (1986); Barbour et al. (1992)), and Chen and Leong (2010); Chen et al. (2011) for some recent results.

Structure and Learning of k-SIIRVs. The main motivation of this work was the problem of learning an unknown k-SIIRV given access to independent samples. Understanding this problem is intimately related to obtaining a refined structural understanding of the space of k-SIIRVs. The connection between structure and distribution learning is the main thrust of this paper.

Distribution learning or *density estimation* is the following task Devroye and Györfi (1985); Kearns et al. (1994); Devroye and Lugosi (2001): Given independent samples from an unknown distribution $\mathbf{P}$ in a family $\mathcal{D}$, and an error tolerance $\epsilon > 0$, output a hypothesis $\mathbf{H}$ such that with high probability the total variation distance $d_{\mathrm{TV}}(\mathbf{H}, \mathbf{P})$ is at most ϵ. The sample and computational complexity of this unsupervised learning problem depends on the *structure* of the underlying family $\mathcal{D}$. The main goals here are: (i) to characterize the *sample complexity* of the learning problem, i.e., to obtain matching information-theoretic upper and lower bounds, and (ii) to design a *computationally efficient* learning algorithm – i.e., an algorithm whose running time is polynomial in the sample (input) size – that uses an information-theoretically optimal sample size.

While density estimation has been studied for several decades, the number of samples required to learn is not yet well understood, even for surprisingly simple and natural classes of univariate discrete distributions. More specifically, there is no known complexity measure of a distribution family $\mathcal{D}$ that *characterizes* the sample complexity of learning an unknown distribution from $\mathcal{D}$. In contrast, the VC dimension of a concept class plays such a role in the PAC model of learning Boolean functions (see, e.g, Blumer et al. (1989); Kearns and Vazirani (1994)).

We remark that the classical information-theoretic quantity of the *metric entropy* van der Vaart and Wellner (1996); Devroye and Lugosi (2001); Tsybakov (2008), i.e., the logarithm of the size of

the smallest ϵ-cover[2] of the distribution class, provides an *upper bound* on the sample complexity of learning. Alas, this upper bound is suboptimal in general – both quantitatively and qualitatively – and in particular for the class of k-SIIRVs, as we show in this paper.

Obtaining a computationally efficient learning algorithm with optimal (or near-optimal) sample complexity is an important goal. In many learning settings, achieving this goal turns out to be quite challenging. More specifically, in many scenarios, both supervised and unsupervised, the only computationally efficient learning algorithms known use a (provably) suboptimal sample size. Intuitively, increasing the sample size (e.g., by a polynomial factor) can make the algorithmic task substantially easier. Characterizing the tradeoff between sample complexity and computational complexity is of fundamental importance in learning theory. In this work, we essentially characterize this tradeoff for the unsupervised problem of learning SIIRVs.

1.2. Our Results

The main technical contribution of this paper is the use of Fourier analytic and geometric tools to obtain a refined structural understanding of the space of k-SIIRVs. As a byproduct of our techniques, we characterize the sample complexity of learning k-SIIRVs (up to constant factors), and moreover we obtain a computationally efficient learning algorithm with near-optimal sample complexity. Our results answer the main open questions of Daskalakis et al. (2012b, 2013).

Along the way we prove several new structural results of independent interest about k-SIIRVs, including: the approximate sparsity of their Fourier transform; tight upper and lower bounds on ϵ-covers (in total variation distance and Kolmogorov distance); and a novel geometric characterization of the space of k-SIIRVs, that is crucial for our sample complexity lower bound. Below, we state our results in detail and elaborate on their context and the connections between them.

Learning SIIRVs via the Fourier Transform. As our first result, we give a sample near-optimal and computationally efficient learning algorithm for k-SIIRVs:

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

Our algorithm outputs a succinct description of the hypothesis $\mathbf{H}$, via its Discrete Fourier Transform (DFT) $\widehat{\mathbf{H}}$, which is supported on a set of small cardinality. The DFT immediately gives a fast evaluation oracle for $\mathbf{H}$. We also show how to use the DFT, in a black-box manner, to obtain an efficient approximate sampler for the target distribution $\mathbf{P}$.

We remark that the sample complexity of our algorithm is optimal up to logarithmic factors. Indeed, even learning a single k-IRV to variation distance ϵ requires $\Omega(k/\epsilon^2)$ samples. For the case of $k = 2$, Daskalakis et al. (2012b) gave a learning algorithm that uses $\widetilde{O}(1/\epsilon^2)$ samples, but runs in quasi-polynomial time, namely $(1/\epsilon)^{\mathrm{polylog}(1/\epsilon)}$. More recently, Daskalakis et al. (2013) studied the case of general k, and gave an algorithm that uses $\mathrm{poly}(k/\epsilon)$ samples and time. Notably, the degree

2. Formally, a subset $\mathcal{D}_\epsilon \subseteq \mathcal{D}$ in a metric space $(\mathcal{D}, d)$ is said to be an *ϵ-cover of $\mathcal{D}$* with respect to the metric $d : \mathcal{X}^2 \to \mathbb{R}_+$, if for every $\mathbf{x} \in \mathcal{D}$ there exists some $\mathbf{y} \in \mathcal{D}_\epsilon$ such that $d(\mathbf{x}, \mathbf{y}) \leq \epsilon$. In this paper, we focus on the total variation distance between distributions.
3. We work in the standard "word RAM" model in which basic arithmetic operations on $O(\log n)$-bit integers are assumed to take constant time.

of this polynomial is quite high: the sample complexity of the Daskalakis et al. (2013) algorithm is $\Omega(k^9/\epsilon^6)$. Theorem 1 gives a nearly-tight upper bound on the sample complexity of this learning problem, 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. If true, this would imply that learning a k-SIIRV is as easy as learning a k-IRV. Surprisingly, we show that this is not the case:

Theorem 2 (Optimal Sample Complexity) *For any $k \in \mathbb{Z}_+$, $\epsilon \leq 1/\mathrm{poly}(k)$, there is an algorithm that learns k-SIIRVs within variation distance ϵ using $O((k/\epsilon^2)\sqrt{\log(1/\epsilon)})$ samples. Moreover, any algorithm for this problem information-theoretically requires $\Omega((k/\epsilon^2)\sqrt{\log(1/\epsilon)})$ samples.*

Theorem 2 precisely characterizes the sample complexity of learning k-SIIRVs (up to constant factors) by giving an upper bound and a matching information-theoretic sample lower bound. The sharp sample complexity bound of $\Theta((k/\epsilon^2)\sqrt{\log(1/\epsilon)})$ is surprising, and cannot be obtained using standard information-theoretic tools (e.g., metric entropy). We elaborate on this issue in Section 1.4.

We remark that the upper bound of Theorem 2 does not specify the running time of the corresponding algorithm. This is because the simplest such algorithm actually runs in time exponential in k. For the important special case of $k = 2$, we obtain a sample–optimal learning algorithm that runs in sample–linear time:

Theorem 3 (Optimal Learning of PBDs (2-SIIRVs)) *For any $\epsilon > 0$, there is an algorithm that learns PBDs within variation distance ϵ using $O((1/\epsilon^2)\sqrt{\log(1/\epsilon)})$ samples and running in time $O((1/\epsilon^2)\sqrt{\log(1/\epsilon)})$.*

Using the Fourier Transform for Distribution Learning. Our learning upper bounds are obtained via an approach which is novel in this context. Specifically, we show that the Fourier transform of k-SIIRVs is *approximately sparse*, and exploit this property to learn the distribution *via learning its Fourier transform in its effective support*. The sparsity of the Fourier transform explains why this family of distributions is learnable with sample complexity independent of n, and moreover it yields the sharp sample-complexity bound. The algorithmic idea of exploiting Fourier sparsity for distribution learning is general, and was subsequently used by the authors in other related settings Diakonikolas et al. (2015c,b).

Structure of k-SIIRVs. Our core structural result is the following simple property of the Fourier transform of k-SIIRVs:

Any k-SIIRV with "large" variance has a Fourier transform with "small" effective support.

One can obtain different versions of the above informal statement depending on the setting and the desired application. See Lemma 7 for a formal statement in the context of the DFT. The Fourier sparsity of k-SIIRVs forms the basis for our upper bounds in this paper. As previously mentioned, this structural property motivates and enables our learning algorithm. Moreover, it is useful in order to obtain sparse ϵ-covers for $\mathcal{S}_{n,k}$, the space of k-SIIRVs, under the total variation distance.

More specifically, using the approximate sparsity of the Fourier transform of SIIRVs combined with analytic arguments, we obtain a computationally efficient algorithm to construct a *proper* ϵ-cover for $\mathcal{S}_{n,k}$, of near-minimum size. In particular, we show:

Theorem 4 (Optimal Covers for k-SIIRVs) *For $\epsilon \leq 1/k$, there exists a proper 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.*

The best previous upper bound on the cover size of 2-SIIRVs is $n^2 + n \cdot (1/\epsilon)^{O(\log^2(1/\epsilon))}$ Daskalakis and Papadimitriou (2009, 2014). For $k > 2$, Daskalakis et al. (2013) gives a *non-proper* cover of size $n \cdot 2^{\mathrm{poly}(k/\epsilon)}$.

Our proper cover upper bound construction provides a smaller search space for essentially any optimization problem over k-SIIRVs. Specifically, Theorem 4 has the following implication in computational game theory: Via a connection established in Daskalakis and Papadimitriou (2007, 2009), the proper cover construction of Theorem 4 (for $k = 2$) yields 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" cannot lead to an FPTAS for this problem. We note that computing an (exact) Nash equilibrium in an anonymous game with a constant number of strategies was recently shown to be intractable Chen et al. (2015). Our cover upper bound is proved in Section 3.

We also prove a matching lower bound on the cover size, showing that our above construction is essentially optimal:

Theorem 5 (Cover Size Lower Bound for k-SIIRVs) *For $\epsilon \leq 1/\mathrm{poly}(k)$, and $n = \Omega(\log(1/\epsilon))$, any ϵ-cover for $\mathcal{S}_{n,k}$ has size at least $n \cdot (1/\epsilon)^{\Omega(k \log(1/\epsilon))}$.*

Before our work, no non-trivial lower bound on the cover size was known. We view the inherent quasi-polynomial dependence on $1/\epsilon$ of the cover size established here as a rather surprising fact. Our cover size lower bound proof relies on a new geometric characterization of the space of k-SIIRVs that we believe is of independent interest, and may find other applications. Our tight lower bound on the sample complexity of learning k-SIIRVs relies critically on this characterization. Our cover size lower bound is proved in Section 4.

1.3. Preliminaries

We record a few definitions that will be used throughout this paper.

Distributions and Metrics. For $m \in \mathbb{Z}_+$, we denote $[m] \stackrel{\text{def}}{=} \{0, 1, \ldots, m\}$. A function $\mathbf{P} : A \to \mathbb{R}$, over a finite set A, is called a *distribution* if $\mathbf{P}(a) \geq 0$ for all $a \in A$, and $\sum_{a \in A} \mathbf{P}(a) = 1$. The function $\mathbf{P}$ is called a *pseudo-distribution* if $\sum_{a \in A} \mathbf{P}(a) = 1$. For a pseudo-distribution $\mathbf{P}$ over $[m]$, $m \in \mathbb{Z}_+$, 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 (pseudo-)distributions $\mathbf{P}$ and $\mathbf{Q}$ supported on a finite set A is $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) \stackrel{\text{def}}{=} \max_{S \subseteq A} |\mathbf{P}(S) - \mathbf{Q}(S)| = (1/2) \cdot \|\mathbf{P} - \mathbf{Q}\|_1$. 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 $d_{\mathrm{K}}(\mathbf{P}, \mathbf{Q}) \stackrel{\text{def}}{=} \sup_{x \in \mathbb{R}} |\mathbf{P}(\leq x) - \mathbf{Q}(\leq x)|$. Note that for any pair of distributions $\mathbf{P}$ and $\mathbf{Q}$ supported on a finite subset of $\mathbb{R}$ we have that $d_{\mathrm{K}}(\mathbf{P}, \mathbf{Q}) \leq d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q})$.

Distribution Learning. Since we are interested in the computational complexity of distribution learning, our algorithms will need to use a *succinct description* of their hypotheses. A simple succinct representation of a discrete distribution is via an evaluation oracle for the probability mass function. For $\epsilon > 0$, an *ϵ-evaluation oracle* for a distribution $\mathbf{P}$ over $[m]$ is a polynomial size circuit C with $O(\log m)$ input bits such that for each input z, the output of the circuit $C(z)$ equals the binary representation of the probability $\mathbf{P}'(z)$, for some pseudo-distribution $\mathbf{P}'$ which has $d_{\mathrm{TV}}(\mathbf{P}', \mathbf{P}) \leq \epsilon$. Another general way to succinctly specify a distribution is to give the code of an efficient algorithm that takes "pure" randomness and transforms it into a sample from the distribution. This is the standard notion of a sampler. An *ϵ-sampler* for $\mathbf{P}$ is a circuit C with $O(\log m + \log(1/\epsilon))$ input bits z and $O(\log m)$ output bits y which is such that when $z \sim U_m$, then $y \sim \mathbf{P}'$, for some distribution $\mathbf{P}'$ which has $d_{\mathrm{TV}}(\mathbf{P}', \mathbf{P}) \leq \epsilon$.

We emphasize that our learning algorithms output *both an ϵ-sampler and an ϵ-evaluation oracle* for the target distribution.

Covers. 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 $d(\mathbf{P}, \mathbf{Q}) \leq \delta$. If $\mathcal{G}$ is not a subset of $\mathcal{F}$, then the cover is called non-proper. 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.

1.4. Our Approach and Techniques

The unifying idea of this work is an analysis of the structure of the Fourier Transform (FT) of k-SIIRVs. The FT is a natural tool to consider in this context. Recall that the FT of a sum of independent random variables is the product of the FT's of the individual variables. Moreover, if two random variables have similar FT's, they also have similar distributions. These two basic facts are the starting point of our analysis. We now provide an overview of the ideas underlying our results, and give a comparison to previous techniques.

Discussion & Previous Approaches for Learning SIIRVs. Let $\mathcal{D}$ be a family of distributions over a domain of size N. How many samples are required to learn an arbitrary $\mathbf{P} \in \mathcal{D}$ within variation distance ϵ? Without any restrictions on $\mathcal{D}$, it is a folklore fact that the sample complexity learning is $\Theta(N/\epsilon^2)$. The optimal learning algorithm in this case is the obvious one: output the empirical distribution. By exploiting the structure of the family $\mathcal{D}$, one may obtain better results.

A very natural type of structure to consider is some sort of "shape constraint" on the probability density function, such as log-concavity or unimodality. There is a long line of work in statistics on this topic (see, e.g., the books Barlow et al. (1972); Groeneboom and Jongbloed (2014)), and more recently in TCS Daskalakis et al. (2012a); Chan et al. (2013, 2014a,b); Acharya et al. (2015). Alas, it turns out that k-SIIRVs do not satisfy any of the shape constraints considered in the literature (see Daskalakis et al. (2013) for a discussion).

A different type of structure, based on the notion of metric entropy Yatracos (1985); Birgé (1986); Devroye and Lugosi (2001), yields the following implication: If a distribution class $\mathcal{D}$ has an $\epsilon/2$-cover of size M, then it is learnable with $O(\log M/\epsilon^2)$ samples.[4] In a celebrated paper in information theory Yang and Barron (1999), Yang and Barron show that, for broad families of

4. We remark that the running time of this method is $\Omega(M/\epsilon^2)$, which is not necessarily polynomial in the sample size.

(continuous) distributions, the metric entropy *characterizes* the sample complexity of learning. For k-SIIRVs, however, this is not the case: Via Theorem 4, the metric entropy method implies a sample upper bound of $O((1/\epsilon^2)\cdot\log n+(k/\epsilon^2)\cdot\log^2(1/\epsilon))$. Note that, since our cover size upper bound is tight, this sample bound is the limit of the metric entropy method for k-SIIRVs. Thus, this method gives a suboptimal sample upper bound for our learning problem, both qualitatively (dependence on n), and quantitatively (dependence on ϵ).

Previous work on learning k-SIIRVs Daskalakis et al. (2012b, 2013) relies on a certain "regularity" lemma about the structure of these distributions: 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 "Gaussian like". In the former case, the distribution can be learned using $O(L/\epsilon^2)$ samples, and in the latter case one can exploit the Gaussian structure to learn with a small number of samples as well. Unfortunately, the sparse case is a bottleneck for this approach, as any algorithm to learn a distribution over support L requires $\Omega(L/\epsilon^2)$ samples. Hence, one needs to exploit the structure of k-SIIRVs beyond the aforementioned.

Our Learning Approach. In this paper, we depart from the aforementioned approaches. We identify a simple condition – the approximate sparsity of the Fourier transform – as the "right" property that determines the sample complexity of our learning problem. The Fourier sparsity explains why the sample complexity of learning k-SIIRVs is independent of n, and allows us to obtain the sharp sample bound as a function of both k and ϵ. We show that this is a more general phenomenon (see Theorem 8): any univariate distribution that has an s-sparse Fourier transform, in a certain well-defined technical sense, is learnable with $\widetilde{O}(s/\epsilon^2)$ samples.

Our computationally efficient learning algorithm proceeds as follows: It starts by drawing an initial set of samples to determine the effective support of the target distribution and its Fourier transform. This is achieved by estimating the mean and variance of our SIIRV. We remark that, for computational purposes, our algorithm uses the Discrete Fourier Transform (DFT). For the appropriate definition of the DFT, we show (Lemma 7) there exists an *explicit* set S of cardinality $|S|=O(k^2\log(k/\epsilon))$ that contains all the "heavy" Fourier coefficients[5]. Our algorithm then draws an additional set of samples to estimate the DFT of the target distribution at the points of the effective support S, and sets the DFT to 0 everywhere else. By exploiting the sparsity in the Fourier domain, we show that the inverse of the empirical DFT achieves total variation distance $\epsilon/2$ after $\widetilde{O}(k/\epsilon^2)$ samples. Note that an explicit description of an accurate hypothesis for our learning problem can have an effective support of size $\Omega(k\sqrt{n})$. While we can easily obtain such a description (by explicitly computing the inverse DFT), this would not lead to a computationally efficient algorithm. We instead output a succinct description of our hypothesis (in time that is independent of n). In particular, our algorithm outputs the empirical DFT at the points of its effective support.

We emphasize that the implicit description of the hypothesis $\mathbf{H}$, via its DFT $\widehat{\mathbf{H}}$, is sufficient to obtain both an approximate evaluation oracle and an approximate sampler for the target k-SIIRV $\mathbf{P}$. Obtaining an approximate evaluation oracle is straightforward: Since $\widehat{\mathbf{H}}$ is supported on the set S, we can compute $\mathbf{H}(i)$ in time $O(|S|)$. To obtain an efficient sampler, we proceed in two steps: We first show how to efficiently compute the CDF of $\mathbf{H}$, using oracle access to the the DFT $\widehat{\mathbf{H}}$. To do this, we express the value of the CDF at any point via a closed form expression involving the values

5. We moreover show that there exists a set of cardinality $O(k\log(k/\epsilon))$ that contains all the "heavy" Fourier coefficients, alas this smaller set is not explicitly known a priori.

of $\widehat{\mathbf{H}}$. Given oracle access to the CDF, we use a simple binary search procedure to sample from a distribution $\mathbf{Q}$ satisfying $d_{\mathrm{TV}}(\mathbf{Q}, \mathbf{H}) \leq \epsilon/2$.

Finally, we note that our above-described Fourier-learning algorithm achieves a near-optimal sample complexity (up to logarithmic factors). The basic idea to obtain the *optimal* sample complexity is to smoothly mollify the DFT instead of truncating it. This removes some artifacts caused by a sharp truncation and yields a hypothesis whose error from the true distribution decays rapidly as we move away from the mean.

Cover Upper Bound. We start by commenting on previous approaches for proving cover upper bounds in this context. The main technique for the 2-SIIRV cover upper bound of Daskalakis and Papadimitriou (2009) is the following lemma (that is deduced in Daskalakis and Papadimitriou (2009) using a result from Roos (2000)): If two 2-SIIRVs agree on their first $\Omega(\log(1/\epsilon))$ moments, then their total variation distance is at most ϵ. First, we show that this moment-matching lemma is quantitatively tight: in the full version we give an example of two 2-SIIRVs over $k+1$ variables that agree on the first k moments and have variation distance $2^{-\Omega(k)}$.

We emphasize however that such a moment-matching technique cannot be generalized to 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 the full version we give an explicit example of two 3-SIIRVs of order $n/2$ that agree exactly on the first $n-1$ moments and have disjoint supports.

In conclusion, moment-based approaches fail to detect periodic structure. On the other hand, this type of structure is easily detectable by considering the Fourier transform. 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, 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.

Cover Lower Bound & Sample Lower Bound. 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 locally invertible. For $k = 2$, we in fact show that the distribution of any 2-SIIRV 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. In addition, for $n = \Theta(\log(1/\epsilon))$, we show that near a particular k-SIIRV not only is the map from parameters to distribution locally a bijection, but that this map is actually surjective onto a ball of reasonable size. In other words, near this particular distribution, the $\Omega(k \log(1/\epsilon))$ parameters of the output distribution 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 geometric understanding with Assouad's lemma Assouad (1983). We note that one might naively expect that such a situation would lead to a lower bound of $\Omega(k \log(1/\epsilon)/\epsilon^2)$, but since the distributions under

consideration have additional structure, it turns out that the best lower bound that can be obtained is $\Omega(k\sqrt{\log(1/\epsilon)}/\epsilon^2)$.

1.5. Related Work

Density estimation is a classical topic in statistics and machine learning with a rich history and extensive literature (see e.g., Barlow et al. (1972); Devroye and Györfi (1985); Silverman (1986); Scott (1992); Devroye and Lugosi (2001)). The reader is referred to Izenman (1991) for a survey of statistical techniques in this context. In recent years, a large body of work in TCS has been studying these questions from a computational perspective; see e.g., Kearns et al. (1994); Freund and Mansour (1999); Arora and Kannan (2001); Cryan et al. (2002); Vempala and Wang (2002); Feldman et al. (2005); Belkin and Sinha (2010); Kalai et al. (2010); Daskalakis et al. (2012a,b, 2013); Chan et al. (2013, 2014a,b); Acharya et al. (2015).

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., Dudley (1974); Makovoz (1986); Blei et al. (2007) and the books Kolmogorov and Tihomirov (1959); Lorentz (1966); Carl and Stephani (1990); Edmunds and Triebel (1996)), geometric approximation algorithms Har-peled (2011), information theory, statistics, and machine learning (see, e.g., Yatracos (1985); Birgé (1986); Hasminskii and Ibragimov (1990); Haussler and Opper (1997); Yang and Barron (1999); Guntuboyina and Sen (2013) and the books van der Vaart and Wellner (1996); Devroye and Lugosi (2001); Tsybakov (2008)).

A preliminary version of this paper (with a different title) was disseminated in Diakonikolas et al. (2015a) (May 4, 2015). Since then, a number of works on SIIRVs and related families have been announced. We briefly summarize their relation to this paper.

Concurrent Work. Concurrent work by Daskalakis et al. (2015b), using different techniques, gives upper bounds on the metric entropy (and learning sample complexity) of Poisson Multinomial Distributions (PMDs), i.e., sums of independent random vectors supported over the standard basis in $\mathbb{R}^k$. While metric entropy upper bounds on PMDs yield similar upper bounds for k-SIIRVs, the implied results for k-SIIRVs are significantly weaker than ours. The Daskalakis et al. (2015b) learning bound has sample complexity exponential in k; and running time doubly exponential in k and super-polynomial in $1/\epsilon$.

Subsequent Work. In subsequent work Diakonikolas et al. (2015c), the authors generalize the techniques of this paper to the family of PMDs. We emphasize that the results of this paper are not subsumed by the results of Diakonikolas et al. (2015c). In particular, Diakonikolas et al. (2015c) give an efficient learning algorithm for PMDs that uses $\log^{O(k)}(1/\epsilon)/\epsilon^2$ samples and runtime, and proves that the optimal cover size for PMDs depends doubly exponentially on k. Daskalakis et al. (2015a) also use the Fourier transform to learn PMDs obtaining sample size $\log^{O(k)}(1/\epsilon)/\epsilon^2$ and running time $(1/\epsilon)^{O(k^2)}$. Our geometric characterization of SIIRVs that forms the basis of our tight sample and cover lower bounds does not appear in any of the aforementioned works.

Structure of this Extended Abstract. In Section 2, we describe our learning algorithms and provide a high-level sketch of their analysis. Section 3 presents the main ideas needed to obtain our proper cover construction. Section 4 describes our geometric characterization that leads to our cover and sample lower bounds. The missing proofs and statements can be found in the full version.

2. Learning Sums of Independent Integer Random Variables

We start by presenting our sample near-optimal and computationally efficient algorithm, establishing Theorem 1. We subsequently sketch the ideas of our sample-optimal upper bound. Our algorithms use the Discrete Fourier Transform, which we now define.

Definition 6 *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 $\|\widehat{F}\|_2 = \sqrt{\frac{1}{M} \sum_{\xi=0}^{M-1} |\widehat{F}(\xi)|^2}$.*

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.

The key ingredient for the analysis of our algorithm is the following lemma, showing that the Fourier transform of $\mathbf{P}$ has appropriately small effective support.

Lemma 7 (Sparsity of DFT for k-SIIRVs) *Let $\mathbf{P} \in \mathcal{S}_{n,k}$ with $\sqrt{\mathrm{Var}_{X \sim \mathbf{P}}[X]} = s$, $1/2 > \delta > 0$, and $M \in \mathbb{Z}_+$ with $M > s$. Let $\widehat{\mathbf{P}}$ be the discrete Fourier transform of $\mathbf{P}$ modulo M. Then, we have*

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

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

Statement (i) of the lemma exhibits an explicit set $\mathcal{L}$ of cardinality $O(Mk^2 s^{-1}\sqrt{\log(1/\delta)})$ 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 $\mathcal{L}' \subseteq \mathcal{L}$).

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 $N = C^3 k/\epsilon^2 \ln^2(k/\epsilon)$ 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 $[\lfloor\widetilde{\mu}\rfloor - (M-1)/2, \lfloor\widetilde{\mu}\rfloor + (M-1)/2]$.

Sparse Fourier Learning. Our algorithmic approach is not specialized to k-SIIRVs, but is applicable more generally. It turns out that by using similar ideas, we can learn *any* probability distribution with these properties. The following simple theorem provides a generalization for integer-valued random variables. However, the approach can also be generalized to higher-dimensions and to continuous distributions.

Theorem 8 *Let $\mathbf{P}$ be an integer-valued random variable and $\epsilon > 0$. Let $S \subset \mathbb{Z}$ and $T \subset \mathbb{R}/\mathbb{Z}$ be known subsets so that: $\sum_{n \in \mathbb{Z} \setminus S} \mathbf{P}(n) \leq \epsilon/3$, and $\int_{\xi \in (\mathbb{R}/\mathbb{Z}) \setminus T} |\widehat{\mathbf{P}}(\xi)|^2 d\xi < \epsilon^2/(9|S|)$. Then, there exists an algorithm which learns $\mathbf{P}$ to total variational distance ϵ using $N = O(|S|\mu(T)/\epsilon^2)$ samples.*

Sample Optimal Algorithm. The basic idea behind our sample optimal upper bound is as follows: In our previous analysis, we made critical use of the fact that essentially all of the mass of the distribution in question lies in an explicit interval of length $O(s\sqrt{\log(1/\epsilon)})$, where s is the standard deviation. By using the Fourier learning idea, we were able to learn a distribution that approximated our target on this support. In order to improve this algorithm, we observe that although it is necessary to move $\Omega(\sqrt{\log(1/\epsilon)})$ standard deviations from the mean before the cumulative distribution function (CDF) drops below ϵ, the CDF has already begun to decay exponentially after only a single standard deviation from the mean.

Unfortunately, applying a sharp threshold to our Fourier transform can lead to effects that fall off relatively slowly with distance. Note that such a thresholding in the Fourier domain is equivalent to convolution with a Sinc function, which has tails proportional to $1/|x|$. In order to correct this issue, we will instead perform our thresholding by multiplication by a function with smooth cutoffs. This corresponds to convolving our distribution with a function of width approximately s with Gaussian tails. This step has the critical effect of causing our expected errors to be much smaller at points further from the mean, since most of our samples (within a few standard deviations of the mean) will have little effect on our output for these points. A careful analysis of the expected error at each point will yield our final bound. For the case of 2-SIIRV, this technique gives a sample-optimal and computationally efficient algorithm. For $k > 2$, our sample optimal algorithm is not computationally efficient. Several technical complications arise in the general case, mostly from the fact that we do not know a priori a good effective support for the Fourier transform.

3. Cover Size Upper Bound and Construction

The algorithm to construct the cover proceeds by an appropriate dynamic programming approach relying on our upper bound existence proof. To bound the size of the cover, we proceed as follows: We start by reducing the problem to the case that the order n of the k-SIIRV is at most $\mathrm{poly}(k/\epsilon)$. In the second and main step, we prove the desired upper bound for the polynomially sparse case. Our proof for the sparse case proceeds by analyzing the Fourier transform of k-SIIRVs.

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 $\widehat{\mathbf{P}}(\xi) = \mathbb{E}_{y \sim \mathbf{P}}[\exp(-2\pi i y \xi)] = \mathbb{E}_{y \sim \mathbf{P}}[e(y\xi)]$. 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 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. Using the above 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.

4. Cover Size and Sample Complexity Lower Bounds

We describe our lower bound construction on the cover size of 2-SIIRVs and our sample complexity lower bound. The more general constructions for k-SIIRVs are deferred to the full version.

Cover Size Lower Bound. Given $\epsilon > 0$ and $n \in \mathbb{Z}$ with $n = \Theta(\log(1/\epsilon))$, 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 2-SIIRV $\mathbf{P_a} \in \mathcal{S}_{n,2}$. We claim that the set of 2-SIIRVs $\left\{\mathbf{P_a}\right\}_{\mathbf{a}\in[s]^n}$ is an ϵ-packing, i.e., for all $\mathbf{a}, \mathbf{b} \in [s]^n$, $\mathbf{a} \neq \mathbf{b}$ implies $d_{\mathrm{TV}}\left(\mathbf{P_a}, \mathbf{P_b}\right) \geq \epsilon$.

We now sketch the proof. For a distribution $\mathbf{P}$ supported on $[n]$, define $r_{\mathbf{P}}(p)$ to be the polynomial $r_{\mathbf{P}}(p) = \mathbb{E}_{X\sim\mathbf{P}}\left[(p-1)^X \cdot p^{n-X}\right]$. If $\mathbf{P}$ is a 2-SIIRV with parameters p_i, it is easy to show that $r_{\mathbf{P}}(p) = \prod_{i=1}^n (p - p_i)$. Since $|(p-1)^i p^{n-i}| < 1$ for all i, by a simple argument, it follows that if $\mathbf{P}, \mathbf{Q} \in \mathcal{S}_{n,2}$ with $d_{\mathrm{TV}}\left(\mathbf{P}, \mathbf{Q}\right) < \epsilon$, then for any $p \in [0,1]$ it holds $|r_{\mathbf{P}}(p) - r_{\mathbf{Q}}(p)| < 2\epsilon$.

Hence, to prove our packing lower bound, it suffices to show that for all $\mathbf{a}, \mathbf{b} \in [s]^n$ with $\mathbf{a} \neq \mathbf{b}$ there exists $p = p_{\mathbf{a},\mathbf{b}} \in [0,1]$ such that $|r_{\mathbf{P_a}}(p) - r_{\mathbf{P_b}}(p)| \geq 2\epsilon$.

Let $\mathbf{a} = (a_1, \ldots, a_n)$, $\mathbf{b} = (b_1, \ldots, b_n)$. Fix $i^* \in [n]$ such that $a_{i^*} \neq b_{i^*}$. Noting that $r_{\mathbf{P_b}}(p_{i^*}^{\mathbf{b}}) = 0$, it suffices to show that $|r_{\mathbf{P_a}}(p_{i^*}^{\mathbf{b}})| \geq 2\epsilon$. If $j = i^*$ we have that, $|p_{i^*}^{\mathbf{b}} - p_j^{\mathbf{a}}| = \frac{|a_{i^*} - b_{i^*}|\sqrt{\epsilon}}{4n} \geq \frac{\sqrt{\epsilon}}{4n}$. If $j \neq i^*$, we have

$$|p_{i^*}^{\mathbf{b}} - p_j^{\mathbf{a}}| = \left|\frac{i^* - j}{n+1} + (b_{i^*} - a_j)\frac{\sqrt{\epsilon}}{4n}\right| \geq \frac{|i^* - j|}{n+1} - \frac{|b_{i^*} - a_j|\sqrt{\epsilon}}{4n} \geq \frac{|i^* - j|}{2n} \ ,$$

where the last inequality uses the fact that $|b_{i^*} - a_j| \leq s$. Therefore, we have that

$$\left|r_{\mathbf{P_a}}(p_{i^*}^{\mathbf{b}})\right| = \prod_{j=1}^{n} \left|p_{i^*}^{\mathbf{b}} - p_j^{\mathbf{a}}\right| \geq \frac{\sqrt{\epsilon}}{4n} \cdot \prod_{j\neq i^*} \frac{|i^* - j|}{2n},$$

which can be shown to be at least 2ϵ by a sequence of elementary inequalities. This completes the cover lower bound sketch.

Sample Complexity Lower Bound. Ideally, we would like to prove our sample lower bound using the set of 2-SIIRVs whose parameters are explicitly described above. We remark, however, that a cover lower bound does not generally imply *any* nontrivial sample lower bound. Moreover, this particular set of distributions is not in a form that allows a direct application of Assouad's lemma. 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 proceed with an indirect approach making essential use of a novel geometric result for the space of 2-SIIRVs. Specifically, we prove the following:

Lemma 9 *(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 $d_{\mathrm{K}}(\mathbf{P},\mathbf{Q}) \leq \delta$ is also 2-SIIRV, i.e., $\mathbf{Q} \in \mathcal{S}_{n,2}$.*

(ii) Let $\mathbf{P}_0 \in \mathcal{S}_{n,2}$ be the 2-SIIRV 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 2-SIIRV with parameters q_i such that $|q_i - p_i| \leq \frac{1}{4(n+1)}$.

Proof [Proof Sketch.] Roughly speaking, the intuition is that the space $\mathcal{S}_{n,2}$ is n-dimensional in a precise sense. We consider the space of cumulative distribution functions (CDFs) 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 2-SIIRV 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 2-SIIRV.

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 $\mathbf{P} \in \mathcal{S}_{n,2}$ from its CDF. We will apply the inverse function theorem for $\mathcal{P}_n$ at the point $\mathbf{p}$ defining the distinct parameters of the 2-SIIRV $\mathbf{P}$ 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., $(\mathrm{Jac}(\mathcal{P}_n)(\mathbf{p}))_{ij} = \frac{\partial(\mathcal{P}_n(\mathbf{p}))_i}{\partial p_j}$. In the full version, we prove the following lemma:

Lemma 10 *For a PBD $\mathbf{P} \in \mathcal{S}_{n,2}$ with parameters $\mathbf{p}$, we have*

$$M(\mathbf{p}) \cdot \mathrm{Jac}(\mathcal{P}_n)(\mathbf{p}) = -\mathrm{diag}(\prod_{j \neq i}(p_i - p_j)) \;, \tag{1}$$

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 $(\mathrm{diag}(x))_{ii} = x_i$.

Given the above lemma, we are ready to prove part (i) of Lemma 9. To this end, consider a 2-SIIRV $\mathbf{P}$ 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 (1) is non-singular. It follows from Lemma 10 that both matrices on the LHS of (1) 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 $\|\mathbf{x} - \mathcal{P}_n(\mathbf{p})\|_\infty \le \delta$.

Let $\mathbf{Q}$ be a distribution over $[n]$ satisfying $d_{\mathrm{K}}(\mathbf{P}, \mathbf{Q}) \le \delta$. Equivalently, if $\mathbf{y} = (\mathbf{Q}(< i))_{i=1}^n \in T_n$ is the CDF of $\mathbf{Q}$, then $\|\mathcal{P}_n(\mathbf{p}) - \mathbf{y}\|_\infty \le \delta$. 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 2-SIIRV with distribution $\mathbf{Q}$. It follows that $\mathbf{Q}$ is a 2-SIIRV 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 9, we use a geometric argument. Recall that the parameters of $\mathbf{P}_0$ are $\mathbf{p}_0 = (\frac{1}{n+1}, \ldots, \frac{n}{n+1})$. Let $S \subseteq T_n$ be the set of vectors $\mathbf{p}$ with $\|\mathbf{p} - \mathbf{p}_0\|_\infty \le \frac{1}{4(n+1)}$. We show in the full version that any $\mathbf{Q}$ in $\mathcal{P}_n(\partial S)$ satisfies $d_{\mathrm{TV}}(\mathbf{P}_0, \mathbf{Q}) \ge \frac{e^{-3n}}{4(n+1)}$, and therefore $d_{\mathrm{K}}(\mathbf{P}_0, \mathbf{Q}) \ge \frac{e^{-3n}}{8(n+1)^2} \ge 2^{-9n}$.

Let B be the set of distributions $\mathbf{Q}$ on $[n]$ so that $d_{\mathrm{K}}(\mathbf{P}_0, \mathbf{Q}) \le 2^{-9n}$. 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). ■

Given Lemma 9, our sample lower bound proceeds as follows: Starting from the 2-SIIRV $\mathbf{P}_0$, we perturb its pdf by a small amount to construct the "hypercube" distributions $\mathbf{P_b}$ satisfying the conditions of Assouad's lemma. Lemma 9 guarantees that, if the perturbation is small enough, all these distributions are indeed 2-SIIRVs.

Acknowledgements. Part of this work was performed while I.D. and A.S. were at the University of Edinburgh, supported in part by EPSRC grant EP/L021749/1 and a Marie Curie Career Integration Grant (CIG). The research of D.K. was supported in part by NSF Award CCF-1553288 (CAREER).

References

J. Acharya, I. Diakonikolas, J. Li, and L. Schmidt. Sample-optimal density estimation in nearly-linear time. *CoRR*, abs/1506.00671, 2015.

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

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

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

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

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

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

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.

A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

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

S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Learning mixtures of structured distributions over discrete domains. In *SODA*, pages 1380–1394, 2013.

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

S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Near-optimal density estimation in near-linear time using variable-width histograms. In *NIPS*, pages 1844–1852, 2014b.

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

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

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

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

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.

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.

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

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

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

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

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

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.

C. Daskalakis, A. De, G. Kamath, and C. Tzamos. A size-free CLT for poisson multinomials and its applications. *CoRR*, abs/1511.03641, 2015a.

C. Daskalakis, G. Kamath, and C. Tzamos. On the structure, covering, and learning of poisson multinomial distributions. In *FOCS*, 2015b. Available as arxiv report abs/1504.08363.

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

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

I. Diakonikolas, D. M. Kane, and A. Stewart. Nearly optimal learning and sparse covers for sums of independent integer random variables. *CoRR*, abs/1505.00662, 2015a.

I. Diakonikolas, D. M. Kane, and A. Stewart. Properly learning poisson binomial distributions in almost polynomial time. *CoRR*, abs/1511.04066, 2015b.

I. Diakonikolas, D. M. Kane, and A. Stewart. The fourier transform of poisson multinomial distributions and its algorithmic applications. *CoRR*, abs/1511.03592, 2015c.

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

R.M Dudley. Metric entropy of some classes of sets with differentiable boundaries. *Journal of Approximation Theory*, 10(3):227 – 236, 1974. ISSN 0021-9045. doi: 10.1016/0021-9045(74)90120-8.

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

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.

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

P. Groeneboom and G. Jongbloed. *Nonparametric Estimation under Shape Constraints: Estimators, Algorithms and Asymptotics*. Cambridge University Press, 2014.

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

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

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.

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

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

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

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

M. Kearns and U. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, Cambridge, MA, 1994.

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.

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

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

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

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

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

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

B. Roos. Binomial approximation to the Poisson binomial distribution: The Krawtchouk expansion. *Theory Probab. Appl.*, 45:328–344, 2000.

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

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

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

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.

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.

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

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