The Fourier Transform of Poisson Multinomial Distributions and Its Algorithmic Applications

Ilias Diakonikolas[*]
CS, USC, USA
diakonik@usc.edu.

Daniel M. Kane[†]
CSE & Math, UCSD, USA
dakane@cs.ucsd.edu.

Alistair Stewart[*]
CS, USC, USA
alistais@usc.edu.

ABSTRACT

An (n, k)-Poisson Multinomial Distribution (PMD) is a random variable of the form $X = \sum_{i=1}^{n} X_i$, where the X_i's are independent random vectors supported on the set of standard basis vectors in $\mathbb{R}^k$. In this paper, we obtain a refined structural understanding of PMDs by analyzing their Fourier transform. As our core structural result, we prove that the Fourier transform of PMDs is *approximately sparse*, i.e., its L_1-norm is small outside a small set. By building on this result, we obtain the following applications:

Learning Theory. We give the first computationally efficient learning algorithm for PMDs under the total variation distance. Our algorithm learns an (n, k)-PMD within variation distance ϵ using a near-optimal sample size of $\widetilde{O}_k(1/\epsilon^2)$, and runs in time $\widetilde{O}_k(1/\epsilon^2) \cdot \log n$. Previously, no algorithm with a $\mathrm{poly}(1/\epsilon)$ runtime was known, even for $k = 3$.

Game Theory. We give the first efficient polynomial-time approximation scheme (EPTAS) for computing Nash equilibria in anonymous games. For normalized anonymous games with n players and k strategies, our algorithm computes a well-supported ϵ-Nash equilibrium in time $n^{O(k^3)} \cdot (k/\epsilon)^{O(k^3 \log(k/\epsilon)/ \log\log(k/\epsilon))^{k-1}}$. The best previous algorithm for this problem [DP08, DP14] had running time $n^{(f(k)/\epsilon)^k}$, where $f(k) = \Omega(k^{k^2})$, for any $k > 2$.

Statistics. We prove a multivariate central limit theorem (CLT) that relates an arbitrary PMD to a discretized multivariate Gaussian with the same mean and covariance, in total variation distance. Our new CLT strengthens the CLT of Valiant and Valiant [VV10, VV11] by removing the dependence on n in the error bound.

Along the way we prove several new structural results of independent interest about PMDs. These include: (i) a robust moment-matching lemma, roughly stating that two PMDs that approximately agree on their low-degree parameter moments are close in variation distance; (ii) near-optimal size proper ϵ-covers for PMDs in total variation distance (constructive upper bound and nearly-matching lower bound). In addition to Fourier analysis, we employ a number of analytic tools, including the saddlepoint method from complex analysis, that may find other applications.

Categories and Subject Descriptors

G.3 [**Mathematics of Computing**]: Probability and Statistics—*Distribution functions*; G.3 [**Mathematics of Computing**]: Probability and Statistics—*Nonparametric statistics*; F.2.2 [**Analysis of Algorithms and Problem Complexity**]: Nonnumerical Algorithms and Problems

Keywords

Computational learning theory, learning distributions, Fourier analysis, central limit theorem, anonymous games

1. INTRODUCTION

1.1 Background and Motivation

The Poisson Multinomial Distribution (PMD) is the discrete probability distribution of a sum of mutually independent categorical random variables over the same sample space. A categorical random variable (k-CRV) describes the result of a random event that takes on one of $k \geq 2$ possible outcomes. Formally, an (n, k)-PMD is any random variable of the form $X = \sum_{i=1}^{n} X_i$, where the X_i's are independent random vectors supported on the set $\{e_1, e_2, \ldots, e_k\}$ of standard basis vectors in $\mathbb{R}^k$.

PMDs comprise a broad class of discrete distributions of fundamental importance in computer science, probability, and statistics. A large body of work in the probability and statistics literature has been devoted to the study of the behavior of PMDs under various structural conditions [Bar88, Loh92, BHJ92, Ben03, Roo99, Roo10]. PMDs generalize the familiar binomial and multinomial distributions, and describe many distributions commonly encountered in computer science (see, e.g., [DP07, DP08, Val08, VV11]). The $k = 2$ case corresponds to the Poisson binomial distribution (PBD), introduced by Poisson [Poi37] as a non-trivial generalization of the binomial distribution.

[*]Part of this work was performed at the University if Edinburgh. Supported by a Marie Curie Career Integration Grant and EPSRC grant EP/L021749/1.

[†]Some of this work was done while visiting the University of Edinburgh.

The full version of this paper is available at [DKS15a].

STOC'16, June 19–21, 2016, Cambridge, MA, USA
© 2016 ACM. 978-1-4503-4132-5/16/06...$15.00
http://dx.doi.org/10.1145/2897518.2897552

Recent years have witnessed a flurry of research activity on PMDs and related distributions, from several perspectives of theoretical computer science, including learning [DDS12, DDO$^+$13, DKS15b, DKT15, DKS15c], property testing [Val08, VV10, VV11], computational game theory [DP07, DP08, BCI$^+$08, DP09, DP14, GT14], and derandomization [GMRZ11, BDS12, De15, GKM15]. More specifically, the following questions have been of interest to the TCS community:

1. Is there a statistically and computationally efficient algorithm for learning PMDs from independent samples in total variation distance?

2. How fast can we compute approximate Nash equilibria in anonymous games with many players and a small number of strategies per player?

3. How well can a PMD be approximated, in total variation distance, by a discretized Gaussian with the same mean and covariance matrix?

The first question is a fundamental problem in unsupervised learning that has received considerable recent attention in TCS [DDS12, DDO$^+$13, DKS15b, DKT15, DKS15c]. The aforementioned works have studied the learnability of PMDs, and related distribution families, in particular PBDs (i.e., $(n, 2)$-PMDs) and sums of independent integer random variables. Prior to this work, no computationally efficient learning algorithm for PMDs was known, even for $k = 3$.

The second question concerns an important class of succinct games previously studied in economics [Mil96, Blo99, Blo05], whose (exact) Nash equilibrium computation was recently shown to be intractable [CDO15]. The connection between computing Nash equilibria in these games and PMDs was established in a sequence of papers by Daskalakis and Papadimitriou [DP07, DP08, DP09, DP14], who leveraged it to gave the first PTAS for the problem. Prior to our work, no efficient PTAS was known, even for anonymous games with 3 strategies per player.

The third question refers to the design of Central Limit Theorems (CLTs) for PMDs with respect to the total variation distance. Despite substantial amount of work in probability theory, the first strong CLT of this form appears to have been shown by Valiant and Valiant [VV10, VV11], motivated by applications in distribution property testing. [VV10, VV11] leveraged their CLT to obtain tight lower bounds for several fundamental problems in property testing. We remark that the error bound of the [VV10] CLT has a logarithmic dependence on the size n of the PMD (number of summands), and it was conjectured in [VV10] that this dependence is unnecessary.

1.2 Our Results

The main technical contribution of this work is the use of Fourier analytic techniques to obtain a refined understanding of the structure of PMDs. As our core structural result, we prove that the Fourier transform of PMDs is *approximately sparse*, i.e., roughly speaking, its L_1-norm is small outside a small set. By building on this property, we are able to obtain various new structural results about PMDs, and make progress on the three questions stated in the previous subsection. In this subsection, we describe our algorithmic and structural contributions in detail.

We start by stating our algorithmic results in learning and game theory, followed by an informal description of our structural results and the connections between them.

Distribution Learning.

We obtain the first statistically and computationally efficient learning algorithm for PMDs with respect to the total variation distance. In particular, we show:

THEOREM 1.1. *For all $n, k \in \mathbb{Z}_+$ and $\epsilon > 0$, there is an algorithm for learning (n, k)-PMDs with the following guarantee: Let $\mathbf{P}$ be an unknown (n, k)-PMD. The algorithm uses $m = O\left(k^{4k} \log^{2k}(k/\epsilon)/\epsilon^2\right)$ samples from $\mathbf{P}$, runs in time*[1]*$O\left(k^{6k} \log^{3k}(k/\epsilon)/\epsilon^2\right) \cdot \log n$, and with probability at least $9/10$ outputs an ϵ-sampler for $\mathbf{P}$.*

We remark that our learning algorithm outputs a succinct description of its hypothesis $\mathbf{H}$, via its Discrete Fourier Transform (DFT), $\widehat{\mathbf{H}}$, which is supported on a small size set. We show that the DFT gives both an efficient ϵ-sampler and an efficient ϵ-evaluation oracle for $\mathbf{P}$.

Our algorithm learns an unknown (n, k)-PMD within variation distance ϵ, with sample complexity $\widetilde{O}_k(1/\epsilon^2)$, and computational complexity $\widetilde{O}_k(1/\epsilon^2) \cdot \log n$. The sample complexity of our algorithm is near-optimal for any fixed k, as $\Omega(k/\epsilon^2)$ samples are necessary, even for $n = 1$. We note that recent work by Daskalakis *et al.* [DKT15] established a similar sample upper bound, alas their algorithm is not computationally efficient. More specifically, it runs in time $(1/\epsilon)^{\Omega(k^{5k} \log^{k+1}(1/\epsilon))}$, which is quasi-polynomial in $1/\epsilon$, even for $k = 2$. For the $k = 2$ case, in recent work [DKS15b] the authors of this paper gave an algorithm with sample complexity and runtime $\widetilde{O}(1/\epsilon^2)$. Prior to this work, no algorithm with a poly$(1/\epsilon)$ sample size and runtime was known, even for $k = 3$.

Computational Game Theory.

We give the first efficient polynomial-time approximation scheme (EPTAS) for computing Nash equilibria in anonymous games with many players and a fixed number of strategies. In anonymous games, all players have the same set of strategies, and the payoff of a player depends on the strategy played by the player and the number of other players who play each of the strategies. In particular, we show:

THEOREM 1.2. *There is an EPTAS for the mixed Nash equilibrium problem for normalized anonymous games with a constant number of strategies. More precisely, there exists an algorithm with the following performance guarantee: for all $\epsilon > 0$, and any normalized anonymous game $\mathcal{G}$ of n players and k strategies, the algorithm runs in time $(kn)^{O(k^3)}(1/\epsilon)^{O(k^3 \log(k/\epsilon)/\log\log(k/\epsilon))^{k-1}}$, and outputs a (well-supported) ϵ-Nash equilibrium of $\mathcal{G}$.*

The previous PTAS for this problem [DP08, DP14] has running time $n^{O(2^{k^2}(f(k)/\epsilon)^{6k})}$, where $f(k) \leq 2^{3k-1}k^{k^2+1}k!$. Our algorithm decouples the dependence on n and $1/\epsilon$, and, importantly, its running time dependence on $1/\epsilon$ is quasi-polynomial. For $k = 2$, an algorithm with runtime poly$(n)\cdot$

[1] 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.

$(1/\epsilon)^{O(\log^2(1/\epsilon))}$ was given in [DP09], which was sharpened to $\mathrm{poly}(n)(1/\epsilon)^{O(\log(1/\epsilon))}$ in our recent work [DKS15b]. Hence, we obtain, for any value of k, the same qualitative runtime dependence on $1/\epsilon$ as in the case $k = 2$.

Similarly to [DP08, DP14], our algorithm proceeds by constructing a *proper* ϵ-cover, in total variation distance, for the space of PMDs. A proper ϵ-cover for $\mathcal{M}_{n,k}$, the set of all (n, k)-PMDs, is a subset C of $\mathcal{M}_{n,k}$ such that any distribution in $\mathcal{M}_{n,k}$ is within total variation distance ϵ from some distribution in C. Our main technical contribution is the efficient construction of a proper ϵ-cover of near-minimum size (see Theorem 1.4). We note that, as follows from Theorem 1.5, the quasi-polynomial dependence on $1/\epsilon$ and the doubly exponential dependence on k in the runtime are unavoidable for *any* cover-based algorithm.

Statistics.

Using our Fourier-based machinery, we prove a strong "size-free" CLT relating the total variation distance between a PMD and an appropriately discretized Gaussian with the same mean and covariance matrix. In particular, we show:

Theorem 1.3. *Let X be an (n, k)-PMD with covariance matrix Σ. Suppose that Σ has no eigenvectors other than $\mathbf{1} = (1, 1, \ldots, 1)$ with eigenvalue less than σ. Then, there exists a discrete Gaussian G so that $d_{\mathrm{TV}}(X, G) \leq \mathrm{poly}(k)/\mathrm{poly}(\sigma)$.*

As mentioned above, Valiant and Valiant [VV10, VV11] proved a CLT of this form and used it as their main technical tool to obtain tight information-theoretic lower bounds for fundamental statistical estimation tasks. This and related CLTs have since been used in proving lower bounds for other problems (see, e.g., [CST14]). The error bound in the CLT of [VV10, VV11] is of the form $\mathrm{poly}(k)/\mathrm{poly}(\sigma) \cdot (1 + \log n)^{2/3}$, i.e., it has a dependence on the size n of the underlying PMD. Our Theorem 1.3 provides a *qualitative* improvement over the aforementioned bound, by establishing that *no* dependence on n is necessary. We note that [VV10] conjectured that such a qualitative improvement may be possible.

Our techniques for proving Theorem 1.3 are orthogonal to those of [VV10, VV11]. While Valiant and Valiant use Stein's method, we prove our strengthened CLT using the Fourier techniques that underly this paper. We view Fourier analysis as the right technical tool to analyze sums of independent random variables. An additional ingredient that we require is the saddlepoint method from complex analysis.

Structure of PMDs.

We now provide a brief intuitive overview of our new structural results, the relation between them, and their connection to our algorithmic results. The unifying theme of our work is a refined analysis of the structure of PMDs, based on their Fourier transform. The Fourier transform is one of the most natural technical tools to consider for analyzing sums of independent random variables, and indeed one of the classical proofs of the (asymptotic) central limit theorem is based on Fourier methods. The basis of our results, both algorithmic and structural, is the following statement:

Informal Lemma (Sparsity of the Fourier Transform of PMDs.) *For any (n, k)-PMD $\mathbf{P}$, and any $\epsilon > 0$ there exists a "small" set $T = T(\mathbf{P}, \epsilon)$, such that the L_1-norm of its Fourier transform, $\widehat{\mathbf{P}}$, outside the set T is at most ϵ.*

We will need two different versions of the above statement for our applications, and therefore we do not provide a formal statement at this stage. The precise meaning of the term "small" depends on the setting: For the continuous Fourier transform, we prove that the product of the volume of the effective support of the Fourier transform times the number of points in the effective support of our distribution is small. In particular, the set T is a scaled version of the dual ellipsoid to the ellipsoid defined by the covariance matrix of $\mathbf{P}$. Hence, roughly speaking, $\widehat{\mathbf{P}}$ has an effective support that is the dual of the effective support of $\mathbf{P}$. (See Lemma 4.3 in Section 4 for the precise statement.)

In the case of the Discrete Fourier Transform (DFT), we show that there exists a discrete set with small cardinality, such that L_1-norm of the DFT outside this set is small. At a high-level, to prove this statement, we need the appropriate definition of the (multidimensional) DFT, which turns out to be non-trivial, and is crucial for the computational efficiency of our learning algorithm. More specifically, we chose the period of the DFT to reflect the shape of the effective support of our PMD. (See Proposition 3.8 in Section 3 for the statement.)

With Fourier sparsity as our starting point, we obtain new structural results of independent interest for PMDs. The first is a "robust" moment-matching lemma, which we now informally state:

Informal Lemma (Parameter Moment Closeness Implies Closeness in Distribution.) *For any pair of (n, k)-PMDs $\mathbf{P}, \mathbf{Q}$, if the "low-degree" parameter moment profiles of $\mathbf{P}$ and $\mathbf{Q}$ are close, then $\mathbf{P}, \mathbf{Q}$ are close in total variation distance.*

See Definition 4.1 for the definition of parameter moments of a PMD. The formal statement of the aforementioned lemma appears as Lemma 4.6 in Section 4.1. Our robust moment-matching lemma is the basis for our proper cover algorithm and our EPTAS for Nash equilibria in anonymous games. Our constructive cover upper bound is the following:

Theorem 1.4 (Optimal Covers for PMDs). *For all $n, k \in \mathbb{Z}_+$, $k > 2$, and $\epsilon > 0$, there exists an ϵ-cover $\mathcal{M}_{n,k,\epsilon} \subseteq \mathcal{M}_{n,k}$, under the total variation distance, of the set $\mathcal{M}_{n,k}$ of (n, k)-PMDs of size $n^{O(k^2)} \cdot (1/\epsilon)^{O(k \log(k/\epsilon)/ \log\log(k/\epsilon))^{k-1}}$. In addition, there exists an algorithm to construct the set $\mathcal{M}_{n,k,\epsilon}$ that runs in time $n^{O(k^3)} \cdot (1/\epsilon)^{O(k^3 \log(k/\epsilon)/ \log\log(k/\epsilon))^{k-1}}$.*

A sparse proper cover quantifies the "size" of the space of PMDs and provides useful structural information that can be exploited in a variety of applications. In addition to Nash equilibria in anonymous games, our efficient proper cover construction provides a smaller search space for approximately solving essentially any optimization problem over PMDs. Using our cover construction, we also obtain the first EPTAS for computing threat points in anonymous games.

Perhaps surprisingly, we prove that our above upper bound is essentially tight:

Theorem 1.5 (Cover Lower Bound). *For any $k > 2$, $\epsilon > 0$ sufficiently small as a function of k, and $n = \Omega_k(\log(1/\epsilon)/ \log\log(1/\epsilon))^{k-1}$, any ϵ-cover for $\mathcal{M}_{n,k}$ has size at least $n^{\Omega(k)} \cdot (1/\epsilon)^{\Omega_k(\log(1/\epsilon)/ \log\log(1/\epsilon))^{k-1}}$.*

We remark that, in previous work [DKS15b], the authors proved a tight cover size bound of $n \cdot (1/\epsilon)^{\Theta(k \log(1/\epsilon))}$ for

(n,k)-SIIRVs, i.e., sums of n independent scalar random variables each supported on $[k]$. While a cover size lower bound for (n,k)-SIIRVs directly implies the same lower bound for (n,k)-PMDs, the opposite is not true. Indeed, Theorem 1.5 shows that covers for (n,k)-PMDs are inherently larger, requiring a doubly exponential dependence on k.

1.3 Our Approach and Techniques

At a high-level, the Fourier techniques of this paper can be viewed as a highly non-trivial generalization of the techniques in our recent paper [DKS15b] on sums of independent scalar random variables. We would like to emphasize that a number of new conceptual and technical ideas are required to overcome the various obstacles arising in the multi-dimensional setting. We start with an intuitive explanation of two key ideas that form the basis of our approach.

Sparsity of the Fourier Transform of PMDs.

Since the Fourier Transform (FT) of a PMD is the product of the FTs of its component CRVs, its magnitude is the product of terms each bounded from above by 1. Note that each term in the product is strictly less than 1 except in a small region, unless the component CRV is trivial (i.e., essentially deterministic). Roughly speaking, to establish the sparsity of the FT of PMDs, we proceed as follows: We bound from above the magnitude of the FT by the FT of a Gaussian with the same covariance matrix as our PMD. This gives us tail bounds for the FT of the PMD in terms of the FT of this Gaussian, and when combined with the concentration of the PMD itself, yields the desired property.

Approximation of the logarithm of the FT.

A key ingredient in our proofs is the approximation of the logarithm of the Fourier Transform (log FT) of PMDs by low-degree polynomials. Observe that the log FT is a sum of terms, which is convenient for the analysis. We focus on approximating the log FT by a low-degree Taylor polynomial within the effective support of the FT. (Note that outside the effective support the log FT can be infinity.) Morally speaking, the log FT is smooth, i.e., it is well-approximated by the first several terms of its Taylor series. Formally however, this statement is in general not true and requires various technical conditions, depending on the setting. One important point to note is that the sparsity of the FT controls the domain in which this approximation will need to hold, and thus help us bound the Taylor error. We will need to ensure that the sizes of the Taylor coefficients are not too large given the location of the effective support, which turns out to be a non-trivial technical hurdle. To ensure this, we need to be very careful about how we perform this Taylor expansion. In particular, the correct choice of the point that we Taylor expand around will be critical for our applications. We elaborate on these difficulties in the relevant technical sections. Finally, we remark that the degree of polynomial approximation we will require depends on the setting: In our cover upper bounds, we will require (nearly) logarithmic degree, while for our CLT degree-2 approximation suffices.

Efficient Learning Algorithm.

The high-level structure of our learning algorithm relies on the sparsity of the Fourier transform, and is similar to the algorithm in our previous work [DKS15b] for learning sums of independent integer random variables. More specifically, our learning algorithm estimates the effective support of the DFT, and then computes the empirical DFT in this effective support. This high-level description would perhaps suffice, if we were only interested in bounding the sample complexity. To obtain a computationally efficient algorithm, it is crucial to use the appropriate definition of the DFT and its inverse.

In more detail, our algorithm works as follows: It starts by drawing poly(k) samples to estimate the mean vector and covariance matrix of our PMD to good accuracy. Using these estimates, we can bound the effective support of our distribution in an appropriate ellipsoid. In particular, we show that our PMD lies (whp) in a fundamental domain of an appropriate integer lattice $L = M\mathbb{Z}^k$, where $M \in \mathbb{Z}^{k\times k}$ is an integer matrix whose columns are appropriate functions of the eigenvalues and eigenvectors of the (sample) covariance matrix. This property allows us to learn our unknown PMD X by learning the random variable $X \pmod{L}$. To do this, we learn its Discrete Fourier transform. Let L^* be the dual lattice to L (i.e., the set of points ξ so that $\xi\cdot x \in \mathbb{Z}$ for all $x \in L$). Importantly, we define the DFT, $\widehat{\mathbf{P}}$, of our PMD $X \sim \mathbf{P}$ on the dual lattice L^*, that is, $\widehat{\mathbf{P}} : L^*/\mathbb{Z}^k \to \mathbb{C}$ with $\widehat{\mathbf{P}}(\xi) = \mathbb{E}[e(\xi\cdot X)]$. A useful property of this definition is the following: the probability that $X \pmod{L}$ attains a given value x is given by the inverse DFT, defined on the lattice L, namely $\Pr[X \pmod{L} = x] = \frac{1}{|\det(M)|}\sum_{\xi\in L^*/\mathbb{Z}^k}\widehat{\mathbf{P}}(\xi)e(-\xi\cdot x)$.

The main structural property needed for the analysis of our algorithm is that there exists an explicit set T with integer coordinates and cardinality $(k\log(1/\epsilon))^{O(k)}$ that contains all but $O(\epsilon)$ of the L_1 mass of $\widehat{\mathbf{P}}$. Given this property, our algorithm draws an additional set of samples of size $(k\log(1/\epsilon))^{O(k)}/\epsilon^2$ from the PMD, and computes the empirical DFT (modulo L) on its effective support T. Using these ingredients, we are able to show that the inverse of the empirical DFT defines a pseudo-distribution that is ϵ-close to our unknown PMD in total variation distance.

Observe that the support of the inverse DFT can be large, namely $\Omega(n^{k-1})$. Our algorithm *does not* explicitly evaluate the inverse DFT at all these points, but outputs a succinct description of its hypothesis $\mathbf{H}$, via its DFT $\widehat{\mathbf{H}}$. We note that this succinct description suffices to efficiently obtain both an approximate evaluation oracle and an approximate sampler for our target PMD $\mathbf{P}$. It is clear that computing the inverse DFT at a single point can be done in time $O(|T|) = (k\log(1/\epsilon))^{O(k)}$, and gives an approximate oracle for the probability mass function of $\mathbf{P}$. By using additional algorithmic ingredients, we show how to use an oracle for the DFT, $\widehat{\mathbf{H}}$, as a black-box to obtain a computationally efficient approximate sampler for $\mathbf{P}$.

Our learning algorithm and its analysis are given in Section 3.

Constructive Proper Cover and Anonymous Games.

The correctness of our learning algorithm easily implies an algorithm to construct a *non-proper* ϵ-cover for PMDs of size $n^{O(k^2)}\cdot(1/\epsilon)^{\log(1/\epsilon))^{O(k)}}$. While this bound is close to best possible, it does not suffice for our algorithmic applications in anonymous games. For these applications, it is crucial to obtain an efficient algorithm that constructs a *proper* ϵ-cover, and in fact one that works in a certain stylized way.

To construct a proper cover, we rely on the sparsity of the continuous Fourier Transform of PMDs. Namely, we show that for any PMD $\mathbf{P}$, with effective support $S \subseteq [n]^k$, there exists an appropriately defined set $T \subseteq [0,1]^k$ such that the contribution of $\overline{T}$ to the L_1-norm of $|\widehat{\mathbf{P}}|$ is at most $\epsilon/|S|$. By using this property, we show that any two PMDs, with approximately the same variance in each direction, that have continuous Fourier transforms close to each other in the set T, are close in total variation distance. We build on this lemma to prove our robust moment-matching result. Roughly speaking, we show that two PMDs, with approximately the same variance in each direction, that are "close" to each other in their low-degree parameter moments are also close in total variation distance. We emphasize that the meaning of the term "close" here is quite subtle: we need to appropriately partition the component CRVs into groups, and approximate the parameter moments of the PMDs formed by each group within a different degree and different accuracy for each degree. (See Lemma 4.6.)

Our algorithm to construct a proper cover, and our EPTAS for Nash equilibria in anonymous games proceed by a careful dynamic programming approach, that is based on our aforementioned robust moment-matching result.

Finally, we note that combining our moment-matching lemma with a recent result in algebraic geometry gives us the following structural result of independent interest: Every PMD is ϵ-close to another PMD that is a sum of at most $O(k + \log(1/\epsilon))^k$ distinct k-CRVs.

The aforementioned algorithmic and structural results are given in Section 4.

Cover Size Lower Bound.

As mentioned above, a crucial ingredient of our cover upper bound is a robust moment-matching lemma, which translates closeness between the low-degree parameter moments of two PMDs to closeness between their Fourier Transforms, and in turn to closeness in total variation distance. To prove our cover lower bound, we follow the opposite direction. We construct an explicit set of PMDs with the property that *any* pair of distinct PMDs in our set have a non-trivial difference in (at least) one of their low-degree parameter moments. We then show that difference in one of the parameter moments implies that there exists a point where the probability generating functions have a non-trivial difference. Notably, our proof for this step is non-constructive making essential use of Cauchy's integral formula. Finally, we can easily translate a pointwise difference between the probability generating functions to a non-trivial total variation distance error. We present our cover lower bound construction in Section 4.4.

Central Limit Theorem for PMDs.

The basic idea of the proof of our CLT will be to compare the Fourier transform of our PMD X to that of the discrete Gaussian G with the same mean and covariance. By taking the inverse Fourier transform, we will be able to conclude that these distributions are pointwise close. A careful analysis using a Taylor approximation and the fact that both $\widehat{X}$ and $\widehat{G}$ have small effective support, gives us a total variation distance error independent of the size n. Alas, this approach results in an error dependence that is exponential in k. To obtain an error bound that scales polynomially with k, we require stronger bounds between X and G at points away from the mean. Intuitively, we need to take advantage of cancellation in the inverse Fourier transform integrals. To achieve this, we will use the saddlepoint method from complex analysis. The proof of our CLT is given in Section 5.

1.4 Related and Prior Work

There is extensive literature on distribution learning and computation of approximate Nash equilibria in various classes of games. We have already mentioned the most relevant references in the introduction.

In recent work, [DKT15] studied the structure and learnability of PMDs. They obtained a non-proper ϵ-cover of size $n^{k^2} \cdot 2^{O(k^{5k}\log(1/\epsilon)^{k+2})}$, and an information-theoretic sample upper bound of $O(k^{5k}\log(1/\epsilon)^{k+2}/\epsilon^2)$. The dependence on $1/\epsilon$ in their cover size is also quasi-polynomial, but is suboptimal as follows from our upper and lower bounds. Importantly, the [DKT15] construction yields a *non-proper* cover. As previously mentioned, a *proper* cover construction is necessary for our algorithmic applications. We note that the learning algorithm of [DKT15] relies on enumeration over a cover, hence runs in time quasi-polynomial in $1/\epsilon$, even for $k = 2$. The techniques of [DKT15] are orthogonal to ours. Their cover upper bound is obtained by a clever black-box application of the CLT of [VV10], combined with a non-robust moment-matching lemma that they deduce from a result of Roos [Roo02]. We remind the reader that our Fourier techniques strengthen both these technical tools: Theorem 1.3 strengthens the CLT of [VV10], and we prove a robust and quantitatively essentially optimal moment-matching lemma.

In recent work [DKS15b], the authors of the current paper used Fourier analytic techniques to study the structure and learnability of sums of independent integer random variables (SIIRVs). The techniques of the current paper can be viewed as a generalization of those in [DKS15b]. Note that our upper bounds for learning and covering PMDs do not subsume the ones in [DKS15b]. In fact, our cover upper and lower bounds in this work show that optimal covers for PMDs are inherently larger than optimal covers for SIIRVs. Moreover, the sample complexity of our SIIRV learning algorithm [DKS15b] is significantly better than that of our PMD learning algorithm in this paper.

1.5 Concurrent and Independent Work

Concurrently and independently to our work, [DDKT16] obtained qualitatively similar results using different techniques. We now provide a statement of the [DDKT16] results in tandem with a comparison to our work.

[DDKT16] give a learning algorithm for PMDs with sample complexity $(k\log(1/\epsilon))^{O(k)}/\epsilon^2)$ and runtime $(k/\epsilon)^{O(k^2)}$. The [DDKT16] algorithm uses the continuous Fourier transform, exploiting its sparsity property, plus additional structural and algorithmic ingredients. The aforementioned runtime is not polynomial in the sample size, unless k is fixed. In contrast, our learning algorithm runs in sample–polynomial time, and, for fixed k, in nearly-linear time. The [DDKT16] learning algorithm outputs an explicit hypothesis, which can be easily sampled. On the other hand, our algorithm outputs a succinct description of its hypothesis (via its DFT), and we show how to efficiently sample from it.

[DDKT16] also prove a size-free CLT, analogous to our Theorem 1.3, with error polynomial in k and $1/\sigma$. Their CLT is obtained by bootstrapping the CLT of [VV10, VV11] using techniques from [DKT15]. As previously mentioned, our proof is technically orthogonal to [VV10, VV11, DDKT16], making use of the sparsity of the Fourier transform combined with tools from complex analysis. It is worth noting that our CLT also achieves a near-optimal dependence in the error as a function of $1/\sigma$ (up to log factors).

Finally, [DDKT16] prove analogues of Theorems 1.2, 1.4, and 1.5 with qualitatively similar bounds to ours. We note that [DDKT16] improve the dependence on n in the cover size to an optimal $n^{O(k)}$, while the dependence on ϵ in their cover upper bound is the same as in [DKT15]. The cover size lower bound of [DDKT16] is qualitatively of the right form, though slightly suboptimal as a function of ϵ. The algorithms to construct proper covers and the corresponding EPTAS for anonymous games in both works have running time roughly comparable to the PMD cover size.

2. PRELIMINARIES

For $n \in \mathbb{Z}_+$, we will denote $[n] \overset{\text{def}}{=} \{1, \ldots, n\}$. For a vector $v \in \mathbb{R}^n$, and $p \geq 1$, we will denote $\|v\|_p \overset{\text{def}}{=} \left(\sum_{i=1}^n |v_i|^p\right)^{1/p}$. We will use the boldface notation $\mathbf{0}$ to denote the zero vector or matrix in the appropriate dimension.

Definition 2.1 ((n,k)-PMD). *For $k \in \mathbb{Z}_+$, let e_j, $j \in [k]$, be the standard unit vector along dimension j in $\mathbb{R}^k$. A* k-Categorical Random Variable *(k-CRV) is a vector random variable supported on the set $\{e_1, e_2, \ldots, e_k\}$. A* k-Poisson Multinomial Distribution of order n, *or (n,k)-PMD, is any vector random variable of the form $X = \sum_{i=1}^n X_i$ where the X_i's are independent k-CRVs. We will denote by $\mathcal{M}_{n,k}$ the set of all (n,k)-PMDs.*

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 $S \subseteq A$, we sometimes write $\mathbf{P}(S)$ to denote $\sum_{a \in S} \mathbf{P}(a)$. A distribution $\mathbf{P}$ supported on a finite domain A can be viewed as the probability mass function of a random variable X, i.e., $\mathbf{P}(a) = \Pr_{X \sim \mathbf{P}}[X = a]$.

The *total variation distance* between pseudo-distributions $\mathbf{P}$ and $\mathbf{Q}$ supported on a finite domain A is $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) = (1/2) \cdot \|\mathbf{P} - \mathbf{Q}\|_1$. For convenience, we will often blur the distinction between a random variable and its distribution.

Let $(\mathcal{X}, d)$ be a metric space. Given $\epsilon > 0$, a subset $\mathcal{Y} \subseteq \mathcal{X}$ is said to be a proper ϵ-*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 $d(\mathbf{x}, \mathbf{y}) \leq \epsilon$. If $\mathcal{Y}$ is not necessarily a subset of $\mathcal{X}$, then we obtain a non-proper ϵ-cover. We will be interested on efficiently constructing sparse covers for PMDs under the total variation distance metric.

Distribution Learning.

Since we are interested in the computational complexity of distribution learning, our algorithms will need to use a *succinct description* of their output hypothesis. A simple succinct representation of a discrete distribution is via an evaluation oracle for the probability mass function. Another general way to succinctly specify a distribution is via a sampler, i.e, an efficient algorithm that takes pure randomness and transforms it into a sample from the distribution. Our learning algorithm outputs both an approximate sampler and an approximate evaluation oracle for the target distribution.

Definition 2.2 (Distribution Learning). *Let $\mathcal{D}$ be a family of distributions. A randomized algorithm $A^{\mathcal{D}}$ is a* distribution learning algorithm for class $\mathcal{D}$, *if for any $\epsilon > 0$, and any $\mathbf{P} \in \mathcal{D}$, on input ϵ and sample access to $\mathbf{P}$, with probability* $9/10$, *algorithm $A^{\mathcal{D}}$ outputs an ϵ-sampler (or an ϵ-evaluation oracle) for $\mathbf{P}$.*

Anonymous Games and Nash Equilibria.

An anonymous game is a triple $(n, k, \{u_\ell^i\}_{i \in [n], \ell \in [k]})$ where $[n]$, $n \geq 2$, is the set of players, $[k]$, $k \geq 2$, a common set of strategies available to all players, and u_ℓ^i the payoff function of player i when she plays strategy ℓ. This function maps the set of partitions $\Pi_{n-1}^k = \{(x_1, \ldots, x_k) \mid x_\ell \in \mathbb{Z}_+ \text{ for all } \ell \in [k] \wedge \sum_{\ell=1}^k x_\ell = n - 1\}$ to the interval $[0,1]$. That is, the payoff of each player depends on her own strategy and only the number of other players choosing each of the k strategies.

We denote by Δ_{n-1}^k the convex hull of the set Π_{n-1}^k, i.e., $\Delta_{n-1}^k = \{(x_1, \ldots, x_k) \mid x_\ell \geq 0 \text{ for all } \ell \in [k] \wedge \sum_{\ell=1}^k x_\ell = n - 1\}$. A *mixed strategy* is an element of $\Delta^k \overset{\text{def}}{=} \Delta_1^k$. A *mixed strategy profile* is a mapping δ from $[n]$ to Δ^k. We denote by δ_i the mixed strategy of player i in the profile δ and δ_{-i} the collection of all mixed strategies but i's in δ. For $\epsilon \geq 0$, a mixed strategy profile δ is a (well-supported) ϵ-*Nash equilibrium* iff for all $i \in [n]$ and $\ell, \ell' \in [k]$ we have: $\mathbb{E}_{x \sim \delta_{-i}}[u_\ell^i(x)] > \mathbb{E}_{x \sim \delta_{-i}}[u_{\ell'}^i(x)] + \epsilon \implies \delta_i(\ell') = 0$.

Multidimensional Fourier Transform.

For $x \in \mathbb{R}$, we will denote $e(x) \overset{\text{def}}{=} \exp(-2\pi i x)$. The *(continuous) Fourier Transform (FT)* of a function $F : \mathbb{Z}^k \to \mathbb{C}$ is the function $\widehat{F} : [0,1]^k \to \mathbb{C}$ defined as $\widehat{F}(\xi) = \sum_{x \in \mathbb{Z}^k} e(\xi \cdot x) F(x)$. For the case that F is a probability mass function, we can equivalently write $\widehat{F}(\xi) = \mathbb{E}_{x \sim F}[e(\xi \cdot x)]$.

For computational purposes, we will also need the Discrete Fourier Transform (DFT) and its inverse. Let $M \in \mathbb{Z}^{k \times k}$ be an integer $k \times k$ matrix. We consider the integer lattice $L = L(M) = M\mathbb{Z}^k \overset{\text{def}}{=} \{p \in \mathbb{Z}^k \mid p = Mq, q \in \mathbb{Z}^k\}$, and its dual lattice $L^* = L^*(M) \overset{\text{def}}{=} \{\xi \in \mathbb{R}^k \mid \xi \cdot x \in \mathbb{Z} \text{ for all } x \in L\}$. Note that $L^* = (M^T)^{-1}\mathbb{Z}^k$. The quotient $\mathbb{Z}^k/L$ is the set of equivalence classes of points in $\mathbb{Z}^k$ such that two points $x, y \in \mathbb{Z}^k$ are in the same equivalence class iff $x - y \in L$. Similarly, the quotient $L^*/\mathbb{Z}^k$ is the set of equivalence classes of points in L^* such that any two points $x, y \in L^*$ are in the same equivalence class iff $x - y \in \mathbb{Z}^k$.

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

Let $X = \sum_{i=1}^n X_i$ be an (n,k)-PMD such that for $1 \leq i \leq n$ and $1 \leq j \leq k$ we denote $p_{i,j} = \Pr[X_i = e_j]$, where $\sum_{j=1}^k p_{i,j} = 1$. To avoid clutter in the notation, we will sometimes use the symbol X to denote the corresponding probability mass function. With this convention, we can write that $\widehat{X}(\xi) = \prod_{i=1}^n \widehat{X_i}(\xi) = \prod_{i=1}^n \sum_{j=1}^k e(\xi_j) p_{i,j}$.

3. EFFICIENTLY LEARNING PMDS

In this section, we describe and analyze our sample near-optimal and computationally efficient learning algorithm for PMDs. This section is organized as follows: We start by giving our main algorithm which, given samples from a PMD $\mathbf{P}$, efficiently computes a succinct description of a hypothesis pseudo-distribution $\mathbf{H}$ such that $d_{\mathrm{TV}}(\mathbf{H}, \mathbf{P}) \leq \epsilon/3$. As previously explained, the succinct description of $\mathbf{H}$ is via its DFT $\widehat{\mathbf{H}}$, which is supported on a discrete set T of cardinality $|T| = (k \log(1/\epsilon))^{O(k)}$. Note that $\widehat{\mathbf{H}}$ provides an ϵ-evaluation oracle for $\mathbf{P}$ with running time $O(|T|)$. We then show how to use $\widehat{\mathbf{H}}$, in a black-box manner, to efficiently obtain an ϵ-sampler for $\mathbf{P}$, i.e., sample from a distribution $\mathbf{Q}$ such that $d_{\mathrm{TV}}(\mathbf{Q}, \mathbf{P}) \leq \epsilon$.

THEOREM 3.1. *For all $n, k \in \mathbb{Z}_+$ and $\epsilon > 0$, the algorithm* `Efficient-Learn-PMD` *has the following performance guarantee: Let $\mathbf{P}$ be an unknown (n, k)-PMD. The algorithm uses $O\left(k^{4k} \log^{2k}(k/\epsilon)/\epsilon^2\right)$ samples from $\mathbf{P}$, runs in time $O\left(k^{6k} \log^{3k}(k/\epsilon)/\epsilon^2 + k^4 \log \log n\right)$, and outputs the DFT $\widehat{\mathbf{H}}$ of a pseudo-distribution $\mathbf{H}$ that, with probability at least $9/10$, satisfies $d_{\mathrm{TV}}(\mathbf{H}, \mathbf{P}) \leq \epsilon/3$.*

Our learning algorithm is described below:

Algorithm `Efficient-Learn-PMD`
Input: sample access to an (n, k)-PMD $X \sim \mathbf{P}$ and $\epsilon > 0$.
Output: A set $T \subseteq (\mathbb{R}/\mathbb{Z})^k$ of cardinality $|T| \leq O(k^2 \log(k/\epsilon))^k$, and the DFT $\widehat{\mathbf{H}} : T \to \mathbb{C}$ of a pseudo-distribution $\mathbf{H}$ such that $d_{\mathrm{TV}}(\mathbf{H}, \mathbf{P}) \leq \epsilon/3$.
Let $C > 0$ be a sufficiently large universal constant.

1. Draw $m_0 = O(k^4)$ samples from X, and let $\widehat{\mu}$ be the sample mean and $\widehat{\Sigma}$ the sample covariance matrix.
2. Compute an approximate spectral decomposition of $\widehat{\Sigma}$, i.e., an orthonormal eigenbasis v_i with corresponding eigenvalues λ_i, $i \in [k]$.
3. Let $M \in \mathbb{Z}^{k \times k}$ be the matrix whose i^{th} column is the closest integer point to the vector $C\left(\sqrt{k \ln(k/\epsilon)\lambda_i + k^2 \ln^2(k/\epsilon)}\right) v_i$.
4. Define $T \subseteq (\mathbb{R}/\mathbb{Z})^k$ to be the set of points $\xi = (\xi_1, \ldots, \xi_k)$ of the form $\xi = (M^T)^{-1} \cdot v + \mathbb{Z}^k$, for some $v \in \mathbb{Z}^k$ with $\|v\|_2 \leq C^2 k^2 \ln(k/\epsilon)$.
5. Draw $m = O\left(k^{4k} \log^{2k}(k/\epsilon)/\epsilon^2\right)$ samples s_i, $i \in [m]$, from X, and output the empirical DFT $\widehat{\mathbf{H}} : T \to \mathbb{C}$, i.e., $\widehat{\mathbf{H}}(\xi) = \frac{1}{m} \sum_{i=1}^{m} e(\xi \cdot s_i)$.

REMARK 3.2. The DFT $\widehat{\mathbf{H}}$ is a succinct description of the pseudo-distribution $\mathbf{H}$, the inverse DFT of $\widehat{\mathbf{H}}$, defined by: $\mathbf{H}(x) = \frac{1}{|\det(M)|} \sum_{\xi \in T} \widehat{\mathbf{H}}(\xi) e(-\xi \cdot x)$, for $x \in \mathbb{Z}^k \cap (\widehat{\mu} + M \cdot (-1/2, 1/2]^k)$, and $\mathbf{H}(x) = 0$ otherwise. Our algorithm **does not** output $\mathbf{H}$ explicitly, but implicitly via its DFT.

Let X be the unknown target (n, k)-PMD. We will denote by $\mathbf{P}$ the probability mass function of X, i.e., $X \sim \mathbf{P}$. Throughout this analysis, we will denote by μ and Σ the mean vector and covariance matrix of X.

First, note that our algorithm is easily seen to have the desired sample and time complexity. Indeed, the algorithm draws m_0 samples in Step 1 and m samples in Step 5, for a total sample complexity of $O(k^{4k} \log^{2k}(k/\epsilon)/\epsilon^2)$. The runtime of the algorithm is dominated by computing the DFT in Step 5 which takes time $O(m|T|) = O(k^{6k} \log^{3k}(k/\epsilon)/\epsilon^2)$. Computing an approximate eigendecomposition can be done in time $O(k^4 \log \log n)$(see, e.g., [PC99]). The remaining part of this section is devoted to proving correctness.

REMARK 3.3. We remark that in Step 4 of the algorithm, the notation $\xi = (M^T)^{-1} \cdot v + \mathbb{Z}^k$ refers to an equivalence class of points. In particular, any pair of distinct vectors $v, v' \in \mathbb{Z}^k$ satisfying $\|v\|_2, \|v'\|_2 \leq C^2 k^2 \ln(k/\epsilon)$, and $(M^T)^{-1} \cdot (v - v') \in \mathbb{Z}^k$ correspond to the same point ξ, and therefore are not counted twice.

Overview of Analysis.

We begin with a brief overview of the analysis. First, we show (Lemma 3.4) that at least $1 - O(\epsilon)$ of the probability mass of X lies in the ellipsoid with center μ and covariance matrix $\widetilde{\Sigma} = O(k \log(k/\epsilon))\Sigma + O(k \log(k/\epsilon))^2 I$. Moreover, with high probability over the samples drawn in Step 1 of the algorithm, the estimates $\widehat{\Sigma}$ and $\widehat{\mu}$ will be good approximations of Σ and μ (Lemma 3.5). By combining these two lemmas, we obtain (Corollary 3.6) that at least $1 - O(\epsilon)$ of the probability mass of X lies in the ellipsoid with center $\widehat{\mu}$ and covariance matrix $\Sigma' = O(k \log(k/\epsilon))\widehat{\Sigma} + O(k \log(k/\epsilon))^2 I$.

By the above, and by our choice of the matrix $M \in \mathbb{Z}^{k \times k}$, we use linear-algebraic arguments to prove (Lemma 3.7) that almost all of the probability mass of X lies in the set $\widehat{\mu} + M(-1/2, 1/2]^k$, a fundamental domain of the lattice $L = M\mathbb{Z}^k$. This lemma is crucial because it implies that, to learn our PMD X, it suffices to learn the random variable $X \pmod{L}$. We do this by learning the DFT of this distribution. This step can be implemented efficiently due to the sparsity property of the DFT (Proposition 3.8): except for points in T, the magnitude of the DFT will be very small. Establishing the desired sparsity property for the DFT is the main technical contribution of this section.

Given the above, it it fairly easy to complete the analysis of correctness. For every point in T we can learn the DFT up to absolute error $O(1/\sqrt{m})$. Since the cardinality of T is appropriately small, this implies that the total error over T is small. The sparsity property of the DFT (Lemma 3.10) completes the proof.

Detailed Analysis.

We now proceed with the detailed analysis of our algorithm. We start by showing that PMDs are concentrated with high probability. More specifically, the following lemma shows that an unknown PMD X, with mean vector μ and covariance matrix Σ, is effectively supported in an ellipsoid centered at μ, whose principal axes are determined by the eigenvectors and eigenvalues of Σ and the desired concentration probability:

LEMMA 3.4. *Let X be an (n, k)-PMD with mean vector μ and covariance matrix Σ. For any $0 < \epsilon < 1$, consider the positive-definite matrix $\widetilde{\Sigma} = k \ln(k/\epsilon)\Sigma + k^2 \ln^2(k/\epsilon) I$. Then, with probability at least $1 - \epsilon/10$ over X, we have that $(X - \mu)^T \cdot \widetilde{\Sigma}^{-1} \cdot (X - \mu) = O(1)$.*

Lemma 3.4 shows that an arbitrary (n, k)-PMD X puts at least $1 - \epsilon/10$ of its probability mass in the ellipsoid $\mathcal{E} = \{x \in \mathbb{R}^k : (x - \mu)^T \cdot (\widetilde{\Sigma})^{-1} \cdot (x - \mu) \leq c\}$, where $c > 0$ is an appropriate universal constant. Note that this ellipsoid depends on the mean vector μ and covariance matrix Σ, that are unknown to the algorithm. To obtain a bounding ellipsoid that is known to the algorithm, we will use the following lemma showing that $\widehat{\mu}$ and $\widehat{\Sigma}$ are good approximations to μ and Σ respectively.

Lemma 3.5. *With probability at least* $19/20$ *over the samples drawn in Step 1 of the algorithm, we have that* $(\widehat{\mu} - \mu)^T \cdot (\Sigma + I)^{-1} \cdot (\widehat{\mu} - \mu) = O(1)$*, and* $2(\Sigma + I) \succeq \widehat{\Sigma} + I \succeq (\Sigma + I)/2$.

We also need to deal with the error introduced in the eigendecomposition of $\widehat{\Sigma}$. Concretely, we factorize $\widehat{\Sigma}$ as $V^T \Lambda V$, for an orthogonal matrix V and diagonal matrix Λ. This factorization is necessarily inexact. By increasing the precision to which we learn $\widehat{\Sigma}$ by a constant factor, we can still have $2(\Sigma + I) \succeq V^T \Lambda V + I \succeq (\Sigma + I)/2$. We could redefine $\widehat{\Sigma}$ in terms of our computed orthonormal eigenbasis, i.e., $\widehat{\Sigma} := V^T \Lambda V$. Thus, we may henceforth assume that the decomposition $\widehat{\Sigma} = V^T \Lambda V$ is exact.

For the rest of this section, we will condition on the event that the statements of Lemma 3.5 are satisfied. By combining Lemmas 3.4 and 3.5 , we show that we can get a known ellipsoid containing the effective support of X, by replacing μ and Σ in the definition of $\mathcal{E}$ by their sample versions. More specifically, we have the following corollary:

Corollary 3.6. *Let* $\Sigma' = k \ln(k/\epsilon) \widehat{\Sigma} + k^2 \ln^2(k/\epsilon) I$*. Then, with probability at least* $1 - \epsilon/10$ *over* X*, we have that* $(X - \widehat{\mu})^T \cdot (\Sigma')^{-1} \cdot (X - \widehat{\mu}) = O(1)$.

Corollary 3.6 shows that our unknown PMD X puts at least $1 - \epsilon/10$ of its probability mass in the ellipsoid $\mathcal{E}' = \{x \in \mathbb{R}^k : (x - \widehat{\mu})^T \cdot (\Sigma')^{-1} \cdot (x - \widehat{\mu}) \leq c\}$, for an appropriate universal constant $c > 0$. Our next step is to relate the ellipsoid $\mathcal{E}'$ to the integer matrix $M \in \mathbb{Z}^{k \times k}$ used in our algorithm. Let $M' \in \mathbb{R}^{k \times k}$ be the matrix with j^{th} column $C \left(\sqrt{k \ln(k/\epsilon) \lambda_j + k^2 \ln^2(k/\epsilon)} \right) v_j$, where $C > 0$ is the constant in the algorithm statement. The matrix $M \in \mathbb{Z}^{k \times k}$ is obtained by rounding each entry of M' to the closest integer point. We note that the ellipsoid $\mathcal{E}'$ can be equivalently expressed as $\mathcal{E}' = \{x \in \mathbb{R}^k : \|(M')^{-1} \cdot (x - \widehat{\mu})\|_2 \leq 1/4\}$. Using the relation between M and M', we show that $\mathcal{E}'$ is enclosed in the ellipsoid $\{x \in \mathbb{R}^k : \|(M)^{-1} \cdot (x - \widehat{\mu})\|_2 < 1/2\}$, which is in turn enclosed in the parallelepiped with integer corner points $\{x \in \mathbb{R}^k : \|(M)^{-1} \cdot (x - \widehat{\mu})\|_\infty < 1/2\}$. This parallelepiped is a fundamental domain of the lattice $L = M\mathbb{Z}^k$. Formally, we have:

Lemma 3.7. *With probability at least* $1 - \epsilon/10$ *over* X*, we have that* $X \in \widehat{\mu} + M(-1/2, 1/2]^k$.

Recall that L denotes the lattice $M\mathbb{Z}^k$. The above lemma implies that it is sufficient to learn the random variable $X \pmod L$. To do this, we will learn its Discrete Fourier transform. Let L^* be the dual lattice to L. Recall that the DFT of the PMD $\mathbf{P}$, with $X \sim \mathbf{P}$, is the function $\widehat{\mathbf{P}} : L^*/\mathbb{Z}^k \to \mathbb{C}$ defined by $\widehat{\mathbf{P}}(\xi) = \mathbb{E}[e(\xi \cdot X)]$. Moreover, the probability that $X \pmod L$ attains a given value x is given by the inverse DFT, namely

$$\Pr\left[X \pmod L = x\right] = \frac{1}{|\det(M)|} \sum_{\xi \in L^*/\mathbb{Z}^k} \widehat{\mathbf{P}}(\xi) e(-\xi \cdot x) \; .$$

The main component of the analysis is the following proposition, establishing that the total contribution to the above sum coming from points $\xi \notin T$ is small. In particular, we prove the following:

Proposition 3.8. *We have that* $\sum_{\xi \in (L^*/\mathbb{Z}^k) \setminus T} |\widehat{\mathbf{P}}(\xi)| < \epsilon/10$.

Our next simple lemma states that the empirical DFT is a good approximation to the true DFT on the set T.

Lemma 3.9. *Letting* $m = (C^5 k^4 \ln^2(k/\epsilon))^k / \epsilon^2$*, with* $19/20$ *probability over the choice of* m *samples in Step 5, we have that* $\sum_{\xi \in T} |\widehat{\mathbf{H}}(\xi) - \widehat{\mathbf{P}}(\xi)| < \epsilon/10$.

Proof. For any given $\xi \in T$, we note that $\widehat{\mathbf{H}}(\xi)$ is the average of m samples from $e(\xi \cdot X)$, a random variable whose distribution has mean $\widehat{\mathbf{P}}(\xi)$ and variance at most $O(1)$. Therefore, we have that

$$\mathbb{E}[|\widehat{\mathbf{H}}(\xi) - \widehat{\mathbf{P}}(\xi)|] \leq O(1)/\sqrt{m}.$$

Summing over $\xi \in T$, and noting that $|T| \leq O(C^2 k^2 \log(k/\epsilon))^k$, we get that the expectation of the quantity in question is less than $\epsilon/400$. Markov's inequality completes the argument. □

Finally, we bound from above the total variation distance between $\mathbf{P}$ and $\mathbf{H}$.

Lemma 3.10. *Assuming that the conclusion of the previous lemma holds, then for any* $x \in \mathbb{Z}^k/L$ *we have that*

$$\left| \Pr[X \equiv x \pmod L] - \frac{1}{|\det(M)|} \sum_{\xi \in T} \widehat{\mathbf{H}}(\xi) e(-\xi \cdot x) \right| \leq \frac{\epsilon}{5|\det(M)|}.$$

It follows that, for each $x \in \widehat{\mu} + M(-1/2, 1/2]^k$, our hypothesis pseudo-distribution $\mathbf{H}(x)$ equals the probability that $X \equiv x \pmod L$ plus an error of at most $\frac{\epsilon}{5|\det(M)|}$. In other words, the pseudo-distribution defined by $\mathbf{H} \pmod L$ differs from $X \pmod L$ by at most $\left(\frac{\epsilon}{5|\det(M)|}\right) |\mathbb{Z}^k/L| = \epsilon/5$. On the other hand, letting $X' \sim \mathbf{P}'$ be obtained by moving a sample from X to its unique representative modulo L lying in $\widehat{\mu} + M(-1/2, 1/2]^k$, we have that $X = X'$ with probability at least $1 - \epsilon/10$. Therefore, $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{P}') \leq \epsilon/10$. Note that $X \pmod L = X' \pmod L$, and so $d_{\mathrm{TV}}(\mathbf{H} \pmod L, \mathbf{P}' \pmod L) < \epsilon/5$. Moreover, $\mathbf{H}$ and $\mathbf{P}'$ are both supported on the same fundamental domain of L, and hence $d_{\mathrm{TV}}(\mathbf{H}, \mathbf{P}') = d_{\mathrm{TV}}(\mathbf{H} \pmod L, \mathbf{P}' \pmod L) < \epsilon/5$. Therefore, assuming that the above high probability events hold, we have that $d_{\mathrm{TV}}(\mathbf{H}, \mathbf{P}) \leq d_{\mathrm{TV}}(\mathbf{H}, \mathbf{P}') + d_{\mathrm{TV}}(\mathbf{P}, \mathbf{P}') \leq 3\epsilon/10$. This completes the proof of Theorem 3.1.

An Efficient Sampler for our Hypothesis.

The learning algorithm `Efficient-Learn-PMD` outputs a succinct description of the hypothesis pseudo-distribution $\mathbf{H}$, via its DFT. This immediately provides us with an efficient evaluation oracle for $\mathbf{H}$, i.e., an ϵ-evaluation oracle for our target PMD $\mathbf{P}$. The running time of this oracle is linear in the size of T, the effective support of the DFT.

We show how to efficiently obtain an ϵ-sampler for our unknown PMD $\mathbf{P}$, using the DFT representation of $\mathbf{H}$ as a black-box. In particular, starting with the DFT of an accurate hypothesis $\mathbf{H}$, we show how to efficiently obtain an ϵ-sampler for the unknown target distribution. We remark that this efficient procedure is not restricted to PMDs, but is more general, applying to all discrete distributions with an approximately sparse DFT (over any dimension) for which an efficient oracle for the DFT is available. We prove:

Theorem 3.11. *Let $M \in \mathbb{Z}^{k\times k}$, $m \in \mathbb{R}^k$, and $S = m + M(-1/2,1/2]^k \cap \mathbb{Z}^k$. Let $\mathbf{H} : S \to \mathbb{R}$ be a pseudo-distribution succinctly represented via its DFT (modulo M), $\widehat{\mathbf{H}}$, which is supported on a set T, i.e., $\mathbf{H}(x) = (1/|\det(M)|)\cdot\sum_{\xi\in T} e(-\xi\cdot x)\widehat{\mathbf{H}}(\xi)$, for $x \in S$, with $\mathbf{0} \in T$ and $\widehat{\mathbf{H}}(\mathbf{0}) = 1$. Suppose that there exists a distribution $\mathbf{P}$ with $d_{\mathrm{TV}}(\mathbf{H},\mathbf{P}) \leq \epsilon/3$. Then, there exists an ϵ-sampler for $\mathbf{P}$, i.e., a sampler for a distribution $\mathbf{Q}$ such that $d_{\mathrm{TV}}(\mathbf{P},\mathbf{Q}) \leq \epsilon$, running in time $O(\log(|\det(M)|)\log(|\det(M)|/\epsilon)\cdot|T|\cdot\mathrm{poly}(k))$.*

Our main learning result, Theorem 1.1, follows by combining Theorem 3.1 with Theorem 3.11.

To prove Theorem 3.11, we first handle the case of one-dimensional distributions, and then appropriately reduce the high-dimensional case to the one-dimensional. Roughly speaking, we obtain the desired sampler by considering an appropriate definition of a Cumulative Distribution Function (CDF) corresponding to $\mathbf{H}$. For the one-dimensional case, the definition of the CDF is clear, and our sampler proceeds as follows: We use the DFT to obtain a closed form expression for the CDF of $\mathbf{H}$, and then we query the CDF using an appropriate binary search procedure to sample from the distribution. One subtle point is that $\mathbf{H}(x)$ is a pseudo-distribution, i.e. it is not necessarily non-negative at all points. Our analysis shows that this does not pose any problems with correctness.

For the case of higher dimensions, we provide an efficient reduction to the 1-dimensional case. In particular, we exploit the fact that the domain is discrete, to define an efficiently computable bijection from the domain to the integers, and consider the corresponding one-dimensional CDF. To achieve this, we efficiently decompose the integer matrix $M \in \mathbb{Z}^{k\times k}$ using a version of the Smith Normal Form, effectively reducing to the case that M is diagonal. For the diagonal case, we can treat the dimensions independently, using the lexicographic ordering.

This completes the sketch of the analysis.

As an application of our learning algorithm, we can also give a simple proof that the space $\mathcal{M}_{n,k}$ of all (n,k)-PMDs has an ϵ-cover of size $n^{O(k^2)}\cdot 2^{O(k\log(1/\epsilon))^{O(k)}}$. Our argument is constructive, yielding an efficient algorithm to construct a non-proper ϵ-cover of this size.

4. EFFICIENT PROPER COVERS AND NASH EQUILIBRIA IN ANONYMOUS GAMES

In this section, we describe our efficient proper cover construction for PMDs, and our EPTAS for computing Nash equilibria in anonymous games. These algorithmic results are based on new structural results for PMDs that we establish. The structure of this section is as follows: In Section 4.1, we show the desired sparsity property of the continuous Fourier transform of PMDs, and use it to prove our robust moment-matching lemma. Our dynamic-programming algorithm for efficiently constructing a proper cover relies on this lemma, and is sketched in Section 4.2. By building on the proper cover construction, in Section 4.3 we sketch our EPTAS for Nash equilibria in anonymous games. Finally, in Section 4.4 we prove our cover size lower bound.

4.1 Low-Degree Parameter Moment Closeness Implies Closeness in Variation Distance

We will require the following notion of a parameter moment for a PMD:

Definition 4.1. *Let $X = \sum_{i=1}^n X_i$ be an (n,k)-PMD such that for $1 \leq i \leq n$ and $1 \leq j \leq k$ we denote $p_{i,j} = \Pr[X_i = e_j]$. For $m = (m_1,\ldots,m_k) \in \mathbb{Z}_+^k$, we define the m^{th}*-parameter moment *of X to be $M_m(X) \stackrel{\mathrm{def}}{=} \sum_{i=1}^n \prod_{j=1}^k p_{i,j}^{m_j}$. We will refer to $|m|_1 = \sum_{j=1}^k m_j$ as the* degree *of the parameter moment $M_m(X)$.*

Our robust moment-matching result (Lemma 4.6) is proved by combining the sparsity of the continuous Fourier transform of PMDs (Lemma 4.3) with very careful Taylor approximations of the logarithm of the Fourier transform (log FT) of our PMDs. For technical reasons related to the convergence of the log FT, we will need one additional property from our PMDs. In particular, we require that each component k-CRV has the same most likely outcome. This assumption is essentially without loss of generality. There exist at most k such outcomes, and we can express an arbitrary PMD as a sum of k independent component PMDs whose k-CRV components satisfy this property. Formally, we have the following definition:

Definition 4.2. *We say that a k-CRV W is j-maximal, for some $j \in [k]$, if for all $\ell \in [k]$ we have $\Pr[W = e_j] \geq \Pr[W = e_\ell]$. We say that an (n,k)-PMD $X = \sum_{i=1}^n X_i$ is j-maximal, for some $j \in [k]$, if for all $1 \leq i \leq n$ X_i is a j-maximal k-CRV.*

Any (n,k)-PMD X can be written as $X = \sum_{i=1}^k X^i$, where X^i is an i-maximal (n_i,k)-PMD, with $\sum_i n_i = n$. For the rest of this intuitive explanation, we focus on two (n,k)-PMDs X, Y that are promised to be i-maximal, for some $i \in [k]$.

To guarantee that $\widehat{X}$, $\widehat{Y}$ have roughly the same effective support, we also assume that they have roughly the same variance in each direction. We will show that if the low-degree parameter moments of X and Y are close to each other, then X and Y are close in total variation distance. We proceed by partitioning the k-CRV components of our PMDs into groups, based on their maximum probability element e_j, with $j \neq i$. The maximum probability of a k-CRV quantifies its maximum contribution to the variance of the PMD in some direction. Roughly speaking, the smaller this contribution is, the fewer terms in the Taylor approximation are needed to achieve a given error. More specifically, we consider three different groups, partitioning the component k-CRVs into ones with small, medium, and large contribution to the variance in some direction. For the PMD (defined by the CRVs) of the first group, we only need to approximate the first 2 parameter moments. For the PMD of the second group, we approximate the low-degree parameter moments up to degree $O_k(\log(1/\epsilon)/\log\log(1/\epsilon))$. Finally, the third group is guaranteed to have very few component k-CRVS, hence we can afford to approximate the individual parameters.

To quantify the above, we need some more notation and definitions. To avoid clutter in the notation, we focus without loss of generality on the case $i = k$, i.e., our PMDs are k-maximal. For a k-maximal (n, k)-PMD, X, let $X = \sum_{i=1}^{n} X_i$, where the X_i is a k-CRV with $p_{i,j} = \Pr[X_i = e_j]$ for $1 \le i \le n$ and $1 \le j \le k$. Observe that $\sum_{j=1}^{k} p_{i,j} = 1$, for $1 \le i \le n$, hence the definition of k-maximality implies that $p_{i,k} \ge 1/k$ for all i. Note that the j^{th} component of the random vector X is a PBD with parameters $p_{i,j}$, $1 \le i \le n$. Let $s_j(X) = \sum_{i=1}^{n} p_{i,j}$ be the expected value of the j^{th} component of X. We can assume that $s_j(X) \ge \epsilon/k$, for all $1 \le j \le k-1$; otherwise, we can remove the corresponding coordinates and introduce an error of at most ϵ in variation distance.

Note that, for $j \ne k$, the variance of the j^{th} coordinate of X is in $[s_j(X)/2, s_j(X)]$. Indeed, the aforementioned variance equals $\sum_{i=1}^{n} p_{i,j}(1 - p_{i,j})$, which is clearly at most $s_j(X)$. The other direction follows by observing that, for all $j \ne k$, we have $1 \ge p_{i,k} + p_{i,j} \ge 2p_{i,j}$, or $p_{i,j} \le 1/2$, where we again used the k-maximality of X. Therefore, by Bernstein's inequality and a union bound, there is a set $S \subseteq [n]^k$ of size $|S| \le O\left(\log(k/\epsilon)^{(k-1)}\right) \cdot \prod_{j=1}^{k-1}(1 + 12 s_j(X)^{1/2})$, so that X lies in S with probability at least $1 - \epsilon$.

We start by showing that the continuous Fourier transform of a PMD is approximately sparse, namely it is effectively supported on a small set T. More precisely, we prove that there exists a set T in the Fourier domain such that the integral of the absolute value of the Fourier transform outside T multiplied by the size of the effective support $|S|$ of our PMD is small.

Lemma 4.3 (Sparsity of the FT of PMDs). *Let X be k-maximal (k, n)-PMD with effective support S. Let*

$$T \stackrel{\text{def}}{=} \{\xi \in [0,1]^k : [\xi_j - \xi_k] < Ck(1+12 s_j(X))^{-1/2} \log^{1/2}(1/\epsilon)\},$$

where $[x]$ is the distance between x and the nearest integer, and $C > 0$ is a sufficiently large universal constant. Then, we have that $\int_{\overline{T}} |\widehat{X}| \ll \epsilon/|S|$.

We now use the sparsity of the Fourier transform to show that if two k-maximal PMDs, with similar variances in each direction, have Fourier transforms that are pointwise sufficiently close to each other in this effective support, then they are close to each other in total variation distance.

Lemma 4.4. *Let X and Y be k-maximal (k, n)-PMDs, satisfying $1/2 \le (1 + s_j(X))/(1 + s_j(Y)) \le 2$ for all j, $1 \le j \le k-1$. Let*

$$T \stackrel{\text{def}}{=} \{\xi \in [0,1]^k : [\xi_j - \xi_k] < Ck(1+12 s_j(X))^{-1/2} \log^{1/2}(1/\epsilon)\},$$

where $[x]$ is the distance between x and the nearest integer, and $C > 0$ is a sufficiently large universal constant. Suppose that for all $\xi \in T$ it holds $|\widehat{X}(\xi) - \widehat{Y}(\xi)| \le \epsilon(Ck\log(k/\epsilon))^{-2k}$. Then, $d_{\mathrm{TV}}(X, Y) \le \epsilon$.

We use this lemma as technical tool for our robust moment-matching lemma. As mentioned in the beginning of the section, we will need to handle separately the component k-CRVs that have a significant contribution to the variance in some direction. This is formalized in the following definition:

Definition 4.5. *Let X be a k-maximal (n, k)-PMD with $X = \sum_{i=1}^{n} X_i$ and $0 < \delta \le 1$. For a given $\ell \in [n]$, we say that a particular component k-CRV X_ℓ, with $p_{\ell,j} = \Pr[X_\ell = e_j]$, is δ-exceptional if there exists a coordinate j, with $1 \le j \le k-1$, such that $p_{\ell,j} \ge \delta \cdot \sqrt{1 + s_j(X)}$. We will denote by $E(\delta, X) \subseteq [n]$ the set of δ-exceptional components of X.*

Recall that the variance of the j^{th} coordinate of X is in $[s_j(X)/2, s_j(X)]$. Therefore, the above definition states that the j^{th} coordinate of X_i has probability mass which is at least a δ-fraction of the standard deviation across the j^{th} coordinate of X. We remark that for any (n, k)-PMD X, at most k/δ^2 of its component k-CRVs are δ-exceptional.

We now have all the necessary ingredients for our robust moment-matching lemma. Roughly speaking, we partition the coordinate k-CRVs of our k-maximal PMDs into three groups. For appropriate values $0 < \delta_1 < \delta_2$, we have: (i) k-CRVs that are *not* δ_1-exceptional, (ii) k-CRVs that are δ_1-exceptional, but *not* δ_2-exceptional, and (iii) δ_2-exceptional k-CRVs. For group (i), we will only need to approximate the first two parameter moments in order to get a good Taylor approximation, and for group (ii) we need to approximate as many as $O_k(\log(1/\epsilon)/\log\log(1/\epsilon))$ degree parameter moments. Group (iii) has $O_k(\log^{3/2}(1/\epsilon))$ coordinate k-CRVs, hence we simply approximate the individual (relatively few) parameters each to high precision. Formally, we have:

Lemma 4.6. *Let X and Y be k-maximal (n, k)-PMDs, satisfying $1/2 \le (1 + s_j(X))/(1 + s_j(Y)) \le 2$ for all j, $1 \le j \le k-1$. Let C be a sufficiently large constant. Suppose that the component k-CRVs of X and Y can be partitioned into three groups, so that $X = X^{(1)} + X^{(2)} + X^{(3)}$ and $Y = Y^{(1)} + Y^{(2)} + Y^{(3)}$, where $X^{(t)}$ and $Y^{(t)}$, $1 \le t \le 3$, are PMDs over the same number of k-CRVs. Additionally assume the following: (i) for $t \le 2$ the random variables $X^{(t)}$ and $Y^{(t)}$ have no δ_t-exceptional components, where $\delta_1 = \delta_1(\epsilon) \stackrel{\text{def}}{=} \epsilon(Ck\log(k/\epsilon))^{-3k-3}$ and $\delta_2 = \delta_2(\epsilon) \stackrel{\text{def}}{=} k^{-1}\log^{-3/4}(1/\epsilon)$, and (ii) there is a bijection between the component k-CRVs of $X^{(3)}$ with those in $Y^{(3)}$, so that corresponding k-CRVs have total variation distance at most $\epsilon/3n_3$, where n_3 is the number of such k-CRVs.*

Finally, suppose that for $t \le 2$, and all vectors $m \in \mathbb{Z}_+^k$ with $m_k = 0$ and $|m|_1 \le K_t$ it holds

$$|M_m(X^{(t)}) - M_m(Y^{(t)})|(2k)^{|m|_1} \le \gamma \stackrel{\text{def}}{=} \epsilon(Ck\log(k/\epsilon))^{-2k-1} ,$$

where $K_1 = 2$ and $K_2 = K_2(\epsilon) = C(\log(1/\epsilon)/\log\log(1/\epsilon) + k)$. Then $d_{\mathrm{TV}}(X, Y) \le \epsilon$.

Remark 4.7. We note that the quantitive statement of Lemma 4.6 is crucial for our algorithm: (i) The set of non δ_1-exceptional components can contain up to n k-CRVs. Since we only need to approximate only the first 2 parameter moments for this set, this only involves poly(n) possibilities. (ii) The set of δ_1-exceptional but not δ_2-exceptional k-CRVs has size $O(k/\delta_1^2)$, which is independent of n. In this case, we approximate the first $O_k(\log(1/\epsilon)/\log\log(1/\epsilon))$ parameter moments, and the total number of possibilities is independent of n and bounded by an appropriate quasipolynomial function of $1/\epsilon$. (ii) The set of δ_2-exceptional components is sufficiently small, so that we can afford to do a brute-force grid over the parameters.

4.2 Efficient Construction of a Proper Cover

Our proper cover construction follows from the following theorem:

THEOREM 4.8. *Let $S_1, S_2, \ldots, S_n$ be sets of k-CRVs. Let $\mathcal{S}$ be the set of (n,k)-PMDs of the form $\sum_{\ell=1}^{n} X_\ell$, where $X_\ell \in S_\ell$. There exists an algorithm that runs in time*

$$n^{O(k^3)} \cdot (k/\epsilon)^{O(k^3 \log(k/\epsilon)/\log\log(k/\epsilon))^{k-1}} \cdot \max_{\ell \in [n]} |S_\ell| \ ,$$

and returns an ϵ-cover of $\mathcal{S}$.

Observe that if we choose each S_i to be a δ-cover for the set of all k-CRVs, with $\delta = \epsilon/n$, we obtain an ϵ-cover for $\mathcal{M}_{n,k}$, the set of all (n,k)-PMDs. It is easy to see that the set of k-CRVs has an explicit δ-cover of size $O(1/\delta)^k$.

The high-level idea is to split each such PMD into its i-maximal PMD components and approximate each to total variation distance $\epsilon' \stackrel{\text{def}}{=} \epsilon/k$. We do this by keeping track of the appropriate data, and using dynamic programming.

4.3 EPTAS for Nash Equilibria in Anonymous Games

THEOREM 4.9. *There exists an algorithm running in time $n^{O(k^3)} \cdot (k/\epsilon)^{O(k^3 \log(k/\epsilon)/\log\log(k/\epsilon))^{k-1}}$ for computing a (well-supported) ϵ-Nash Equilibrium in a normalized n-player, k-strategy anonymous game.*

We now sketch the proof of Theorem 4.9. We compute a well-supported ϵ-Nash equilibrium, using a procedure similar to [DP14]. We start by using a dynamic program very similar to that of our Theorem 4.8 in order to construct an $\epsilon/10$-cover. We iterate over this $\epsilon/10$-cover. For each element of the cover, we compute a set of possible $\epsilon/5$-best responses. Finally, we again use the dynamic program of Theorem 4.8 to check if we can construct this element of the cover out of best responses. If we can, then we have found an ϵ-Nash equilibrium. Since there exists an $\epsilon/5$-Nash equilibrium in our cover, this procedure must produce an output.

In more detail, to compute the aforementioned best responses, we use a modification of the algorithm in Theorem 4.8, which produces output at the penultimate step. The reason for this modification is the following: For the approximate Nash equilibrium computation, we need the data produced by the dynamic program, not just the cover of PMDs. Using this data, we can subtract the data corresponding to each candidate best response. This allows us to approximate the distribution of the sum of the other players strategies, which we need in order to calculate the players expected utilities.

As an additional application of our proper cover construction, we give an EPTAS for computing threat points in anonymous games [BCI+08].

We note that, by combining our moment-matching lemma with recent results in algebraic geometry, we show that any PMD is close to another PMD with few distinct CRV components.

4.4 Cover Size Lower Bound for PMDs

Our lower bound on the cover size is restated below:

THEOREM 4.10. *(Cover Size Lower Bound for (n,k)-PMDs) Let $k > 2$, $k \in \mathbb{Z}_+$, and ϵ be sufficiently small as a function of k. For $n = \Omega((1/k)\cdot\log(1/\epsilon)/\log\log(1/\epsilon))^{k-1}$ any ϵ-cover of $\mathcal{M}_{n,k}$ under the total variation distance must be of size $n^{\Omega(k)} \cdot (1/\epsilon)^{\Omega((1/k)\cdot\log(1/\epsilon)/\log\log(1/\epsilon))^{k-1}}$.*

Theorem 4.10 follows easily from the following theorem:

THEOREM 4.11. *Let $k > 2$, $k \in \mathbb{Z}_+$, and ϵ be sufficiently small as a function of k. Let*

$$n = \Omega((1/k) \cdot \log(1/\epsilon)/\log\log(1/\epsilon))^{k-1}.$$

There exists a set $\mathcal{S}$ of (n,k)-PMDs so that for $x, y \in \mathcal{S}$, $x \neq y$ implies that $d_{\mathrm{TV}}(x,y) \geq \epsilon$, and

$$|\mathcal{S}| \geq (1/\epsilon)^{\Omega((1/k)\cdot\log(1/\epsilon)/\log\log(1/\epsilon))^{k-1}}.$$

In the rest of this section, we sketch the proof of Theorem 4.11. Let $k > 2$, $k \in \mathbb{Z}_+$, and ϵ be sufficiently small as a function of k. Let $n = \Theta((1/k) \cdot \log(1/\epsilon)/\log\log(1/\epsilon))^{k-1}$.

We express an (n,k)-PMD X as a sum of independent k-CRVs X_s, where s ranges over some index set. For $1 \leq j \leq k-1$, we will denote $p_{s,j} = \Pr[X_s = e_j]$. Note that $\Pr[X_s = e_k] = 1 - \sum_{j=1}^{k-1} p_{s,j}$.

We construct our lower bound set $\mathcal{S}$ explicitly as follows. Let $0 < c < 1$ be an appropriately small universal constant. We define the parameters $a \stackrel{\text{def}}{=} \lfloor c\ln(1/\epsilon)/2k\ln\ln(1/\epsilon)) \rfloor$ and $t \stackrel{\text{def}}{=} \lfloor \epsilon^{-c} \rfloor$. We define the set $\mathcal{S}$ to have elements indexed by a function $f : [a]^{k-1} \to [t]$, where the function f corresponds to the PMD

$$X^f \stackrel{\text{def}}{=} \sum_{s \in [a]^{k-1}} X^f_s \ ,$$

and the k-CRV X^f_s, $s = (s_1, \ldots, s_{k-1}) \in [a]^{k-1}$, has the following parameters:

$$p^f_{s,j} = \frac{s_j + \delta_{j,1}\epsilon^{3c} f(s)}{\ln^k(1/\epsilon)}, \qquad (1)$$

for $1 \leq j \leq k-1$. (Note that we use $\delta_{i,j}$ to denote the standard Kronecker delta function, i.e., $\delta_{i,j} = 1$ if and only if $i = j$).

Let $\mathcal{F} = \{f \mid f : [a]^{k-1} \to [t]\}$ be the set of all functions from $[a]^{k-1}$ to $[t]$. Then, we have that

$$\mathcal{S} \stackrel{\text{def}}{=} \{X^f : f \in \mathcal{F}\}.$$

That is, each PMD in $\mathcal{S}$ is the sum of a^{k-1} many k-CRVs, and there are t possibilities for each k-CRV. Therefore,

$$|\mathcal{S}| = t^{a^{k-1}} = (1/\epsilon)^{\Omega((1/k)\cdot\log(1/\epsilon)/\log\log(1/\epsilon))^{k-1}}.$$

Observe that all PMDs in $\mathcal{S}$ are k-maximal. In particular, for any $f \in \mathcal{F}$, $s \in [a]^{k-1}$, and $1 \leq j \leq k-1$, the above definition implies that

$$p^f_{s,j} \leq \frac{1}{k} \cdot \frac{1}{\ln^k(1/\epsilon)}. \qquad (2)$$

An important observation, that will be used throughout our proof, is that for each k-CRV X^f_s, only the first out of the $k-1$ parameters $p^f_{s,j}$, $1 \leq j \leq k-1$, depends on the function f. More specifically, the effect of the function f on

$p^f_{s,1}$ is a very small perturbation of the numerator. Note that the first summand in the numerator of (1) is a positive integer, while the summand corresponding to f is at most $\epsilon^{2c} = o(1)$. We emphasize that this perturbation term is an absolutely crucial ingredient of our construction. As will become clear from the proof below, this term allows us to show that distinct PMDs in $\mathcal{S}$ have a parameter moment that is substantially different.

The proof proceeds in two main conceptual steps that we explain in detail below.

First Step.

In the first step, we show that for any two distinct PMDs in $\mathcal{S}$, there exists a parameter moment in which they differ by a non-trivial amount. For $m \in \mathbb{Z}_+^{k-1}$, we recall that the m^{th} parameter moment of a k-maximal PMD $X = \sum_{s \in S} X_s$ is defined to be $M_m(X) \stackrel{\text{def}}{=} \sum_{s \in S} \prod_{j=1}^{k-1} p_{s,j}^{m_j}$. In Lemma 4.12 below, we show that for any distinct PMDs $X^f, X^g \in \mathcal{S}$, there exists $m \in [a]^{k-1}$ such that their m^{th} parameter moments differ by at least poly(ϵ).

LEMMA 4.12. *If $f, g : [a]^{k-1} \to [t]$, with $f \neq g$, then there exists $m \in [a]^{k-1}$ so that*

$$|M_m(X^f) - M_m(X^g)| \geq \epsilon^{4c} .$$

We now give a brief intuitive overview of the proof. It is clear that, for $f \neq g$, the PMDs X^f and X^g have distinct parameters. Indeed, since $f \neq g$, there exists an $s \in [a]^{k-1}$ such that $f(s) \neq g(s)$, which implies that the k-CRVs X^f_s and X^g_s have $p^f_{s,1} \neq p^g_{s,1}$.

We start by pointing out that if two arbitrary PMDs have distinct parameters, there exists a parameter moment where they differ. This implication uses the fact that PMDs are determined by their moments, which can be established by showing that the Jacobian matrix of the moment function is non-singular. Lemma 4.12 is a a robust version of this fact, that applies to PMDs in $\mathcal{S}$, and is proved by crucially exploiting the structure of the set $\mathcal{S}$.

Our proof of Lemma 4.12 proceeds as follows: We start by approximating the parameter moments $M_m(X^f)$, $X^f \in \mathcal{S}$, from above and below, using the definition of the parameters of X^f. This approximation step allows us to express the desired difference $M_m(X^f) - M_m(X^g)$ (roughly) as the product of two terms: the first term is always positive and has magnitude poly(ϵ), while the second term is $L \cdot (f - g)$, for a certain linear transformation (matrix) L. We show that L is the tensor product of matrices L_i, where each L_i is a Vandermonde matrix on distinct integers. Hence, each L_i is invertible, which in turn implies that L is invertible. Therefore, since $f \neq g$, we deduce that $L \cdot (f - g) \neq \mathbf{0}$. Noting that the elements of this vector are integers, yields the desired lower bound.

Second Step.

In the second step of the proof, we show that two PMDs in $\mathcal{S}$ that have a parameter moment that differs by a non-trivial amount, must differ significantly in total variation distance. In particular, we prove:

LEMMA 4.13. *Let $f, g : [a]^{k-1} \to [t]$, with $f \neq g$. If $|M_m(X^f) - M_m(X^g)| \geq \epsilon^{4c}$ for some $m \in [a]^{k-1}$, then $d_{\mathrm{TV}}(X^f, X^g) \geq \epsilon$.*

We establish this lemma in two sub-steps: We first show that if the m^{th} parameter moments of two PMDs in $\mathcal{S}$ differ by a non-trivial amount, then the corresponding probability generating functions (PGF) must differ by a non-trivial amount at a point. An intriguing property of our proof of this claim is that it is non-constructive: we prove that there exists a point where the PGF's differ, but we do not explicitly find such a point. Our non-constructive argument makes essential use of Cauchy's integral formula. We are then able to directly translate a distance lower bound between the PGFs to a lower bound in total variation distance.

By putting together Lemmas 4.12 and 4.13, it follows that any two distinct elements of $\mathcal{S}$ are ϵ-separated in total variation distance. This completes the proof of Theorem 4.11.

5. A SIZE–FREE CLT FOR PMDS

In this section, we sketch the main proof ideas of our new CLT thereby establishing Theorem 1.3. For the purposes of this section, we define a discrete Gaussian in k dimensions to be a probability distribution supported on $\mathbb{Z}^k$ so that the probability of a point x is proportional to $e^{Q(x)}$, for some quadratic polynomial Q. The statement of our CLT is:

THEOREM 5.1. *Let X be an (n, k)-PMD with covariance matrix Σ. Suppose that Σ has no eigenvectors other than $\mathbf{1} = (1, 1, \ldots, 1)$ with eigenvalue less than σ. Then, there exists a discrete Gaussian G so that*

$$d_{TV}(X, G) \leq O(k^{7/2}\sqrt{\log^3(\sigma)/\sigma}).$$

We note that our phrasing of the theorem above is slightly different than the CLT statement of [VV10]. More specifically, we work with (n, k)-PMDs directly, while [VV10] work with projections of PMDs onto $k - 1$ coordinates. Also, our notion of a discrete Gaussian is not the same as the one discussed in [VV10]. At the end of the section, we show how our statement can be rephrased to be directly comparable to the [VV10] statement.

The basic idea of the proof will be to compare the Fourier transform of X to that of the discrete Gaussian with density proportional to the pdf of $\mathcal{N}(\mu, \Sigma)$ (where μ is the expectation of X). By taking the inverse Fourier transform, we will conclude that these distributions are pointwise close. A careful analysis of this combined with the claim that both X and G have small effective support will yield our result.

We provide a summary of the main steps of the proof. We start by bounding the effective support of X under our assumptions (Lemma 5.2 and Corollary 5.3). Then, we describe the effective support of its Fourier transform (Lemma 5.4). We further show that the effective support of the distribution X and the Fourier transform of the discrete Gaussian G are similar (see Lemmas 5.6 and 5.7). We then obtain an estimate of the error between the Fourier transforms of X and a Gaussian with the same mean and covariance (Lemma 5.5). The difference between the distributions of X and G at a point, as given by the inverse Fourier transform, is approximately equal to the integral of this error over the effective support of the Fourier transform of X and G. If we take bounds on the size of this integral naively, we get a weaker result than Theorem 5.1, concretely that $d_{\mathrm{TV}}(X, G) \leq O(\log(\sigma))^k \sigma^{-1/2}$ (Proposition 5.8). Finally, we are able to show the necessary bound on this integral by using the saddlepoint method.

We already have a bound on the effective support of a general PMD (Lemma 3.4). Using this lemma, we obtain simpler bounds that hold under our assumptions.

Lemma 5.2. *Let X be an (n, k)-PMD with mean μ and covariance matrix Σ, where all non-trivial eigenvalues of Σ are at least σ, then for any $\epsilon > \exp(-\sigma/k)$, with probability $1 - \epsilon$ over X we have that*

$$(X-\mu)^T(\Sigma+I)^{-1}(X-\mu) = O(k\log(k/\epsilon)).$$

Specifically, if we take $\epsilon = 1/\sigma$, we have the following:

Corollary 5.3. *Let X be as above, and let S be the set of points $x \in \mathbb{Z}^k$ where $(x-\mu)^T\mathbf{1} = 0$ and*

$$(x-\mu)^T(\Sigma+I)^{-1}(x-\mu) \leq (Ck\log(\sigma)) ,$$

for some sufficiently large constant C. Then, $X \in S$ with probability at least $1 - 1/\sigma$, and

$$|S| = \sqrt{\det(\Sigma+I)} \cdot O(\log(\sigma))^{k/2}.$$

Next, we proceed to describe the Fourier support of X. In particular, we show that $\widehat{X}$ has a relatively small effective support, T. Our Fourier sparsity lemma in this section is somewhat different than in previous section, but the ideas are similar.

Lemma 5.4. *Let $T \stackrel{\text{def}}{=} \{\xi \in \mathbb{R}^k \mid \xi \cdot \Sigma\xi \leq Ck\log(\sigma)\}$, for C some sufficiently large constant. Then, we have that:*

(i) For all $\xi \in T$, the entries of ξ are contained in an interval of length $2\sqrt{Ck\log(\sigma)/\sigma}$.

(ii) Letting $T' = T \cap \{\xi \in \mathbb{R}^k \mid \xi_1 \in [0,1]\}$, it holds $\mathrm{Vol}(T') = \det(\Sigma+I)^{-1/2} \cdot O(C\log(\sigma))^{k/2}$.

(iii) $\int_{[0,1]^k\setminus(T+\mathbb{Z}^k)} |\widehat{X}(\xi)|d\xi \leq 1/(\sigma|S|)$.

The previous lemma establishes that the contribution to the Fourier transform of X coming from points outside of T is negligibly small. We next claim that, for $\xi \in T$, it is approximated by a Gaussian.

Lemma 5.5. *For $\xi \in T$, we have that*

$$\widehat{X}(\xi) = \exp\left(2\pi i\mu\cdot\xi - 2\pi^2\xi\cdot\Sigma\xi + O(C^{3/2}k^{7/2}\sqrt{\log^3(\sigma)/\sigma})\right).$$

We now define G to be the discrete Gaussian supported on the set of points in $\mathbb{Z}^k$ whose coordinates sum to n, so that for such a point p we have:

$$\begin{aligned} G(p) &= (2\pi)^{-(k-1)/2}\det(\Sigma')^{-1/2}\exp((p-\mu)\cdot\Sigma^{-1}(p-\mu)/2) \\ &= \int_{\xi,\sum\xi_j=0} e(-p\cdot\xi)\exp(2\pi i(\xi\cdot\mu) - 2\pi^2\xi\cdot\Sigma\xi) \\ &= \int_{\xi,\xi_1\in[0,1]} e(-p\cdot\xi)\exp(2\pi i(\xi\cdot\mu) - 2\pi^2\xi\cdot\Sigma\xi) , \end{aligned}$$

where $\Sigma' = \Sigma + \mathbf{1}\mathbf{1}^T$ restricted to the space of vectors whose coordinates sum to 0.

We let $\widehat{G}$ equal

$$\widehat{G}(\xi) := \exp(2\pi i(\xi\cdot\mu) - 2\pi^2\xi\cdot\Sigma\xi).$$

Next, we claim that G and X have similar effective supports and subsequently that $\widehat{G}$ and $\widehat{X}$ do as well. Firstly, the effective support of the distribution of G is similar to that of X, namely S:

Lemma 5.6. *The sum of the absolute values of G at points not is S is at most $1/\sigma$.*

Secondly, the effective support of the Fourier Transform of G is similar to that of X, namely T:

Lemma 5.7. *The integral of $|\widehat{G}(\xi)|$ over ξ with $\xi_1 \in [0,1]$ and ξ not in T is at most $1/(|S|\sigma)$.*

Using the above ingredients, we can prove a weaker version of Theorem 5.1:

Proposition 5.8. *We have $d_{\mathrm{TV}}(X, G) \leq O(\log(\sigma))^k\sigma^{-1/2}$.*

The proof of Theorem 5.1 is substantially the same as the above. The one obstacle that we face is that above we are only able to prove L^∞ bounds on the difference between X and G, and these bounds are too weak for our purposes. What we would like to do is to prove stronger bounds on the difference between X and G at points p far from μ. In order to do this, we will need to take advantage of cancellation in the inverse Fourier transform integrals. To achieve this, we will use the saddle point method from complex analysis.

Comparison to the [VV10] CLT.

We note that the above statement of Theorem 5.1 is not immediately comparable to the CLT of [VV10]. More specifically, we work with PMDs directly, while [VV10] works with projections of PMDs onto $k-1$ coordinates. Also, our notion of a discrete Gaussian is not the same as the one discussed in [VV10]. However, it is not difficult to relate the two results. We establish the following corollary of Theorem 5.1:

Corollary 5.9. *Let X be an (n, k)-PMD, and X' be obtained by projecting X onto its first $k-1$ coordinates. Let Σ' be the covariance matrix of X'. Suppose that Σ' has no eigenvectors with eigenvalue less than σ'. Let G' be the distribution obtained by sampling from $\mathcal{N}(\mathbb{E}[X'], \Sigma')$ and rounding to the nearest point in $\mathbb{Z}^k$. Then, we have that*

$$d_{\mathrm{TV}}(X', G') \leq O(k^{7/2}\sqrt{\log^3(\sigma')/\sigma'}).$$

6. CONCLUSIONS AND OPEN PROBLEMS

In this work, we used Fourier analytic techniques to obtain a number of structural results on PMDs. As a consequence, we gave a number of applications in distribution learning, statistics, and game theory. We believe that our techniques are of independent interest and may find other applications.

Several interesting open questions remain:

- What is the precise complexity of learning PMDs? Our bound is nearly-optimal when the dimension k is fixed. The case of high dimension is not well-understood, and seems to require different ideas.

- Is there an efficient *proper* learning algorithm, i.e., an algorithm that outputs a PMD as its hypothesis? This question is still open even for $k = 2$; see [DKS15c] for some recent progress.

- What is the optimal error dependence in Theorem 1.3 as a function of the dimension k?

- Is there a fully-polynomial time approximation scheme (FPTAS) for computing ϵ-Nash equilibria in anonymous games? We remark that cover-based algorithms cannot lead to such a result, because of the quasi-polynomial cover size lower bounds in this paper, as well as in our previous work [DKS15b] for the case $k = 2$. Progress in this direction requires a deeper understanding of the relevant fixed points.

7. REFERENCES

[Bar88] A. D. Barbour. Stein's method and poisson process convergence. *Journal of Applied Probability*, 25:pp. 175–184, 1988.

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

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

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

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

[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.

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

[CST14] X. Chen, R. A. Servedio, and L.Y. Tan. New algorithms and lower bounds for monotonicity testing. In *FOCS*, pages 286–295, 2014.

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

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

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

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

[DKS15a] I. Diakonikolas, D. M. Kane, and A. Stewart. The fourier transform of poisson multinomial distributions and its algorithmic applications. *CoRR*, abs/1511.03592, 2015.

[DKS15b] I. Diakonikolas, D. M. Kane, and A. Stewart. Optimal learning via the fourier transform for sums of independent integer random variables. *CoRR*, abs/1505.00662, 2015.

[DKS15c] I. Diakonikolas, D. M. Kane, and A. Stewart. Properly learning poisson binomial distributions in almost polynomial time. *CoRR*, 2015.

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

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

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

[DP09] C. Daskalakis and C. Papadimitriou. On Oblivious PTAS's for Nash Equilibrium. In *STOC*, pages 75–84, 2009.

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

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

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

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

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

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

[PC99] V. Y. Pan and Z. Q. Chen. The complexity of the matrix eigenproblem. In *Proceedings of the Thirty-first Annual ACM Symposium on Theory of Computing*, pages 507–516, 1999.

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

[Roo99] B. Roos. On the rate of multivariate poisson convergence. *Journal of Multivariate Analysis*, 69(1):120 – 134, 1999.

[Roo02] B. Roos. Multinomial and krawtchouk approximations to the generalized multinomial distribution. *Theory of Probability & Its Applications*, 46(1):103–117, 2002.

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

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

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

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