The Fourier Transform of Poisson Multinomial Distributions and its Algorithmic Applications

The Fourier Transform of Poisson Multinomial Distributions and its Algorithmic Applications

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

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

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

November 13, 2015

Abstract

We study Poisson Multinomial Distributions – a fundamental family of discrete distributions that generalize the binomial and multinomial distributions, and are commonly encountered in computer science. Formally, 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 $\{e_1, e_2, \ldots, e_k\}$ 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., roughly speaking, its L_1-norm is small outside a small set. By building on this result, we obtain the following applications:

Learning Theory. We design the first computationally efficient learning algorithm for PMDs with respect to the total variation distance. Our algorithm learns an arbitrary (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 completely 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.

*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.
‡Supported by EPSRC grant EP/L021749/1.

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.

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$^+$13, DKS15a, DKT15, DKS15b], property testing [Val08, VV10, VV11], computational game theory [DP07, DP08, BCI$^+$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, DKS15a, DKT15, DKS15b]. 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 the case of $k = 3$.

The second question concerns an important class of succinct games previously studied in the economics literature [Mil96, Blo99, Blo05], whose (exact) Nash equilibrium computation was recently shown to be intractable [CDO15]. The formal 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 this 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 computational game theory, followed by an informal description of our structural results and the connections between them.

Distribution Learning. As our main learning result, 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 (Efficiently Learning PMDs)**.** *For all $n, k \in \mathbb{Z}_+$ and $\epsilon > 0$, there is an algorithm for learning (n,k)-PMDs with the following performance 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, however 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 [DKS15a] 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 $\mathrm{poly}(1/\epsilon)$ sample size and runtime was known, even for $k = 3$.

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

Computational Game Theory. As our second algorithmic contribution, we give the first efficient polynomial-time approximation scheme (EPTAS) for computing Nash equilibria in anonymous games with many players and a small 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 (EPTAS for Nash in Anonymous Games)**.** *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 $f(k) \le 2^{3k-1}k^{k^2+1}k!$. 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

[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.

$\mathrm{poly}(n)(1/\epsilon)^{O(\log^2(1/\epsilon))}$ was given in [DP09], which was sharpened to $\mathrm{poly}(n)(1/\epsilon)^{O(\log(1/\epsilon))}$ in the recent work of the authors [DKS15a]. 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. Our cover upper and lower bounds and our Nash approximation algorithm are given in Section 4.

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.

We remark that 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. We hope that our new CLT will be of broader use as an analytic tool to the TCS community. Our CLT is proved in Section 5.

Structure of PMDs. We now provide a brief intuitive overview of our new structural results for PMDs, the relation between them, and their connection to our algorithmic results mentioned above. 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 essentially 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 4.2 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 2.2 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 $|\mathcal{M}_{n,k,\epsilon}|\leq n^{O(k^2)}\cdot(1/\epsilon)^{O(k\log(k/\epsilon)/\log\log(k/\epsilon))^{k-1}}$. 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. As another corollary of our cover construction, we obtain the first EPTAS for computing threat points in anonymous games.

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

Theorem 1.5 (Cover Lower Bound for PMDs)**.** *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 [DKS15a], 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, Theorems 1.4 and 1.5 show 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 [DKS15a] 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. (See, for example, Lemma 3.10.) 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 Fourier Transform. 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 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.

We are now ready to give an overview of the ideas in the proofs of each of our results.

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 [DKS15a] 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. In order 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 $\mathrm{poly}(k)$ 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 emphasize that this succinct description suffices to efficiently obtain both an approximate evaluation oracle and an approximate sampler for our target PMD $\mathbf{P}$. Indeed, 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 (see Section 3.3) 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 upper bound is close to being best possible (see Section 4.5), 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 in Section 4.1.)

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 4.5.

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 full 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.

Daskalakis *et al.* [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 upper bound on the learning sample complexity 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 [DKS15a], the authors used Fourier analytic techniques to study the structure and learnability of sums of independent integer random variables (SIIRVs). The techniques of this paper can be viewed as a (highly nontrivial) generalization of those in [DKS15a]. We also note that the upper bounds we obtain in this paper for learning and covering PMDs do not subsume the ones in [DKS15a]. 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 [DKS15a] is significantly better than that of our PMD learning algorithm in this paper.

1.5 Organization In Section 3, we describe and analyze our learning algorithm for PMDs. Section 4 contains our proper cover upper bound construction, our cover size lower bound, and the related approximation algorithm for Nash equilibria in anonymous games. Finally, Section 5 contains the proof of our CLT.

2 Preliminaries

In this section, we record the necessary definitions and terminology that will be used throughout the technical sections of this paper.

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

Poisson Multinomial Distributions. We start by defining our basic object of study:

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.

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

Definition 2.2 (m^{th}-parameter moment of a PMD)**.** 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 $M_m(X) \stackrel{\text{def}}{=} \sum_{i=1}^n \prod_{j=1}^k p_{i,j}^{m_j}$. We will refer to $|m|_1 = \sum_{j=1}^k m_j$ as the *degree* of the parameter moment $M_m(X)$.

(Pseudo-)Distributions and Total Variation Distance. 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 two pseudo-distributions $\mathbf{P}$ and $\mathbf{Q}$ supported on a finite domain A is $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) \stackrel{\text{def}}{=} \max_{S \subseteq A} |\mathbf{P}(S) - \mathbf{Q}(S)| = (1/2) \cdot \|\mathbf{P} - \mathbf{Q}\|_1 = (1/2) \cdot \sum_{a \in A} |\mathbf{P}(a) - \mathbf{Q}(a)|$. If X and Y are two random variables ranging over a finite set, their total variation distance $d_{\mathrm{TV}}(X, Y)$ is defined as the total variation distance between their distributions. For convenience, we will often blur the distinction between a random variable and its distribution.

Covers. 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.) There may exist many ϵ-covers of $\mathcal{X}$, but one is typically interested in one with the minimum cardinality. The *ϵ-covering number* of $(\mathcal{X}, d)$ is the minimum cardinality of any ϵ-cover of $\mathcal{X}$. Intuitively, the covering number of a metric space captures the "size" of the space. In this work, we will be interested on efficiently constructing sparse covers for PMDs under the total variation distance metric.

Distribution Learning. We now define the notion of distribution learning we use in this paper. Note that an explicit description of a discrete distribution via its probability mass function scales linearly with the support size. 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:

Definition 2.3 (Evaluation Oracle)**.** Let $\mathbf{P}$ be a distribution over $[n]^k$. An *evaluation oracle* for $\mathbf{P}$ is a polynomial size circuit C with $m = O(k \log n)$ input bits $z \in [n]^k$ such that for each $z \in [n]^k$, the output of the circuit $C(z)$ equals the binary representation of the probability $\mathbf{P}(z)$. For $\epsilon > 0$, an *ϵ-evaluation oracle* for $\mathbf{P}$ is an evaluation oracle for some pseudo-distribution $\mathbf{P'}$ which has $d_{\mathrm{TV}}(\mathbf{P'}, \mathbf{P}) \leq \epsilon$.

One of the most general ways to succinctly specify a distribution is to give the code of an efficient algorithm that takes "pure" randomness and transforms it into a sample from the distribution. This is the standard notion of a sampler:

Definition 2.4 (Sampler)**.** Let $\mathbf{P}$ be a distribution over $[n]^k$. An *ϵ-sampler* for $\mathbf{P}$ is a circuit C with $m = O(k \log n + \log(1/\epsilon))$ input bits z and $m' = O(k \log n)$ output bits y which is such that when $z \sim U_m$ then $y \sim \mathbf{P'}$, for some distribution $\mathbf{P'}$ which has $d_{\mathrm{TV}}(\mathbf{P'}, \mathbf{P}) \leq \epsilon$.

We can now give a formal definition of distribution learning:

Definition 2.5 (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 $\mathbf{P}$.

Remark 2.6. *We emphasize that our learning algorithm in Section 3 outputs* both an ϵ-sampler and an ϵ-evaluation oracle *for the target distribution.*

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, it is assumed that 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 $\Delta^k \stackrel{\text{def}}{=} \Delta_1^k$. 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: $\mathbb{E}_{x \sim \delta_{-i}}[u_\ell^i(x)] > \mathbb{E}_{x \sim \delta_{-i}}[u_{\ell'}^i(x)] + \epsilon \implies \delta_i(\ell') = 0$. Note that given a mixed strategy profile δ, we can compute a player's expected payoff in time $\mathrm{poly}(n^k)$ by straightforward dynamic programming.

Note that the mixed strategy δ_i of player $i \in [n]$ defines the k-CRV X_i, i.e., a random vector supported in the set $\{e_1, \ldots, e_k\}$, such that $\Pr[X_i = e_\ell] = \delta_i(\ell)$, for all ℓ. Hence, if $(X_1, \ldots, X_n)$ is a mixed strategy profile, the expected payoff of player $i \in [n]$ for using pure strategy $\ell \in [k]$ is $\mathbb{E}\left[u_\ell^i\left(\sum_{j \neq i, j \in [n]} X_j\right)\right]$.

Multidimensional Fourier Transform. Throughout this paper, we will make essential use of the (continuous and the discrete) multidimensional Fourier transform. For $x \in \mathbb{R}$, we will denote $e(x) \stackrel{\text{def}}{=} \exp(-2\pi i x)$. 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 $\widehat{F}(\xi) = \mathbb{E}_{x \sim F}\left[e(\xi \cdot x)\right]$.

For computational purposes, we will also need the Discrete Fourier Transform (DFT) and its inverse, whose definition is somewhat more subtle. Let $M \in \mathbb{Z}^{k \times k}$ be an integer $k \times k$ matrix. We consider the integer lattice $L = L(M) = M\mathbb{Z}^k \stackrel{\text{def}}{=} \{p \in \mathbb{Z}^k \mid p = Mq, q \in \mathbb{Z}^k\}$, and its dual lattice $L^* = L^*(M) \stackrel{\text{def}}{=} \{\xi \in \mathbb{R}^k \mid \xi \cdot x \in \mathbb{Z} \text{ for all } x \in L\}$. Note that $L^* = (M^T)^{-1}\mathbb{Z}^k$, and that L^* is not necessarily integral. The quotient $\mathbb{Z}^k/L$ is the set of equivalence classes of points in $\mathbb{Z}^k$ such that two points $x, y \in \mathbb{Z}^k$ are in the same equivalence class 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 $\widehat{F}_M(\xi) = \sum_{x \in \mathbb{Z}^k} e(\xi \cdot x) F(x)$. (We will remove the subscript M when it is clear from the context.) Similarly, for the case that F is a probability mass function, we can equivalently write $\widehat{F}(\xi) = \mathbb{E}_{x \sim F}\left[e(\xi \cdot x)\right]$. The *inverse DFT* of a function $\widehat{G} : L^*/\mathbb{Z}^k \to \mathbb{C}$ 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 $1 \le i \le n$ and $1 \le j \le k$ 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 $\widehat{X}(\xi) = \prod_{i=1}^n \widehat{X_i}(\xi) = \prod_{i=1}^n \sum_{j=1}^k e(\xi_j) p_{i,j}$.

Basics from Linear Algebra. We remind the reader a few basic definitions from linear algebra that we will repeatedly use throughout this paper. The Frobenius norm of $A \in \mathbb{R}^{m \times n}$ is $\|A\|_F \stackrel{\text{def}}{=} \sqrt{\sum_{i,j} A_{i,j}^2}$. The spectral norm (or induced L_2-norm) of $A \in \mathbb{R}^{m \times n}$ is defined as $\|A\|_2 \stackrel{\text{def}}{=} \max_{x : \|x\|_2 = 1} \|Ax\|_2 = \sqrt{\lambda_{\max}(A^T A)}$. We note that for any $A \in \mathbb{R}^{m \times n}$, it holds $\|A\|_2 \le \|A\|_F$. A symmetric matrix $A \in \mathbb{R}^{n \times n}$ is called positive semidefinite (PSD), denoted by $A \succeq \mathbf{0}$, if $x^T A x \ge 0$ for all $x \in \mathbb{R}^n$, or equivalently all the eigenvalues of A are nonnegative. Similarly, a symmetric matrix $A \in \mathbb{R}^{n \times n}$ is called positive definite (PD), denoted by $A \succ \mathbf{0}$, if $x^T A x > 0$ for all $x \in \mathbb{R}^n$, $x \neq \mathbf{0}$, or equivalently all the eigenvalues of A are strictly positive. For two symmetric matrices $A, B \in \mathbb{R}^{n \times n}$ we write $A \succeq B$ to denote that the difference $A - B$ is PSD, i.e., $A - B \succeq \mathbf{0}$. Similarly, we write $A \succ B$ to denote that the difference $A - B$ is PD, i.e., $A - B \succ \mathbf{0}$.

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: In Section 3.1, we give 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 $d_{\mathrm{TV}}(\mathbf{H}, \mathbf{P}) \le \epsilon/3$. 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 $O(|T|)$. In Section 3.2, we 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 $d_{\mathrm{TV}}(\mathbf{Q}, \mathbf{P}) \le \epsilon$. Finally, in Section 3.3 we show how a nearly–tight cover upper bound can easily be deduced from our learning algorithm.

3.1 Main Learning Algorithm In this subsection, we give an algorithm `Efficient-Learn-PMD` establishing the following theorem:

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 in the following pseudo-code:

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 $\|v\|_2 \leq C^2k^2\ln(k/\epsilon)$.
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., $\widehat{\mathbf{H}}(\xi) = \frac{1}{m}\sum_{i=1}^m e(\xi\cdot s_i)$.

   / 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 the algorithm `Efficient-Learn-PMD` 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 the correctness of our algorithm.

Remark 3.2. We remark that in Step 4 of our 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 $\|v\|_2, \|v'\|_2 \leq C^2k^2\ln(k/\epsilon)$, 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 3.3) 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 3.4). By combining these two lemmas, we obtain (Corollary 3.5) 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 3.6) 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 Discrete Fourier transform 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 3.14) 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 3.3. *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 $(X - \mu)^T \cdot \widetilde{\Sigma}^{-1} \cdot (X - \mu) = O(1)$.*

Proof. Let $X = \sum_{i=1}^n X_i$, where the X_i's are independent k-CRVs. We can write $\mu = \mathbb{E}[X] = \sum_{i=1}^n \mu_i$, where $\mu_i = \mathbb{E}[X_i]$. Note that for any unit vector $u \in \mathbb{R}^k$, $\|u\|_2 = 1$, we have that the scalar random variable $u \cdot (X - \mu)$ is a sum of independent, mean 0, bounded random variables. Indeed, we have that $u \cdot (X - \mu) = \sum_{i=1}^n u \cdot (X_i - \mu_i)$, and that $\mathbb{E}[u \cdot (X_i - \mu_i)] = u \cdot (\mathbb{E}[X_i] - \mu_i) = 0$. Moreover, we can write

$$|u \cdot (X_i - \mu_i)| \leq \|u\|_2 \cdot \|X_i - \mu_i\|_2 \leq \|X_i\|_2 + \|\mu_i\|_2 \leq 1 + \sqrt{k}\|\mu_i\|_1 \leq 2\sqrt{k}\,,$$

where we used the Cauchy-Schwartz inequality twice, the triangle inequality, and the fact that a k-CRV X_i with mean μ_i by definition satisfy $\|X_i\|_2 = 1$, and $\|\mu_i\|_1 = 1$.

Let ν be the variance of $u \cdot (X - \mu)$. By Bernstein's inequality, we obtain that for $t = \sqrt{2\nu \ln(10k/\epsilon)} + 2\sqrt{k} \ln(10k/\epsilon)$ it holds

$$\Pr\left[|u \cdot (X - \mu)| > t\right] \leq \exp\left(-\frac{t^2/2}{\nu + 2\sqrt{k}t/3}\right) \leq \frac{\epsilon}{10k}\,. \tag{1}$$

Let Σ, the covariance matrix of X, have an orthonormal eigenbasis $u_j \in \mathbb{R}^k$, $\|u_j\|_2 = 1$, with corresponding eigenvalues ν_j, $j \in [k]$. Since Σ is positive-semidefinite, we have that $\nu_j \geq 0$, for all $j \in [k]$. We consider the random variable $u_j \cdot (X - \mu)$. In addition to being a sum of independent,

mean 0, bounded random variables, we claim that $\mathrm{Var}[u_j \cdot (X-\mu)] = \nu_j$. First, it is clear that $\mathrm{Var}[u_j \cdot (X-\mu)] = \mathbb{E}\left[(u_j \cdot (X-\mu))^2\right]$. Moreover, note that for any vector $v \in \mathbb{R}^k$, we have that $\mathbb{E}[(v \cdot (X-\mu))^2] = v^T \cdot \Sigma \cdot v$. For $v = u_j$, we thus get $\mathbb{E}\left[(u_j \cdot (X-\mu))^2\right] = u_j^T \cdot \Sigma \cdot u_j = \nu_j$.

Applying (1) for $u_j \cdot (X-\mu)$, with $t_j = \sqrt{2\nu_j \ln(10k/\epsilon)} + 2\sqrt{k}\ln(10k/\epsilon)$, yields that for all $j \in [k]$ we have

$$\Pr\left[|u_j \cdot (X-\mu)| > t_j\right] \leq \exp\left(-\frac{t_j^2/2}{\nu_j + 2\sqrt{k}t_j/3}\right) \leq \frac{\epsilon}{10k}.$$

By a union bound, it follows that

$$\Pr\left[\forall j \in [k] : |u_j \cdot (X-\mu)| \leq t_j\right] \geq 1 - \epsilon/10\ .$$

We condition on this event.

Since u_j and ν_j are the eigenvectors and eigenvalues of Σ, we have that $\Sigma = U^T \cdot \mathrm{diag}\,(\nu_j) \cdot U$, where U has j^{th} column u_j. We can thus write

$$\widetilde{\Sigma} = U^T \cdot \mathrm{diag}\left(k\nu_j \ln(k/\epsilon) + k^2 \ln^2(k/\epsilon)\right) \cdot U.$$

Therefore, we have:

$$\begin{aligned}
(X-\mu)^T \cdot \widetilde{\Sigma}^{-1} \cdot (X-\mu) &= \left\| \mathrm{diag}\left(\sqrt{k\nu_j \ln(k/\epsilon) + k^2 \ln^2(k/\epsilon)}\right)^{-1} \cdot U^T \cdot (X-\mu) \right\|_2^2 \\
&\leq k \left\| \mathrm{diag}\left(\sqrt{k\nu_j \ln(k/\epsilon) + k^2 \ln^2(k/\epsilon)}\right)^{-1} \cdot U^T \cdot (X-\mu) \right\|_\infty^2 \\
&= \left(\max_{j \in [k]} \frac{|u_j \cdot (X-\mu)|}{\sqrt{\nu_j \ln(k/\epsilon) + k \ln^2(k/\epsilon)}} \right)^2 \\
&= O(1)\ ,
\end{aligned}$$

where the last inequality follows from our conditioning, the definition of t_j, and the elementary inequality $\sqrt{a+b} \geq (\sqrt{a} + \sqrt{b})/\sqrt{2}$, $a, b \in \mathbb{R}_+$. This completes the proof of Lemma 3.3. □

Lemma 3.3 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. This is the ellipsoid centered at μ, whose principal semiaxes are parallel to the u_j's, i.e., the eigenvectors of Σ, or equivalently of $(\widetilde{\Sigma})^{-1}$. The length of the principal semiaxis corresponding to u_j is determined by the corresponding eigenvalue of $(\widetilde{\Sigma})^{-1}$, and is equal to $c^{-1/2} \cdot \sqrt{k\nu_j \ln(k/\epsilon) + k^2 \ln^2(k/\epsilon)}$.

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 (see Appendix A for the simple proof) showing that $\widehat{\mu}$ and $\widehat{\Sigma}$ are good approximations to μ and Σ respectively.

Lemma 3.4. *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 3.4 are satisfied. By combining Lemmas 3.3 and 3.4 , 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 3.5. *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 $(X - \widehat{\mu})^T \cdot (\Sigma')^{-1} \cdot (X - \widehat{\mu}) = O(1)$.*

Proof. By Lemma 3.4, it holds that $2(\Sigma + I) \succeq \widehat{\Sigma} + I \succeq (\Sigma + I)/2$. Hence, we have that

$$2\left(k \ln(k/\epsilon)\Sigma + k^2 \ln^2(k/\epsilon) I\right) \succeq k \ln(k/\epsilon)\widehat{\Sigma} + k^2 \ln^2(k/\epsilon) I \succeq \frac{1}{2}\left(k \ln(k/\epsilon)\Sigma + k^2 \ln^2(k/\epsilon) I\right) .$$

In terms of Σ' and $\widetilde{\Sigma}$, this is $2\widetilde{\Sigma} \succeq \Sigma' \succeq \frac{1}{2}\widetilde{\Sigma}$. By standard results, taking inverses reverses the positive semi-definite ordering (see e.g., Corollary 7.7.4 (a) in [HJ85]). Hence,

$$2\widetilde{\Sigma}^{-1} \succeq \Sigma'^{-1} \succeq \frac{1}{2}\widetilde{\Sigma}^{-1} .$$

Combining the above with Lemma 3.3, with probability at least $1 - \epsilon/10$ over X we have that

$$(X - \mu)^T \cdot \Sigma'^{-1} \cdot (X - \mu) \leq 2(X - \mu)^T \cdot \widetilde{\Sigma}^{-1} \cdot (X - \mu) = O(1) . \tag{2}$$

Since $\Sigma' \succeq \frac{1}{2}\widetilde{\Sigma} \succeq \Sigma + I$, and therefore $(\Sigma')^{-1} \preceq (\Sigma + I)^{-1}$, Lemma 3.4 gives that

$$(\widehat{\mu} - \mu)^T \cdot \Sigma'^{-1} \cdot (\widehat{\mu} - \mu) \leq (\widehat{\mu} - \mu)^T \cdot (\Sigma + I)^{-1} \cdot (\widehat{\mu} - \mu) = O(1) . \tag{3}$$

We then obtain:

$$\begin{aligned}(X - \widehat{\mu})^T \cdot (\Sigma')^{-1} \cdot (X - \widehat{\mu}) &= ((X - \mu) - (\widehat{\mu} - \mu))^T \cdot (\Sigma')^{-1} \cdot ((X - \mu) - (\widehat{\mu} - \mu)) \\ &= (X - \mu)^T \cdot \Sigma'^{-1} \cdot (X - \mu) + (\widehat{\mu} - \mu)^T \cdot \Sigma'^{-1} \cdot (\widehat{\mu} - \mu) \\ &\quad - (X - \mu)^T \cdot \Sigma'^{-1} \cdot (\widehat{\mu} - \mu) - (\widehat{\mu} - \mu)^T \cdot \Sigma'^{-1} \cdot (X - \mu) .\end{aligned}$$

Equations (2) and (3) yield that the first two terms are $O(1)$. Since Σ'^{-1} is positive-definite, $x^T \Sigma'^{-1} y$ as a function of vectors $x, y \in \mathbb{R}^k$ is an inner product. So, we may apply the Cauchy-Schwartz inequality to bound each of the last two terms from above by

$$\sqrt{((X - \mu)^T \cdot \Sigma'^{-1} \cdot (X - \mu)) \cdot ((\widehat{\mu} - \mu)^T \cdot \Sigma'^{-1} \cdot (\widehat{\mu} - \mu))} = O(1) ,$$

where the last equality follows from (2) and (3). This completes the proof of Corollary 3.5. □

Corollary 3.5 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 $c > 0$. This is the ellipsoid centered at $\widehat{\mu}$, whose principal semiaxes are parallel to the v_j's, the eigenvectors of $\widehat{\Sigma}$, and the length of the principal semiaxis parallel to v_j is equal to $c^{-1/2} \cdot \sqrt{k\lambda_j \ln(k/\epsilon) + k^2 \ln^2(k/\epsilon)}$.

Our next step is to 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 $\mathcal{E}' = \{x \in \mathbb{R}^k : \|(M')^{-1}\cdot(x-\widehat{\mu})\|_2 \le 1/4\}$. 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 3.6. *With probability at least $1-\epsilon/10$ over X, we have that $X \in \widehat{\mu} + M(-1/2, 1/2]^k$.*

Proof. Let $M' \in \mathbb{R}^{k\times k}$ be the matrix with columns $C\left(\sqrt{k\ln(k/\epsilon)\lambda_i + k^2\ln^2(k/\epsilon)}\right)v_i$, where $C > 0$ is the constant in the algorithm statement. Note that,

$$M'(M')^T = C^2 k\ln(k/\epsilon)\widehat{\Sigma} + C^2k^2\ln^2(k/\epsilon)I = C^2\Sigma' \,. \tag{4}$$

For a large enough constant C, Corollary 3.5 implies that with probability at least $1-\epsilon/10$,

$$\|(M')^{-1}\cdot(X-\widehat{\mu})\|_2^2 = C^{-2}(X-\widehat{\mu})^T\cdot(\Sigma')^{-1}\cdot(X-\widehat{\mu}) \le 1/16 \,. \tag{5}$$

Note that the above is an equivalent description of the ellipsoid $\mathcal{E}'$. Our lemma will follow from the following claim:

Claim 3.7. *For any $x \in \mathbb{R}^k$, it holds*

$$\|M^{-1}x\|_2 < 2\|(M')^{-1}x\|_2 \text{ and } \|M^Tx\|_2 < 2\|(M')^Tx\|_2 \,. \tag{6}$$

Proof. By construction, M and M' differ by at most 1 in each entry, and therefore $M-M'$ has Frobenius norm (and, thus, induced L_2-norm) at most k. For any $x \in \mathbb{R}^k$, we thus have that

$$\|(M')^Tx\|_2 = \sqrt{x^T\cdot M'\cdot(M')^T\cdot x} \ge \sqrt{x^T\cdot(C^2k^2I)\cdot x} > 2k\|x\|_2 \,,$$

and therefore

$$\|M^Tx\|_2 \ge \|(M')^Tx\|_2 - \|(M-M')^T\|_2\|x\|_2 > \frac{1}{2}\|(M')^Tx\|_2.$$

Similarly, we get $\|M^Tx\|_2 \le \|(M')^Tx\|_2 + \|(M-M')^T\|_2\|x\|_2 < 2\|(M')^Tx\|_2$. In terms of the PSD ordering, we have:

$$\frac{1}{4}M'(M')^T \prec MM^T \prec 4M'(M')^T \,. \tag{7}$$

Since $M'(M')^T \succeq I$, both $M'(M')^T$ and MM^T are positive-definite, and so M and M' are invertible. Taking inverses in Equation (7) reverses the ordering, that is:

$$\frac{1}{4}(M'^{-1})^TM'^{-1} \prec (M^{-1})^TM^{-1} \prec 4(M'^{-1})^TM'^{-1} \,.$$

The claim now follows. □

Hence, Claim 3.7 implies that with probability at least $1-\epsilon/10$, we have:

$$\|M^{-1}(X-\widehat{\mu})\|_\infty \le \|M^{-1}(X-\widehat{\mu})\|_2 < 2\|(M')^{-1}(X-\widehat{\mu})\|_2 < 1/2 \,,$$

where the last inequality follows from (5). In other words, with probability at least $1-\epsilon/10$, X lies in $\widehat{\mu} + M(-1/2, 1/2]^k$, which was to be proved. □

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

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

The main 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* $\sum_{\xi \in (L^*/\mathbb{Z}^k) \setminus T} |\widehat{\mathbf{P}}(\xi)| < \epsilon/10$.

To prove this proposition, we will need a number of intermediate claims and lemmas. We start with the following claim, showing that for every point $\xi \in \mathbb{R}^k$, there exists an integer shift whose coordinates lie in an interval of length strictly less than 1:

Claim 3.9. *For each* $\xi \in \mathbb{R}^k$, *there exists* $a \in \mathbb{Z}_+$ *with* $0 \le a \le k$, *and* $b \in \mathbb{Z}^k$ *such that* $\xi - b \in \left[\frac{a}{k+1}, \frac{a+k}{k+1}\right]^k$.

Proof. Consider the k fractional parts of the coordinates of ξ, i.e., $\xi_i - \lfloor \xi_i \rfloor$, for $1 \le i \le k$. Now consider the $k+1$ intervals $I_{a'} = \left(\frac{a'-1}{k+1}, \frac{a'}{k+1}\right)$, for $1 \le a' \le k+1$. By the pigeonhole principle, there is an a' such that $\xi_i - \lfloor \xi_i \rfloor \notin I_{a'}$, for all i, $1 \le i \le k$. We define $a = a'$ when $a' < k+1$, and $a = 0$ when $a' = k+1$.

For any i, with $1 \le i \le k$, since $\xi_i - \lfloor \xi_i \rfloor \notin I_{a'}$, we have that $\xi_i - \lfloor \xi_i \rfloor \in \left[0, \frac{a-1}{k+1}\right] \cup \left[\frac{a}{k+1}, 1\right]$ (taking the first interval to be empty if $a = 0$). Hence, by setting one of $b_i = \lfloor \xi_i \rfloor$, or $b_i = \lfloor \xi_i \rfloor - 1$, we get $\xi_i - b_i \in \left[\frac{a}{k+1}, \frac{a+k}{k+1}\right]$. This completes the proof. □

The following lemma gives a "Gaussian decay" upper bound on the magnitude of the DFT, at points ξ whose coordinates lie in an interval of length less than 1. Roughly speaking, the proof of Proposition 3.8 proceeds by applying this lemma for all $\xi \notin T$.

Lemma 3.10. *Fix* $\delta \in (0,1)$. *Suppose that* $\xi \in \mathbb{R}^k$ *has coordinates lying in an interval* I *of length* $1 - \delta$. *Then,* $|\widehat{\mathbf{P}}(\xi)| = \exp(-\Omega(\delta^2 \xi^T \cdot \Sigma \cdot \xi))$.

Proof. Since $\mathbf{P}$ is a PMD, we have $X = \sum_{i=1}^n X_i$, where $X_i \sim \mathbf{P}_i$ for independent k-CRV's $\mathbf{P}_i$, we have that $|\widehat{\mathbf{P}}(\xi)| = \prod_{i=1}^n |\widehat{\mathbf{P}_i}(\xi)|$. Note also that $\xi^T \cdot \Sigma \cdot \xi = \mathrm{Var}[\xi \cdot X] = \sum_{i=1}^n \mathrm{Var}[\xi \cdot X_i]$. It therefore suffices to show that for each $i \in [n]$ it holds

$$|\widehat{\mathbf{P}_i}(\xi)| = \exp(-\Omega(\delta^2 \mathrm{Var}[\xi \cdot X_i])) .$$

Let $X_i' \sim \mathbf{P}_i$ be an independent copy of X_i, and $Y_i = X_i - X_i'$. We note that $\mathrm{Var}[\xi \cdot X_i] = (1/2)\mathbb{E}[(\xi \cdot Y_i)^2]$, and that $|\widehat{\mathbf{P}_i}(\xi)|^2 = \mathbb{E}[e(\xi \cdot Y_i)]$. Since Y_i is a symmetric random variable, we have that

$$|\widehat{\mathbf{P}_i}(\xi)|^2 = \mathbb{E}[e(\xi \cdot Y_i)] = \mathbb{E}[\cos(2\pi \xi \cdot Y_i)] .$$

We will need the following technical claim:

Claim 3.11. *Fix* $0 < \delta < 1$. *For all* $x \in \mathbb{R}$ *with* $|x| \le 1 - \delta$, *it holds* $1 - \cos(2\pi x) \ge \delta^2 x^2$.

Proof. When $0 \le |x| \le 1/4$, $\sin(2\pi x)$ is concave, since its second derivative is $-4\pi^2 \sin(2\pi x) \le 0$. So, we have $\sin(2\pi x) \ge (1-4x)\sin(0) + 4x\sin(\pi/2) = 4x$. Integrating the latter inequality, we obtain that $(1-\cos(2\pi x))/2\pi \ge 2x^2$, i.e., $(1-\cos(2\pi x)) \ge 4\pi x^2$. Thus, for $0 \le |x| \le 1/4$, we have $1-\cos(2\pi x)) \ge 4\pi x^2 \ge \delta^2 x^2$.

When $1/4 \le |x| \le 3/4$, we have $(1-\cos(2\pi x)) \ge 1 \ge \delta^2 x^2$. Finally, when $3/4 \le |x| \le 1-\delta$, we have $0 \le 1-|x| \le \delta \le 1/4$, and therefore $1-\cos(2\pi x) = 1-\cos(2\pi(1-|x|)) \ge 1-\cos(2\pi\delta) \ge 4\pi\delta^2 \ge \delta^2 x^2$. This establishes the proof of the claim. □

Since $\xi \cdot Y_i$ is by assumption supported on the interval $[-1+\delta, 1-\delta]$, we have that $|\widehat{\mathbf{P}_i}(\xi)|^2$ is

$$\mathbb{E}[\cos(2\pi\xi \cdot Y_i)] = \mathbb{E}[1 - \Omega(\delta^2(\xi \cdot Y_i)^2)] \le \exp(-\Omega(\delta^2 \mathbb{E}[(\xi \cdot Y_i)^2])) = \exp(-\Omega(\delta^2 \mathrm{Var}[\xi \cdot X_i])) \,.$$

This completes the proof of Lemma 3.10. □

We are now ready to prove the following crucial lemma, which shows that the DFT of $\mathbf{P}$ is effectively supported on the set T.

Lemma 3.12. *For integers $0 \le a \le k$, we have that*

$$\sum_{\xi \in L^* \cap \left[\frac{a}{k+1}, \frac{a+k}{k+1}\right]^k \setminus (T + \mathbb{Z}^k)} |\widehat{\mathbf{P}}(\xi)| < \frac{\epsilon}{10(k+1)} \,.$$

We start by providing a brief overview of the proof. First note that Claim 3.9 and Lemma 3.10 together imply that for $\xi \in [a/(k+1), (a+k)/(k+1)]^k$, if $\xi^T \cdot \Sigma \cdot \xi \ge k^2 \log(1/\epsilon')$, then $|\widehat{\mathbf{P}}(\xi)| \le \epsilon'$, for any $\epsilon' > 0$. Observe that the set $\{\xi \in \mathbb{R}^k : \xi^T \cdot \Sigma \cdot \xi \le k^2 \log(1/\epsilon')\}$ is not an ellipsoid, because Σ is singular. However, by using the fact that M and M' are close to each other, – more specifically, using ingredients from the proof of Lemma 3.6 – we are able to bound its intersection with $[a/(k+1), (a+k)/(k+1)]^k$ by an appropriate ellipsoid of the form $\{\xi^T \cdot (M \cdot M^T) \cdot \xi \le r\}$.

The gain here is that $M^T\xi \in \mathbb{Z}^k$. This allows us to provide an upper bound on the cardinality of the set of lattice points in L^* in one of these ellipsoids. Note that, in terms of $v = M^T\xi$, these are the integer points that lie in a sphere of some radius $r > r_0 = C^2k^2\ln(k/\epsilon)$. Now, if we consider the set $2^t r_0 \le \xi^T \cdot (MM^T) \cdot \xi < 2^{t+1} r_0$, we have both an upper bound on the magnitude of the DFT and on the number of integer points in the set. By summing over all values of $t \ge 0$, we get an upper bound on the error coming from points outside of T.

Proof of Lemma 3.12. Since $\xi \notin T + \mathbb{Z}^k$, we have that $\|M^T\xi\|_2 > C^2k^2\ln(k/\epsilon)$. Since the coordinates of ξ lie in $\left[\frac{a}{k+1}, \frac{a+k}{k+1}\right]$, an interval of length $1 - 1/(k+1)$, we may apply Lemma 3.10 to obtain:

$$\begin{aligned}
|\widehat{\mathbf{P}}(\xi)| &= \exp\left(-\Omega\left(k^{-2}\xi^T \cdot \Sigma \cdot \xi\right)\right) \\
&= \exp\left(-\Omega\left(k^{-2}\xi^T \cdot (\widehat{\Sigma} - I) \cdot \xi\right)\right) && \text{(by Lemma 3.4)} \\
&= \exp\left(-\Omega\left(k^{-2}\left(\frac{\xi^T M'(M')^T \xi}{C^2 k \log(k/\epsilon)} - k\log(k/\epsilon)\|\xi\|_2^2\right)\right)\right) && \text{(by Equation (4))} \\
&= \exp\left(-\Omega\left(k^{-2}\left(\frac{\|(M')^T\xi\|_2^2}{C^2 k \log(k/\epsilon)} - k^2\log(k/\epsilon)\|\xi\|_\infty^2\right)\right)\right) \\
&= \exp\left(-\Omega\left(k^{-2}\left(\frac{\|M^T\xi\|_2^2}{C^2 k \log(k/\epsilon)} - k^2\log(k/\epsilon)\|\xi\|_\infty^2\right)\right)\right) && \text{(by Equation (6))} \\
&= \exp\left(-\Omega\left(C^{-2}k^{-3}\log^{-1}(k/\epsilon)\|M^T\xi\|_2^2 - \log(k/\epsilon)\right)\right) && \text{(since } \|\xi\|_\infty \le 2\text{)} \\
&= \exp\left(-\Omega\left(C^{-2}k^{-3}\log^{-1}(k/\epsilon)\|M^T\xi\|_2^2\right)\right) \,. && \text{(since } \|M^T\xi\|_2 > C^2k^2\ln(k/\epsilon)\text{)}
\end{aligned}$$

Next note that for $\xi \in L^*$ we have that $M^T\xi \in \mathbb{Z}^k$. Thus, letting $v = M^T\xi$, it suffices to show that

$$\sum_{v\in\mathbb{Z}^k, \|v\|_2 \geq C^2k^2\ln(k/\epsilon)} \exp\left(-\Omega\left(C^{-2}k^{-3}\log^{-1}(k/\epsilon)\|v\|_2^2\right)\right) < \frac{\epsilon}{10(k+1)} \,. \tag{8}$$

Although the integer points in the above sum are not in the sphere $\|v\|_2 \leq C^2k^2\ln(k/\epsilon)$, they lie in some sphere $\|v\|_2 \leq 2^{t+1}C^2k^2\ln(k/\epsilon)$, for some integer $t > 0$. The number of integral points in one of these spheres is less than that of the appropriate enclosing cube. Namely, we have that

$$\begin{aligned} & \#\left\{v \in \mathbb{Z}^k, \|v\|_2 \leq 2^{t+1}C^2k^2\ln(k/\epsilon)\right\} \\ \leq\ & \#\left\{v \in \mathbb{Z}^k, \|v\|_\infty \leq 2^{t+1}C^2k^2\ln(k/\epsilon)\right\} \\ =\ & \left(1 + 2\lfloor 2^{t+1}C^2k^2\ln(k/\epsilon)\rfloor\right)^k \,. \end{aligned} \tag{9}$$

Inequality (8) is obtained by bounding the LHS from above as follows:

$$\begin{aligned} & \sum_{t=0}^{\infty} \exp\left(-\Omega(C^{-2}k^{-3}\log^{-1}(k/\epsilon)(2^tC^2k^2\ln(k/\epsilon))^2)\right) \cdot \#\left\{v \in \mathbb{Z}^k, \|v\|_2 \leq 2^{t+1}C^2k^2\ln(k/\epsilon)\right\} \\ =& \sum_{t=0}^{\infty} \exp\left(-\Omega(Ck\log(k/\epsilon)4^t)\right) \cdot \#\left\{v \in \mathbb{Z}^k, \|v\|_2 \leq 2^{t+1}C^2k^2\ln(k/\epsilon)\right\} \\ \leq& \sum_{t=0}^{\infty} \exp\left(-\Omega(Ck\log(k/\epsilon)4^t)\right) \cdot \left(2^{t+2}C^2k^2\log(k/\epsilon)\right)^k \qquad \text{(by Equation (9))} \\ =& \sum_{t=0}^{\infty} \exp\left(-\Omega(Ck\log(k/\epsilon)4^t)\right) \exp\left(O(k(t + \log k + \log\log(k/\epsilon)))\right) \\ \leq& \exp\left(-\ln((k+1)/10\epsilon)\right) \cdot \sum_{t=0}^{\infty} \exp(-k\sqrt{C}(4^t - t))) \\ \leq& \frac{\epsilon^2}{10(k+1)} \cdot \sum_{t=0}^{\infty} \exp(-\sqrt{C}(4^t - t)) \\ <& \frac{\epsilon}{10(k+1)} \,. \end{aligned}$$

This completes the proof of Lemma 3.12. □

We are now prepared to prove Proposition 3.8.

Proof of Proposition 3.8. Let T_a be the set of points $\xi \in \mathbb{R}^k$ which have a lift with all coordinates in the interval $\left[\frac{a}{k+1}, \frac{a+k}{k+1}\right]^k$, for some integer $0 \leq a \leq k$. By Claim 3.9, we have that $\bigcup_a T_a = (L^*/\mathbb{Z}^k)$. By Lemma 3.12, for all $0 \leq a \leq k$,

$$\sum_{\xi\in T_a\setminus T} |\widehat{\mathbf{P}}(\xi)| < \frac{\epsilon}{10(k+1)} \,,$$

and so we have:

$$\sum_{\xi\in(L^*/\mathbb{Z}^k)\setminus T} |\widehat{\mathbf{P}}(\xi)| \leq \sum_{a=0}^{k} \sum_{\xi\in T_a\setminus T} |\widehat{\mathbf{P}}(\xi)| < \epsilon/10 \,.$$

□

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

Lemma 3.13. *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 3.14. *Assuming that the conclusion of the previous lemma holds, then for any $x \in \mathbb{Z}^k/L$ we have that*

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

Proof. We note that

$$\begin{aligned}
&\left| \Pr[X \equiv x \pmod{L}] - \frac{1}{|\det(M)|} \sum_{\xi \in T} \widehat{\mathbf{H}}(\xi) e(-\xi \cdot x) \right| \\
=&\left| \frac{1}{|\det(M)|} \sum_{\xi \in L^*/\mathbb{Z}^k} \widehat{\mathbf{P}}(\xi) e(-\xi \cdot x) - \frac{1}{\det(M)} \sum_{\xi \in T} \widehat{\mathbf{H}}(\xi) e(-\xi \cdot x) \right| \\
\leq&\frac{1}{|\det(M)|} \sum_{\xi \in L^*/\mathbb{Z}^k, \xi \notin T} |\widehat{\mathbf{P}}(\xi)| - \frac{1}{|\det(M)|} \sum_{\xi \in T} |\widehat{\mathbf{P}}(\xi) - \widehat{\mathbf{H}}(\xi)| \\
\leq&\frac{\epsilon}{5|\det(M)|},
\end{aligned}$$

where the last line follows from Proposition 3.8 and Lemma 3.13. □

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 analysis and the proof of Theorem 3.1.

3.2 An Efficient Sampler for our Hypothesis The learning algorithm of Section 3.1 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.

Note that we can explicitly output the hypothesis $\mathbf{H}$ by computing the inverse DFT at all the points of the support of $\mathbf{H}$. However, in contrast to the effective support of $\widehat{\mathbf{H}}$, the support of $\mathbf{H}$ can be large, and this explicit description would not lead to a computationally efficient algorithm. In this subsection, 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 $\mathbf{H}$, represented via its DFT, we show how to efficiently obtain an ϵ-sampler for the unknown target distribution. We remark that the efficient procedure of this subsection 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.

In particular, we prove the following theorem:

Theorem 3.15. *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 $d_{\mathrm{TV}}(\mathbf{H}, \mathbf{P}) \le \epsilon/3$. Then, there exists an ϵ-sampler for $\mathbf{P}$, i.e., a sampler for a distribution $\mathbf{Q}$ such that $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) \le \epsilon$, running in time $O(\log(|\det(M)|)\log(|\det(M)|/\epsilon) \cdot |T| \cdot \mathrm{poly}(k))$.*

We remark that the ϵ-sampler in the above theorem statement can be described as a randomized algorithm that takes as input M, T, $\widehat{\mathbf{H}}(\xi)$, for $\xi \in T$, and the Smith normal form decomposition of M(see Lemma 3.21).

We start by observing that our main learning result, Theorem 1.1, follows by combining Theorem 3.1 with Theorem 3.15. Indeed, note that the matrix M in the definition of our PMD algorithm in Section 3.1 satisfies $|\det(M)| \le O(\sqrt{\det(\widetilde{\Sigma})}) \le (nk\log(1/\epsilon))^{O(k)}$. Also recall that $|T| = O(k^{2k}\log^k(k/\epsilon))$. Since M has largest entry n, by Lemma 3.21, we can compute its Smith normal form decomposition in time $\mathrm{poly}(k)\log n$. Hence, for the case of PMDs, we obtain the following corollary, establishing Theorem 1.1:

Corollary 3.16. *For all $n, k \in \mathbb{Z}_+$ and $\epsilon > 0$, there is an algorithm with 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 $O\left(k^{6k}\log^{3k}(k/\epsilon)/\epsilon^2 \cdot \log n\right)$, and with probability at least $9/10$ outputs an ϵ-sampler for $\mathbf{P}$. This ϵ-sampler runs (i.e., produces a sample) in time $\mathrm{poly}(k)O(k^{2k}\log^{k+1}(k/\epsilon)) \cdot \log^2 n$.*

This section is devoted to the proof of Theorem 3.15. We first handle the case of one-dimensional distributions, and then appropriately reduce the high-dimensional case to the one-dimensional.

Remark 3.17. We remark that the assumption that $\widehat{\mathbf{H}}(\mathbf{0}) = 1$ in our theorem statement, ensures that $\sum_{x\in S}\mathbf{H}(x) = 1$, and so, for any distribution $\mathbf{P}$ over S the total variational distance $d_{\mathrm{TV}}(\mathbf{H}, \mathbf{P}) \stackrel{\mathrm{def}}{=} \frac{1}{2}\sum_{x\in S}|\mathbf{H}(x) - \mathbf{P}(x)|$ is well behaved in the sense that $d_{\mathrm{TV}}(\mathbf{H}, \mathbf{P}) = \sum_{x:\mathbf{P}(x)>\mathbf{H}(x)}(\mathbf{P}(x) - \mathbf{H}(x)) = \sum_{x:\mathbf{P}(x)<\mathbf{H}(x)}(\mathbf{H}(x) - \mathbf{P}(x))$. This fact will be useful in the correctness of our sampler.

We start by providing some high-level intuition. Roughly speaking, we obtain the desired sampler by considering an appropriate definition of a Cumulative Distribution Function (CDF) corresponding to $\mathbf{H}$. For the 1-dimensional case (i.e., the case $k = 1$ in our theorem statement), 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, by using the aforementioned remark.

For the case of two or more dimensions ($k \geq 2$), we essentially provide a computationally efficient reduction to the 1-dimensional case. In particular, we exploit the fact that the underlying domain is discrete, to define an efficiently computable bijection from the domain to the integers, and consider the corresponding 1-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 intuitively treat the dimensions independently, using the lexicographic ordering.

Our first lemma handles the 1-dimensional case, assuming the existence of an efficient oracle for the CDF:

Lemma 3.18. *Given a pseudo-distribution $\mathbf{H}$ supported on $[a, b] \cap \mathbb{Z}$, $a, b \in \mathbb{Z}$, with CDF $c_{\mathbf{H}}(x) = \sum_{i:a \leq i \leq x} \mathbf{H}(i)$ (which satisfies $c_{\mathbf{H}}(b) = 1$), and oracle access to a function $c(x)$ so that $|c(x) - c_{\mathbf{H}}(x)| < \epsilon/(10(b - a + 1))$ for all x, we have the following: If there is a distribution $\mathbf{P}$ with $d_{\mathrm{TV}}(\mathbf{H}, \mathbf{P}) \leq \epsilon/3$, there is a sampler for a distribution $\mathbf{Q}$ with $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) \leq \epsilon$, using $O(\log(b+1-a)+\log(1/\epsilon))$ uniform random bits as input, and running in time $O((D+1)(\log(b+1-a))+\log(1/\epsilon))$, where D is the running time of evaluating the CDF $c(x)$.*

Proof. We begin our analysis by producing an algorithm that works when we are able to exactly sample $c_{\mathbf{H}}(x)$.

We can compute an inverse to the CDF $d_{\mathbf{H}} : [0, 1] \to [a, b] \cap \mathbb{Z}$, at $y \in [0, 1]$, using binary search, as follows:

1. We have an interval $[a', b']$, initially $[a - 1, b]$, with $c_{\mathbf{H}}(a') \leq y \leq c_{\mathbf{H}}(b')$ and $c_{\mathbf{H}}(a') < c_{\mathbf{H}}(b')$.
2. If $b' - a' = 1$, output $d_{\mathbf{H}}(y) = b'$.
3. Otherwise, find the midpoint $c' = \lfloor (a' + b')/2 \rfloor$.
4. If $c_{\mathbf{H}}(a') < c_{\mathbf{H}}(c')$ and $y \leq c_{\mathbf{H}}(c')$, repeat with $[a', c']$; else repeat with $[c', b]$.

The function $d_{\mathbf{H}}$ can be thought of as some kind of inverse to the CDF $c_{\mathbf{H}} : [a - 1, b] \cap \mathbb{Z} \to [0, 1]$ in the following sense:

Claim 3.19. *The function $d_{\mathbf{H}}$ satisfies: For any $y \in [0, 1]$, it holds $c_{\mathbf{H}}(d_{\mathbf{H}}(y)-1) \leq y \leq c_{\mathbf{H}}(d_{\mathbf{H}}(y))$ and $c_{\mathbf{H}}(d_{\mathbf{H}}(y) - 1) < c_{\mathbf{H}}(d_{\mathbf{H}}(y))$.*

Proof. Note that if we don't have $c_{\mathbf{H}}(a') < c_{\mathbf{H}}(c')$ and $y \leq c_{\mathbf{H}}(c')$, then $c_{\mathbf{H}}(c') < y \leq c_{\mathbf{H}}(b')$. So, Step 4 gives an interval $[a', b']$ which satisfies $c_{\mathbf{H}}(a') \leq y \leq c_{\mathbf{H}}(b')$ and $c_{\mathbf{H}}(a') < c_{\mathbf{H}}(b')$. The initial interval $[a - 1, b]$ satisfies these conditions since $c_{\mathbf{H}}(a - 1) = 0$ and $c_{\mathbf{H}}(b) = 1$. By induction, all $[a', b']$ in the execution of the above algorithm have $c_{\mathbf{H}}(a') \leq y \leq c_{\mathbf{H}}(b')$ and $c_{\mathbf{H}}(a') < c_{\mathbf{H}}(b')$. Since this is impossible if $a' = b'$, and Step 4 always recurses on a shorter interval, we eventually have $b' - a' = 1$. Then, the conditions $c_{\mathbf{H}}(a') \leq y \leq c_{\mathbf{H}}(b')$ and $c_{\mathbf{H}}(a') < c_{\mathbf{H}}(b')$ give the claim. □

Computing $d_{\mathbf{H}}(y)$ requires $O(\log(b-a+1))$ evaluations of $c_{\mathbf{H}}$, and $O(\log(b-a+1))$ comparisons of y. For the rest of this proof, we will use $n = b - a + 1$ to denote the support size.

Consider the random variable $d_{\mathbf{H}}(Y)$, for Y uniformly distributed in $[0,1]$, whose distribution we will call $\mathbf{Q}'$. When $d_{\mathbf{H}}(Y) = x$, we have $c_{\mathbf{H}}(x-1) \leq Y \leq c_{\mathbf{H}}(x)$, and so when $\mathbf{Q}'(x) > 0$, we have $\mathbf{Q}'(x) \leq \Pr[c_{\mathbf{H}}(x-1) \leq Y \leq c_{\mathbf{H}}(x)] = c_{\mathbf{H}}(x) - c_{\mathbf{H}}(x-1) = \mathbf{H}(x)$. So, when $\mathbf{H}(x) > 0$, we have $\mathbf{H}(x) \geq \mathbf{Q}'(x)$. But when $\mathbf{H}(x) \leq 0$, we have $\mathbf{Q}'(x) = 0$, since then $c_{\mathbf{H}}(x) < c_{\mathbf{H}}(x-1)$ and no y has $c_{\mathbf{H}}(x-1) \leq y \leq c_{\mathbf{H}}(x)$. So, we have $d_{\mathrm{TV}}(\mathbf{Q}', \mathbf{H}) = \sum_{x:\mathbf{H}(x)<0} -\mathbf{H}(x) \leq d_{\mathrm{TV}}(\mathbf{H}, \mathbf{P}) \leq \epsilon/3$.

We now show how to effectively sample from $\mathbf{Q}'$. The issue is how to simulate a sample from the uniform distribution on $[0,1]$ with uniform random bits. We do this by flipping coins for the bits of Y lazily. We note that we will only need to know more than m bits of Y if Y is within 2^{-m} of one of the values of $c_{\mathbf{H}}(x)$ for some x. By a union bound, this happens with probability at most $n2^{-m}$ over the choice of Y. Therefore, for $m > \log_2(10n/\epsilon)$, the probability that this will happen is at most $\epsilon/10$ and can be ignored.

Therefore, the random variable $d_{\mathbf{H}}(Y')$, for Y' uniformly distributed on the multiples of 2^{-r} in $[0,1)$ for $r = O(\log n + \log(1/\epsilon))$, has distribution $\mathbf{Q}'$ that satisfies $d_{\mathrm{TV}}(\mathbf{Q}, \mathbf{Q}') \leq \epsilon/10$. This means that $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}') \leq d_{\mathrm{TV}}(\mathbf{P}, \mathbf{H}) + d_{\mathrm{TV}}(\mathbf{H}, \mathbf{Q}) + d_{\mathrm{TV}}(\mathbf{Q}, \mathbf{Q}') \leq 9\epsilon/10$. That is, we obtain an ϵ-sampler that uses $O(\log n + \log(1/\epsilon))$ coin flips, $O(\log n)$ calls to $c_{\mathbf{H}}(x)$, and has the desired running time.

We now need to show how this can be simulated without access to $c_{\mathbf{H}}$, and instead only having access to its approximation $c(x)$. The modification required is rather straightforward. Essentially, we can run the same algorithm using $c(x)$ in place of $c_{\mathbf{H}}(x)$. We note that all comparisons with Y will produce the same result, unless the chosen Y is between $c(x)$ and $c_{\mathbf{H}}(x)$ for some value of x. Observe that because of our bounds on their difference, the probability of this occurring for any given value of x is at most $\epsilon/(10n)$. By a union bound, the probability of it occurring for any x is at most $\epsilon/10$. Thus, with probability at least $1 - \epsilon/10$ our algorithm returns the same result that it would have had it had access to $c_{\mathbf{H}}(x)$ instead of $c(x)$. This implies that the variable sampled by this algorithm has variation distance at most $\epsilon/10$ from what would have been sampled by our other algorithm. Therefore, this algorithm samples a $\mathbf{Q}$ with $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) \leq \epsilon$. □

We next show that we can efficiently compute an appropriate CDF using the DFT. For the 1-dimensional case, this follows easily via a closed form expression. For the high-dimensional case, we first obtain a closed form expression for the case that the matrix M is diagonal. We then reduce the general case to the diagonal case, by using a Smith normal form decomposition.

Proposition 3.20. *For* $\mathbf{H}$ *as in Theorem 3.15, we have the following:*

(i) If $k = 1$*, there is an algorithm to compute the CDF* $c_{\mathbf{H}} : [a,b] \cap \mathbb{Z} \to [0,1]$ *with* $c_{\mathbf{H}}(x) = \sum_{i:a\leq i\leq x} \mathbf{H}(i)$ *to any precision* $\delta > 0$*, where* $a = m - \lceil M/2 \rceil + 1$ *and* $b = m + \lfloor M/2 \rfloor$*,* $M \in \mathbb{Z}_+$*. The algorithm runs in time* $O(|T| \log(1/\delta))$*.*

(ii) If $M \in \mathbb{Z}^{k\times k}$ *is diagonal, there is an algorithm which computes the CDF to any precision* $\delta > 0$ *under the lexicographic ordering* $\leq_{\mathrm{lex}}$*, i.e.,* $c_{\mathbf{H}}(x) = \sum_{y\in T: y\leq_{\mathrm{lex}} x} \mathbf{H}(y)$*. The algorithm runs in time* $O(k^2|T| \log(1/\delta))$*.*

(iii) For any $M \in \mathbb{Z}^{k\times k}$*, there is an explicit ordering* $\leq_g$ *for which we can compute the CDF* $c_{\mathbf{H}}(x) = \sum_{y\in T: y\leq_g x} \mathbf{H}(y)$ *to any precision* $\delta > 0$*. This computation can be done in time* $O(k^2|T| \log(1/\delta) + \mathrm{poly}(k))$*.*

In cases (ii) and (iii), we can also compute the embedding of the corresponding ordering onto the integers $[|\det M|] = \{1, 2, ..., |\det M|\}$*, i.e., we can give a monotone bijection* $f : S \to [|\det M|]$ *for which we can efficiently compute* f *and* f^{-1} *(i.e., with the same running time bound we give for computing* $c_{\mathbf{H}}(x)$*).*

Proof. Recall that the PMF of $\mathbf{H}$ at $x \in S$ is given by the inverse DFT:

$$\mathbf{H}(x) = \frac{1}{|\det M|} \sum_{\xi \in T} e(-\xi \cdot x) \widehat{\mathbf{H}}(\xi) \ . \tag{10}$$

Proof of (i): For (i), the CDF is given by:

$$\begin{aligned} c_{\mathbf{H}}(x) &= \frac{1}{M} \sum_{i : a \le i \le x} \sum_{\xi \in T} e(-\xi x) \widehat{\mathbf{H}}(\xi) \\ &= \frac{1}{M} \sum_{\xi \in T} \widehat{\mathbf{H}}(\xi) \sum_{i : a \le i \le x} e(-\xi x) \end{aligned}$$

When $\xi \neq 0$, the term $\sum_{i:a \le i \le x} e(-\xi \cdot x)$ is a geometric series. By standard results on its sum, we have:

$$\sum_{i : a \le i \le x} e(-\xi x) = \frac{e(-\xi a) - e(-\xi(x+1))}{1 - e(-\xi)} \ .$$

When $\xi = 0$, $e(-\xi) = 1$, and we get $\sum_{a \le i \le x} e(-\xi x) = i + 1 - a$. In this case, we also have $\widehat{\mathbf{H}}(\xi) = 1$. Putting this together we have:

$$c_{\mathbf{H}}(x) = \frac{1}{M} \left(i + 1 - a + \sum_{\xi \in T \setminus \{0\}} \widehat{\mathbf{H}}(\xi) \frac{e(-\xi a) - e(-\xi(x+1))}{1 - e(-\xi)} \right) \ . \tag{11}$$

Hence, we obtain a closed form expression for the CDF that can be approximated to desired precision in time $O(|T| \log(1/\delta))$.

Proof of (ii): For (ii), we can write $M = \mathrm{diag}(M_i)$, $1 \le i \le k$, and $S = \prod_{i=1}^{k} ([a_i, b_i] \cap \mathbb{Z})$, where $a_i = m_i - \lceil |M_i|/2 \rceil + 1$ and $b_i = m_i + \lfloor |M_i|/2 \rfloor$. With our lexicographic ordering, we have:

$$\begin{aligned} c_{\mathbf{H}}(x) &= \sum_{y \in S : y \le_{\mathrm{lex}} x} \mathbf{H}(y) \\ &= \sum_{y_1 = a_1}^{x_1 - 1} \sum_{y_2 = a_2}^{b_2} \cdots \sum_{y_k = a_k}^{b_k} \mathbf{H}(y) \\ &\quad + \sum_{y_2 = a_2}^{x_2 - 1} \sum_{y_3 = a_3}^{b_3} \cdots \sum_{y_k = a_k}^{b_k} \mathbf{H}(x_1, y_2, \ldots, y_k) \\ &\quad \cdots \\ &\quad + \sum_{y_k = a_k}^{x_k - 1} \mathbf{H}(x_1, \ldots, x_{k-1}, y_k) \\ &\quad + \mathbf{H}(x) \ . \end{aligned}$$

To avoid clutter in the notation, we define $c_{\mathbf{H},i}(x)$ to be one of these sums, i.e.,

$$\begin{aligned}
c_{\mathbf{H},i}(x) &\stackrel{\text{def}}{=} \sum_{y_i=a_i}^{x_i-1} \sum_{y_{i+1}=a_{i+1}}^{b_{i+1}} \cdots \sum_{y_k=a_k}^{b_k} \mathbf{H}(x_1,\ldots,x_{i-1},y_i,\ldots,y_k) \\
&= \frac{1}{|\det(M)|} \cdot \sum_{y_i=a_i}^{x_i-1} \sum_{y_{i+1}=a_{i+1}}^{b_{i+1}} \cdots \sum_{y_k=a_k}^{b_k} \sum_{\xi\in T} \widehat{\mathbf{H}}(\xi) e\left(-\sum_{j=1}^{i-1}\xi_j x_j - \sum_{j=i}^{k}\xi_j y_j\right) \\
&= \frac{1}{|\det(M)|} \cdot \sum_{\xi\in T} \widehat{\mathbf{H}}(\xi) e\left(-\sum_{j=1}^{i-1}\xi_j x_j\right) \left(\sum_{y_i=a_i}^{x_i-1} e(-\xi_i y_i)\right) \prod_{j=i+1}^{k} \sum_{y_i=a_i}^{b_i} e(-\xi_j y_j) \\
&= \frac{1}{|\det(M)|} \cdot \sum_{\xi\in T} \widehat{\mathbf{H}}(\xi) e\left(-\sum_{j=1}^{i-1}\xi_j x_j\right) s_i(a_i, x_i-1) \prod_{j=i+1}^{k} s_j(a_j,b_j)\ ,
\end{aligned}$$

where $s_i(a_i', b_i') := \sum_{y_i=a_i'}^{b_i'} e(-\xi_i y_i)$. As before, this is a geometric series, so either $\xi_i = 0$, when we have $s_i = b_i' + 1 - a_i'$, or $s_i = \frac{e(-\xi_i a_i') - e(-\xi(b_i'+1))}{1-e(-\xi)}$.

We can thus evaluate $c_{\mathbf{H},i}$ in $O(|T|k)$ arithmetic operations and so compute $c_{\mathbf{H}}(x)$ to desired accuracy in $O(|T|k^2 \log(1/\delta))$ time. We also note that $f : S \to \{0,1,...,(\prod_i |M_i|)-1\}$, defined by $f(x) := \sum_i (x_i - a_i) \prod_{j=1}^{i} M_j$ is a strictly monotone bijection, and that f and f^{-1} can be computed in time $O(k)$. Now, $c_{\mathbf{H}}(f^{-1}(y)$ is the CDF of the distribution on $y \in \{0,1,...,(\prod_i |M_i|)-1\}$ whose PMF is given by $\mathbf{H}(f^{-1}(y))$.

Proof of (iii): We will reduce (iii) to (ii). To do this, we use Smith normal form, a canonical factorization of integer matrices:

Lemma 3.21 (See [Sto00], [Sto96])**.** *Given any integer matrix $M \in \mathbb{Z}^{k\times k}$, we can factorize M as $M = U \cdot D \cdot V$, where $U, D, V \in \mathbb{Z}^{k\times k}$ with D diagonal and U,V unimodular, i.e., $|\det(U)| = |\det(V)| = 1$, and therefore $U^{-1}, V^{-1} \in \mathbb{Z}^{k\times k}$. This factorization can be computed in time*

$$\text{poly}(k) \log \max_{i,j} |M_{i,j}|.$$

Note that the Smith normal form satisfies additional conditions on D than those in Lemma 3.21, but we are only interested in finding such a decomposition where D is a diagonal integer matrix.

We note that the integer lattices $M\mathbb{Z}^k$ and $UD\mathbb{Z}^k$ are identical, since if $x = Mb$ for $b \in \mathbb{Z}^k$, then $x = UDc$ for $c = Vb \in \mathbb{Z}^k$. For any $\xi \in (M^T)^{-1}\mathbb{Z}^k$, $\xi = (U^T)^{-1}(D^T)^{-1}(V^T)^{-1}b$, for $b \in \mathbb{Z}$. Then, if we take $\nu = U^T\xi$, we have $\nu \in (DV)^{-1}\mathbb{Z}^k = D^{-1}\mathbb{Z}^k$.

Hence, we can re-write (10) as follows:

$$\mathbf{H}(x) = \frac{1}{|\det M|} \sum_{\nu\in U^T T} e\left(-\nu U^{-1} x\right) \widehat{\mathbf{H}}((U^T)^{-1}\nu)\ .$$

Since $U^T T \subseteq D^{-1}\mathbb{Z}^k$, substituting $y = U^{-1}x$ almost gives us the conditions which would allow us to apply (ii). The issue is that for $x \in m + M(-1/2,1/2]^k$, we have $U^{-1}X \in U^{-1}m + DV(-1/2,1/2]^k$, but we do not necessarily have $U^{-1}X \in U^{-1}m + D(-1/2,1/2]^k$. The following claim gives the details of how to change the fundamental domain:

Claim 3.22. *Given a non-singular $M \in \mathbb{Z}^{k\times k}$ and $m, x \in \mathbb{R}^k$, then $x' = x + MR\left(M^{-1}(m-x)\right)$ is the unique $x' \in m + M(-1/2,1/2]^k$ with $x - x' \in M\mathbb{Z}^k$, where $R(x)$ is x with each coordinate rounded to the nearest integer, rounding half integers up, i.e., $(R(x))_i := \frac{1}{2} + \lceil x_i - \frac{1}{2} \rceil$.*

So we take $y = g(x) \stackrel{\text{def}}{=} U^{-1}x + DR\left((UD)^{-1}m - (UD)^{-1}x\right)$, which using Claim 3.22 has $g(x) \in U^{-1}m + D(-1/2,1/2]^k$. We need the inverse function of $g : m + M(-1/2,1/2]^k \to U^{-1}m + D(-1/2,1/2]^k$. Note that $g^{-1}(y) = Uy + D^{-1}R(U^{-1}m - y)$, which again by Claim 3.22 is in $m + M(-1/2,1/2]^k$.

So, if $y = g(x)$, since $|\det(M)| = |\det(U)| \cdot |\det(D)| \cdot |\det(V)| = |\det(D)|$, we have:

$$\mathbf{H}(g^{-1}(y)) = \frac{1}{|\det(D)|} \sum_{\nu \in U^T T} e(\nu \cdot y)\widehat{\mathbf{H}}((U^T)^{-1}\nu) \; . \tag{12}$$

Now, we can take $\mathbf{H}(g^{-1}(y)$ to be a function of y supported on $U^{-1}m + D(-1/2,1/2]^k$ with a sparse DFT modulo D supported on $U^T T \subseteq D^{-1}\mathbb{Z}^k$. At this point, we can apply the algorithm of (ii), which gives a way to compute the CDF of $\mathbf{H}(g^{-1}(y))$ with respect to the lexicographic ordering on y. Note that g and g^{-1} can be computed in time $\mathrm{poly}(k)$, or more precisely the running time of matrix multiplication and inversion.

For the ordering on $x \in S$, which has $x_1 \leq_g x_2$ when $g(x_1) \leq_{\text{lex}} g(x_2)$, we can compute $c_{\mathbf{H}}(x) = \sum_{y \in S : y \leq_g x} \mathbf{H}(y)$ by applying the algorithm in (ii) to the function given in (12) applied at $g(x)$. So, we can compute $c_{\mathbf{H}}(x)$ in time $O(k^2|T| + \mathrm{poly}(k))$. Again, the function given by $f(g(x))$, where f is as in (ii) is a monotone bijection from S to $\{0, 1, \ldots, \det(M) - 1\}$, and we can calculate this function and its inverse in time $\mathrm{poly}(k)$. □

Now we can prove the main theorem of this subsection.

Proof of Theorem 3.15. By Proposition 3.20 (iii), there is a bijection f which takes the support S of $\mathbf{H}$ to the integers $\{0, 1, \ldots, |S| - 1\}$, and we can efficiently calculate the CDF of the distribution considered on this set of integers. So, we can apply Lemma 3.18 to this CDF on this distribution. This gives us an ϵ-sampler for this distribution, which we can then apply f^{-1} to each sample to get an ϵ-sampler for $\mathbf{H}$. To find the time it takes to compute each sample, we need to substitute $D = O(\mathrm{poly}(k) + k^2|T| \log(|\det(M)|/\epsilon))$ from the running time of the CDF in Proposition 3.20 (iii) into the bound in Lemma 3.18, yielding

$$O(\log(|\det(M)|) \log(|\det(M)|/\epsilon) \cdot |T| \cdot \mathrm{poly}(k))$$

time. This completes the proof. □

3.3 Using our Learning Algorithm to Obtain a Cover As an application of our learning algorithm in Section 3.1, we provide a simple proof that the space $\mathcal{M}_{n,k}$ of all (n,k)-PMDs has an ϵ-cover under the total variation distance 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.

Two remarks are in order: (i) the non-proper cover construction in this subsection does not suffice for our algorithmic applications of Section 4. These applications require the efficient construction of a *proper* ϵ-cover plus additional algorithmic ingredients. (ii) The upper bound on the cover size obtained here is nearly optimal, as follows from our lower bound in Section 4.5.

The idea behind using our algorithm to obtain a cover is quite simple. In order to determine its hypothesis, $\mathbf{H}$, our algorithm `Efficient-Learn-PMD` requires the following quantities:

- A vector $\widehat{\mu} \in \mathbb{R}^k$ and a PSD matrix $\widehat{\Sigma} \in \mathbb{R}^{k\times k}$ satisfying the conclusions of Lemma 3.4.

- Values of the DFT $\widehat{\mathbf{H}}(\xi)$, for all $\xi \in T$, so that $\sum_{\xi \in T} |\widehat{\mathbf{H}}(\xi) - \widehat{\mathbf{P}}(\xi)| < \epsilon/10$. Recall that $T \stackrel{\text{def}}{=} \left\{\xi = (M^T)^{-1} v \;\middle|\; \left(v \in \mathbb{Z}^k\right) \wedge \left(\|v\|_2 \leq C^2 k^2 \ln(k/\epsilon)\right)\right\}$.

Given this information, the analysis in Section 3.1 carries over immediately. The algorithm `Efficient-Learn-PMD` works by estimating the mean and covariance using samples, and then taking $\widehat{\mathbf{H}}(\xi)$ to be the sample Fourier transform. If we instead guess the values of these quantities using an appropriate discretization, we obtain an ϵ-cover for $\mathcal{M}_{n,k}$. More specifically, we discretize the above quantities as follows:

- To discretize the mean vector, we consider a 1-cover $\mathcal{Y}_1$ of the set $\mathcal{Y} \stackrel{\text{def}}{=} \{y = (y_1, \ldots, y_k) \in \mathbb{R}^k : (y_i \geq 0) \wedge (\sum_{i=1}^k y_i = k)\}$ with respect to the L_2 norm. It is easy to construct such a cover with size $|\mathcal{Y}_1| \leq O(n)^k$.

- To discretize the covariance matrix, we consider a 1/2-cover $\mathcal{S}_{1/2}$ of the set of matrices $\mathcal{S} \stackrel{\text{def}}{=} \{A \in \mathbb{R}^{k \times k} : (A \succeq \mathbf{0}) \wedge \|A\|_2 \leq n\}$, with respect to the spectral norm $\|\cdot\|_2$. Note that we can construct such a 1/2-cover with size $|\mathcal{S}_{1/2}| \leq (4n+1)^{k(k+1)/2}$. This is because any maximal 1/4-packing of this space (i.e., a maximal set of matrices in $\mathcal{S}$ with pairwise distance under the spectral norm at least 1/4) is such a 1/2-cover. Observe that for any such maximal packing, the balls of radius 1/4 centered at these points are disjoint and contained in the ball (under the spectral norm) about the origin of radius $(n+1/4)$. Since the ball of radius $(n+1/4)$ has volume $(4n+1)^{k(k+1)/2}$ times as much as the ball of radius 1/4, a simple volume argument completes the proof.

- Finally, to discretize the Fourier transform, we consider a δ-cover $\mathcal{C}_\delta$ of the unit disc on the complex plane $\mathbb{C}$, with respect to the standard distance on $\mathbb{C}$, where

$$\delta \stackrel{\text{def}}{=} \epsilon(2C^2k^2\log(k/\epsilon))^{-k}/10 = \epsilon/(10t) \leq \epsilon/(10|T|).$$

  We note that $t \stackrel{\text{def}}{=} (2C^2k^2\log(k/\epsilon))^k$ is an upper bound on $|T|$.

We claim that there is an ϵ-cover of the space of (n,k)-PMDs indexed by $\mathcal{Y}_1 \times \mathcal{S}_{1/2} \times \mathcal{C}_\delta^t$. Such a cover is clearly of the desired size. The cover is constructed as follows: We let $\widehat{\mu}$ and $\widehat{\Sigma}$ be the selected elements from $\mathcal{Y}_1$ and $\mathcal{S}_{1/2}$, respectively. We use these elements to define the matrix $M \in \mathbb{Z}^{k \times k}$ as in the algorithm description. We then use our elements of $\mathcal{C}_\delta$ as the values of $\widehat{\mathbf{H}}(\xi)$ for $\xi \in T$ (noting that $|T| \leq t$).

We claim that for any (n,k)-PMD $\mathbf{P}$ there exists a choice of parameters, so that the returned distribution $\mathbf{H}$ is within total variation distance ϵ of $\mathbf{P}$. We show this as follows: Let μ and Σ be the true mean and covariance matrix of $\mathbf{P}$. We have that $\mu \in \mathcal{Y}$ and that $\Sigma \in \mathcal{S}$. Therefore, there exist $\widehat{\mu} \in \mathcal{Y}_1$ and $\widehat{\Sigma} \in \mathcal{S}_{1/2}$ so that $|\mu - \widehat{\mu}|_2 \leq 1$ and $I/2 \succeq \Sigma - \widehat{\Sigma} \succeq -I/2$. It is easy to see that these conditions imply the conclusions of Lemma 3.4. Additionally, we can pick elements of $\mathcal{C}_\delta$ in order to make $|\widehat{\mathbf{H}}(\xi) - \widehat{\mathbf{P}}(\xi)| < \epsilon/(10|T|)$ for each $\xi \in T$. This will give that $\sum_{\xi \in T} |\widehat{\mathbf{H}}(\xi) - \widehat{\mathbf{P}}(\xi)| < \epsilon/10$. In particular, the hypothesis $\mathbf{H}$ indexed by this collection of parameters will be within variation distance ϵ of $\mathbf{P}$. Hence, the set we have constructed is an ϵ-cover, and our proof is complete.

4 Efficient Proper Covers and Nash Equilibria in Anonymous Games

In this section, we give 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 given in Section 4.2. By building on the proper cover construction, in Section 4.3 we give our EPTAS for Nash equilibria in anonymous games. In Section 4.4, we combine our moment-matching lemma with recent results in algebraic geometry, to show that any PMD is close to another PMD with few distinct CRV components. Finally, in Section 4.5 we prove out cover size lower bound.

4.1 Low-Degree Parameter Moment Closeness Implies Closeness in Variation Distance In this subsection, we establish the sparsity of the continuous Fourier transform of PMDs, and use it to prove our robust moment-matching lemma, translating closeness in the low-degree parameter moments to closeness in total variation distance.

At a high-level, our robust moment-matching lemma (Lemma 4.6) is proved by combining the sparsity of the continuous Fourier transform of PMDs (Lemma 4.2) 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 4.1. 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 $1 \leq i \leq n$ and $1 \leq j \leq k$. Observe that $\sum_{j=1}^{k} p_{i,j} = 1$, for $1 \leq i \leq n$, hence the definition of k-maximality implies that $p_{i,k} \geq 1/k$ for all i. Note that the j^{th} component of the random vector X is a PBD with parameters $p_{i,j}$, $1 \leq i \leq n$. 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 $s_j(X) \geq \epsilon/k$, for all $1 \leq j \leq k-1$; otherwise, we can remove the corresponding coordinates and introduce an error of at most ϵ in variation distance.

Note that, for $j \neq k$, 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 $j \neq k$, we have $1 \geq p_{i,k} + p_{i,j} \geq 2p_{i,j}$, or $p_{i,j} \leq 1/2$, 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

$$|S| \leq \prod_{j=1}^{k-1} \left(1 + 12 s_j(X)^{1/2} \ln(2k/\epsilon)\right) \leq O\left(\log(k/\epsilon)^{(k-1)}\right) \cdot \prod_{j=1}^{k-1} (1 + 12 s_j(X)^{1/2}) \,,$$

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

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

where $[x]$ is the distance between x and the nearest integer, and $C > 0$ is a sufficiently large universal constant. Then, we have that

$$\int_{\overline{T}} |\widehat{X}| \ll \epsilon/|S|.$$

Proof. To prove the lemma, we will need the following technical claim:

Claim 4.3. *For all $\xi = (\xi_1, \ldots, \xi_k) \in [0,1]^k$, for all $1 \leq i \leq n$ and $1 \leq j \leq k-1$, it holds:*

$$|\widehat{X}_i(\xi)| = \exp(-\Omega(p_{i,j}[\xi_j - \xi_k]^2/k)) \,.$$

Proof. The claim follows from the following sequence of (in-)equalities:

$$\begin{aligned}
|\widehat{X_i}(\xi)|^2 =& \Big(\sum_{j=1}^{k} p_{i,j} e(\xi_j)\Big)\Big(\sum_{j'=1}^{k} p_{i,j'} e(-\xi_{j'})\Big) = \sum_{j,j'} p_{i,j} p_{i,j'} e(\xi_j - \xi_{j'}) \\
=& \sum_{j,j'} p_{i,j} p_{i,j'} \cos(2\pi(\xi_j - \xi_{j'})) = 1 - \sum_{j \neq j'} p_{i,j} p_{i,j'} \big(1 - \cos(2\pi(\xi_j - \xi_{j'}))\big) \\
=& 1 - \sum_{j \neq j'} p_{i,j} p_{i,j'} \big(1 - \Omega([\xi_j - \xi_{j'}])^2\big) \qquad \text{(by Claim 3.11 with } \delta = \tfrac{1}{2}) \\
=& \exp\left(-\Omega\left(\sum_{j \neq j'} p_{i,j'} p_{i,j} [\xi_j - \xi_{j'}]^2\right)\right) \\
=& \exp\left(-\Omega\left(\sum_{j<k} p_{i,j} p_{i,k} [\xi_j - \xi_k]^2\right)\right) \\
=& \exp\left(-\Omega\left(p_{i,j}[\xi_j - \xi_k]^2/k\right)\right) \,,
\end{aligned}$$

where the last lines uses the fact $p_{i,k} \geq 1/k$, which follows from k-maximality. □

As a consequence of Claim 4.3, we have that

$$|\widehat{X}(\xi)| = \prod_{i=1}^{n} |\widehat{X_i}(\xi)| = \exp(-\Omega(s_j(X)[\xi_j - \xi_k]^2/k)) \; . \tag{13}$$

Let $\overline{T} = [0,1]^k \setminus T$ be the complement of T. To bound $\int_{\overline{T}} |\widehat{X}|$, we proceed as follows: For $\ell \in \mathbb{Z}_+$, we define the sets

$$\overline{T}_\ell = \left\{ \xi : \max([\xi_j - \xi_k]/(Ck(1+12s_j(X))^{-1/2}\log^{1/2}(1/\epsilon))) \in [2^\ell, 2^{\ell+1}] \right\} \; ,$$

and observe that $\overline{T} \subseteq \cup_{\ell \in \mathbb{Z}_+} \overline{T}_\ell$. Now, Equation (13) implies that for $\xi \in \overline{T}_\ell$ it holds $|\widehat{X}(\xi)| \leq \epsilon^{10k2^\ell}$, where we used the assumption that the constant C is sufficiently large. It is easy to see that the volume of $\overline{T}_\ell$ is at most $O(2^\ell k \log^{1/2}(1/\epsilon))^k \prod_{j<k}(1+12s_j(X))^{-1/2}$. We can bound $\int_{\overline{T}} |\widehat{X}|$ from above by the sum over ℓ of the maximum value of $|\widehat{X}|$ within $\overline{T}_\ell$ times the volume of $\overline{T}_\ell$, namely

$$\int_{\overline{T}} |\widehat{X}| \leq \sum_{\ell=0}^{\infty} \int_{\overline{T}_\ell} |\widehat{X}| \leq \sum_{\ell=0}^{\infty} \left(\sup_{\xi \in \overline{T}_\ell} |\widehat{X}(\xi)| \right) \mathrm{Vol}(\overline{T}_\ell) \leq \epsilon^k \prod_{j<k} (1+12s_j(X))^{-1/2} \ll \epsilon/|S|.$$

This completes the proof of Lemma 4.2. □

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 $1/2 \leq (1+s_j(X))/(1+s_j(Y)) \leq 2$ for all j, $1 \leq j \leq k-1$. Let $T \stackrel{\text{def}}{=} \left\{ \xi \in [0,1]^k : [\xi_j - \xi_k] < Ck(1+12s_j(X))^{-1/2}\log^{1/2}(1/\epsilon) \right\}$, 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 $|\widehat{X}(\xi) - \widehat{Y}(\xi)| \leq \epsilon(Ck\log(k/\epsilon))^{-2k}$. Then, $d_{\mathrm{TV}}(X,Y) \leq \epsilon$.*

Proof. We start with an intuitive explanation of the proof. Since $1+s_j(X)$, $1+s_j(Y)$ are within a factor of 2 for $1 \leq j \leq k-1$, it follows from the above that X and Y are both effectively supported on a set $S \subseteq [n]^k$ of size $|S| \leq O\left(\log(k/\epsilon)\right)^{k-1} \cdot \prod_{j=1}^{k-1} \left(1+12s_j(X)^{1/2}\right)$. Therefore, to prove the lemma, it is sufficient to establish that $\|X-Y\|_\infty \leq O(\epsilon/|S|)$.

We prove this statement in two steps by analyzing the continuous Fourier transforms $\widehat{X}$ and $\widehat{Y}$. The first step of the proof exploits the fact that the Fourier transforms of X and Y are each essentially supported on the set T of the lemma statement. Recalling the assumption that $(1+s_j(X))/(1+s_j(Y)) \in [1/2,2]$, $1 \leq j \leq k-1$, an application of Lemma 4.2 yields that $\int_{\overline{T}} |\widehat{X}|$ and $\int_{\overline{T}} |\widehat{Y}|$ are both at most $\epsilon/|S|$. Thus, we have that

$$\int_{\overline{T}} |\widehat{X} - \widehat{Y}| \leq \int_{\overline{T}} |\widehat{X}| + \int_{\overline{T}} |\widehat{Y}| \ll \epsilon/|S| \; .$$

In the second step of the proof, we use the assumption that the absolute difference $|\widehat{X}(\xi) - \widehat{Y}(\xi)|$, $\xi \in T$, is small, and the fact that $\int_{\overline{T}} |\widehat{X}|$ and $\int_{\overline{T}} |\widehat{Y}|$ are individually small, to show that $\|\widehat{X} - \widehat{Y}\|_1 \leq O(\epsilon/|S|)$. The straightforward inequality $\|X-Y\|_\infty \leq \|\widehat{X} - \widehat{Y}\|_1$ combined with the concentration of X, Y completes the proof.

Given the aforementioned, in order to bound $\|\widehat{X} - \widehat{Y}\|_1$, it suffices to show that the integral over T is small. By the assumption of the lemma, we have that $|\widehat{X}(\xi) - \widehat{Y}(\xi)|$ is point-wise at most

$\epsilon(Ck\log(k/\epsilon))^{-2k}$ over T. We obtain an upper bound on $\int_T |\widehat{X}-\widehat{Y}|$ by multiplying this quantity by the volume of T. Note that the volume of T is at most $O(k\log^{1/2}(1/\epsilon))^{k-1}\prod_{j<k}(1+12s_j(X))^{-1/2})$. Hence,

$$\int_T |\widehat{X}-\widehat{Y}| \leq \epsilon\cdot\Theta(\log(k/\epsilon))^{-k}\cdot\prod_{j<k}(1+12s_j(X))^{-1/2} \ll \epsilon/|S|.$$

Combining the above, we get that $\|X-Y\|_\infty \leq \|\widehat{X}-\widehat{Y}\|_1 = O(\epsilon/|S|)$, which implies that the L_1 distance between X and Y over S is $O(\epsilon)$. The contribution of $\overline{S}$ to the L_1 distance is at most ϵ, since both X and Y are in S with probability at least $1-\epsilon$. This completes the proof of Lemma 4.4. □

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 $0<\delta\leq 1$. 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 $1\leq j\leq k-1$, such that $p_{\ell,j}\geq\delta\cdot\sqrt{1+s_j(X)}$. 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. To see this, we observe that the number of δ-exceptional components is at most $k-1$ times the number of (δ,j)-exceptional components, i.e., the k-CRVs X_i which are δ-exceptional for the same value of j. We claim that for any j, $1\leq j\leq k-1$, the number of (δ,j)-exceptional components is at most $1/\delta^2$. Indeed, let $E_j\subseteq[n]$ denote the corresponding set. Then, we have that $\sum_{i\in E_j}p_{i,j}^2\geq\delta^2|E_j|s_j(X)=\delta^2|E_j|\sum_{i=1}^n p_{i,j}$. Noting that $\sum_{i\in E_j}p_{i,j}^2\leq\sum_{i=1}^n p_{i,j}^2\leq\sum_{i=1}^n p_{i,j}$, we get that $\delta^2|E_j|\leq 1$, thus yielding the claim.

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 $1/2\leq(1+s_j(X))/(1+s_j(Y))\leq 2$ for all j, $1\leq j\leq k-1$. 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 $Y^{(t)}$, $1\leq t\leq 3$, are PMDs over the same number of k-CRVs. Additionally assume the following: (i) for $t\leq 2$ 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

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

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

Proof. First, note that $d_{\mathrm{TV}}(X, Y) \le \sum_{t=1}^{3} d_{\mathrm{TV}}(X^{(t)}, Y^{(t)})$, so it suffices to show that $d_{\mathrm{TV}}(X^{(t)}, Y^{(t)}) = \epsilon/3$, for $t = 1, 2, 3$. This holds trivially for $t = 3$, by assumption. To prove the statement for $t = 1, 2$, by Lemma 4.4, it is sufficient to show that $\widehat{X^{(t)}}$ and $\widehat{Y^{(t)}}$ are point-wise close on the set T, namely that for all $\xi \in T$ it holds $|\widehat{X^{(t)}}(\xi) - \widehat{Y^{(t)}}(\xi)| \le \epsilon(Ck\log(k/\epsilon))^{-2k}$. To show this, we show separately that $\widehat{X^{(t)}}$ is close to $\widehat{Y^{(t)}}$ for each $t = 1, 2$.

Let $X^{(t)} = \sum_{i \in A_t} X_i$, where $A_t \subseteq [n]$ with $|A_t| = n_t$. We have the following formula for the Fourier transform of $X^{(t)}$:

$$\begin{aligned}
\widehat{X^{(t)}}(\xi) &= \prod_{i \in A_t} \sum_{j=1}^{k} e(\xi_j) p_{i,j} \\
&= e(n_t \xi_k) \prod_{i \in A_t} \left(1 - \sum_{j=1}^{k-1} (1 - e(\xi_j - \xi_k)) p_{i,j}\right) \\
&= e(n_t \xi_k) \exp\left(\sum_{i \in A_t} \log\left(1 - \sum_{j=1}^{k-1} (1 - e(\xi_j - \xi_k)) p_{i,j}\right)\right) \\
&= e(n_t \xi_k) \exp\left(-\sum_{i \in A_t} \sum_{\ell=1}^{\infty} \frac{1}{\ell} \left(\sum_{j=1}^{k-1} (1 - e(\xi_j - \xi_k)) p_{i,j}\right)^{\ell}\right) \\
&= e(n_t \xi_k) \exp\left(-\sum_{m \in \mathbb{Z}_+^{k-1}} \binom{|m|_1}{m} \frac{1}{|m|_1} M_m(X^{(t)}) \prod_{j=1}^{k-1} (1 - e(\xi_j - \xi_k))^{m_j}\right). \qquad (14)
\end{aligned}$$

An analogous formula holds for $\widehat{Y^{(t)}}$. To prove the lemma, we will show that, for all $\xi \in T$, the two corresponding expressions inside the exponential of (14) agree for $X^{(t)}$ and $Y^{(t)}$, up to a sufficiently small error.

We first deal with the terms with $|m|_1 \le K_t$, $t = 1, 2$. By the statement of the lemma, for any two such terms we have that $|M_m(X^{(t)}) - M_m(Y^{(t)})| \le (2k)^{-|m|_1} \cdot \epsilon(Ck\log(k/\epsilon))^{-2k}$. Hence, for any $\xi \in [0, 1]^k$, the contribution of these terms to the difference is at most

$$\epsilon(Ck\log(k/\epsilon))^{-2k-1} \sum_{m \in \mathbb{Z}_+^{k-1},\, |m|_1 \le K_t} \binom{|m|_1}{m} (2k)^{-|m|_1} 2^{|m|_1} \le K_t \epsilon (Ck\log(k/\epsilon))^{-2k-1} \le \epsilon(Ck\log(k/\epsilon))^{-2k}\,.$$

To deal with the remaining terms, we need the following technical claim:

Claim 4.7. *Let $X^{(t)}$ be as above. For $t \le 2$ and $m \in \mathbb{Z}_+^{k-1}$ with $|m|_1 \ge 2$, we have that*

$$|M_m(X^{(t)})| \prod_{j=1}^{k-1} \left((1 + s_j(X))^{-1/2} \log^{1/2}(1/\epsilon)\right)^{m_j} \le \log^{|m|_1/2}(1/\epsilon) \cdot \delta_t^{|m|_1 - 2}.$$

Proof. By definition we have that $M_m(X^{(t)}) = \sum_{i \in A_t} \prod_{j=1}^{k-1} p_{i,j}^{m_j}$. Thus, the claim is equivalent to showing that

$$\sum_{i \in A_t} \prod_{j=1}^{k-1} \left(p_{i,j} \cdot (1 + s_j(X))^{-1/2}(X) \right)^{m_j} \leq \delta_t^{|m|_1 - 2}.$$

Since, by definition, $X^{(t)}$ does not contain any δ_t-exceptional k-CRV components, we have that for all $i \in A_t$ and all $j \in [k-1]$ it holds $p_{i,j} \cdot (1 + s_j(X))^{-1/2}(X) \leq \delta_t$. Now observe that decreasing any component of m by 1 decreases the left hand side of the above by a factor of at least δ_t. Therefore, it suffices to prove the desired inequality for $|m|_1 = 2$, i.e., to show that

$$\sum_{i \in A_t} \left(p_{i,j_1} (1 + s_{j_1}(X))^{-1/2} \right) \left(p_{i,j_2} (1 + s_{j_2}(X))^{-1/2} \right) \leq 1.$$

Indeed, the above inequality holds true, as follows from an application of the Cauchy-Schwartz inequality, and the fact that

$$\sum_{i \in A_t} p_{i,j}^2 \leq \sum_{i=1}^{n} p_{i,j} = s_j(X) \leq s_j(X) + 1.$$

This completes the proof of Claim 4.7. □

Now, for $\xi \in T$, the contribution to the exponent of (14), coming from terms with $|m|_1 > K_t$, is at most

$$\sum_{\ell > K_t} \sum_{m \in \mathbb{Z}_+^{k-1} : |m|_1 = \ell} \left(\binom{\ell}{m} \frac{1}{\ell} M_m(X^{(t)}) \prod_{j=1}^{k-1} O([\xi_j - \xi_k])^{m_j} \right) \leq \sum_{\ell > K_t} k^\ell \cdot \log^{\ell/2}(1/\epsilon) \cdot \delta_t^{\ell - 2} \,. \tag{15}$$

Equation (15) requires a few facts to be justified. First, we use the multinomial identity $\sum_{m \in \mathbb{Z}_+^{k-1} : |m|_1 = \ell} \binom{\ell}{m} = (k-1)^\ell$. We also require the fact that

$$|1 - e(\xi_j - \xi_k)| \leq O([\xi_j - \xi_k]) \,,$$

$\xi \in [0,1]^T$, and recall that $[\xi_j - \xi_k] < Ck(1 + 12 s_j(X))^{-1/2} \log^{1/2}(1/\epsilon)$, for $\xi \in T$. Combining the above with Claim 4.7 gives (15).

Finally, we claim that

$$\sum_{\ell > K_t} k^\ell \cdot \log^{\ell/2}(1/\epsilon) \cdot \delta_t^{\ell - 2} \leq \epsilon (Ck \log(k/\epsilon))^{-2k} \,,$$

where the last inequality holds for both $t = 1, 2$, as can be readily verified from the definition of $\delta_1, \delta_2, K_1, K_2$. Combining with the bounds for smaller $|m|_1$, we get that the absolute difference between $\widehat{X^{(t)}}$ and $\widehat{Y^{(t)}}$ on T is at most $\epsilon (Ck \log(k/\epsilon))^{-2k}$. Therefore, Lemma 4.2 implies that $d_{\mathrm{TV}}(X^{(t)}, Y^{(t)}) \leq \epsilon$. A suitable definition of C in the statement of the lemma to make this $\epsilon/3$ completes the proof of Lemma 4.6. □

Remark 4.8. 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 As a warm-up for our proper cover algorithm, we use the structural lemma of the previous section to show the following upper bound on the cover size of PMDs.

Proposition 4.9. *For all $n, k \in \mathbb{Z}_+$, $k > 2$, and $\epsilon > 0$, there exists an ϵ-cover of the set of (n,k)-PMDs of size $n^{O(k^3)}(1/\epsilon)^{O(k\log(1/\epsilon)/\log\log(1/\epsilon))^{k-1}}$.*

Remark 4.10. We remark that, for the sake of simplicity, we have not optimized the dependence of our cover upper bound on the parameter n. With a slightly more careful argument, one can easily obtain a cover size upper bound $n^{O(k^2)}(1/\epsilon)^{O(k\log(1/\epsilon)/\log\log(1/\epsilon))^{k-1}}$. On the other hand, the asymptotic dependence of our upper bound on the error parameter ϵ is optimal. In Section 4.5, we show a lower bound of $(1/\epsilon)^{\Omega_k(\log(1/\epsilon)/\log\log(1/\epsilon))^{k-1}}$.

Proof of Proposition 4.9. Let X be an arbitrary (n,k)-PMD. We can write X as $\sum_{i=1}^{k} X^i$, where X^i is an i-maximal $(n^{(i)},k)$-PMD, where $\sum_{i=1}^{k} n^{(i)} = n$. By the subadditivity of the total variation distance for independent random variables, it suffices to show that the set of i-maximal (n,k)-PMDs has an ϵ/k-cover of size $n^{O(k^2)}(1/\epsilon)^{O(k\log(k/\epsilon)/\log\log(k/\epsilon))^{k-1}}$.

To establish the aforementioned upper bound on the cover size of i-maximal PMDs, we focus without loss of generality on the case $i = k$. The proof proceeds by an appropriate application of Lemma 4.6 and a counting argument. The idea is fairly simple: for a k-maximal (n,k)-PMD X, we start by approximating the means $s_j(X)$, $1 \le j \le k-1$, within a factor of 2, and then impose an appropriate grid on its low-degree parameter moments.

We associate to such a k-maximal (n,k)-PMD X the following data, and claim that if X and Y are two k-maximal PMDs with the same data, then their total variational distance is at most $\epsilon' \stackrel{\text{def}}{=} \epsilon/k$. An ϵ'-cover for the set of k-maximal (n,k)-PMDs can then be obtained by taking one representative X for each possible setting of the data in question. Let us denote $\delta_1' \stackrel{\text{def}}{=} \delta_1(\epsilon')$, $\delta_2' \stackrel{\text{def}}{=} \delta_2(\epsilon')$, $\gamma' \stackrel{\text{def}}{=} \gamma(\epsilon')$, $K_1' \stackrel{\text{def}}{=} K_1 = 2$, and $K_2' \stackrel{\text{def}}{=} K_2(\epsilon')$, where the functions $\delta_1(\epsilon)$, $\delta_2(\epsilon)$, $\gamma(\epsilon)$, and $K_2(\epsilon)$ are defined in the statement of Lemma 4.6.

In particular, for any X, we partition the coordinates of $[n]$ into the sets $A_1 = \overline{E(\delta_1', X)}$, $A_2 = \overline{E(\delta_2', X)} \setminus A_1$, and $A_3 = E(\delta_2', X)$. We use these subsets to define $X^{(1)}$, $X^{(2)}$ and $X^{(3)}$ on n_1, n_2, n_3 k-CRVs respectively.

Now, to X we associate the following data:

- n_1, n_2, n_3.
- The nearest integer to $\log_2(s_j(X)+1)$ for each j, $1 \le j \le k-1$.
- The nearest integer multiple of $\gamma'/(2k)^{|m|_1}$ to each of the $M_m(X^{(1)})$ for $|m|_1 \le 2$.
- The nearest integer multiple of $\gamma'/(2k)^{|m|_1}$ to $M_m(X^{(2)})$ for $|m|_1 \le K_2'$.
- Rounding of each of the $p_{i,j}$ for $i \in A_3$ to the nearest integer multiple of $\epsilon'/(kn_3)$.

First, note that if X and Y are k-maximal (n,k)-PMDs with the same associated data, then they are partitioned as $X = X^{(1)} + X^{(2)} + X^{(3)}$, $Y = Y^{(1)} + Y^{(2)} + Y^{(3)}$, where $X^{(t)}, Y^{(t)}$, $t \le 2$, have no δ_t'-exceptional variables and have the same number of component k-CRVs. Furthermore, $1 + s_j(X)$ and $1 + s_j(Y)$ differ by at most a factor of 2 for each j. We also must have that $|M_m(X^{(t)}) - M_m(Y^{(t)})|(2k)^{|m|_1} \le \gamma'$ for $|m|_1 \le K_t'$, and there is a bijection between the variables in $X^{(3)}$ and those in $Y^{(3)}$ so that corresponding variables differ by at most $\epsilon'/(kn_3)$ in each parameter (and, thus, differ by at most ϵ'/n_3 in total variation distance). Lemma 4.6 implies that if X and

Y have the same data, then $d_{\mathrm{TV}}(X,Y) \leq \epsilon'$. Hence, this does provide an ϵ'-cover for the set of k-maximal (n,k)-PMDs.

We are left to prove that this cover is of the appropriate size. To do that, we need to prove a bound on the number of possible values that can be taken by the above data. We have at most n choices for each n_i, and $O(\log(n))$ choices for each of the k rounded values of $\log_2(s_j(X)+1)$ (since each is an integer between 0 and $\log_2(n)+1$). $X^{(1)}$ has $O(k^2)$ parameter moments with $|m|_1 \leq 2$, and there are at most $O(kn/\gamma')$ options for each of them (since each parameter moment is at most n). There are $O((k+K_2')^{k-1})$ parameter moments of $X^{(2)}$ that need to be considered. By Claim 4.7, each such parameter moment has magnitude at most $O(k/{\delta_1'}^2)$, and, by our aforementioned rounding, needs to be evaluated to additive accuracy at worst $\gamma'/(2k)^{K_2'}$. Finally, note that $n_3 = |A_3| \leq k/{\delta_2'}^2$, since the coordinates of A_3 are δ_2'-exceptional under X. Each of the corresponding $O(k^2/{\delta_2'}^2)$ parameters $p_{i,j}$ for $i \in A_3$ need to be approximated to precision $\epsilon'/(kn_3)$. We remark that the number of such parameters is less than $O(k\log(1/\epsilon')/\log\log(1/\epsilon'))^{k-1}$, since $k \geq 3$. Putting this together, we obtain that the number of possible values for this data is at most $n^{O(k^2)}(1/\epsilon')^{O(k\log(1/\epsilon')/\log\log(1/\epsilon'))^{k-1}}$. This completes the proof of Proposition 4.9. □

The proof of Proposition 4.9 can be made algorithmic using Dynamic Programming, yielding an efficient construction of a proper ϵ-cover for the set of all (n,k)-PMDs.

Theorem 4.11. *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 $\delta = \epsilon/n$, by the subadditivity of the total variation distance for independent random variables, 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 $O(1/\delta)^k$. This gives the following corollary:

Corollary 4.12. *There exists an algorithm that, on input $n,k \in \mathbb{Z}_+$, $k>2$, and $\epsilon>0$, computes a proper ϵ-cover for the set $\mathcal{M}_{n,k}$ and runs in time $n^{O(k^3)} \cdot (k/\epsilon)^{O(k^3\log(k/\epsilon)/\log\log(k/\epsilon))^{k-1}}$.*

Proof of Theorem 4.11. 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, along the lines of Proposition 4.9, and using dynamic programming.

For the sake of readability, we start by introducing the notation that is used throughout this proof. We use X to denote a generic (n,k)-PMD, and X^i, $1 \leq i \leq k$, to denote its i-maximal PMD components. For an (n,k)-PMD and a vector $m = (m_1,\ldots,m_k) \in \mathbb{Z}_+^k$, we denote its m^{th} parameter moment by $M_m(X) = \sum_{\ell=1}^n \sum_{j=1}^k p_{\ell,j}^{m_j}$. Throughout this proof, we will only consider parameter moments of i-maximal PMDs, in which case the vector m of interest will by construction satisfy $m_i = 0$, i.e., $m = (m_1,\ldots,m_{i-1},0,m_{i+1},\ldots,m_k)$.

In the first step of our algorithm, we guess approximations to the quantities $1+s_j(X^i)$ to within a factor of 2, where X^i is intended to be the i-maximal PMD component of our final PMD X. We represent these guesses in the form of a matrix $G = (G_{i,j})_{1\leq i\neq j\leq k}$. Specifically, we take $G_{i,j} = (2^{a_{i,j}}+3)/4$ for each integer $a_{i,j} \geq 0$, where each $a_{i,j}$ is bounded from above by $O(\log n)$. For each fixed guess G, we proceed as follows: For $h \in [n]$, we denote by $\mathcal{S}_h$ the set of all (h,k)-PMDs of the form $\sum_{\ell=1}^h X_\ell$, where $X_\ell \in S_\ell$. For each $h \in [n]$, we compute the set of all possible (distinct) data $D_G(X)$, where $X \in \mathcal{S}_h$. The data $D_G(X)$ consists of the following:

- The number of i-maximal k-CRVs of X, for each i, $1 \leq i \leq k$.
- Letting X^i denote the i-maximal PMD component of X, we partition the k-CRV components of X^i into three sets based on whether or not they are δ_1'-exceptional or δ_2'-exceptional *with respect to our guess matrix* G for $1 + s_j(X^i)$. Formally, we have the following definition:

  Definition 4.13. Let X^i be an i-maximal (h_i, k)-PMD with $X^i = \sum_{\ell \in A^i} X_\ell$ and $0 < \delta \leq 1$. We say that a particular component k-CRV X_ℓ, $\ell \in A^i$, is δ-exceptional with respect to $G = (G_{i,j})$, if there exists a coordinate $j \neq i$, $1 \leq j \leq k$, such that $p_{\ell,j} \geq \delta \cdot \sqrt{G_{i,j}}$. We will denote by $E(\delta, G) \subseteq A^i$ the set of δ-exceptional coordinates of X^i.

  With this notation, we partition A^i into the following three sets: $A_1^i = \overline{E(\delta_1', G)}$, $A_2^i = \overline{E(\delta_2', G)} \setminus A_1^i$, and $A_3^i = E(\delta_2', G)$. For each i, $1 \leq i \leq k$, we store the following information:

  - $n_1^i = |A_1^i|$, $n_2^i = |A_2^i|$, and $n_3^i = |A_3^i|$.
  - Approximations $\widetilde{s}_{j,i}$ of the quantities $s_j(X^i)$, for each $j \neq i$, $1 \leq j \leq k$ to within an additive error of $(h/4n)$.
  - Approximations of the parameter moments $M_m\left((X^i)^{(1)}\right)$, for all $m = (m_1, \ldots, m_k) \in \mathbb{Z}_+^k$ with $m_i = 0$ and $|m|_1 \leq 2$, to within an additive $\left(\gamma'/(2k)^{|m|_1}\right) \cdot (n_1^i/n)$.
  - Approximations of the parameter moments $M_m\left((X^i)^{(2)}\right)$, for all $m = (m_1, \ldots, m_k) \in \mathbb{Z}_+^k$ with $m_i = 0$ and $|m|_1 \leq K_2'$ to within an additive $\left(\gamma'/(2k)^{|m|_1}\right) \cdot (n_2^i \cdot {\delta_1'}^2/2k)$.
  - Rounding of each of the parameters $p_{\ell,j}$, for each k-CRV X_ℓ, $\ell \in A_3^i$, to the nearest integer multiple of $\epsilon' {\delta_2'}^2/2k^2$.

Note that $D_G(X)$ can be stored as a vector of counts and moments. In particular, for the data associated with k-CRVs in A_3^i, $1 \leq i \leq k$, we can store a vector of counts of the possible roundings of the parameters using a sparse representation.

We emphasize that our aforementioned approximate description needs to satisfy the following property: for independent PMDs X and Y, we have that $D_G(X + Y) = D_G(X) + D_G(Y)$. This property is crucial, as it allows us to store only one PMD as a representative for each distinct data vector. This follows from the fact that, if the property is satisfied, then $D_G(X + Y)$ only depends on the data associated with X and Y.

To ensure this property is satisfied, for a PMD $X = \sum_{\ell=1}^n X_\ell$, where X_ℓ is a k-CRV, we define $D_G(X) = \sum_{\ell=1}^n D_G(X_\ell)$. We now need to define $D_G(W)$ for a k-CRV W. For $D_G(W)$, we store the following information:

- The value of i for which W is i-maximal.
- Whether or not W is δ_1'-exceptional and δ_2'-exceptional with respect to G.
- $s_j(W) = \Pr[W = j]$ rounded down to a multiple of $1/4n$, for each $j \neq i$, $1 \leq j \leq k$.
- If W is not δ_1'-exceptional with respect to G, the nearest integer multiple of $\gamma'/(n(2k)^{|m|_1})$ to $M_m(W)$ for each $m \in \mathbb{Z}_+^k$, with $m_i = 0$ and $|m|_1 \leq 2$.
- If W is δ_1'-exceptional but not δ_2'-exceptional with respect to G, the nearest integer multiple of $\left(\gamma'/(2k)^{|m|_1}\right) \cdot ({\delta_1'}^2/2k)$ to $M_m(W)$, for each $m \in \mathbb{Z}_+^k$ with $m_i = 0$ and $|m|_1 \leq K_2'$.
- If W is δ_2'-exceptional with respect to G, we store roundings of each of the probabilities $\Pr[W = j]$ to the nearest integer multiple of $\epsilon' {\delta_2'}^2/2k$.

Given the above detailed description, we are ready to describe our dynamic programming based algorithm. Recall that for each h, $1 \le h \le n$, we compute sets of all possible (distinct) data $D_G(X)$, where $X \in \mathcal{S}_h$. We do the computation by a dynamic program that works as follows:

- At the beginning of step h, we have a set $\mathcal{D}_{h-1}$ of all possibilities of $D_G(X)$ for PMDs of the form $X = \sum_{\ell=1}^{h-1} X_\ell$, where $X_\ell \in S_\ell$, that have $1 + \widetilde{s}_{j,i}^{D_G(X)} \le 2G_{i,j}$. Moreover, for each $D \in \mathcal{D}_{h-1}$, we have a representative PMD Y_D, given in terms of its k-CRVs, that satisfies $D_G(Y_D) = D$.

- To compute $\mathcal{D}_h$, we proceed as follows: We start by computing $D_G(X_h)$, for each $X_h \in S_h$. Then, we compute a list of possible data for $\mathcal{D}_h$ as follows: For each $D \in \mathcal{D}_{h-1}$ and $X_h \in S_h$, we compute the data $D + D_G(X_h)$, and the k-CRVs of the PMD $Y_D + X_h$ that has this data, since $D_G(Y_D + X_h) = D + D_G(X_h)$. We then remove duplicate data from this list, arbitrarily keeping one PMD that can produce the data. We then remove data D where $1 + \widetilde{s}_{j,i}^{D} \ge 2G_{i,j}$. This gives our set of possible data $\mathcal{D}_h$. Now, we note that $\mathcal{D}_h$ contains all possible data of PMDs of the form $\sum_{\ell=1}^{h} X_\ell$, where each $X_\ell \in S_\ell$, that have $1 + \widetilde{s}_{j,i}^{D_G(X)} \le 2G_{i,j}$, and for each distinct possibility, we have an explicit PMD that has this data.

- After step n, for each $D \in \mathcal{D}_n$, we output the data and the associated explicit PMD, if the following condition is satisfied:

  Condition 4.14. *For each $i, j \in \mathbb{Z}_+$, with $1 \le i \ne j \le k$, it holds (a) $G_{i,j} \le 1 + \max\{0, \widetilde{s}_{j,i}^{D_G(X)} - 1/4\}$ and (b) $1 + \widetilde{s}_{j,i}^{D_G(X)} \le 2G_{i,j}$, where $\widetilde{s}_{j,i}^{D_G(X)}$ is the approximation to $s_j(X^i)$ in $D_G(X)$, and $G_{i,j}$ is the guess for $1 + s_j(X^i)$ in G.*

We claim that the above computation, performed for all values of G, outputs an ϵ-cover of the set $\mathcal{S}$. This is formally established using the following claim:

Claim 4.15. *(i) For any $X, Y \in \mathcal{S}$, if $D_G(X) = D_G(Y)$ and $D_G(X)$ satisfies Condition 4.14, then $d_{\mathrm{TV}}(X, Y) \le \epsilon$.*

(ii) For any $X \in \mathcal{S}$, there exists a G such that $D_G(X)$ satisfies for $i, j \in \mathbb{Z}_+$, with $1 \le i \ne j \le k$, (a) $G_{i,j} \le 1 + \max\{0, \widetilde{s}_{j,i}^{D_G(X)} - 3/4\}$ and (b) $1 + \widetilde{s}_{j,i}^{D_G(X)} \le 2G_{i,j}$, hence also Condition 4.14.

Remark 4.16. Note that Condition (a) in statement (ii) of the claim above is slightly stronger than that in Condition 4.14. This slightly stronger condition will be needed for the anonymous games application in the following section.

Proof. To prove (i), we want to use Lemma 4.6 to show that that for all $i \in [k]$, the i-maximal components of X and Y are close, i.e., that $d_{\mathrm{TV}}(X^i, Y^i) \le \epsilon/k$. To do this, we proceed as follows:

We first show that $\frac{1}{2} \le (1 + s_j(X^i))/(1 + s_j(Y^i)) \le 2$. By the definition of $\widetilde{s}_{j,i}^{D_G(X)}$, for each i-maximal k-CRV $X_\ell \in A^i$, we have $s_j(X_\ell) - \frac{1}{4n} \le \widetilde{s}_{j,i}^{D_G(X)} \le s_j(X_\ell)$. Thus, for $X^i = \sum_{\ell \in A^i} X_\ell$, we have that $s_j(X^i) - (1/4) \le \widetilde{s}_{j,i}^{D_G(X)} \le s_j(X^i)$. Since $s_j(X^i) \ge 0$, we have that $\max\{0, s_j(X^i) - 1/4\} \le \widetilde{s}_{j,i}^{D_G(X)} \le s_j(X^i)$. Combining this with Condition 4.14 yields that

$$G_{i,j} \le 1 + s_j(X^i) \le 2G_{i,j}. \tag{16}$$

Since an identical inequality holds for Y, we have that $\frac{1}{2} \le (1 + s_j(X^i))/(1 + s_j(Y^i)) \le 2$.

We next show that the set of coordinates A_1^i for X does not contain any δ_1' exceptional variables for X^i. For all $\ell \in A_1^i$, since ℓ is not δ_1'-exceptional with respect to G, using (16), we have that

$p_{\ell,j} \le \delta_1' \cdot \sqrt{G_{i,j}} \le \sqrt{1 + s_j(X^i)}$. Similarly, it follows that A_2^i for X does not contain any δ_2'-exceptional variables. The same statements also directly follow for Y.

Now, we obtain bounds on the size of the A_t^i's, $t = 1, 2, 3$. We trivially have $|A_1^i| \le n$. From (16), we have that all variables in A_2^i are δ_1'-exceptional with respect to $G_{i,j}$. If we denote by $E_j \subseteq A_2^i$ the set of $\ell \in A_2^i$ with $p_{\ell,j} \ge \delta_1' \sqrt{G_{i,j}}$, then using (16), we have that

$$s_j(X^i) = \sum_{\ell \in A^i} p_{\ell,j} \ge \sum_{\ell \in A^i} p_{\ell,j}^2 \ge \sum_{\ell \in E_j} p_{\ell,j}^2 \ge \delta_1'^2 |E_j| G_{i,j} \ge \delta_1'^2 |E_j| (1 + s_j(X^i))/2 \ge s_j(X^i) \cdot \delta_1'^2 |E_j|/2.$$

Thus, $|E_j| \le 2/\delta_1'^2$. Since $A_2^i = \bigcup_{j=1}^k E_j$, we have $|A_2^i| \le 2k/\delta_1'^2$. Similarly, we have $|A_3^i| \le 2k/\delta_2'^2$.

For $\ell \in A_1^i$, $D_G(X_\ell)$ contains an approximation to $M_m(X_\ell)$ for each $m \in \mathbb{Z}_+^k$ with $m_i = 0$ and $|m|_1 \le 2$, to within accuracy $\gamma'/(2n(2k)^{|m|_1})$. Since $|A_1^i| \le n$, we have that $D_G(X^i)$ contains an approximation to $M_m\left((X^i)^{(1)}\right)$ to within $\gamma'/(2(2k)^{|m|_1})$. Since a similar bound holds for $(Y^i)^{(1)}$, and $D_G\left((Y^i)^{(1)}\right) = D_G\left((X^i)^{(1)}\right)$, we have that $|M_m\left((X^i)^{(1)}\right) - M_m\left((Y^i)^{(1)}\right)| \le \gamma'/(2k)^{|m|_1}$.

Similarly, for $\ell \in A_2^i$, $D_G(X_\ell)$ contains an approximation to $M_m(X_\ell)$ for each $m \in \mathbb{Z}_+^k$ with $m_i = 0$ and $|m|_1 \le K_2'$ to within accuracy $(1/2) \cdot \left(\gamma'/(2k)^{|m|_1}\right) \cdot (\delta_1'^2/2k)$. Since $|A_2^i| \le 2k/\delta_2'^2$, $D_G(X^i)$ contains an approximation to $M_m\left((X^i)^{(2)}\right)$ to within $\gamma'/(2(2k)^{|m|_1})$. Since a similar bound holds for $(Y^i)^{(2)}$ and $D_G\left((Y^i)^{(2)}\right) = D_G\left((X^i)^{(2)}\right)$, we have that $|M_m\left((X^i)^{(2)}\right) - M_m\left((Y^i)^{(2)}\right)| \le \gamma'/(2k)^{|m|_1}$.

Finally, for $\ell \in A_3^i$, $D_G(X_\ell)$ contains an approximation to $p_{\ell,j}$ for all $j \ne i$ to within $\epsilon' \delta_2'^2/4k^2$. The counts of variables with these approximations are the same in $(X^i)^{(3)}$ and $(Y^i)^{(3)}$. So, there is bijection f from $A_3^i(X)$ to $A_3^i(Y)$ such that an $\ell' = f(\ell)$ has $D_G(X_\ell) = D_G(Y_{\ell'})$. Then, we have that $d_{\mathrm{TV}}(X_\ell, Y_{\ell'}) \le \sum_{j \ne i} \epsilon' \delta_2'^2/2k^2 \le \epsilon' \delta_2'^2/2k \le \epsilon'/|A_3^i|$.

We now have all the necessary conditions to apply Lemma 4.6, yielding that $d_{\mathrm{TV}}(X^i, Y^i) \le \epsilon/k$. By the sub-additivity of total variational distance, we have $d_{\mathrm{TV}}(X, Y) \le \epsilon$, proving statement (i) of the claim.

To prove (ii), it suffices to show that for any i, j there is a $G_{i,j}$ that satisfies the inequalities claimed. Recall that $G_{i,j}$ takes values of the form $(2^a + 3)/4$ for an integer $a \ge 0$. For $a = 0$, $G_{i,j} = 1$ and the inequality $G_{i,j} \le 1 + \max\{0, \widetilde{s}_{j,i}^{D_G(X)} - 3/4\}$ is satisfied for any value of $\widetilde{s}_{j,i}^{D_G(X)}$. When $a \ge 1$, $G_{i,j} > 1$, so the inequality $G_{i,j} \le 1 + \max\{0, \widetilde{s}_{j,i}^{D_G(X)} - 3/4\}$ is only satisfied when $G_{i,j} \le 1 + \widetilde{s}_{j,i}^{D_G(X)} - 3/4$, i.e., when $\widetilde{s}_{j,i}^{D_G(X)} \ge G_{i,j} - 1/4 = (2^a + 2)/4 = (2^{a-1} + 1)/2$. The second inequality in (i) is satisfied when $\widetilde{s}_{j,i}^{D_G(X)} \le 2G_{i,j} - 1 = 2 \cdot (2^a + 1)/4 = (2^a + 1)/2$.

Summarizing, for $a = 0$, we need that $\widetilde{s}_{j,i}^{D_G(X)} \in [0, 1]$, and for $a \ge 1$, we need that $\widetilde{s}_{j,i}^{D_G(X)} \in [(2^{a-1} + 1)/2, (2^a + 1)/2]$. So, there is a $G_{i,j} = (2^{a_{i,j}} + 3)/4$ for which the required inequalities are satisfied. Thus, there is a G for which we get the necessary inequalities for all $1 \le i, j \le k$ with $i \ne j$. This completes the proof of (ii). □

We now bound the running time:

Claim 4.17. *For a generic (n, k)-PMD X, the number of possible values taken by $D_G(X)$ considered is at most $n^{O(k^3)}(k/\epsilon)^{O(k^3 \log(1/\epsilon)/\log\log(1/\epsilon))^{k-1}}$.*

Proof. For a fixed G, we consider the number of possibilities for $D_G(X^i)$ for each $1 \le i \le k$.

For each $j \ne i$, we approximate $s_j(X^i)$ up to an additive $1/(4n)$. Since $0 \le s_j(X_i) \le n$, there are at most $4n^2$ possibilities. For all such j we have $O(n^{2k})$ possibilities.

We approximate the parameter moments of $M_m\left((X^i)^{(1)}\right)$ as an integer multiple of $\gamma'/(n(2k)^{|m|_1})$ for all m with $m_1 \le 2$. For each such m, we have $0 \le M_m\left((X^i)^{(1)}\right) \le n$, so there are at

most $n^2(2k)^{|m|_1}/\gamma' = n^2(k\log(1/\epsilon))^{O(k)}(1/\epsilon)$ possibilities. There are $O(k^2)$ such m, so we have $n^{O(k^2)} \cdot (k\log(1/\epsilon)^{O(k^3)}(1/\epsilon)^{O(k^2)}$ possibilities.

We approximate the parameter moments of $M_m\left((X^i)^{(2)}\right)$ as a multiple of $\left(\gamma'/(2k)^{|m|_1}\right)\cdot(\delta_1'^2/2k)$ for each m with $|m|_1 \le K_2'$. The number of k-CRVs in $(X^i)^{(2)}$ is $|A_2^i| \le 2k/\delta_1'^2$ from the proof of Claim 4.15. So, for each m, we have $0 \le M_m\left((X^i)^{(2)}\right) \le |A_2^i|$, and there are at most $(2k)^{K_2'+2}/(\gamma'\delta_1'^2\delta_2'^2) = k^{O(k+\ln(k/\epsilon)/\ln\ln(k/\epsilon))}\ln(1/\epsilon)^{O(k)}/\epsilon = (k/\epsilon)^{O(k)}$ possibilities. Since there are at most

$$K_2'^{k-1} = O((\ln(k/\epsilon)/\ln\ln(k/\epsilon)+k)^{k-1}$$

such moments, there are $(k/\epsilon)^{O(k\ln(k/\epsilon)/\ln\ln(k/\epsilon)+k^2)^{k-1}}$ possibilities.

We approximate each X_ℓ for $\ell \in A_3^i$ as a k-CRV whose probabilities are multiples of $\epsilon\delta_2'^2/2k^2$. So, there are $(2k^2/(\epsilon\delta_2')^k = (k/\epsilon)^{O(k)}$ possible k-CRVs. Since there may be $|A_3^i| \le 2k/\delta_2'^2 = 2k^2\log^{3/2}(k/\epsilon)$ such k-CRVs, there are $(k/\epsilon)^{O(k^3\log^{3/2}(k/\epsilon))}$ possibilities.

Multiplying these together, for every G, there are at most $n^{O(k^2)}(k/\epsilon)^{O(k\ln(k/\epsilon)/\ln\ln(k/\epsilon)+k^2)^{k-1}}$ possible values of $D_G(X^i)$. Hence, there are at most $n^{O(k^3)}(k/\epsilon)^{O(k^3\ln(k/\epsilon)/\ln\ln(k/\epsilon))^{k-1}}$ possible values of $D_G(X)$ for a given G. Finally, there are $O(\log n)^{k^2}$ possible values of $G_{i,j}$, since $G_{i,j} = (2^{a_{i,j}}+3)/4$, for integers $a_{i,j}$, and we do not need to consider $G_{i,j} > n$. Therefore, the number of possible values of $D_G(X)$ is at most $n^{O(k^3)} \cdot (k/\epsilon)^{O(k^3\ln(k/\epsilon)/\ln\ln(k/\epsilon))^{k-1}}$. □

The runtime of the algorithm is dominated by the runtime of the substep of each step h, where we calculate $D + D_G(X_h)$ for all $D \in \mathcal{D}_{h-1}$ and $X_h \in S_h$. Note that D and $D_G(X_h)$ are vectors with $O(K_2'^k) = O(\log(k/\epsilon)/\log\log(k/\epsilon)+k)^k$ non-zero coordinates. So, the runtime of step h is at most

$$|\mathcal{D}_{h-1}| \cdot |S_h| \cdot O((K_2')^k) = |S_h| \cdot n^{O(k^3)} \cdot (k/\epsilon)^{O(k^3\ln(k/\epsilon)/\ln\ln(k/\epsilon))^{k-1}},$$

by Claim 4.17. The overall runtime of the algorithm is thus $n^{O(k^3)} \cdot (k/\epsilon)^{O(k^3\ln(k/\epsilon)/\ln\ln(k/\epsilon))^{k-1}} \cdot \max_h |S_h|$. This completes the proof of Theorem 4.11. □

4.3 An EPTAS for Nash Equilibria in Anonymous Games In this subsection, we describe our EPTAS for computing Nash equilibria in anonymous games:

Theorem 4.18. *There exists an $n^{O(k^3)}\cdot(k/\epsilon)^{O(k^3\log(k/\epsilon)/\log\log(k/\epsilon))^{k-1}}$-time algorithm for computing a (well-supported) ϵ-Nash Equilibrium in an n-player, k-strategy anonymous game.*

This subsection is devoted to the proof of Theorem 4.18.

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 4.11 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 4.11 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 4.11, 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.

Recall that a mixed strategy profile for a k-strategy anonymous game can be represented as a set of k-CRVs, $\{X_i\}_{i\in[k]}$, where the k-CRV X_i describes the mixed strategy for player i. Recall that a mixed strategy profile is an ϵ-approximate Nash equilibrium, if for each player i we have $\mathbb{E}[u^i_{X_i}(X_{-i})] \geq \mathbb{E}[u^i_\ell(X_{-i})] - \epsilon$, for $\ell \in [k]$, where $X_{-i} = \sum_{j\in[n]\setminus\{i\}} X_j$ is the distribution of the sum of other players strategies. A strategy profile is an ϵ-well-supported Nash equilibrium if for each player i, $\mathbb{E}[u^i_{\ell'}(X_{-i})] \geq \mathbb{E}[u^i_\ell(X_{-i})] - \epsilon$ for each $\ell \in [k]$ and $e_{\ell'}$ in the support of X_i. If this holds for one player i, then we call X_i an ϵ(-well-supported) best response to X_{-i}.

Lemma 4.19. *Suppose that X_i is a δ-best response to X_{-i} for player i. Then, if an $n-1$ PMD Y_{-i} has $d_{\mathrm{TV}}(X_{-i}, Y_{-i}) \leq \epsilon$, X_i is a $(\delta+2\epsilon)$-best response to Y_{-i}. If, additionally, a k-CRV Y_i has $\Pr[Y_i = e_j] = 0$ for all j with $\Pr[X_i = e_j] = 0$, then Y_i is a $(\delta+2\epsilon)$-best response to Y_{-i}.*

Proof. Since $u^i_\ell(x) \in [0,1]$ for $\ell \in [k]$ and any x in the support of X_{-i}, we have that $\mathbb{E}[u^i_\ell(X_{-i})] - \mathbb{E}[u^i_\ell(Y_{-i})] \leq d_{\mathrm{TV}}(X_{-i}, Y_{-i}))$. Similarly, we have $\mathbb{E}[u^i_\ell(X_{-i})] - \mathbb{E}[u^i_\ell(Y_{-i})] \leq d_{\mathrm{TV}}(X_{-i}, Y_{-i})$. Thus, for all $e_{\ell'}$ in the support of X_i and all $\ell \in [k]$, we have

$$\mathbb{E}[u^i_{\ell'}(Y_{-i})] \geq \mathbb{E}[u^i_{\ell'}(X_{-i})] - \epsilon \geq \mathbb{E}[u^i_\ell(X_{-i})] - \epsilon - \delta \geq \mathbb{E}[u^i_\ell(Y_{-i})] - 2\epsilon - \delta.$$

That is, X_i is a $(\delta+2\epsilon)$-best response to Y_{-i}. Since the support of Y_i is a subset if the support of X_i, Y_i is also a $(\delta+2\epsilon)$-best response to Y_{-i}. □

We note that by rounding the entries of an actual Nash Equilibrium, there exists an $\epsilon/5$-Nash equilibrium where all the probabilities of all the strategies are integer multiples of $\epsilon/(10kn)$:

Claim 4.20. *There is an $\epsilon/5$-well-supported Nash equilibrium $\{X_i\}$, where the probabilities $\Pr[X_i = e_j]$ are multiples of $\epsilon/(10kn)$, for all $1 \leq i \leq n$ and $1 \leq j \leq k$.*

Proof. By Nash's Theorem, there is a Nash equilibrium $\{Y_i\}$. We construct $\{X_i\}$ from $\{Y_i\}$ as follows: If Y_i is ℓ-maximal, then for every $j \neq \ell$, we set $\Pr[X_i = e_j]$ to be $\Pr[Y_i = e_j]$ rounded down to a multiple of $\epsilon/(10kn)$ and $\Pr[X_i = e_\ell] = 1 - \sum_{j\neq\ell} \Pr[X_j = e_j]$. Now, we have $d_{\mathrm{TV}}(X_i, Y_i) \leq \epsilon/(10n)$ and the support of X_i is a subset of the support of Y_j. By the sub-additivity of total variational distance, for every i we have $d_{\mathrm{TV}}(X_{-i}, Y_{-i}) \leq \epsilon/10$. Since $\{Y_i\}$ is a Nash equilibrium, for all players i, Y_i is a 0-best response to Y_{-i}. By Lemma 4.19, we have that X_i is a $2\epsilon/10$-best response to X_{-i} for all players i. Hence, $\{X_i\}$ is an $\epsilon/5$-well supported Nash equilibrium. □

Let S be the set of all k-CRVs whose probabilities are multiples of $\epsilon/(10kn)$. We will require a modification of the algorithm from Theorem 4.11 (applied with $S_i \stackrel{\text{def}}{=} S$ for all i, and $\epsilon \stackrel{\text{def}}{=} \epsilon/5$), which produces output at both step n and step $n-1$. Specifically, in addition to outputting a subset $V_{G,n} \subseteq \mathcal{D}_{G,n}$ of the data of possible (n,k)-PMDs that satisfy conditions (a) and (b) of Claim‘4.15 (ii), we output the subset $V_{G,n-1} \subseteq \mathcal{D}_{G,n-1}$ of the data of possible $(n-1,k)$-PMDs that satisfy the slightly weaker conditions (a) and (b) of Condition 4.14 .

In more detail, we need the following guarantees about the output of our modified algorithm:

Claim 4.21. *For every PMD $X = \sum_{i=1}^n X_i$ and $X_{-j} = \sum_{i\in[n]\setminus j} X_i$, for some $1 \leq j \leq n$, and any $X_i \in S$, for $1 \leq i \leq n$, we have:*

- *There is a guess G, such that $D_G(X) \in V_{G,n}$.*
- *For any G such that $D_G(X) \in V_{G,n}$, we also have $D_G(X) - D_G(X_j) = D_G(X_{-j}) \in V_{G,n-1}$.*
- *If $D_G(X) \in V_{G,n}$, for any PMD Y with $D_G(Y) = D_G(X)$ or $D_G(Y) = D_G(X_{-j})$, we have $d_{\mathrm{TV}}(X,Y) \leq \epsilon/5$ or $d_{\mathrm{TV}}(X_{-j}, Y) \leq \epsilon/5$ respectively.*

Proof. By Claim 4.15 (ii), there is a G such that $D_G(X)$ satisfies conditions (a) and (b) and so $D_G(X) \in V_{G,n}$.

We note that by the correctness of the dynamic program, since X_{-j} is a sum of $n-1$ many k-CRVs in S, we have $D_G(X_{-j}) \in \mathcal{D}_{G,n-1}$. To show that it is in $V_{G,n-1}$, we need to show that all $\widetilde{s}_{h,i}^{D_G(X_{-j})}$ satisfy Condition 4.14, for all $1 \le i, h \le k$ and $h \neq i$. We know that $\widetilde{s}_{h,i}^{D_G(X)}$ satisfies the stronger conditions (a) and (b) of Claim 4.15 (ii). All we need to show is that

$$\widetilde{s}_{h,i}^{D_G(X)} - 1/2 \le \widetilde{s}_{h,i}^{D_G(X_{-j})} \le \widetilde{s}_{h,i}^{D_G(X)}.$$

This condition is trivial unless X_j is i-maximal. If it is, we note that $\Pr[X_j = e_h] \le \Pr[X_j = e_i]$, and so $\widetilde{s}_{j,i}^{D_G(X_j)} \le \Pr[X_j = e_h] \le 1/2$. Thus, $D_G(X_{-j}) = D_G(X) - D_G(X_j) \in V_{G,n-1}$.

We now have that both $D_G(X)$ and $D_G(X_{-j})$ satisfy Condition 4.14. Therefore, Claim 4.15 (i) yields the third claim. □

We note that we can calculate the expected utilities efficiently to sufficient precision:

Claim 4.22. *Given an anonymous game $(n, k, \{u_\ell^i\}_{i\in[n],\ell\in[k]})$ with each utility given to within an additive $\epsilon/2$ using $O(\log(1/\epsilon))$ bits, and given a PMD X in terms of its constituent k-CRVs X_i, we can approximate the expected utility $\mathbb{E}[u_\ell^i(\sum_{j\neq i} X_i)]$ for any player i and pure strategy ℓ to within ϵ in time $O(n^{k+1} \cdot k\log(n) \cdot \mathrm{polylog}(1/\epsilon))$.*

Proof. We can compute the probability mass function of $X_{-i} = \sum_{j\neq i} X_j$ by using the FFT on $[n]^k$. We calculate the DFT of each X_i, $\widehat{X}_i$, calculate the DFT of X_{-i}, $\widehat{X_{-i}}(\xi) = \prod_{j\neq i} \widehat{X}_j$, and finally compute the inverse DFT. To do this within $\epsilon/2$ total variational error needs time $O(n^{k+1} \cdot k\log(n)\mathrm{polylog}(1/\epsilon))$, since we need to use the FFT algorithm $n+1$ times. We then use this approximate pmf to compute the expectation $\mathbb{E}[u_\ell^i(X_{-i})] = \sum_x u_\ell^i(x)X_i(x)$. This takes time $O(n^k)$ and gives error ϵ. □

Henceforth, we will assume that we can compute these expectations exactly, but it should be clear that computing them to within a suitably small $O(\epsilon)$ error suffices.

Proof of Theorem 4.18. We use the modified dynamic programming algorithm given above to produce an $\epsilon/5$-cover with explicit sets $V_{G,n}$, $V_{G,n-1}$ of data and PMDs which produce each output data.

Then, for each G and for each $D \in V_{G,n}$, we try to construct an ϵ-Nash equilibrium whose associated PMD X has $D_G(X) = D$. Firstly, for each player i we compute a set $S_i \subseteq S$ of best responses to X. To do this, we check each $X_i \in S$ individually. We first check if $D_G(X) - D_G(X_i) \in V_{G,n-1}$. If it is not, then Claim 4.21 implies that there is no set of strategies for the other players $X_j \in S$, for $j \neq i$, such that $D_G(\sum_{i=1}^n X_i) = D$. In this case, we do not put this $X_i \in S_i$. If we do have $D_{-i} := D - D_G(X_i) \in V_{G,n-1}$, then we recall that the algorithm gives us an explicit $Y_{D_{-i}}$ such that $D(Y_{D_{-i}}) = D_{-i}$. Now, we calculate the expected utilities $\mathbb{E}[u_\ell^i(Y_{D_{-i}})]$ for each $1 \le \ell \le k$. If X_i is a $3\epsilon/5$-best response to $Y_{D_{-i}}$, then we add it to S_i.

When we have calculated the set of best responses S_i for each player, we use the algorithm from Theorem 4.11 with these S_i's and this guess G. If the set of data it outputs contains D, then we output the explicit PMD $X := Y_D$ that does so in terms of its constituent CRVs $X = \sum_{i=1}^n X_i$ and terminate.

To prove correctness, we first show that $\{X_i\}$ is an ϵ-Nash equilibrium, and second that that the algorithm always produces an output. We need to show that X_i is an ϵ-best response to

$X_{-i} = \sum_{j \in [n] \setminus i} X_j$. When we put X_i in S_i, we checked that X_i was a $3\epsilon/5$-best response to $Y_{D_{-i}}$, where $D_{-i} = D - D_G(X_i)$. But note that

$$D_G(Y_{D_{-i}}) = D - D_G(X_i) = D_G(X) - D_G(X_i) = D_G(X_{-i}).$$

Since $D \in V_{G,n}$ and $D_G(Y_{D_{-i}}) = D_G(X_{-i})$, Claim 4.21 yields that $d_{\mathrm{TV}}(X_{-i}, Y_{D_{-i}}) \leq \epsilon/5$. So, by Lemma 4.19, we indeed have that X_i is an ϵ-best response to X_{-i}. Since this holds for all X_i, X is an ϵ-Nash equilibrium.

By Claim 4.20, there exists an $\epsilon/5$-Nash equilibrium $\{X_i'\}$, with each $X_i' \in S$. By Claim 4.21, we have that for $X' = \sum_{i=1}^n X_i'$, there is a guess G with $D_G(X') \in V_{G,n}$. So, if the algorithm does not terminate successfully first, it eventually considers G and $D := D_G(X')$. We next show that that the algorithm puts X_i' in S_i. For each $1 \leq i \leq n$, $X'_{-i} = \sum_{j \in [n] \setminus i} X_j'$ has $D_G(X'_{-j}) \in V_{G,n-1}$ by Claim 4.21, since $D_G(X') \in V_{G,n}$. So, $D_{-i} = D - D_G(X_i') = D_G(X') - D_G(X_i') = D_G(X'_{-i})$, and we have $D_{-i} \in V_{G,n-1}$. Hence, the algorithm will put X_i' in S_i if X_i' is an $4\epsilon/5$-best response to $Y_{D_{-i}}$. By Claim 4.21, since $D_G(X) \in V_{G,n}$ and $D_G(X_{-i}) = D_G(Y_{D_{-i}})$, we have $d_{\mathrm{TV}}(Y_{D_{-i}}, X_{-i}) \leq \epsilon/5$. Since $\{X_i'\}$ is an $\epsilon/5$-Nash equilibrium, X'_{-i} is an $\epsilon/5$-best response to X'_{-i}. Since $d_{\mathrm{TV}}(Y_{D_{-i}}, X_{-i}) \leq \epsilon/5$, by Lemma 4.19, this implies that X_i' is a $3\epsilon/5$-best response to $Y_{D_{-i}}$. Thus, the algorithm puts X_i' in S_i. Since each X_i' satisfies $X_i' \in S_i$, by Theorem 4.11, the algorithm from that theorem outputs a set of data that includes $D_G(X') = D$. Therefore, if the algorithm does not terminate successfully first, when it considers G and D, it will produce an output. This completes the proof of Theorem 4.18. □

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

Definition 4.23. The threat point of an anonymous game $(n, k, \{u_\ell^i\}_{i \in [n], \ell \in [k]})$ is the vector θ with

$$\theta_i = \min_{X_{-i} \in \mathcal{M}_{n-1,k}} \max_{1 \leq j \leq k} \mathbb{E}[u_j^i(X_{-i})].$$

Intuitively, If all other players cooperate to try and punish player i, then they can force her expected utility to be θ_i but no lower, so long as player i is trying to maximize it. This notion has applications in finding Nash equilibria in repeated anonymous games.

Corollary 4.24. *Given an anonymous game $(n, k, \{u_\ell^i\}_{i \in [n], \ell \in [k]})$ with $k > 2$, we can compute a $\tilde{\theta}$ with $\|\theta - \tilde{\theta}\|_\infty \leq \epsilon$ in time $n^{O(k^3)} \cdot (k/\epsilon)^{O(k^3 \log(k/\epsilon)/\log\log(k/\epsilon))^{k-1}}$. Additionally, for each player i, we obtain strategies $X_{i,j}$ for all other players $j \neq i$ such that $\max_{1 \leq \ell \leq k} \mathbb{E}[u_\ell^i(\sum_{j \neq i} X_{i,j})] \leq \theta_i + \epsilon$.*

Proof. Using the dynamic programming algorithm of Theorem 4.11, we can construct an ϵ-cover $\mathcal{C}$ of $\mathcal{M}_{n-1,k}$. For each player i, we then compute $\tilde{\theta}_i = \min_{X_{-i} \in \mathcal{C}} \max_{1 \leq j \leq k} \mathbb{E}[u_j^i(X_{-i})]$ by brute force. Additionally, we return the k-CRVs $X_{i,j}$ that the algorithm gives us as the explicit parameters of the PMD X_{-i} which achieves this minimum, i.e., with $\tilde{\theta}_i = \max_{1 \leq j \leq k} \mathbb{E}[u_j^i(X_{-i})]$. The running time of this algorithm is dominated by the dynamic programming step.

We now show correctness. Let $Y_{-i} \in \mathcal{M}_{n-1,k}$ be such that $\theta_i = \max_{1 \leq j \leq k} \mathbb{E}[u_j^i(X_{-i})]$. Then, there exists a $Y'_{-i} \in \mathcal{C}$ with $d_{\mathrm{TV}}(Y_{-i}, Y'_{-i}) \leq \epsilon$, and so we have $|\mathbb{E}[u_j^i(Y_{-i})] - \mathbb{E}[u_j^i(Y'_{-i})]| \leq \epsilon$. Therefore,

$$\theta_i = \max_{1 \leq j \leq k} \mathbb{E}[u_j^i(Y_{-i})] \geq \max_{1 \leq j \leq k} \mathbb{E}[u_j^i(Y'_{-i})] - \epsilon \geq \tilde{\theta}_i - \epsilon\,.$$

Similarly, there is an $X_{-i} \in \mathcal{C}$ with $\tilde{\theta}_i = \max_{1 \leq j \leq k} \mathbb{E}[u_j^i(X_{-i})] \leq \theta_i$. And so we have $|\tilde{\theta}_i - \theta_i| \leq \epsilon$, as required. Additionally, for the $\sum_{j \neq i} X_{i,j} = X_{-i}$ we have $\max_{1 \leq \ell \leq k} \mathbb{E}[u_\ell^i(\sum_{j \neq i} X_{i,j})] = \tilde{\theta}_i \leq \theta_i + \epsilon$. □

4.4 Every PMD is close to a PMD with few distinct parameters In this section, we prove our structural result that states that any PMD is close to another PMD which is the sum of k-CRVs with a small number of distinct parameters.

Theorem 4.25. *Let $n, k \in \mathbb{Z}_+$, $k > 2$, and $\epsilon > 0$. For any (n,k)-PMD X, there is an (n,k)-PMD Y such that $d_{\mathrm{TV}}(X,Y) \leq \epsilon$ satisfying the following property: We can write $Y = \sum_{i=1}^{n} Y_i$ where each k-CRV Y_i is distributed as one of*

$$O\left((\log(k/\epsilon)/(\log\log(k/\epsilon)) + k)\right)^k$$

distinct k-CRVs.

The main geometric tool used to prove this is the following result from [GRW15]:

Lemma 4.26 (Theorem 14 from [GRW15])**.** *Let $f(x)$ be a multivariate polynomial with variables $x_{i,j}$, for $1 \leq i \leq n$ and $1 \leq j \leq k$, which is symmetric up to permutations of the i's, i.e., such that for any permutation $\sigma \in \mathbb{S}_n$, we have that, for $(x_\sigma)_{i,j} := x_{\sigma(i),j}$, for all $1 \leq i \leq n$ and $1 \leq j \leq k$, $f(x_\sigma) = f(x)$. Let $w \in \mathbb{Z}^k_{>0}$. Suppose that f has weighted w degree at most d, i.e., each monomial $\prod_{i,j} x_{i,j}^{a_{i,j}}$ has $\sum_{i,j} w_j a_{i,j} \leq d$. Suppose that the minimum of $f(x)$ is attained by some $x' \in \mathbb{R}^n$, i.e., that $f(x') = \min_{x \in \mathbb{R}^{n\times k}} f(x)$. Then, there is a point x^* with $f(x^*) = \min_{x\in\mathbb{R}^{n\times k}} f(x)$, such that the number of distinct $y \in \mathbb{R}^k$ of the form $y_j = x^*_{i,j}$, for some i, is at most $\prod_{j=1}^{k} \left\lfloor \frac{d}{w_j} \right\rfloor$.*

Proof of Theorem 4.25. Firstly we're going to divide our PMD into i-maximal PMDs. We assume wlog that X is k-maximal below.

We divide this PMD X into component PMDs $X^{(1)}$, $X^{(2)}$, $X^{(3)}$ according to whether these are δ_1 and δ_2, as in the proof of Proposition 4.9. We want to show that there exists a $Y^{(1)}$,$Y^{(2)}$ such that $X^{(1)}$ and $Y^{(1)}$ agree on the first 2 moments, $X^{(2)}$ and $Y^{(2)}$ agree on the first K_2 moments, but each has few distinct CRVs. Then $Y = Y^{(1)} + Y^{(2)} + X^{(3)}$ is close to X by Lemma 4.6 (because the first moments agree, i.e., we have $s_j(X) = s_j(Y)$).

We are going to use Lemma 4.26 to show that we can satisfy some polynomial equations $p_l(x) = 0$ by setting f to be a sum of squares $f(x) = \sum_l p_l(x)^2$. Then if the polynomial equations have a simultaneous solution at x, f attains its minimum of 0 at x. Some of these p_l's are going to be symmetric in terms of i. For the rest, we are going to have identical equations that hold for each individual i, so f overall will be symmetric.

We have $X^{(t)}$ for $t = 1, 2$, and we want to construct a $Y^{(t)}$ with few distinct k-CRVs. That is, we want to find $p_{i,j}$, the probability that $Y_i = p_{i,j}$, for $1 \leq i \leq n$, $1 \leq j \leq k$. These $p_{i,j}$'s have to satisfy certain inequalities to ensure each $Y_i^{(t)}$ is a non-δ_t exceptional k-CRV and $p_{i,1} \leq p_{i,2} \leq \ldots \leq p_{i,k}$. To do this, we will need to introduce variables whose square is the slack in each of these inequalities.

The free variables of these equations will be $p_{i,1}, \ldots, p_{i,k}, x_{i,1}, \ldots, x_{i,3k}$. The equations we consider are as follows:

The following two equations mean that $Y_i^{(t)}$ is a k-CRV with the necessary properties: For each $1 \leq i \leq n$ and $1 \leq j \leq k-1$,

$$p_{i,j} = x_{i,j}^2 \tag{17}$$

$$p_{i,j} + x_{i,j+k}^2 = p_{i,k} \tag{18}$$

and

$$p_{i,j} + x_{i,j+2k}^2 = \delta_t \sqrt{1 + s_j(X)}\,. \tag{19}$$

For each $1 \leq i \leq n$,

$$\sum_{j=1}^{k} p_{i,j} = 1 \ . \tag{20}$$

We need an equation that the m^{th} moment of $Y^{(t)}$ is identical to the m^{th} moment of $X^{(t)}$, i.e.,

$$\left(\sum_i \prod_j p_{i,j}^{m_j} \right) - M_m(X^{(t)}) = 0 \ , \tag{21}$$

for each moment m with $|m|_1 \leq K_t$.

If these equations have a solution for real $p_{i,j}$'s and $x_{i,j}$'s, then the $p_{i,j}$'s satisfy all the inequalities we need. We square all these expressions and sum them giving f. Note that the slack variables $x_{i,j}$ only appear in monomials of degree 4 in f. We set the weights w_j of the $p_{i,j}$ to be 1 and the weights of the $x_{i,j}$ to be $K_t/2$. Then, f has w degree $2K_t$: (21) has degree K_t in terms of $p_{i,j}$, so when we square it to put it in f, it has degree $2K_t$. So we have that, for $d = 2K_t$, $\prod_{j=1}^{k} \left\lfloor \frac{d}{w_j} \right\rfloor = (2K_t)^k 4^{3k} = O(K_t)^k$ Now f is symmetric in terms of the n different values of i, so we can apply Lemma 4.26, which yields that there is a minimum with $O(K_t)^k$ distinct $(k+1)$-vectors provided that there is any minimum.

However, note that if we set $p'_{i,j} = \Pr[X_i = e_j]$ and define the $x'_{i,j}$ appropriately, we obtain an x' such that $f(x') = 0$. Since f is a sum of squares $f(x) \geq 0$. So, there is an x^* with $f(x^*) = 0$, but such that x^* has $O(K_t)^k$ distinct $4k$-vectors $(p^*_{i,1}, \ldots, p^*_{i,k}, x^*_{i,1}, \ldots, x^*_{i,k})$.

Using the $p^*_{i,j}$'s in this solution, we have a $Y^{(t)}$ with $O(K_t)^k$ distinct CRVs. So, the Y which is $O(\epsilon)$ close to X has $(O(K_1)^k + O(K_2)^k + k(\log 1/\epsilon)^2)$ distinct k-CRVs. Overall, we have that any PMD is $O(k\epsilon)$-close to one with

$$k \cdot O(K_2)^k = O\left((\log(1/\epsilon)/(\log\log(1/\epsilon)) + k) \right)^k$$

distinct constituent k-CRVs. Thus, every PMD is ϵ-close to one with $k \cdot O\left((\log(k/\epsilon)/(\log\log(k/\epsilon)) + k) \right)^k$ distinct constituent k-CRVs. This completes the proof. □

4.5 Cover Size Lower Bound for PMDs In this subsection, we prove our lower bound on the cover size of PMDs, which is restated below:

Theorem 4.27. *(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 4.27 will follow from the following theorem:

Theorem 4.28. *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 $|\mathcal{S}| \geq (1/\epsilon)^{\Omega((1/k)\cdot\log(1/\epsilon)/\log\log(1/\epsilon))^{k-1}}$.*

The proof of Theorem 4.28 is quite elaborate and is postponed to the following subsection. We now show how Theorem 1.5 follows from it. Let $n_0 \stackrel{\text{def}}{=} \Theta((1/k) \cdot \log(1/\epsilon)/\log\log(1/\epsilon))^{k-1}$. By Theorem 4.28, there exists a set $\mathcal{S}_{n_0}$ of size $(1/\epsilon)^{\Omega(n_0)}$ consisting of (n_0,k)-PMDs that are ϵ-far from each other.

We construct $(n/n_0)^{\Omega(k)}$ appropriate "shifts" of the set $\mathcal{S}_{n_0}$, by selecting appropriate sets of $n - n_0$ deterministic component k-CRVs. These sets shift the mean vector of the corresponding

PMD, while the remaining n_0 components form an embedding of the set $\mathcal{S}_{n_0}$. We remark that the PMDs corresponding to different shifts have disjoint supports. Therefore, any ϵ-cover must contain disjoint ϵ-covers for each shift, which is isomorphic to $\mathcal{S}_{n_0}$. Therefore, any ϵ-cover must be of size

$$(n/n_0)^{\Omega(k)} \cdot (1/\epsilon)^{\Omega((1/k)\cdot\log(1/\epsilon)/\log\log(1/\epsilon))^{k-1}} = n^{\Omega(k)} \cdot (1/\epsilon)^{\Omega((1/k)\cdot\log(1/\epsilon)/\log\log(1/\epsilon))^{k-1}} ,$$

where the last inequality used the fact that $n_0^k = o((1/\epsilon)^{n_0})$, if the parameter ϵ is sufficiently small as a function of k. This completes the proof. The following subsection is devoted to the proof of Theorem 4.28.

4.5.1 Proof of Theorem 4.28. 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 $1 \le j \le k-1$, 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 integer 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

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

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

$$p_{s,j}^f = \frac{s_j + \delta_{j,1}\epsilon^{3c} f(s)}{\ln^k(1/\epsilon)}, \tag{22}$$

for $1 \le j \le k-1$. (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 $1 \le j \le k-1$, the above definition implies that

$$p_{s,j}^f \le \frac{1}{k} \cdot \frac{1}{\ln^k(1/\epsilon)}. \tag{23}$$

An important observation, that will be used throughout our proof, is that for each k-CRV X_s^f, only the first out of the $k-1$ parameters $p_{s,j}^f$, $1 \le j \le k-1$, depends on the function f. More specifically, the effect of the function f on $p_{s,1}^f$ is a very small perturbation of the numerator. Note that the first summand in the numerator of (22) 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 4.29 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 $\text{poly}(\epsilon)$.

Lemma 4.29. *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_s^f and X_s^g have $p_{s,1}^f \neq p_{s,1}^g$.

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 4.29 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 4.29 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 $\text{poly}(\epsilon)$, 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.

Proof of Lemma 4.29. We begin by approximating the m^{th} parameter moment of X^f. We have that

$$\begin{aligned} M_m(X^f) &= \sum_{s \in [a]^{k-1}} \prod_{j=1}^{k-1} \left(\frac{s_j + \delta_{j,1} \epsilon^{3c} f(s)}{\ln^k(1/\epsilon)} \right)^{m_j} \\ &= \ln^{-k\|m\|_1}(1/\epsilon) \sum_{s \in [a]^{k-1}} (s_1 + \epsilon^{3c} f(s))^{m_1} \prod_{j=2}^{k-1} s_j^{m_j} . \end{aligned}$$

Note that in the expression $(s_1 + \epsilon^{3c} f(s))^{m_1} = \sum_{i=0}^{m_1} \binom{m_1}{i} s_1^{m_1 - i} (\epsilon^{3c} f(s))^i$, the ratio of the $(\epsilon^{3c} f(s))^{i+1}$ term to the $(\epsilon^{3c} f(s))^i$ term is $(m_1 - i)\epsilon^{3c} f(s)/s_1 i \leq a\epsilon^{2c} \leq 1/2$. So, we have

$$\begin{aligned} (s_1 + \epsilon^{3c} f(s))^{m_1} &= \sum_{i=0}^{m_1} \binom{m_1}{i} s_1^{m_1 - i} (\epsilon^{3c} f(s))^i \\ &\leq s_1^{m_1} + m_1 s_1^{m_1 - 1} \epsilon^{3c} f(s) + (m_1(m_1 - 1)/2) s_1^{m_1 - 2} (\epsilon^{3c} f(s))^2 \sum_{i=0}^{m_1 - 2} 2^{-i} \\ &\leq s_1^{m_1} + m_1 s_1^{m_1 - 1} \epsilon^{3c} f(s) + a^a \epsilon^{4c} . \end{aligned}$$

We can therefore write

$$\begin{aligned}
M_m(X^f) &= \ln^{-k\|m\|_1}(1/\epsilon) \sum_{s\in[a]^{k-1}} \left(s_1 + \epsilon^{3c} f(s)\right)^{m_1} \prod_{j=2}^{k-1} s_j^{m_j} \\
&\leq \ln^{-k\|m\|_1}(1/\epsilon) \sum_{s\in[a]^{k-1}} \left(s_1^{m_1} + m_1 s_1^{m_1-1}\epsilon^{3c} f(s) + a^a \epsilon^{4c}\right) \left(\prod_{j=2}^{k-1} s_j^{m_j}\right) \\
&\leq \ln^{-k\|m\|_1}(1/\epsilon) \left(\left(\sum_{s\in[a]^{k-1}} \prod_{j=1}^{k-1} s_j^{m_j}\right) + \left(\epsilon^{3c} \sum_{s\in[a]^{k-1}} m_1 f(s) s_1^{m_1-1} \prod_{j=2}^{k-1} s_j^{m_j}\right) + a^{ka}\epsilon^{4c}\right) .
\end{aligned}$$

Note that $a^{ka} = \exp(ak \ln a) \leq \exp(ak \ln\ln 1/\epsilon) \leq \exp(c \ln \epsilon/2) = (1/\epsilon)^{c/2}$, and so finally we have

$$M_m(X^f) \leq \ln^{-k\|m\|_1}(1/\epsilon) \left(\left(\sum_{s\in[a]^{k-1}} \prod_{j=1}^{k-1} s_j^{m_j}\right) + \left(\epsilon^{3c} \sum_{s\in[a]^{k-1}} m_1 f(s) s_1^{m_1-1} \prod_{j=2}^{k-1} s_j^{m_j}\right) + \epsilon^{7c/2}\right) ,$$

and that

$$M_m(X^f) \geq \ln^{-k\|m\|_1}(1/\epsilon) \left(\left(\sum_{s\in[a]^{k-1}} \prod_{j=1}^{k-1} s_j^{m_j}\right) + \left(\epsilon^{3c} \sum_{s\in[a]^{k-1}} m_1 f(s) s_1^{m_1-1} \prod_{j=2}^{k-1} s_j^{m_j}\right)\right) .$$

An analogous formula holds for the parameter moments of X^g and therefore

$$M_m(X^f) - M_m(X^g) = \ln^{-k\|m\|_1}(1/\epsilon) \left(\epsilon^{3c} \sum_{s\in[a]^{k-1}} m_1 s_1^{m_1-1} \prod_{j=2}^{k-1} s_j^{m_j} (f(s) - g(s)) + O(\epsilon^{7c/2})\right) .$$

We need to show that, for at least one value of $m \in \mathbb{Z}_+^{k-1}$, the integer

$$\sum_{s\in[a]^{k-1}} \prod_{j=1}^{k-1} s_j^{m_j-1} (f(s) - g(s))$$

is non-zero, since $\log^{-k\|m\|_1}(1/\epsilon) \cdot \epsilon^{3c} m_1 \prod_{j=1}^{k-1} s_j > 0$ for all s and m.

We observe that these integers are the coordinates of $L(f-g)$, where $L : \mathbb{R}^{[a]^{k-1}} \to \mathbb{R}^{[a]^{k-1}}$ is the linear transformation with

$$L(h)_m \stackrel{\text{def}}{=} \sum_{s\in[a]^{k-1}} \prod_{j=1}^{k-1} s_j^{m_j-1} (h(s)) ,$$

for $m \in [a]^{k-1}$. It should be noted that L is the tensor product of the linear transformations $L_i : \mathbb{R}^m \to \mathbb{R}^m$, with

$$L_i(h)_{m_i} = \sum_{s_i\in[a]} s_i^{m_i-1} h(s) ,$$

for $1 \leq i \leq k-1$. Moreover, each $L_i(h)$ is given by the Vandermonde matrix on the distinct integers $1, 2, \ldots, a$, which is non-singular. Since each L_i is invertible, the tensor product L is also invertible. Therefore, $L(f-g)$ is non-zero. That is, there exists an $m \in [a]^{k-1}$ with $(L(f-g))_m \neq 0$, and so

$$M_m(X^f) - M_m(X^g) = \log^{-k\|m\|_1}(1/\epsilon) \cdot \epsilon^{3c} \cdot m_1 \prod_{j=1}^{k-1} s_j (L(f-g))_m \neq 0 .$$

Since $m_1 \prod_{j=1}^{k-1} s_j(L(f-g))_m$ is an integer, $|m_1 \prod_{j=1}^{k-1} s_j(L(f-g))_m| \geq 1$. So, we get

$$|M_m(X^f) - M_m(X^g)| \geq \ln^{-k\|m\|_1}(1/\epsilon)\epsilon^{3c} .$$

Finally, we note that $\ln^{-k\|m\|_1}(1/\epsilon) = \exp(-k\|m\|_1 \ln\ln 1/\epsilon) \geq \exp(-ka \ln\ln 1/\epsilon) \geq \epsilon^{c/2}$. We therefore conclude that $|M_m(X^f) - M_m(X^g)| \geq \epsilon^{4c}$, as required. □

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 4.30. *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.

For a random variable $W = (W_1, \ldots, W_k)$ taking values in $\mathbb{Z}^k$ and $z = (z_1, \ldots, z_{k-1}) \in \mathbb{C}^{k-1}$, we recall the definition of the probability generating function: $P(W, z) \stackrel{\text{def}}{=} \mathbb{E}\left[\prod_{i=1}^{k-1} z_i^{W_i}\right]$. For a PMD X^f, we have that

$$P(X^f, z) = \mathbb{E}\left[\prod_{s\in[a]^{k-1}} \prod_{i=1}^{k-1} z_i^{X^f_{s,i}}\right] = \prod_{s\in[a]^{k-1}} \left(1 + \sum_{i=1}^{k-1} p^f_{s,i}(z_i - 1)\right) .$$

We start by establishing the following crucial claim:

Claim 4.31. *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 there exists $z^* \in \mathbb{C}^{k-1}$ with $\|z^*\|_\infty \leq 2$ such that $|P(X^f, z^*) - P(X^g, z^*)| \geq \epsilon^{5c}$.*

Before we proceed with the formal proof, we provide an intuitive explanation of the argument. The proof of Claim 4.31 proceeds as follows: We start by expressing $\ln\big(P(X^f, z)\big)$, the logarithm of the PGF of a PMD $X^f \in \mathcal{S}$, as a Taylor series whose coefficients depend on its parameter moments $M_m(X^f)$. We remark that appropriate bounds on the parameter moments of $X^f \in \mathcal{S}$ imply that this series is in fact absolutely convergent in an appropriate region R. Note that, using the aforementioned Taylor expansion, we can express each parameter moment $M_m(X^f)$, $m \in \mathbb{Z}_+^{k-1}$, as a partial derivative of $\ln\big(P(X^f, z)\big)$. Hence, if X^f and X^g are distinct PMDs in $\mathcal{S}$, the difference $|M_m(X^f) - M_m(X^g)|$ can also be expressed as the absolute value of the partial derivative of the difference between the PGFs $\left|\ln\big(P(X^f, z)\big) - \ln\left(P(X^g, z)\right)\right|$. We then use Cauchy's integral formula to express this partial derivative as an integral, which we can further be absolutely bounded from above by the difference $\left|\ln\big(P(X^f, z^*)\big) - \ln\left(P(X^g, z^*)\right)\right|$, for some point $z^* \in R$. Finally, we use the fact that $\ln\big(P(X^f, z)\big)$ is absolutely bounded for all $z \in R$ to complete the proof of the claim.

Proof of Claim 4.31. For $z \in \mathbb{C}^{k-1}$, let $w \in \mathbb{C}^{k-1}$ be defined by $w_i = z_i - 1$, $1 \le i \le k-1$. When $|w_i| \le 5$, for all $1 \le i \le k-1$, we take logarithms and obtain:

$$\begin{aligned}
\ln(P(X^f, z)) &= \sum_{s \in [a]^{k-1}} \ln\left(1 + \sum_{i=1}^{k-1} p_{s,i}^f w_i\right) \\
&= \sum_{s \in [a]^{k-1}} \sum_{\ell=1}^{\infty} \frac{(-1)^{\ell+1}}{\ell} \left(\sum_{i=1}^{k-1} p_{s,i}^f w_i\right)^{\ell} \\
&= \sum_{s \in [a]^{k-1}} \sum_{\ell=1}^{\infty} \frac{(-1)^{\ell+1}}{\ell} \sum_{\|m\|_1 = \ell} \binom{\ell}{m} \prod_{i=1}^{k-1} (p_{s,i}^f)^{m_i} \prod_{i=1}^{k-1} w_i^{m_i} \\
&= \sum_{\substack{a \in \mathbb{Z}_{\ge 0}^{k-1} \\ a \neq \bar{0}}} \frac{(-1)^{\|m\|_1+1}}{\|m\|_1} \binom{\|m\|_1}{m} \left(\prod_{i=1}^{k-1} w_i^{m_i}\right) \sum_{s \in [a]^{k-1}} \left(\prod_{i=1}^{k-1} (p_{s,i}^f)^{m_i}\right) \\
&= \sum_{\substack{a \in \mathbb{Z}_{\ge 0}^{k-1} \\ a \neq \bar{0}}} \frac{(-1)^{\|m\|_1+1}}{\|m\|_1} \binom{\|m\|_1}{m} \left(\prod_{i=1}^{k-1} w_i^{m_i}\right) M_m(X^f).
\end{aligned}$$

We note that

$$|M_m(X^f)| = \left| \sum_{s \in [a]^{k-1}} \prod_{i=1}^{k-1} (p_{s,i}^f)^{m_i} \right| \le a^{k-1} \ln^{-(k-1)\|m\|_1}(1/\epsilon) < (10k)^{-\|m\|_1} ,$$

where the first inequality follows from (23). Therefore, for $\|z\|_\infty \le 4$ and so $\|w\|_\infty \le 5$, we obtain that

$$|\ln(P(X^f, z))| \le \sum_{\ell=1}^{\infty} \frac{1}{\ell} \sum_{\substack{a \in \mathbb{Z}_+^{k-1} \\ \|m\|_1 = \ell}} \binom{\|m\|_1}{m} \|w\|_\infty^\ell (10k)^{-\ell} \le \sum_{\ell=1}^{\infty} (k-1)^\ell 5^\ell (10k)^{-\ell} \le \sum_{\ell=1}^{\infty} 2^{-\ell} = 1 . \quad (24)$$

This suffices to show that all the series above are absolutely convergent when $\|z\|_\infty \le 4$, and thus that their manipulations are valid, and that we get the principal branch of the logarithm. To summarize, for all $z \in \mathbb{C}^{k-1}$ with $\|z\|_\infty \le 4$, and $w \in \mathbb{C}^{k-1}$ with $w_i = z_i - 1$, $i \in [k-1]$, we have:

$$\ln(P(X^f, z)) = \sum_{\substack{a \in \mathbb{Z}_{\ge 0}^{k-1} \\ a \neq \bar{0}}} \frac{(-1)^{\|m\|_1+1}}{\|m\|_1} \binom{\|m\|_1}{m} \left(\prod_{i=1}^{k-1} w_i^{m_i}\right) M_m(X^f) . \quad (25)$$

In particular, we have that the $\prod_{i=1}^{k-1} w_i^{m_i}$ coefficient of $\ln(P(X^f, z))$ is an integer multiple of $M_m(X^f)/\|m\|_1$. This expansion is a Taylor series in the w_i's, so this coefficient is equal to a partial derivative, which we can extract by Cauchy's integral formula. Now, suppose that X^f, X^g are

distinct elements of $\mathcal{S}$. We have that:

$$\begin{aligned}
|M_m(X^f) - M_m(X^g)|/\|m\|_1 &\leq |M_m(X^f) - M_m(X^g)|\binom{\|m\|_1}{m}/\|m\|_1 \\
&= \left(\prod_{i=1}^{k-1} 1/m_i!\right)\left|\frac{\partial^{\|m\|_1}\ln(P(X^f,z)) - \ln(P(X^g,z))}{\partial w_1^{m_1}\cdots\partial w_{k-1}^{m_{k-1}}}\right| \\
&= \left|(1/2\pi i)^{k-1}\oint_\gamma \ldots \oint_\gamma (\ln(P(X^f,z)) - \ln(P(X^g,z)))/\prod_{i=1}^{k-1} w_i^{m_i} dw_1 \ldots dw_{k-1}\right| \\
&\leq \max_{|w_1|=1,\ldots,|w_{k-1}|=1} |\ln(P(X^f,z)) - \ln(P(X^g,z)))/\prod_{i=1}^{k-1} w_i^{m_i}| \\
&= \max_{|w_1|=1,\ldots,|w_{k-1}|=1} |\ln(P(X^f,z)) - \ln(P(X^g,z)))|\ ,
\end{aligned}$$

where the third line above follows from Cauchy's integral formula, and γ is the path round the unit circle.

Now suppose that there exists an $m \in [a]^{k-1}$, i.e., with $\|m\|_1 \leq (k-1)a$, such that it holds $|M_m(X^f) - M_m(X^g)| \geq \epsilon^{4c}$. By the above, this implies that there is some $w^* = (w_1^*, \ldots, w_{k-1}^*)$ with $|w_i^*| = 1$ for all i so that for the corresponding z^*,

$$|\ln(P(X^f,z^*)) - \ln(P(X^g,z^*))| \geq \epsilon^{4c}/\|m\|_1 \geq \epsilon^{4c}/(ka)\ . \tag{26}$$

Note that $\|z^*\|_\infty \leq \|w^*\|_\infty + 1 = 2$. Hence, $z^* \in R$. Applying (24), at this z^*, we have $|\ln(P(X^f,z^*))| \leq 1$ and $|\ln(P(X^g,z^*))| \leq 1$. Therefore, by Equation (26), for this z^* with $\|z^*\|_\infty \leq 2$, we have that

$$|P(X^f,z^*) - P(X^g,z^*)| = \Omega\left(\epsilon^{4c}/(ka)\right) \geq \epsilon^{5c}\ ,$$

where the last inequality follows from our definition of a. This completes the proof of Claim 4.31. □

We are now ready to translate a lower bound on the distance between the PGFs to a lower bound on total variation distance. Namely, we prove the following:

Claim 4.32. *If there exists $z^* \in \mathbb{C}^{k-1}$ with $\|z^*\|_\infty \leq 2$ such that $|P(X^f,z^*) - P(X^g,z^*)| \geq \epsilon^{5c}$, then $d_{\mathrm{TV}}(X^f, X^g) \geq \epsilon$.*

The main idea of the proof of Claim 4.32 is this: By Equation (24), we know that $|\ln(P(X^f,z))| \leq 1$, for all $z \in R$. We use this fact to show that the contribution to the value of the PGF $P(X^f,z^*)$ coming from the subset of the probability space $\{X^f > T\}$ is at most $O(2^{-T})$. On the other hand, the contribution to the difference $|\ln(P(X^f,z^*)) - \ln(P(X^g,z^*))|$ coming from the set $\{X^f \leq T, X^g \leq T\}$ can be easily bounded from above by $2^T \cdot d_{\mathrm{TV}}(X^f, X^g)$. The claim follows by selecting an appropriate value of $T = \Theta(\log(1/\epsilon))$, balancing these two terms.

Proof. First note that exponentiating Equation (24) at $z = (4,4,\ldots,4)$, and using the definition of the PGF we get:

$$\mathbb{E}\left[4^{\sum_{i=1}^{k-1} X_i^f}\right], \mathbb{E}\left[4^{\sum_{i=1}^{k-1} X_i^g}\right] \leq e\ .$$

Therefore, for any z with $\|z\|_\infty \leq 2$ and any $T \in \mathbb{Z}_+$ we have that

$$\left|\sum_{x,|x|_1\geq T}\prod_{i=1}^{k-1} z_i^{x_i}\Pr[X^f = x]\right| \leq (1/2)^T \sum_{x,|x|_1\geq T}\prod_{i=1}^{k-1} 4^{|x|_1}\Pr[X^f = x] \leq e(1/2)^T.$$

A similar bound holds for X^g. By assumption, there exists such a z^* so that

$$\begin{aligned}
\epsilon^{5c} &\leq \left| P(X^f, z^*) - P(X^g, z^*) \right| \\
&= \left| \sum_x \prod_{i=1}^{k-1} (z^*)_i^{x_i} \left(\Pr[X^f = x] - \Pr[X^g = x] \right) \right| \\
&= \left| \sum_{|x|_1 < T} \prod_{i=1}^{k-1} (z^*)_i^{x_i} \left(\Pr[X^f = x] - \Pr[X^g = x] \right) \right| + 2e(1/2)^T \\
&\leq 2^T \sum_{|x|_1 < T} \left| \Pr[X^f = x] - \Pr[X^g = x] \right| + 2e(1/2)^T \\
&\leq 2^T d_{\mathrm{TV}}(X^f, X^g) + 2e/2^T \, .
\end{aligned}$$

Taking $T = \lceil 5c \log_2(1/\epsilon) \rceil$, we get

$$d_{\mathrm{TV}}(X^f, X^g) \geq \Omega(\epsilon^{10c}) \geq \epsilon.$$

This completes the proof Claim 4.32. □

Lemma 4.30 follows by combining Claims 4.31 and 4.32. By putting together Lemmas 4.29 and 4.30, it follows that any two distinct elements of $\mathcal{S}$ are ϵ-separated in total variation distance. This completes the proof of Theorem 4.28, establishing the correctness of our lower bound construction. □

5 A Size–Free Central Limit Theorem for PMDs

In this section, we prove 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 formal statement of our CLT is the following:

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*

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

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.

Proof of Theorem 5.1. We note that unless $\sigma > k^7$ that there is nothing to prove, and thus we will assume this throughout the rest of the proof.

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 be able to 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 start by providing 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 5.5). We further show that the effective support of the distribution X and the Fourier transform of the discrete Gaussian G are similar (see Lemmas 5.9 and 5.10). We then obtain an estimate of the error between the Fourier transforms of X and a Gaussian with the same mean and covariance (Lemma 5.8). 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 5.11). 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 3.3). 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*

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

Proof. From Lemma 3.3, we have that $(X - \mu)^T (k \ln(k/\epsilon)\Sigma + k^2 \ln^2(k/\epsilon) I)^{-1} (X - \mu) = O(1)$ with probability at least $1 - \epsilon/10$.

By our assumptions on Σ, we can write $\Sigma = U^T \mathrm{diag}(\lambda_i) U$, for an orthogonal matrix U with k^{th} column $\mathbf{1}/\sqrt{k}$ and $\lambda_i \geq \sigma$ for $1 \leq i \leq k-1$, and $\lambda_k = 0$.

By our assumptions on ϵ, we have that $\sigma \geq k \ln(1/\epsilon)$, and so for $1 \leq i \leq k-1$, we have $\lambda_i + 1 \geq \frac{1}{2}(\lambda_i + k \ln(1/\epsilon))$.

Since $\mathbf{1}^T (X - \mu) = 0$, we have that $(U^T (X - \mu))_k = 0$, and so we can write

$$\begin{aligned}
(X - \mu) \cdot (\Sigma + I)^{-1} (X - \mu) &= (U^T (X - \mu))^T \mathrm{diag}(1/(\lambda_i + 1)) U^T (X - \mu) \\
&\leq (U^T (X - \mu))^T \mathrm{diag}(2/(\lambda_i + k \ln(1/\epsilon))) U^T (X - \mu) \\
&= 2k \ln(k/\epsilon) \cdot (X - \mu)^T \left(k \ln(k/\epsilon)\Sigma + k^2 \ln^2(k/\epsilon) I \right)^{-1} (X - \mu) \\
&= O(k \ln(k/\epsilon)).
\end{aligned}$$

□

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

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

Proof. Noting that $\ln(k\sigma) = O(\log \sigma)$ since $\sigma > k$, by Lemma 5.2, applied with $\epsilon = 1/\sigma$, it follows that $x \in S$ with probability $1 - 1/\sigma$.

The remainder of the claim is a standard counting argument where we need to bound the number of integer lattice points within a continuous region (in this case, an ellipsoid). We deal with this by way of the standard technique of erecting a unit cube about each of the lattice points and bounding the volume of the union. Note that for $x \in \mathbb{Z}^k$, the cubes $x + (-1/2, 1/2)^k$ each have

volume one and are disjoint. Thus, if we define S' to be the set of y such that there exists an $x \in S$ with $\|y - x\|_\infty < \frac{1}{2}$, then $|S| = \mathrm{Vol}(S')$. For any $y \in S'$, there is an $x \in S$ with:

$$\begin{aligned}
(y-\mu)\cdot(\Sigma+I)^{-1}(y-\mu) &= (x-\mu)\cdot(\Sigma+I)^{-1}(x-\mu) + (y-x)\cdot(\Sigma+I)^{-1}(y-x) + \\
&\quad 2(y-x)\cdot(\Sigma+I)^{-1}(x-\mu) \\
&\leq O((x-\mu)\cdot(\Sigma+I)^{-1}(x-\mu) + (y-x)\cdot(\Sigma+I)^{-1}(y-x)) \\
&\leq O(Ck\log(\sigma) + (y-x)\cdot I(y-x)) \\
&\leq O(Ck\log(\sigma) + k\|y-x\|_\infty^2) \\
&\leq O(Ck\log(\sigma) + k) \\
&= O(Ck\log(\sigma))\,.
\end{aligned}$$

That is, S' is contained in the ellipsoid $(y-\mu)\cdot(\Sigma+I)^{-1}(y-\mu) \leq O(Ck\log(\sigma))$. The corollary follows by bounding the volume of this ellipsoid. We have the following simple claim:

Claim 5.4. *The volume of the ellipsoid $x^T A^{-1} x \leq ck$ for a symmetric $k \times k$ matrix A and $c > 0$ is $\sqrt{\det(A)} \cdot O(c)^{k/2}$.*

Proof. We can factorize $A = U^T \mathrm{diag}(\lambda_i) U^T$ for some orthogonal matrix U. Then, the ellipsoid is the set of x with $\|\mathrm{diag}(1/\sqrt{ck\lambda_i})U^T x\|_2 \leq 1$. The volume of the ellipsoid is

$$\left|\det(\mathrm{diag}(1/\sqrt{ck\lambda_i})U^T)^{-1}\right| V_k = \sqrt{\det(A)}\cdot(ck)^{k/2} V_k,$$

where V_k is the volume of the unit sphere. By standard results, $V_k = \pi^{k/2}/\Gamma(1+k/2) = \Omega(k)^{-k/2}$, using Stirling's approximation

$$\Gamma(1+k/2) = \sqrt{2\pi/(1+k/2)}((1+k/2)/e)^{1+k/2}(1+O(2/(k+2)).$$

Therefore, the volume is $O(\sqrt{\det(A)}\cdot(c)^{k/2})$. □

As a consequence, the volume of the ellipsoid $(y-\mu)\cdot(\Sigma+I)^{-1}(y-\mu) \leq O(Ck\log(\sigma))$ is $\sqrt{\det(\Sigma+I)}\cdot O(C\log\sigma)^{k/2}$. Thus, we conclude that $|S| \leq \mathrm{Vol}(S') \leq \sqrt{\det(\Sigma+I)}\cdot O(\log\sigma)^{k/2}$. This completes the proof of the corollary. □

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. The proof will similarly need Lemma 3.10.

Lemma 5.5. *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) $\int_{[0,1]^k \setminus (T+\mathbb{Z}^k)} |\widehat{X}(\xi)| d\xi \leq 1/(\sigma|S|)$.

Proof. We define $\tilde{\xi}$ to be the projection of ξ onto the plane where the coordinates sum to 0, i.e., $\tilde{\xi} = \xi + \alpha\mathbf{1}$ for some $\alpha \in \mathbb{R}$ and $\tilde{\xi} \cdot \mathbf{1} = 0$. Then, we have that $\xi \cdot \Sigma\xi \geq \sigma|\tilde{\xi}|_2^2$. Hence, for $\xi \in T$, we have that $|\tilde{\xi}|_\infty \leq |\tilde{\xi}|_2 \leq \sqrt{Ck\log(\sigma)/\sigma}$. This implies that for any i, j it holds

$$|\xi_i - \xi_j| = |\tilde{\xi}_i - \tilde{\xi}_j| \leq 2|\tilde{\xi}|_\infty \leq 2\sqrt{Ck\log(\sigma)/\sigma}.$$

This proves (i).

In particular, for any $\xi \in T$, and all i, we have $|\xi_1 - \xi_i| \leq 2\sqrt{Ck\log(\sigma)/\sigma} \leq 2C$. And so if $\xi \in T'$, then $\xi_i \in [-2\sqrt{C}, 1+2\sqrt{C}]$. Thus, for $\xi \in T'$, it holds $\xi \cdot \xi \leq O(C)$. Thus, for $\xi \in T'$, we have $\xi \cdot (\Sigma + I) \cdot \xi \leq O(Ck\log(\sigma))$. By Claim 5.4, we get that $\mathrm{Vol}(T') \leq \det(\Sigma + I)^{-1/2} O(C\log(\sigma))^{k/2}$. This proves (ii).

By Claim 3.9, for every $\xi \in [0,1)^k$, there is an interval I_ξ of length $1 - 1/(k+1)$ such that $\xi' = \xi + b$, for some $b \in \mathbb{Z}^k$, has coordinates in I_ξ. Let T_m be the set of ξ such that there is such a ξ' with

$$2^{m+1}Ck\log(\sigma) \geq \xi' \cdot \Sigma\xi' \geq 2^m Ck\log(\sigma)\ ,$$

and $\xi_i = \xi'_i - \lfloor \xi'_i \rfloor$ for all $1 \leq i \leq k$. Then, for every $\xi \in [0,1]^k$, we either have $\xi \in T + \mathbb{Z}^k$ or else $\xi \in T_m$ for some $m \geq 0$. Hence,

$$\int_{[0,1]^k \setminus (T+\mathbb{Z}^k)} |\widehat{X}(\xi)| d\xi \leq \sum_{m=0}^{\infty} \mathrm{Vol}(T_m) \sup_{\xi \in T_m} |\widehat{X}(\xi)|\ . \tag{27}$$

To bound the RHS above, we need bounds on the volume of each T_m. These can be obtained using a similar argument to (ii) along with some translation.

Claim 5.6. *We have that* $\mathrm{Vol}(T_m) \leq \det(\Sigma + I)^{-1/2} \cdot O(2^{m+1}C\log(\sigma))^{k/2}$.

Proof. Let U_m be the set of ξ such that there is a ξ' with $2^{m+1}Ck\log(\sigma) \geq \xi' \cdot \Sigma\xi'$ and $\xi_i = \xi'_i - \lfloor \xi'_i \rfloor$ for all $1 \leq i \leq k$. Note that $T_m \subseteq U_m$. Let U'_m be the set of ξ' with $\xi_1 \in [0,1]$ and $2^{m+1}Ck\log(\sigma) \geq \xi' \cdot \Sigma\xi'$. Note that for any ξ'' with $2^{m+1}Ck\log(\sigma) \geq \xi'' \cdot \Sigma\xi''$, we have that $\xi'' + \lambda\mathbf{1}$ also satisfies $2^{m+1}Ck\log(\sigma) \geq (\xi'' + \lambda\mathbf{1}) \cdot \Sigma(\xi'' + \lambda\mathbf{1})$ for any $\lambda \in \mathbb{R}$. In particular $\xi' = \xi'' - (\lfloor \xi''_1 \rfloor)\mathbf{1}) \in U'_m$. Note that $\xi''_i - \lfloor \xi''_i \rfloor = \xi'_i$ for all i. So U_m is the set of ξ such that there is a $\xi' \in U'_m$ with $\xi_i = \xi'_i - \lfloor \xi'_i \rfloor$. Then, $\mathrm{Vol}(U_m) \leq \mathrm{Vol}(U'_m)$ since $U_m = \cup_{b \in \mathbb{Z}^n}(U'_m \cap \prod_{i=1}^k [b_i, b_i + 1)) - b$, and so $\mathrm{Vol}(U_m) \leq \sum_{b \in \mathbb{Z}^n} \mathrm{Vol}(U'_m \cap \prod_{i=1}^k [b_i, b_i + 1))) = \mathrm{Vol}(U'_m)$.

Note that by Lemma 5.5 (ii) applied with $C := 2^{m+1}C$ gives the bound

$$\mathrm{Vol}(U'_m) \leq \det(\Sigma + I)^{-1/2} \cdot O(2^{m+1}C\log(\sigma))^{k/2}.$$

Therefore, we have $\mathrm{Vol}(T_m) \leq \mathrm{Vol}(U_m) \leq \mathrm{Vol}(U'_m) \leq \det(\Sigma + I)^{-1/2} \cdot O(2^{m+1}C\log(\sigma))^{k/2}$. This completes the proof. □

Next, we obtain bounds on $\sup_{\xi \in T_m} |\widehat{X}(\xi)|$ by using Lemma 3.10.

Claim 5.7. *For* $\xi \in T_m$, *it holds* $|\widehat{X}(\xi)| \leq \exp(-\Omega(C2^m \log(\sigma)/k))$. *If additionally we have* $m \leq 4\log_2 k$, *then* $|\widehat{X}(\xi)| = \exp(-\Omega(C2^m k\log(\sigma)))$.

Proof. Note that ξ' has coordinates in an interval of length $1 - 1/k$, so we may apply Lemma 3.10, yielding

$$|\widehat{X}(\xi)| = |\widehat{X}(\xi')| \leq \exp(-\Omega(\xi'^T \cdot \Sigma \cdot \xi'/k^2)) = \exp(-\Omega(C2^m \log(\sigma)/k)).$$

To get the stronger bound, we need to show that for small m, all the coordinates of ξ' are in a shorter interval. As before, we consider $\tilde{\xi'}$, the projection of ξ' onto the set of x with $x \cdot \mathbf{1} = 0$. Similarly, we have $\xi' \cdot \Sigma \xi' \geq \sigma |\tilde{\xi'}|_2^2$. So, for any i, j, it holds

$$|\xi'_i - \xi'_j| \leq |\tilde{\xi'}_i - \tilde{\xi'}_j| \leq 2|\tilde{\xi'}|_\infty \leq 2|\tilde{\xi'}|_2 \leq \sqrt{\xi' \cdot \Sigma \xi'/\sigma} \leq \sqrt{C 2^{m+1} k \log(\sigma)/\sigma}.$$

For $m \leq \log_2(\sigma/Ck\log(\sigma)) - 3$, we have that the coordinates of ξ lie in an interval of length $1/2$. Now, Lemma 3.10 gives that

$$|\widehat{X}(\xi)| = |\widehat{X}(\xi')| \leq \exp(-\Omega(\xi'^T \cdot \Sigma \cdot \xi')) = \exp(-\Omega(Ck 2^m \log(\sigma)/k)).$$

Finally, note that $4\log_2 k \leq \log_2(\sigma/Ck\log(\sigma)) - 3$, when $\sigma \geq Ck^3$. This completes the proof of the claim. □

Using the above, we can write

$$\begin{aligned}
\int_{[0,1]^k \setminus (T+\mathbb{Z}^k)} |\widehat{X}(\xi)| d\xi &\leq \sum_{m=0}^{\infty} \mathrm{Vol}(T_m) \sup_{\xi \in T_m} |\widehat{X}(\xi)| \\
&\leq \det(\Sigma + I)^{-1/2} \cdot O(C\log(\sigma))^{k/2} \sum_{m=0}^{\infty} 2^{mk/2} \sup_{\xi \in T_m} |\widehat{X}(\xi)| \,.
\end{aligned}$$

We divide this sum into two pieces:

$$\begin{aligned}
\sum_{m=0}^{4\log_2 k} 2^{mk/2} \sup_{\xi \in T_m} |\widehat{X}(\xi)| &\leq \sum_{m=0}^{\log_2(\sigma/Ck\log(\sigma))-3} 2^{mk/2} \exp(-\Omega(C 2^m k \log(\sigma))) \\
&\leq \sum_{m=0}^{4\log_2 k} \exp(-\Omega(C(2^m - m) k \log(\sigma))) \\
&\leq \sum_{m=0}^{4\log_2 k} 2^{-m} \exp(-\Omega(Ck\log(\sigma))) \\
&\leq \exp(-\Omega(Ck\log(\sigma))) = \sigma^{-\Omega(Ck)} \,,
\end{aligned}$$

and

$$\begin{aligned}
\sum_{m=4\log_2 k}^{\infty} 2^{mk/2} \sup_{\xi \in T_m} |\widehat{X}(\xi)| &\leq \sum_{m=4\log_2 k}^{\infty} 2^{mk/2} \exp(-\Omega(C 2^m \log(\sigma)/k)) \\
&\leq \sum_{m=4\log_2 k}^{\infty} \exp(-\Omega(C(2^m - m)\log(\sigma)/k)) \\
&\leq \sum_{m=4\log_2 k}^{\infty} \exp(-\Omega(C(k^2 + m)\log(\sigma)/k)) \\
&\leq \sum_{m=4\log_2 k}^{\infty} 2^{-m} \exp(-\Omega(Ck\log(\sigma))) \leq \sigma^{-\Omega(Ck)} \,.
\end{aligned}$$

We thus have $\int_{[0,1]^k \setminus (T+\mathbb{Z}^k)} |\widehat{X}(\xi)| d\xi \leq \det(\Sigma + I)^{-1/2} O(C\log(\sigma))^{k/2} \log(\sigma)^{-O(Ck)}$. □

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

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

This also holds for complex ξ, under the assumption that the coordinate-wise complex and real parts of ξ are in T, i.e., such that $\mathrm{Re}(\xi) \cdot \Sigma \mathrm{Re}(\xi), \mathrm{Im}(\xi) \cdot \Sigma \mathrm{Im}(\xi) \leq O(Ck\log(\sigma))$.

Proof. Recall that $\widehat{X}(\xi) = \prod_{i=1}^n \sum_{j=1}^k e(\xi_j)p_{ij}$. Let m_i be the element of $[k]$ so that p_{im_i} is as large as possible for each i. In particular, $p_{im_i} \geq 1/k$. We will attempt to approximate the above product by approximating the log of $\sum_{j=1}^k e(\xi_j)p_{ij}$ by its Taylor series expansion around the point $(\xi_{m_i}, \xi_{m_i}, \ldots, \xi_{m_i})$. In particular, by Taylor's Theorem, we find that

$$\sum_{j=1}^k e(\xi_j)p_{ij} = \exp\left(2\pi i\left(\xi_{m_i} + \sum_{j=1}^k p_{ij}(\xi_j - \xi_{m_i})\right) - 2\pi^2\left(\sum_{j=1}^k p_{ij}(\xi_j - \xi_{m_i})^2\right) + 2\pi^2\left(\sum_{j=1}^k p_{ij}(\xi_j - \xi_{m_i})\right)^2 + E_i\right),$$

where E_i is the third directional derivative in the $\xi - (\xi_{m_i}, \ldots, \xi_{m_i})$ direction of $\log(\widehat{X_i}(\xi))$ at some point $\tilde{\xi}$ along the line between ξ and $(\xi_{m_i}, \ldots, \xi_{m_i})$. Note that the above is exactly

$$\exp\left(2\pi i(\xi \cdot \mathbb{E}[X_i]) - 2\pi^2(\xi \cdot \mathrm{Cov}(X_i)\xi) + E_i\right).$$

Thus, taking a product over i, we find that

$$\widehat{X}(\xi) = \exp\left(2\pi i \mu \cdot \xi - 2\pi^2 \xi \cdot \Sigma \xi + \sum_{i=1}^n E_i\right).$$

We remark that the coefficients of this Taylor series are (up to powers of $-2\pi i$) the cumulants of X.

Since the coordinates of $\tilde{\xi}$ lie in an interval of length at most $1/2$, we have that $\sum_{j=1}^k e(\tilde{\xi}_j)p_{ij}$ is bounded away from 0. Therefore, we get that

$$|E_i| = O\Bigg(\sum_{j=1}^k p_{ij}|\tilde{\xi}_j - \xi_{m_i}|^3 + \sum_{j_1,j_2=1}^k p_{ij_1}p_{ij_2}|\tilde{\xi}_{j_1} - \xi_{m_i}|^2|\tilde{\xi}_{j_2} - \xi_{m_i}| + \sum_{j_1,j_2,j_3=1}^k p_{ij_1}p_{ij_2}p_{ij_3}|\tilde{\xi}_{j_1} - \xi_{m_i}||\tilde{\xi}_{j_2} - \xi_{m_i}||\tilde{\xi}_{j_3} - \xi_{m_i}|\Bigg).$$

Next note that

$$\mathrm{Var}(X_i \cdot \tilde{\xi}) \geq p_{ij}p_{im_i}|\tilde{\xi}_j - \xi_{m_i}|^2 \geq p_{ij}|\tilde{\xi}_j - \xi_{m_i}|^2/k.$$

Additionally, note that

$$Ck\log(\sigma) \geq \xi \cdot \Sigma\xi = \mathrm{Var}(X \cdot \xi) = \sum_{i=1}^n \mathrm{Var}(X_i \cdot \xi).$$

Therefore,

$$\sum_{i=1}^{n} p_{ij}|\tilde{\xi}_j - \xi_{m_i}|^2 \leq \sum_{i=1}^{n} p_{ij}|\xi_j - \xi_{m_i}|^2 \leq Ck^2 \log(\sigma).$$

Thus, since $|\tilde{\xi}_j - \xi_{m_i}| = O(\sqrt{Ck\log(\sigma)/\sigma})$ for all i, j we have that

$$\sum_{i=1}^{n}\sum_{j=1}^{k} p_{ij}|\tilde{\xi}_j - \xi_{m_i}|^3 \leq O(C^{3/2}k^{7/2}\sqrt{\log^3(\sigma)/\sigma}).$$

We have that

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j_1,j_2=1}^{k} p_{ij_1}p_{ij_2}|\tilde{\xi}_{j_1} - \xi_{m_i}||\tilde{\xi}_{j_2} - \xi_{m_i}| &\leq \sum_{i=1}^{n}\left(\sum_{j=1}^{k} p_{i,j}|\tilde{\xi}_j - \xi_{m_i}|\right)^2 \\ &\leq \sum_{i=1}^{n}\left(\sum_{j=1}^{k} p_{i,j}\right)\left(\sum_{j=1}^{k} p_{i,j}|\tilde{\xi}_j - \xi_{m_i}|^2\right) \\ &= O(Ck^3\log(\sigma)).\end{aligned}$$

Therefore,

$$\sum_{i=1}^{n}\sum_{j_1,j_2=1}^{k} p_{ij_1}p_{ij_2}|\tilde{\xi}_{j_1} - \xi_{m_i}|^2|\tilde{\xi}_{j_2} - \xi_{m_i}|$$

and

$$\sum_{j_1,j_2,j_3=1}^{k} p_{ij_1}p_{ij_2}p_{ij_3}|\tilde{\xi}_{j_1} - \xi_{m_i}||\tilde{\xi}_{j_2} - \xi_{m_i}||\tilde{\xi}_{j_3} - \xi_{m_i}|$$

are both

$$O(C^{3/2}k^{7/2}\sqrt{\log^3(\sigma)/\sigma}).$$

□

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

Proof. For this it suffices to prove a tail bound for G analogous to that satisfied by X. In particular, assuming that Σ has unit eigenvectors v_i with eigenvalues λ_i, it suffices to prove that $|(G-\mu)\cdot v_i| < \sqrt{\lambda_i}t$ except with probability at most $\exp(-\Omega(t^2))$. Recall that

$$G(p) = (2\pi)^{-(k-1)/2}\det(\Sigma')^{-1/2}\exp((p-\mu)\cdot\Sigma^{-1}(p-\mu)/2).$$

Let $\tilde{G}$ be the continuous probability density defined by

$$\tilde{G}(x) = (2\pi)^{-k/2}\det(\Sigma')^{-1/2}\exp((x-\mu)\cdot\Sigma'^{-1}(x-\mu)/2).$$

Note that for any p with $(p-\mu)\cdot\mathbf{1} = 0$, and $x \in [-1/2,1/2]^k$, we have that

$$G(p) = O\left(\tilde{G}(p+x) + \tilde{G}(p-x)\right).$$

Therefore, we have that

$$G(p) = O\left(\int_{x\in p+[-1/2,1/2]^k}\tilde{G}(x)\right).$$

Applying this formula for each p with $(p-\mu)\cdot v_i \geq \sqrt{\lambda_i}t$ and noting that $(x-\mu)\cdot v_i \geq (p-\mu)\cdot v_i - \sqrt{k} \geq \sqrt{\lambda_i}t - \sqrt{k}$ yields

$$\Pr(|(G-\mu)\cdot v_i| > \sqrt{\lambda_i}t) = O(\Pr(|(\tilde{G}-\mu)\cdot v_i| > \sqrt{\lambda_i}t - \sqrt{k} = \exp(-\Omega(t^2)).$$

Taking a union bound over $1 \leq i \leq k$ yields our result. □

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

Lemma 5.10. *The integral of $|\widehat{G}(\xi)|$ over ξ with $\xi_1 \in [0,1]$ and ξ not in T is at most $1/(|S|\sigma)$.*

Proof. We consider the integral over $\xi \in T_m$, where

$$T_m := \{\xi : \xi_1 \in [0,1] \mid \xi\cdot\Sigma\xi \in [2^m Ck\log(\sigma), 2^{m+1}Ck\log(\sigma)]\}.$$

We note that it has volume $2^{mk}k^{O(k)}\log^{O(k)}(\sigma)/|S|$, and that within T_m it holds $|\widehat{G}(\xi)| = \exp(-\Omega(Ck\log(\sigma)2^m))$. From this it is easy to see that the integral over T_m is at most $2^{-m-1}/(|S|\sigma)$. Summing over m yields the result. □

We now have all that is necessary to prove a weaker version of our main result.

Proposition 5.11. *We have the following:*

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

Proof. First, we bound the L^∞ of the difference. In particular, we note that for any p with integer coordinates summing to n we have that

$$X(p) = \int_{\xi\in\mathbb{R}^k,\xi_1\in[0,1],\xi_i\in[\xi_1-1/2,\xi_1+1/2]} e(-p\cdot\xi)\widehat{X}(\xi)d\xi\ ,$$

and

$$G(p) = \int_{\xi,\xi_1\in[0,1]} e(-p\cdot\xi)\exp(2\pi i(\xi\cdot\mu) - 2\pi^2\xi\cdot\Sigma\xi).$$

We note that in both cases the integral for ξ not in T is at most $1/(|S|\sigma)$. To show this, we need to note that any ξ with $\xi_1 \in [0,1], \xi_i \in [\xi_1 - 1/2, \xi_1 + 1/2]$ equivalent to a point of T modulo $\mathbb{Z}^k$ must lie in T itself. This is because the element of T must have all its coordinates differing by at most $1/4$, and thus must differ from ξ by an integer multiple of $(1, 1, \ldots, 1)$. Therefore, we have that

$$\begin{aligned}
|X(p) - G(p)| &= \left| \int_{\xi \in T, \xi_1 \in [0,1]} e(-p \cdot \xi)(\widehat{X}(\xi) - \widehat{G}(\xi)) d\xi \right| + O(1/(|S|\sigma)) \\
&\leq \int_{\xi \in T, \xi_1 \in [0,1]} |\widehat{X}(\xi) - \widehat{G}(\xi)| d\xi + O(1/(|S|\sigma)) \\
&\leq \int_{\xi \in T, \xi_1 \in [0,1]} O(C^{3/2} k^{7/2} \sqrt{\log^3(\sigma)/\sigma}) d\xi + O(1/(|S|\sigma)) \\
&\leq \det(\Sigma + I)^{-1/2} (C^{3/2} \sqrt{\log(\sigma)/\sigma}) O(C \log(\sigma))^{k/2}.
\end{aligned}$$

Therefore, the sum of $|X(p) - G(p)|$ over $p \in S$ is at most

$$(C^{3/2} \sqrt{\log(\sigma)/\sigma}) O(C^2 \log^2(\sigma))^{k/2}.$$

The sum over $p \notin S$ is at most $O(1/\sigma)$. This completes the proof. □

The proof of the main theorem 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.

Proof of Theorem 5.1. For $p \in S$ we have as above that

$$|X(p) - G(p)| = \left| \int_{\xi \in T, \xi_1 \in [0,1]} e(-p \cdot \xi)(\widehat{X}(\xi) - \widehat{G}(\xi)) d\xi \right| + O(1/(|S|\sigma)).$$

Let $\xi_0 \in \mathbb{R}^k$ be such that $\xi_0 . \mathbf{1} = 0$ and so that $\Sigma \xi_0 = (\mu - p)/(2\pi)$ (i.e., take $\xi_0 = (\Sigma + \mathbf{1}\mathbf{1}^T)^{-1} \mu - p)/(2\pi)$). We think of the integral above as an iterated contour integral. By deforming the contour associated with the innermost integral, we claim that it is the same as the sum of the integrals over ξ with $\mathrm{Re}(\xi) \in T, \mathrm{Re}(\xi_1) \in [0,1]$ and $\mathrm{Im}(\xi) = \xi_0$ and the integral over $\mathrm{Re}(\xi) \in \delta T, \mathrm{Re}(\xi_1) \in [0,1]$ and $\mathrm{Im}(\xi) = t\xi_0$ for some $t \in [0,1]$ (the extra pieces that we would need to add at $\mathrm{Re}(\xi_1) = 0$ and $\mathrm{Re}(\xi_1) = 1$ cancel out).

Claim 5.12. $\int_{\xi \in \delta T, \xi_1 \in [0,1]} e(-p \cdot \xi)(\widehat{X}(\xi) - \widehat{G}(\xi)) d\xi$ *equals*

$$\int_{\xi \in T, \xi_1 \in [0,1]} e(-p \cdot (\xi + i\xi_0))(\widehat{X}(\xi + i\xi_0) - \widehat{G}(\xi + i\xi_0)) d\xi$$

plus

$$\int_{\xi \in \delta T, \xi_1 \in [0,1]} \int_{t=0}^{1} e(-p \cdot (\xi + it\xi_0))(\widehat{X}(\xi + it\xi_0) - \widehat{G}(\xi + it\xi_0)) d(t\xi_0) \cdot d\xi.$$

Proof. We write $f(\xi) = e(-p \cdot \xi)(\widehat{X}(\xi) - \widehat{G}(\xi))$. Let O be an orthogonal matrix with kth column $\xi_0/\|\xi_0\|_2$. Then, we change variables from ξ to $\nu = O^T\xi$, yielding

$$\int_{\xi \in T'} f(\xi) d\xi = \int_{\nu \in O^T T'} f(O^T \nu) d\nu$$

We can consider this as an iterated integral where ν_i is integrated from $a_i(\nu_1, \ldots, \nu_{i-1})$ to $b_i(\nu_1, \ldots, \nu_{i-1})$.

$$\int_{\nu \in O^T T'} f(O^T \nu) d\nu = \int_{a_1}^{b_1} \int_{a_2(\nu_1)}^{b_2(\nu_1)} \cdots \int_{a_k(\nu_1,\ldots,\nu_{k-1})}^{b_k(\nu_1,\ldots,\nu_{k-1})} f(O^T \nu) d\nu_k \ldots d\nu_2 d\nu_1 \ .$$

We consider the innermost integral. The function $f(O^T\nu)$ is a linear combination of exponentials and so is holomorphic on all of $\mathbb{C}^n$. Let $\mathcal{C}$ be the contour which consists of three straight lines, from $a_k(\nu_1, \ldots, \nu_{k-1})$ via $a_k(\nu_1, \ldots, \nu_{k-1}) + i\|\xi_0\|_2$ and $b_k(\nu_1, \ldots, \nu_{k-1}) + i\|\xi_0\|_2$ to $b_k(\nu_1, \ldots, \nu_{k-1})$. Then, by standard facts of complex analysis, we have:

$$\begin{aligned}
&\int_{a_k(\nu_1,\ldots,\nu_{k-1})}^{b_k(\nu_1,\ldots,\nu_{k-1})} f(O^T \nu) d\nu_k \\
&= \int_{\mathcal{C}} f(O^T \nu) d\nu_k \\
&= \int_0^1 f(O^T(\nu_1, \ldots, \nu_{k-1}, a_k(\nu_1, \ldots, \nu_{k-1}) + i\|\xi_0\|_2 t)) i\|\xi_0\|_2 dt \\
&+ \int_{a_k(\nu_1,\ldots,\nu_{k-1})}^{b_k(\nu_1,\ldots,\nu_{k-1})} f(O^T(\nu + i\|\xi_0\|_2 e_k)) d\nu_k \\
&+ \int_0^1 f(O^T(\nu_1, \ldots, \nu_{k-1}, b_k(\nu_1, \ldots, \nu_{k-1}) + i\|\xi_0\|_2(1-t'))) i\|\xi_0\|_2 dt'
\end{aligned}$$

The middle part of this path gives the first term in the statement of the claim:

$$\begin{aligned}
&\int_{a_1}^{b_1} \int_{a_2(\nu_1)}^{b_2(\nu_1)} \cdots \int_{a_k(\nu_1,\ldots,\nu_{k-1})}^{b_k(\nu_1,\ldots,\nu_{k-1})} f(O^T(\nu + i\|\xi_0\|_2 e_k)) d\nu_k \ldots d\nu_2 d\nu_1 \\
&= \int_{\nu \in O^T T'} f(O^T(\nu + i\|\xi_0\|_2 e_k)) \\
&= \int_{\xi \in T'} f(\xi + i\xi_0) \ .
\end{aligned}$$

A change of variables allows us to express the sum of the contributions from the first and third part of the path:

$$\begin{aligned}
&\int_0^1 f(O^T(\nu_1, \ldots, \nu_{k-1}, a_k(\nu_1, \ldots, \nu_{k-1}) + i\|\xi_0\|_2 t)) i\|\xi_0\|_2 dt \\
&- \int_0^1 f(O^T(\nu_1, \ldots, \nu_{k-1}, b_k(\nu_1, \ldots, \nu_{k-1}) + i\|\xi_0\|_2 t')) i\|\xi_0\|_2 dt' \ .
\end{aligned}$$

Changing variables to replace $(\nu_1, \ldots, \nu_{k-1}, a_k(\nu_1, \ldots, \nu_{k-1}))$ or $(\nu_1, \ldots, \nu_{k-1}, b_k(\nu_1, \ldots, \nu_{k-1}))$ with $\xi \in \delta T$ or $\xi \in T \cap \{0, 1\}$ we get an appropriate integral of $\pm i f(\xi + it\xi_0)$. We note that the volume form for ξ_0 assigns to a surface element the volume of the projection of that element in the ξ_0 direction. Multiplying by $\|\xi_0\|_2$ and the appropriate sign yields exactly the measure $\xi_0 \cdot d\xi$. Thus,

we are left with an integral of $f(\xi + it\xi_0)d(t\xi_0) \cdot d\xi$. However, it should be noted that the measures $\xi_0 \cdot d\xi$ are opposite on $\xi_1 = 0$ and $\xi_1 = 1$ boundaries (as $d\xi$ is the outward pointing normal). Since $f(\xi+it\xi_0) = f(\xi+\mathbf{1}+it\xi_0)$, the integrals over these regions cancel, leaving exactly with the claimed integral. $\square$

In order to estimate this difference, we use Lemma 5.8, which still applies. Furthermore, we note that

$$\xi_0 \cdot \Sigma\xi_0 = (p-\mu) \cdot \Sigma^{-1}(p-\mu)/(4\pi^2) = O(Ck\log(\sigma)) \ ,$$

because $p \in S$.

Therefore, we have that $|X(p) - G(p)|$ is $O(1/(|S|\sigma))$ plus

$$O(k^{7/2}\sqrt{\log^3(\sigma)/\sigma}) \int_{\mathcal{C}} \left|\exp\left(2\pi i(\mu-p)\cdot\xi - 2\pi^2\xi\cdot\Sigma\xi\right)\right| d\xi.$$

Now, when $\mathrm{Im}(\xi) = \xi_0$, we have that

$$\exp\left(2\pi i(\mu-p)\cdot\xi - 2\pi^2\xi\cdot\Sigma\xi\right) = \exp\left(-(p-\mu)\Sigma^{-1}(p-\mu)/2 - 2\pi^2\mathrm{Re}(\xi)\cdot\Sigma\mathrm{Re}(\xi)\right).$$

Integrating, we find that the difference over this region is at most times

$$O(k^{7/2}\sqrt{\log^3(\sigma)/\sigma}) \int \exp\left(-(p-\mu)\Sigma^{-1}(p-\mu)/2 - 2\pi^2\mathrm{Re}(\xi)\cdot\Sigma\mathrm{Re}(\xi)\right) d\xi = O(k^{7/2}\sqrt{\log^3(\sigma)/\sigma})G(p).$$

The contribution from the part of the contour where $\mathrm{Re}(\xi)$ is on the boundary of T is also easy to bound after noting that both $|\widehat{X}(\xi)|$ and $|\widehat{G}(\xi)|$ are $O(\sigma^{-k})$. We furthermore claim that the total volume of the region of integration is $O(\sqrt{k}\mathrm{Vol}(T))$. Together, these would imply that the total integral over this region is $O(1/(\sigma|S|))$. To do this, we note that the total volume of the region being integrated over is at most the volume of the projection of T in the direction perpendicular to ξ_0 times the length of ξ_0. In order to analyze this we consider each slice of T given by $\xi \cdot \mathbf{1} = \alpha$ separately. Noting that $|\alpha| \leq 2$ for all $\xi \in \delta T$, it suffices to consider only a single slice. In particular, since for all such α, we have that $\tilde{\xi} \in T$, it suffices to consider the slice $\alpha = 0$. Along this slice, we have that T is an ellipsoid. Note that $\xi_0 \cdot \Sigma\xi_0 = (p-\mu)\cdot(\Sigma + \mathbf{1}\mathbf{1}^T)^{-1}(p-\mu) \leq Ck\log(\sigma)$. Therefore $\xi_0 \in T$.

Next, we claim that if E is any ellipsoid in at most k dimensions, and if v is a vector with $v \in E$, then the product of the length of v times the volume of the projection of E perpendicular to v is at most $O(\sqrt{k}\mathrm{Vol}(E))$. This follows after noting that the claim is invariant under affine transformations, and thus it suffices to consider E the unit ball for which it is easy to verify.

Therefore, for $p \in S$, we have that

$$|X(p) - G(p)| \leq O(1/|S|\sigma) + O(k^{7/2}\sqrt{\log^3(\sigma)/\sigma})G(p).$$

From this it is easy to see that it is also

$$|X(p) - G(p)| \leq O(1/|S|\sigma) + O(k^{3/2}\sqrt{\log^3(\sigma)/\sigma})X(p).$$

Summing over $p \in S$ gives a total difference of at most

$$O(k^{7/2}\sqrt{\log^3(\sigma)/\sigma}).$$

Combining this with the fact that the sum of $X(p)$ and $G(p)$ for p not in S is at most $1/\sigma$ gives us that

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

This completes the proof of Theorem 5.1. $\square$

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. First, we need to relate our PMD (supported on integer vectors whose coordinates sum to n) to theirs (which are projections of PMDs onto $k-1$ coordinates). In particular, we need to show that this projection does not skew minimum eigenvalue in the wrong direction. This is done in the following simple proposition:

Proposition 5.13. *Let X be an (n,k)-PMD, and X' be obtained by projecting X onto its first $k-1$ coordinates. Let Σ and Σ' be the covariance matrices of X and X', respectively, and let σ and σ' be the second smallest and smallest eigenvalues respectively of Σ and Σ'. Then, we have $\sigma \geq \sigma'$.*

Proof. Note that, since $\mathbf{1}$ is in the kernel of Σ, σ is the minimum of v orthogonal to $\mathbf{1}$ of $\frac{v^T\Sigma v}{v^T v}$. Whereas, σ' is the minimum over $w \in \mathbb{R}^{k-1}\backslash\{\mathbf{0}\}$ of $\frac{w^T\Sigma' w}{w^T w}$. This is the same as the minimum over w in $\mathbb{R}^k$ with k^{th} coordinate equal to 0 of $\frac{w^T\Sigma w}{w^T w}$.

Let the minimization problem defining σ be obtained by some particular v orthogonal to $\mathbf{1}$. In particular, a v so that $\sigma = \frac{v^T\Sigma v}{v^T v}$. Let w be the unique vector of the form $v + a\mathbf{1}$ so that w has last coordinate 0. Then, we have that

$$\sigma' \geq \frac{w^T\Sigma w}{w^T w} = \frac{v^T\Sigma v}{w^T w} \geq \frac{v^T\Sigma v}{v^T v} = \sigma.$$

This completes the proof. □

Next, we need to relate the two slightly different notions of discrete Gaussian.

Proposition 5.14. *Let G be a Gaussian in $\mathbb{R}^k$ with covariance matrix Σ, which has no eigenvalue smaller than σ. Let G' be the discrete Gaussian obtained by rounding the values of G to the nearest lattice point. Let G'' be the discrete distribution obtained by assigning each integer lattice point mass proportional to the probability density function of G. Then, we have that*

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

Proof. We note that the probability density function of G is proportional to $\exp(-(x\cdot\Sigma^{-1}x)/2)dx$. Suppose that y is another vector with $\|x-y\|_\infty < 1$. We would like to claim that the probability density function at y is approximately the same as at x. In particular, we write $y = x + z$ and note that

$$\begin{aligned} y\cdot\Sigma^{-1}y &= x\cdot\Sigma^{-1}x + 2z\cdot\Sigma^{-1}x + z\cdot\Sigma^{-1}z \\ &= x\cdot\Sigma^{-1}x + 2(\Sigma^{-1/2}x)\cdot(\Sigma^{-1/2}z) + O(|z|_2^2\sigma^{-1}) \\ &= x\cdot\Sigma^{-1}x + O(|\Sigma^{-1/2}x|_2\sqrt{k/\sigma} + k\sigma^{-1}) \\ &= x\cdot\Sigma^{-1}x + O(\sqrt{(k/\sigma)x\cdot\Sigma^{-1}x} + k\sigma^{-1}). \end{aligned}$$

Note that, for lattice points x, $G'(x)$ is the average over y in a unit cube about x of the pdf of G at y, while $G''(x)$ is just the pdf of G at x. These quantities are within a $1+O(\sqrt{(k/\sigma)x\cdot\Sigma^{-1}x}+k\sigma^{-1})$ multiple of each other by the above so long as the term in the "O" is $o(1)$. Therefore, for all x with $x\cdot\Sigma^{-1}x \ll k\log(\sigma)$, we have that $G'(x) = G''(x)(1+O(k\sqrt{\log(\sigma)/\sigma}))$. We note however

that G' has only a $1/\sigma$ probability of x being outside of this range. Furthermore, we claim that $G''(x) = O(G'(x))$ for all x. To see this, note that for any v with $\|v\|_\infty \leq 1/2$, we have

$$G(x) = e^{-v^T\Sigma^{-1}v/2}\sqrt{G(x+v)G(x-v)} \leq e^{-(k/2\sigma)}\max\{G(x+v), G(x-v)\}.$$

We assume that $\sigma \geq k^2$ or else we have nothing to prove. Then, we have $G(x) = O(G(x+v) + G(x-v))$, and by considering the integral that defines G', we have $G''(x) = O(G'(x))$. Thus, G'' similarly has $O(1/\sigma)$ mass outside of the range $x \cdot \Sigma^{-1}x \ll k\log(\sigma)$. Therefore, the L_1 difference inside the range is $O(k\sqrt{\log(\sigma)/\sigma})$ and the L_1 error from outside is $O(1/\sigma)$. This completes the proof. □

Armed with these propositions, we have the following corollary of Theorem 5.1:

Corollary 5.15. *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*

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

Proof. Let Σ be the covariance matrix of X. Since X is a PMD, $\mathbf{1}$ is an eigenvector of Σ with eigenvalue 0. By Proposition 5.13, the other eigenvalues of Σ are at least σ'. Theorem 5.1 now yields that $d_{\mathrm{TV}}(X, G) \leq O(k^{7/2}\sqrt{\log^3(\sigma')/\sigma'})$, where G is a discrete Gaussian in k dimensions (as defined in the context of the theorem statement). Let G'' be the discrete Gaussian obtained by projecting G onto the first $k-1$ coordinates. Then, we have that $d_{\mathrm{TV}}(X', G'') \leq O(k^{7/2}\sqrt{\log^3(\sigma')/\sigma'})$. From the proof of Theorem 5.1, G is proportional to the pdf of $\mathcal{N}(\mathbb{E}[X], \Sigma)$. Note that G'' is proportional to the pdf of $\mathcal{N}(\mathbb{E}[X'], \Sigma')$. Then, by Proposition 5.14, it follows that $d_{\mathrm{TV}}(G', G'') \leq O(k\sqrt{\log(\sigma')/\sigma'})$. So, by the triangle inequality, we have $d_{\mathrm{TV}}(X', G') \leq O(k^{7/2}\sqrt{\log^3(\sigma')/\sigma'})$, as required. □

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.

[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. Optimal learning via the fourier transform for sums of independent integer random variables. *CoRR*, abs/1505.00662, 2015.

[DKS15b] 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.

[GRW15] P. Gorlach, C Riener, and T. Weisser. Deciding positivity of multisymmetric polynomials. *Journal of Symbolic Computation*, 2015. Also available as arxiv report http://arxiv.org/abs/1409.2707.

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

[HJ85] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

[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.

[Sto96] A. Storjohann. Near optimal algorithms for computing smith normal forms of integer matrices. In *Proceedings of the 1996 international symposium on Symbolic and algebraic computation*, pages 267–274, 1996.

[Sto00] A. Storjohann. *Algorithms for matrix canonical forms*. PhD thesis, Diss., Technische Wissenschaften ETH Zürich, Nr. 13922, 2001, 2000.

[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.

Appendix

A Proof of Lemma 3.4

Lemma 3.4 follows directly from the following statement:

Lemma A.1. *If we take $O(k^4/\epsilon^2)$ samples from an (n,k)-PMD and let $\widehat{\mu}$ and $\widehat{\Sigma}$ be the sample mean and sample covariance matrix, then with probability $19/20$, for any $y \in \mathbb{R}^k$, we have:*

$$|y^T(\widehat{\mu}-\mu)| \le \epsilon\sqrt{y^T(\Sigma+I)y}\ ,$$

and

$$|y^T(\widehat{\Sigma}-\Sigma)y| \le \epsilon y^T(\Sigma+I)y\ .$$

The above lemma and its proof follow from a minor modification of an analogous lemma in [DKT15]. We include the proof here for the sake of completeness. We will use the following simple lemma:

Lemma A.2 (Lemma 21 from [DKT15])**.** *For any vector $y \in \mathbb{R}^k$, given sample access to an (n,k)-PMD $\mathbf{P}$ with mean μ and covariance matrix Σ, there exists an algorithm which can produce estimates $\widehat{\mu}$ and $\widehat{\Sigma}$, such that with probability at least $19/20$: $|y^T(\widehat{\mu}-\mu)| \le \epsilon\sqrt{y^T\Sigma y}$ and $|y^T(\widehat{\Sigma}-\Sigma)y| \le \epsilon y^T\Sigma y\sqrt{1+\frac{y^Ty}{y^T\Sigma y}}$. The sample and time complexity are $O(1/\epsilon^2)$.*

Proof of Lemma A.1. The proof will follow by applying Lemma A.2 to k^2 carefully chosen vectors simultaneously using the union bound. Using the resulting guarantees, we show that the same estimates hold for any direction, at a cost of rescaling ϵ by a factor of k. Let S be the set of k^2 vectors $\{v_i\}$, for $1 \le i \le k$, and $\{\frac{1}{\sqrt{\lambda_i+1}}v_i+\frac{1}{\sqrt{\lambda_j}}v_j\}$, for each $i \ne j$, where the v_i's are an orthonormal eigenbasis for Σ with eigenvalues λ_i. From Lemma A.2 and a union bound, with probability $9/10$, for all $y \in S$, we have

$$|y^T(\widehat{\mu}-\mu)| \le (\epsilon/k)\sqrt{y^T\Sigma y}\ ,$$

and

$$|y^T(\widehat{\Sigma}-\Sigma)y| \le (\epsilon/3k)(y^T\Sigma y)\sqrt{1+\frac{y^Ty}{y^T\Sigma y}}\ .$$

We claim that the latter implies that:

$$|y^T(\widehat{\Sigma}-\Sigma)y| \le (\epsilon/3k)(y^T(\Sigma+I)y)\ .$$

Note that if $y^T\Sigma y = 0$, we must have $y^T\widehat{\Sigma}y = 0$, since then y^TX is a constant for a PMD random variable X. Otherwise,

$$(y^T\Sigma y)\sqrt{1+\frac{y^Ty}{y^T\Sigma y}} = \sqrt{(y^T\Sigma y)(y^T\Sigma y+y^Yy)} = \sqrt{(y^T\Sigma y)(y^T(\Sigma+I)y)} \le (y^T(\Sigma+I)y)\ .$$

The claim about the accuracy of $\widehat{\Sigma}$ now follows from Lemma A.2.

We now prove that the mean estimator $\widehat{\mu}$ is accurate. Consider an arbitrary vector y, which can be decomposed into a linear composition of the eigenvectors $y = \sum_i \alpha_i v_i$.

Then,

$$y^T(\widehat{\mu} - \mu) = \sum_i \alpha_i v_i^T(\widehat{\mu} - \mu) \leq (\epsilon/k) \sum_i |\alpha_i| \sqrt{\lambda_i + 1} \leq (\epsilon/k)\sqrt{k} \sqrt{k \sum_i \alpha_i^2 (\lambda_i + 1)} \, ,$$

but $\sum_i \alpha_i^2(\lambda_i + 1) = y^T(\Sigma + I)y$, so we have $y^T(\widehat{\mu} - \mu) \leq \epsilon \sqrt{y^T(\Sigma + I)y}$ as required. □

We are now ready to complete the proof of the desired lemma.

Lemma 3.4. *With probability* 19/20, *we have that* $(\widehat{\mu} - \mu)^T(\Sigma + I)^{-1}(\widehat{\mu} - \mu) = O(1)$, $2(\Sigma + I) \geq \widehat{\Sigma} + I \geq (\Sigma + I)/2$.

Proof. We apply Lemma A.1 with $\epsilon := 1/2$. For all y, we have $|y^T(\widehat{\Sigma} - \Sigma)y| \leq \epsilon y^T(\Sigma + I)y$, that is

$$\frac{1}{2} y^T(\Sigma + I)y \leq y^T(\widehat{\Sigma} + I)y \leq \frac{3}{2} y^T(\Sigma + I)y \, .$$

Thus, we have $\frac{1}{2}(\Sigma + I) \leq \widehat{\Sigma} + I \leq \frac{3}{2}(\Sigma + I)$ as required.

Note that since $\Sigma + I$ is positive definite, it is non-singular. Setting $y = \frac{1}{(\widehat{\mu} - \mu)^T(\Sigma + I)^{-1}(\widehat{\mu} - \mu)} (\Sigma + I)^{-1}(\widehat{\mu} - \mu)$, we have $y^T\widehat{\mu} - \mu) = 1$ and $y^T(\Sigma + I)y = 1/(\widehat{\mu} - \mu)^T(\Sigma + I)^{-1}(\widehat{\mu} - \mu)$. So, Lemma A.1 gives us:

$$|y^T(\widehat{\mu} - \mu)| \leq \frac{1}{2} \sqrt{y^T(\Sigma + I)y}.$$

Substituting the above:

$$1 \leq \frac{1}{2} \sqrt{1/(\widehat{\mu} - \mu)^T(\Sigma + I)^{-1}(\widehat{\mu} - \mu)} \, .$$

Therefore, we have $(\widehat{\mu} - \mu)^T(\Sigma + I)^{-1}(\widehat{\mu} - \mu) \leq 1/4$, as required. □