

A Size-Free CLT for Poisson Multinomials and its Applications

Constantinos Daskalakis*
EECS, MIT
costis@mit.edu

Anindya De
Northwestern University
anindya.de1@northwestern.edu

Gautam Kamath†
EECS, MIT
g@csail.mit.edu

Christos Tzamos‡
EECS, MIT
tzamos@mit.edu

June 17, 2016

Abstract

An (n, k) -Poisson Multinomial Distribution (PMD) is the distribution of the sum of n independent random vectors supported on the set $\mathcal{B}_k = \{e_1, \dots, e_k\}$ of standard basis vectors in \mathbb{R}^k . We show that any (n, k) -PMD is $\text{poly}(\frac{k}{\sigma})$ -close in total variation distance to the (appropriately discretized) multi-dimensional Gaussian with the same first two moments, removing the dependence on n from the Central Limit Theorem of Valiant and Valiant [VV11]. Interestingly, our CLT is obtained by bootstrapping the Valiant-Valiant CLT itself through the structural characterization of PMDs shown in recent work [DKT15]. In turn, our stronger CLT can be leveraged to obtain an efficient PTAS for approximate Nash equilibria in anonymous games, significantly improving the state of the art [DP08], and matching qualitatively the running time dependence on n and $1/\varepsilon$ of the best known algorithm for two-strategy anonymous games [DP09]. Our new CLT also enables the construction of covers for the set of (n, k) -PMDs, which are proper and whose size is shown to be essentially optimal. Our cover construction combines our CLT with the Shapley-Folkman theorem and recent sparsification results for Laplacian matrices [BSS12]. Our cover size lower bound is based on an algebraic geometric construction. Finally, leveraging the structural properties of the Fourier spectrum of PMDs we show that these distributions can be learned from $O_k(1/\varepsilon^2)$ samples in $\text{poly}_k(1/\varepsilon)$ -time, removing the quasi-polynomial dependence of the running time on $1/\varepsilon$ from [DKT15].

*Supported by a Microsoft Research Faculty Fellowship, and NSF Award CCF-0953960 (CAREER) and CCF-1551875. This work was done in part while the author was visiting the Simons Institute for the Theory of Computing.

†Supported by NSF Award CCF-0953960 (CAREER) and ONR grant N00014-12-1-0999. This work was done in part while the author was an intern at Microsoft Research Cambridge and visiting the Simons Institute for the Theory of Computing.

‡Supported by NSF Award CCF-0953960 (CAREER), ONR grant N00014-12-1-0999, and a Simons Award for Graduate Students in Theoretical Computer Science. This work was done in part while the author was visiting the Simons Institute for the Theory of Computing.

1 Introduction

The Poisson Multinomial Distribution (PMD) is the multi-dimensional generalization of the more familiar Poisson Binomial Distribution (PBD). To illustrate its meaning, consider a city of n people and k newspapers. Suppose that person i has his own proclivity to buy each newspaper, so that his purchase each day can be modeled as a random vector X_i – also called a Categorical Random Variable (CRV) – taking values in the set $\mathcal{B}_k = \{e_1, \dots, e_k\}$ of standard basis vectors in \mathbb{R}^k .¹ If people buy their newspapers independently, the total circulation of newspapers is the sum $X = \sum_i X_i$. The distribution of X is a (n, k) -PMD, and we need $n \cdot (k - 1)$ parameters to describe it. When $k = 2$, the distribution is called an n -PBD. When people have identical proclivities to buy the different newspapers, the distribution degenerates to the more familiar Multinomial (general k) or Binomial ($k = 2$) distribution.² In other words, n -PBDs are distributions of sums of n independent, not necessarily identically distributed Bernoullis, while (n, k) -PMDs are their multi-dimensional generalization, where we are summing independent categorical random variables. As such, these distributions are one of the most widely studied multi-dimensional families of distributions.

In Probability theory, a large body of literature aims at approximating PMDs via simpler distributions. The Central Limit Theorem (CLT) informs us that the limiting behavior of an appropriately normalized PMD, as $n \rightarrow \infty$, is a multi-dimensional Gaussian, under conditions on the eigenvalues of the summands' covariance matrices; see e.g. [VdV00]. The rate of convergence in the CLT is quantified by multi-dimensional Berry-Esseen theorems. As PMDs are discrete, while Gaussians are continuous distributions, such theorems typically bound the maximum difference in probabilities assigned by the two distributions to convex subsets of \mathbb{R}^k . Again, these bounds degrade as the PMD's covariance matrix tends to singularity; see e.g. [Ben05, CST14]. Similarly, approximations of PMDs via multivariate Poisson [Bar88, DP88], multinomial [Loh92], and other discrete distributions has been intensely studied, often using Stein's method.

In theoretical computer science, PMDs are commonly used in the analysis of randomized algorithms, often through large deviation inequalities. They have also found applications in algorithmic problems where one is looking for a collection of random vectors optimizing a certain probabilistic objective, or satisfying probabilistic constraints. For example, understanding the behavior of PMDs has led to polynomial-time approximation schemes for anonymous games [Mil96, Blo99, Blo05, Kal05, DP07, DP08, DP09], despite the PPAD-completeness of their exact equilibria [CDO15]. Anonymous games are games where a large number n of players share the same k strategies, and each player's utility only depends on his own choice of strategy and the number of other players that chose each of the k strategies. In particular, the expected payoff of each player depends on the PMD resulting from the mixed strategies of the other players. It turns out that understanding the behavior of PMDs provides a handle on the computation of approximate Nash equilibria. One of our main contributions is to advance the state of the art for computing approximate Nash equilibria in anonymous games. We will come to this contribution shortly.

A New CLT. Recently Valiant and Valiant have used PMDs to obtain sample complexity lower bounds for testing symmetric properties of distributions [VV11]. The workhorse in their lower bounds is a new CLT bounding the total variation distance between a (n, k) -GMD and a multidimensional Gaussian with the same mean vector and covariance matrix. Since they are

¹Of course, we can always add a dummy newspaper to account for the possibility that somebody may decide not to buy a newspaper.

²It is customary to project Binomial and Poisson Binomial distributions to one of their coordinates. In multiple dimensions, it will be convenient to call a distribution resulting from the projection of a PMD to all but one coordinates a Generalized Multinomial distribution (GMD).

comparing a discrete to a continuous distribution under the total variation distance, they need to discretize the Gaussian by rounding its coordinates to their closest point in the integer lattice. If X is distributed according to some (n, k) -GMD with mean vector μ and covariance matrix Σ , and Y is distributed according to the multi-dimensional Gaussian $\mathcal{N}(\mu, \Sigma)$, [VV11] shows that:

$$d_{\text{TV}}(X, \lfloor Y \rfloor) \leq \frac{k^{4/3}}{\sigma^{1/3}} \cdot 2.2 \cdot (3.1 + 0.83 \log n)^{2/3}, \quad (1)$$

where σ^2 is the minimum eigenvalue of Σ and $\lfloor Y \rfloor$ denotes the rounding of Y to the closest point in the integer lattice. The dependence of the bound on the dimension k and the minimum eigenvalue σ^2 is necessary, and quite typical of Berry-Esseen type bounds. Answering a question raised in [VV11], we prove a qualitatively stronger CLT by showing that the explicit dependence of the bound on n can be removed (hence, the CLT is “size-free”).

Theorem 1 (Size-free CLT). *Suppose that X is distributed according to some (n, k) -GMD with mean μ and covariance matrix Σ , and $Y \sim \mathcal{N}(\mu, \Sigma)$. There exists some constant $C > 0$ such that*

$$d_{\text{TV}}(X, \lfloor Y \rfloor) \leq C \frac{k^{7/2}}{\sigma^{1/10}}, \quad (2)$$

where σ^2 is the minimum eigenvalue of Σ .

Interestingly, Theorem 1 is proven by bootstrapping the Valiant-Valiant CLT itself. Indeed, this CLT was used as one of the key ingredients in a recent structural characterization of PMDs [DKT15], where it was shown that any (n, k) -Poisson multinomial random vector is ε -close in total variation distance to the sum of an (appropriately discretized) Gaussian and a $(\text{poly}(k/\varepsilon), k)$ -Poisson multinomial random vector; see Theorem 6. In turn, we prove Theorem 1 by using Theorem 6 as a black box.

We start with an invocation of the structural characterization for some $\varepsilon = \text{poly}(k/\sigma)$. With a judicious such choice of ε , the structural result approximates an arbitrary (n, k) -Poisson multinomial random vector X (to within $\text{poly}(k/\sigma)$ in total variation distance) by the sum $G + P$ of a discretized Gaussian G and a $(o(\sigma), k)$ -Poisson multinomial random vector P . As P has too few components, namely $o(\sigma)$, we show that G must account for the variance of X , which is at least σ^2 in all directions. Next, since G has variance $\Omega(\sigma^2)$ in all directions and P has variance $o(\sigma^2)$, we can show that G swamps P , in that $d_{\text{TV}}(G, G + P)$ is small, using Proposition 6. So $d_{\text{TV}}(X, G)$ is also small by triangle inequality. The remaining step is to argue that G can be replaced by a discretized multidimensional Gaussian with the same first two moments as X . This is done in two parts. First, since X and G are close in total variation distance, we can argue that their first two moments are close using Proposition 8. Then, we relate G to a discretized Gaussian with the same mean and covariance as X using Lemma 2, which bounds the total variation distance between two Gaussians with similar moments. Finally, we need to argue that the resulting Gaussian can be trivially discretized to the integer lattice, obviating the need for a more sophisticated structure preserving rounding.

For more details on our proof’s approach, see Section 3.

In the remainder of this section we discuss the algorithmic applications of our CLT, concluding with our improved algorithms for learning PMDs using Fourier analysis.

Anonymous Games. We have already discussed anonymous games earlier in this section, where we have also explained their relation to PMDs. In particular, the expected utility u_i of some

player i in a n -player k -strategy anonymous game only depends on his own choice of mixed strategy X_i and the $(n-1, k)$ -Poisson multinomial random vector $\sum_{j \neq i} X_j$ aggregating the mixed strategies of his opponents. It is therefore natural to expect that a better understanding of the structure of PMDs could lead to improved algorithms for computing Nash equilibria in these games. Indeed, earlier work [DP08, DP15] has exploited this connection to obtain algorithms for approximate Nash equilibria, whose running time is

$$n^{O\left(2^{k^2} \cdot \left(\frac{f(k)}{\varepsilon}\right)^{6 \cdot k}\right)}, \text{ where } f(k) \leq 2^{3k-1} k^{k^2+1} k!$$

While clearly of theoretical interest, this bound shows that anonymous games are one of the few classes of games where *approximate* equilibria can be efficiently computed, while *exact* equilibria are PPAD-hard [CDO15], even for n -player 7-strategy anonymous games. Exploiting our CLT we obtain a significant improvement over [DP08].

Theorem 2 (Equilibria in Anonymous Games). *An ε -approximately well supported Nash equilibrium of an n -player k -strategy anonymous games whose utilities are in $[0, 1]$ can be computed in time.³*

$$n^{O(k^2)} \cdot 2^{O(k^{5k} \cdot \log^{k+2}(1/\varepsilon))}. \quad (3)$$

The salient feature of Theorem 2 is the polynomial dependence of the running time on n and its quasi-polynomial dependence on ε^{-1} . In terms of these dependencies our algorithm matches the best known algorithm for 2-strategy anonymous games [DP09], where much more is known given the single-dimensional nature of $(n, 2)$ -PMDs.

Moreover, the recent hardness results for anonymous games [CDO15] establish that not only finding an exact but also a 2^{n^a} -approximate Nash equilibrium is PPAD-hard. An interesting corollary of Theorem 2 is that this cannot be pushed to $\text{poly}(1/n)$ -approximations, unless PPAD can be solved in quasi-polynomial time.

Corollary 1 (Non-PPAD Hardness of FPTAS). *Unless $\text{PPAD} \subseteq \text{Quasi-PTIME}$, it is not PPAD-hard to find a $\text{poly}(1/n)$ -approximately well supported Nash equilibrium in anonymous games, for any $\text{poly}(\cdot)$.*

It is interesting to contrast this corollary with normal-form games where it is known that computing inverse polynomial approximations *is* PPAD-hard [DGP09, CDT09].

From a technical standpoint, our algorithm for anonymous games uses the structural understanding of PMDs as follows. Since every player views the aggregate strategies of the other players as a PMD, one approach would be to guess each player’s view using a cover as developed in [DKT15]. However, this approach gives a runtime which is exponential in n , since it requires us to enumerate the cover for each player. An alternative approach is to guess the overall PMD which occurs at a Nash equilibrium, and guess appropriate “corrections” that allow us to infer each player’s view. To do this, we must find an alternative PMD which approximately matches the PMD at Nash in the following sense:

- The PMD that results by removing the CRV corresponding to a player should be close to the view that the player observes;

³As it is customary in Nash equilibrium algorithms, approximate Nash equilibria are defined with respect to additive approximations and the player utilities are normalized to $[0, 1]$ to make these approximations meaningful.

- A player’s CRV must only assign probability to strategies which are approximate best responses to his view.

It turns out that these conditions can be satisfied by using a careful dynamic program together with the structural understanding provided by [DKT15] and the CLT of Theorem 1. According to this structural result, we can partition the players into a “sparse” and a “Gaussian” component. Moreover, our CLT implies that matching the first two moments of the Gaussian suffices to approximate this component. This allows us to perform guesses at a different granularity for the sparse and Gaussian components. Roughly speaking, our dynamic program guesses a succinct representation of the two components and tries to compute CRVs which obey this representation and satisfy the conditions outlined above.

For more details on our PTAS, refer to Section 4.

Proper Covers. The second application of our CLT is to obtain proper covers for the set $\mathcal{S}_{n,k}$ of (n,k) -PMDs. A proper ε -cover of $\mathcal{S}_{n,k}$, in total variation distance, is a subset $\mathcal{S}_{n,k,\varepsilon} \subseteq \mathcal{S}_{n,k}$ such that for all $(X_1, \dots, X_n) \in \mathcal{S}_{n,k}$ there exists some $(Y_1, \dots, Y_n) \in \mathcal{S}_{n,k,\varepsilon}$ such that $d_{\text{TV}}(\sum_i X_i, \sum_i Y_i) \leq \varepsilon$. We show the following:

Theorem 3 (Proper Cover). *For all $n, k \in \mathbb{N}$, and $\varepsilon > 0$, there exists a proper ε -cover, in total variation distance, of the set of all (n,k) -PMDs whose size is*

$$n^{O(k)} \cdot \min \left\{ 2^{\text{poly}(k/\varepsilon)}, 2^{O(k^{5k} \log^{k+2}(1/\varepsilon))} \right\}. \quad (4)$$

Moreover, we can efficiently enumerate this cover in time polynomial in its size.

It is important to contrast Theorem 3 with Theorem 2 in [DKT15], which provides a non-proper cover whose size is similar, albeit with a leading factor of $n^{O(k^2)}$. Instead, our cover is proper, which is important for approximation algorithms that require searching over PMDs. Its dependence on n is also optimal, as the number of (n,k) -PMDs whose summands are deterministic is already $n^{\Omega(k)}$. Moreover, we provide a lower bound for the dependence on $1/\varepsilon$, establishing that the quasi-polynomial dependence is also essentially optimal.

Theorem 4 (Cover Size Lower Bound). *For any $n, k \in \mathbb{Z}$, $\varepsilon > 0$ such that $n > 2 \log^k(1/\varepsilon)$, there exist (n,k) -PMDs Z_1, \dots, Z_s such that for $1 \leq i < j \leq s$, $d_{\text{TV}}(Z_i, Z_j) \geq \varepsilon$ and $s = \Omega_k(n^{k-1} \cdot 2^{\tilde{\Omega}(\log^{k-1}(1/\varepsilon))})$. The $\tilde{\Omega}$ in the exponent hides factors of $\text{poly}(\log \log(1/\varepsilon))$ and dependence on k .*

We describe our proper cover construction in two parts. First, we give details on how to construct a non-proper cover of size $n^{O(k)}$. The main tool we use is the existence of spectral sparsifiers for Laplacian matrices. Our non-proper cover sparsifies the non-proper cover of [DKT15], showing how its leading factor of $n^{O(k^2)}$ can be reduced to $n^{O(k)}$. Roughly speaking, the factor of $n^{O(k^2)}$ was due to spectrally approximating all possible covariance matrices Σ , whose $O(k^2)$ entries are bounded by n . These covariance matrices corresponded to covariance matrices of (n,k) -PMDs, and the cover maintained for each such Σ some Σ' such that $|v^T(\Sigma - \Sigma')v| \leq \text{poly}(\varepsilon/k) \cdot v^T \Sigma v, \forall v$. (We call this guarantee a “ $\text{poly}(\varepsilon/k)$ -spectral approximation.”) The realization leading to our sparsification result is that covariance matrices of PMDs are in fact graph Laplacians. Indeed, a (n,k) -PMD, $X = \sum_i X_i$, has covariance matrix, $\text{cov}(X) = \sum_i \text{cov}(X_i)$, corresponding to the sum of the covariance matrices of its summands. Now the covariance matrix of a k -CRV, X_i , is actually the Laplacian of a graph that has one node j per dimension, along with an edge from node j to node j' of weight $\mathbf{E}[X_{ij}] \cdot \mathbf{E}[X_{ij}']$; and the covariance matrix of a (n,k) -PMD is the Laplacian of the graph with the sum of the weights from each constituent k -CRV— see Observation 1. We show

that Laplacians corresponding to (n, k) -PMDs can be $\text{poly}(\varepsilon/k)$ -spectrally covered with a set of covariance matrices of size $n^{O(k)} \cdot \left(\frac{k}{\varepsilon}\right)^{O(k^3)}$.

We appeal to recent results in spectral sparsification of Laplacian matrices [ST11, SS11, BSS12, BSST13]. In particular, we use the result of Batson, Spielman, and Srivastava [BSS12] (Theorem 8) to argue that the underlying graph can be sparsified to linearly many edges in the dimension k . We do this in the hopes that we would have fewer parameters in the covariance matrix to guess. Unfortunately, the [BSS12] sparsification theorem has polynomial dependence in the accuracy. So applying it with a $\text{poly}(\varepsilon/k)$ -approximation error, which is what we need, gives a meaningless result (namely no sparsification at all). Instead, we only use this theorem to get a rough $O(1)$ -spectral cover of (n, k) -PMD covariance matrices. Around every covariance matrix in this rough cover we grow a local $\text{poly}(\varepsilon/k)$ -spectral cover. Roughly speaking, as the $O(1)$ -spectral cover provides multiplicative approximation to the variance in every direction v , every covariance matrix in this cover gives us a multiplicative handle on the eigenvalues of the matrices approximated by it. This is sufficient information to cover these matrices to $\text{poly}(\varepsilon/k)$ -spectral error with a “local” spectral cover of size $(k/\varepsilon)^{O(k^2)}$ —see Lemma 6. Putting everything together, we get a $\text{poly}(\varepsilon/k)$ -spectral cover of all covariance matrices of (n, k) -PMDs of size $n^{O(k)} \cdot \left(\frac{k}{\varepsilon}\right)^{O(k^3)}$ —see Section 5.1.3. As covering these matrices was the bottleneck in the size of the non-proper cover, this completes the construction of a non-proper cover whose size is (4).

Further details on our non-proper construction are provided in Section 5.

We then show how to convert each element of this improper cover back to a PMD. We bypass the difficulty involved with a non-convex optimization problem by exploiting the “almost convexity” of the Minkowski sum as guaranteed by the Shapley-Folkman lemma. The cover provided by Theorem 7 is non-proper. It utilizes the structural result of [DKT15] (see Theorem 6) to cover the set of (n, k) -PMDs by hypotheses which take the form of the convolution of a discretized multidimensional Gaussian with a $(\text{poly}(k/\varepsilon), k)$ -PMD. The benefit of this class of hypotheses is that they have only $\text{poly}(k/\varepsilon)$ parameters. This allows us to efficiently enumerate over them, resulting in a cover size of (4). To convert this cover into a proper one, we need an algorithm which, given a convolution of a discretized Gaussian with some $(\kappa \triangleq \text{poly}(k/\varepsilon), k)$ -PMD, finds a (n, k) -PMD that is $O(\varepsilon)$ -close to this distribution, if such a PMD exists. As the (κ, k) -PMD is already a PMD, this boils down to answering whether a given discretized Gaussian with parameters (μ, Σ) is $O(\varepsilon)$ -close to a $(n - \kappa, k)$ -PMD. To answer this question, we exploit our new CLT (Theorem 1) and the fact that the discretized Gaussians that arise in the cover have an extra property: all their non-zero eigenvalues are at least $\text{poly}(k/\varepsilon)$ -large. Exploiting this we argue that (i) if there exists an $(n - \kappa, k)$ -PMD that is close to the discretized Gaussian with parameters (μ, Σ) , then its mean μ' should be close to μ and its covariance matrix Σ' should be spectrally close to Σ ; and (ii) if we can find any $(n - \kappa, k)$ -PMD with these properties, then it will be close to the discretized Gaussian. With (i) and (ii), our task becomes a convex geometry question: Let \mathcal{M} be all possible first two moments $(\mathbf{E}[Y], \mathbf{cov}(Y))$, of k -CRVs Y whose parameters have been finely discretized. As the first two moments of a $(n - \kappa, k)$ -PMD are sums of the first two moments of its constituent k -CRVs, we can reduce our problem to finding a point in the Minkowski sum $\mathcal{M}^{\oplus n - \kappa}$ that (spectrally) approximates the target (μ, Σ) . We write an LP to find a point in the convex hull of $\mathcal{M}^{\oplus n - \kappa}$ with this property, and the Shapley-Folkman theorem to “round” it into a point in $\mathcal{M}^{\oplus n - \kappa}$ that is only a little worse. The Shapley-Folkman theorem comes in handy because \mathcal{M} lives in $\mathbb{R}^{O(k^2)}$, i.e. much smaller dimension than $n - \kappa$. The whole approximation can be carried out in time $n^{O(k)}$ —see Lemma 8.

Details on this conversion process are provided in Section 6.

Our lower bound is described further in Section 7. Our technique shows a lower bound on

the metric entropy of a polynomial map of the moments of PMDs using an extension of Bézout’s theorem and other tools from algebraic geometry.

Learning. Finally, we give a new learning algorithm for PMDs:

Theorem 5. *For all $n, k \in \mathbb{N}$ and $\varepsilon > 0$, there is a learning algorithm for (n, k) -PMDs with the following properties: Let $X = \sum_{i=1}^n X_i$ be any (n, k) -Poisson multinomial random vector. The algorithm uses $\frac{\text{poly}(k, \log(1/\varepsilon))^k}{\varepsilon^2}$ samples from X , runs in time⁴ $\text{poly}(\frac{k}{\varepsilon})^{k^2}$ and with probability at least 9/10 outputs a (succinct description of a) random vector \tilde{X} such that $d_{\text{TV}}(X, \tilde{X}) \leq \varepsilon$.*

This improves the learning algorithm from [DKT15] by eliminating the superpolynomial dependence on ε in the running time that was obtained in that paper. Our algorithm exploits properties of the continuous Fourier transform of a PMD, as opposed to recent work by Diakonikolas, Kane and Stewart on learning univariate sums of independent integer random variables, which uses the discrete Fourier transform [DKS16b]. They also apply similar discrete Fourier techniques in their simultaneous work on PMDs [DKS16a].

We note that such Fourier-based learning algorithms may simply output a description of the Fourier transform of a distribution. This allows one to compute the PMF of the distribution at any point of interest, but it is not obvious how to sample from such a description. Our algorithm outputs an explicit description of a distribution, which allows one to efficiently (i.e., in time independent of n) draw samples from the distribution. In contrast, they output the Fourier transform of a distribution and describe how to sample from it.

For more details on our learning algorithm, refer to Section 8.

1.1 Comparison of Results with [DKS16a]

Simultaneous to our work, Diakonikolas, Kane, and Stewart also studied Poisson Multinomial distributions [DKS16a]. In this section, we describe and compare their results with ours. While both papers independently prove many qualitatively similar results, the techniques are quite different, and thus both may be of independent interest.

Both papers prove new CLTs, which manage to remove the dependence on n which is found in the CLT of [VV11], while the dependence on k and $1/\sigma$ remains polynomial. Additionally, both works improve upon the previous best covers for PMDs [DKT15]. First, both manage to reduce the size of the cover – interestingly, the two improvements seem to be orthogonal. Our result improves the dependence on n from n^{k^2} to $n^{O(k)}$, while theirs improves the dependence on k and $1/\varepsilon$ from $(1/\varepsilon)^{O(k^{5k} \log^{k+1}(1/\varepsilon))}$ to $(1/\varepsilon)^{O(k \log(k/\varepsilon)/\log \log(k/\varepsilon))^{k-1}}$.⁵ Furthermore, both papers describe how to efficiently achieve a proper cover of this size. These cover sizes are asymptotically optimal, as shown by lower bounds in both papers. In particular, the double-exponential dependence in k is necessary. Both works also consider the problem of finding approximate Nash equilibria in anonymous games. The complexity of both algorithms is roughly comparable to the PMD cover size. Finally, both papers study the learning of PMDs, obtaining algorithms with sample complexity $\text{poly}(k, \log(1/\varepsilon))^k/\varepsilon^2$. The runtime of our algorithm is $\text{poly}(k/\varepsilon)^{k^2}$, and the runtime of their algorithm is $\text{poly}(k, \log(1/\varepsilon))^k/\varepsilon^2 \cdot \log n$, both in the standard word RAM model.

⁴We work in the standard “word RAM” model in which basic arithmetic operations on $O(\log n)$ -bit integers are assumed to take constant time.

⁵We note that this upper bound holds for $k > 2$: for $k = 2$, [DKS16b] proves the tight cover size bound of $n \cdot (1/\varepsilon)^{\Theta(\log(1/\varepsilon))}$.

2 Preliminaries

2.1 Definitions

We more formally define several of the distribution classes we consider.

Definition 1. A k -Categorical Random Variable (k -CRV) is a random variable that takes values in $\{e_1, \dots, e_k\}$ where e_j is the k -dimensional unit vector along direction j . $\pi(i)$ is the probability of observing e_i .

Definition 2. An (n, k) -Poisson Multinomial Distribution ((n, k) -PMD) is given by the law of the sum of n independent but not necessarily identical k -CRVs. An (n, k) -PMD is parameterized by a nonnegative matrix $\pi \in [0, 1]^{n \times k}$ each of whose rows sum to 1 is denoted by M^π , and is defined by the following random process: for each row $\pi(i, \cdot)$ of matrix π interpret it as a probability distribution over the columns of π and draw a column index from this distribution. Finally, return a row vector recording the total number of samples falling into each column (the histogram of the samples).

We note that a sample from an (n, k) -PMD is redundant – given $k-1$ coordinates of a sample, we can recover the final coordinate by noting that the sum of all k coordinates is n . For instance, while a Binomial distribution is over a support of size 2, a sample is 1-dimensional since the frequency of the other coordinate may be inferred given the parameter n . With this inspiration in mind, we define the Generalized Multinomial Distribution, which is the primary object of study in [VV11].

Definition 3. A Truncated k -Categorical Random Variable is a random variable that takes values in $\{0, e_1, \dots, e_{k-1}\}$ where e_j is the $(k-1)$ -dimensional unit vector along direction j , and 0 is the $(k-1)$ dimensional zero vector. $\rho(0)$ is the probability of observing the zero vector, and $\rho(i)$ is the probability of observing e_i .

Definition 4. An (n, k) -Generalized Multinomial Distribution ((n, k) -GMD) is given by the law of the sum of n independent but not necessarily identical truncated k -CRVs. A GMD is parameterized by a nonnegative matrix $\rho \in [0, 1]^{n \times (k-1)}$ each of whose rows sum to at most 1 is denoted by G^ρ , and is defined by the following random process: for each row $\rho(i, \cdot)$ of matrix ρ interpret it as a probability distribution over the columns of ρ – including, if $\sum_{j=1}^k \rho(i, j) < 1$, an “invisible” column 0 – and draw a column index from this distribution. Finally, return a row vector recording the total number of samples falling into each column (the histogram of the samples).

For both (n, k) -PMDs and (n, k) -GMDs, we will refer to n and k as the *size* and *dimension*, respectively.

We note that a PMD corresponds to a GMD where the “invisible” column is the zero vector, and thus the definition of GMDs is more general than that of PMDs. However, whenever we refer to a GMD in this paper, it will explicitly have a non-zero invisible column.

While we will approximate the Multinomial distribution with Gaussian distributions, it does not make sense to compare discrete distributions with continuous distributions, since the total variation distance is always 1. As such, we must discretize the Gaussian distributions. We will use the notation $\lfloor x \rfloor$ to say that x is rounded to the nearest integer (with ties being broken arbitrarily). If x is a vector, we round each coordinate independently to the nearest integer.

Definition 5. The k -dimensional Discretized Gaussian Distribution with mean μ and covariance matrix Σ , denoted $\lfloor \mathcal{N}(\mu, \Sigma) \rfloor$, is the distribution with support \mathbb{Z}^k obtained by sampling according to the k -dimensional Gaussian $\mathcal{N}(\mu, \Sigma)$, and then rounding each coordinate to the nearest integer.

As seen in the definition of an (n, k) -GMD, we have one coordinate which is equal to n minus the sum of the other coordinates. We define a similar notion for a discretized Gaussian. However, we go one step further, to take care of when there are several such Gaussians which live in disjoint dimensions. By this, we mean that given two Gaussians, the set of directions in which they have a non-zero variance are disjoint. Without loss of generality (because we can simply relabel the dimensions), we assume all of a Gaussian's non-zero variance directions are consecutive, i.e., the covariance matrix is all zeros, except for a single block on the diagonal. Therefore, when we add the covariance matrices, the result is block diagonal. The resulting distribution is described in the following definition.

Definition 6. *The structure preserving rounding of a multidimensional Gaussian Distribution takes as input a multi-dimensional Gaussian $\mathcal{N}(\mu, \Sigma)$ with Σ in block-diagonal form. It chooses one coordinate as a “pivot” in each block, samples from the Gaussian ignoring these pivots and rounds each value to the nearest integer. Finally, the pivot coordinate of each block is set by taking the difference between the sum of the means and the sum of the values sampled within the block.*

Finally, we formally define the notion of a cover.

Definition 7. *An ε -cover for a set of distributions \mathcal{S} is a set of distributions \mathcal{S}' such that for any distribution $X \in \mathcal{S}$, there exists some distribution $Y \in \mathcal{S}'$ such that $d_{\text{TV}}(X, Y) \leq \varepsilon$. A cover is proper if $\mathcal{S}' \subseteq \mathcal{S}$.*

2.2 Probability Metrics

To compare probability distributions, we will require the total variation and Kolmogorov distances:

Definition 8. *The total variation distance between two probability measures P and Q on a σ -algebra F is defined by*

$$d_{\text{TV}}(P, Q) = \sup_{A \in F} |P(A) - Q(A)| = \frac{1}{2} \|P - Q\|_1.$$

Unless explicitly stated otherwise, in this paper, when two distributions are said to be ε -close, we mean in total variation distance.

Definition 9. *The Kolmogorov distance between two probability measures P and Q with CDFs F_P and F_Q is defined by*

$$d_{\text{K}}(P, Q) = \sup_{x \in \mathbb{R}} |F_P(x) - F_Q(x)|.$$

We note that Kolmogorov distance is, in general, weaker than total variation distance. In particular, total variation distance between two distributions is lower bounded by the Kolmogorov distance.

Fact 1. $d_{\text{K}}(P, Q) \leq d_{\text{TV}}(P, Q)$

2.3 Miscellaneous Lemmata

We will use the following tools for bounding total variation distance between various random variables.

Lemma 1 (Data Processing Inequality for Total Variation Distance). *Let X, X' be two random variables over a domain Ω . Fix any (possibly randomized) function F on Ω (which may be viewed as a distribution over deterministic functions on Ω) and let $F(X)$ be the random variable such that a draw from $F(X)$ is obtained by drawing independently x from X and f from F and then outputting $f(x)$ (likewise for $F(X')$). Then we have*

$$d_{\text{TV}}(F(X), F(X')) \leq d_{\text{TV}}(X, X').$$

Proposition 1 (Berry-Esseen theorem [Ber41, Ess42, She10]). *Let X_1, \dots, X_n be independent random variables, with $E[X_i] = 0, E[X_i^2] = \sigma_i^2 > 0, E[|X_i|^3] = \rho_i < \infty$, and define $X = \sum_{i=1}^n X_i, \sigma^2 = \sum_{i=1}^n \sigma_i^2, \rho = \sum_{i=1}^n \rho_i$. Then for an absolute constant $C_0 \leq 0.56$,*

$$d_K(X, \mathcal{N}(0, \sigma^2)) \leq \frac{C_0 \rho}{\sigma^3}.$$

Proposition 2 (Proposition 32 in [VV10]). *Given two k -dimensional Gaussians $\mathcal{N}_1 = \mathcal{N}(\mu_1, \Sigma_1), \mathcal{N}_2 = \mathcal{N}(\mu_2, \Sigma_2)$ such that for all $i, j \in [k]$, $|\Sigma_1(i, j) - \Sigma_2(i, j)| \leq \alpha$, and the minimum eigenvalue of Σ_1 is at least $\sigma^2 \geq \alpha$,*

$$d_{\text{TV}}(\mathcal{N}_1, \mathcal{N}_2) \leq \frac{\|\mu_1 - \mu_2\|_2}{\sqrt{2\pi\sigma^2}} + \frac{k\alpha}{\sqrt{2\pi e}(\sigma^2 - \alpha)}.$$

In addition, we prove the following general purpose lemma showing that two multivariate Gaussians with spectrally-close moments are close in total variation distance. This is intended to be a multivariate version of Proposition B.4 of [DDO⁺13], which proves a similar statement for univariate Gaussians. The proof appears in Section A.

Lemma 2. *Suppose there exist two k -dimensional Gaussians, $X \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $Y \sim \mathcal{N}(\mu_2, \Sigma_2)$, such that for all unit vectors v ,*

$$\begin{aligned} |v^T(\mu_1 - \mu_2)| &\leq \varepsilon s_v, \\ |v^T(\Sigma_1 - \Sigma_2)v| &\leq \frac{\varepsilon s_v^2}{2\sqrt{k}}; \end{aligned}$$

where $s_v^2 = \max\{v^T \Sigma_1 v, v^T \Sigma_2 v\}$. Then $d_{\text{TV}}(X, Y) \leq \varepsilon$.

2.4 Results on PMDs from [DKT15]

Our work builds upon recent structural results on PMDs [DKT15]. We recall some of the key results which we will refer to in this paper.

Two key parameters used in this paper are $c = c(\varepsilon, k) = \text{poly}(\varepsilon/k)$ and $t = t(\varepsilon, k) = \text{poly}(k/\varepsilon)$, set as $c = \left(\frac{\varepsilon^2}{k^5}\right)^{1+\delta_c}$ and $t = \left(\frac{k^{19}}{c\varepsilon^6}\right)^{1+\delta_t}$, for constants $\delta_c, \delta_t > 0$.

The main tool from this paper we will use is the structural characterization, stating that every PMD is close to the sum of an appropriately discretized Gaussian and a “sparse” PMD.

Theorem 6 (Theorem 5 from [DKT15]). *For parameters c and t as described above, every (n, k) -Poisson multinomial random vector is ε -close to the sum of a Gaussian with a structure preserving rounding and a (tk^2, k) -Poisson multinomial random vector. For each block of the Gaussian, the minimum non-zero eigenvalue of Σ_i is at least $\frac{tc}{2k^4}$.*

Finally, we will also use their rounding procedure, which relates a PMD to a nearby PMD with all parameters either equal to or sufficiently far from 0 and 1:

Lemma 3 (Lemma 1 from [DKT15]). *For any $c \leq \frac{1}{2k}$, given access to the parameter matrix ρ for an (n, k) -PMD M^ρ , we can efficiently construct another (n, k) -PMD $M^{\hat{\rho}}$, such that, for all i, j , $\hat{\rho}(i, j) \notin (0, c)$, and $d_{\text{TV}}(M^\rho, M^{\hat{\rho}}) < O\left(c^{1/2} k^{5/2} \log^{1/2}\left(\frac{1}{ck}\right)\right)$.*

3 A Size-Free CLT

We overview our proof of Theorem 1. Recall that the Central Limit Theorem of Valiant and Valiant, (1), has a poly-logarithmic dependence on the size parameter of the GMD. Their work raised the question whether this CLT could be made size-independent, and we resolve this conjecture by showing that it can be. This qualitative improvement comes at a quantitative loss in the polynomial dependence of the bound on the parameters k and σ^2 .

Our CLT builds off of the structural result of [DKT15], Theorem 6, which we use as a black box. This structural result says that every (n, k) -PMD is ε -close to the sum of an appropriately discretized Gaussian and a $(\text{poly}(k/\varepsilon), k)$ -PMD. We note that the statement of Theorem 6 does not tell us anything about the moments of this Gaussian and sparse PMD, while our new CLT requires that the discretized Gaussian has the same moments as the original PMD. We prove this CLT in two steps. First, we show that the original PMD X and the discretized Gaussian from the cover G are close in total variation distance, i.e., we show that we can “drop” the sparse PMD component from Theorem 6 in the relevant approximation regime. Then, we bound the distance between the discretized Gaussian from the cover, G , and a discretized Gaussian with the same mean and covariance as the original PMD, G_X . The proof is concluded by combining these two bounds using the triangle inequality.

To bound the distance between the original PMD X and the discretized Gaussian from the cover G , we start by invoking Theorem 6 with parameter $\varepsilon = \text{poly}(k/\sigma)$. This tells us that the PMD is close to the sum of a discretized Gaussian with a structure preserving rounding G and a “sparse” PMD P , which has size parameter at most some $\text{poly}(\sigma) = o(\sigma)$. We first show that the structure preserving rounding only has a single block in its structure. This is proved by contradiction. If there were multiple blocks in the structure, there would exist some direction v in which G contributes 0 variance. Since P is sparse, it can contribute at most $o(\sigma)$ variance when projected in direction v . However, we know that X had at least σ^2 variance in direction v . By projecting both X and P in direction v and applying Berry-Esseen’s theorem, we can show that such a large discrepancy in the variance implies large Kolmogorov distance between the projections, see Proposition 5. This acts as a certificate demonstrating a large total variation distance, contradicting our invocation of Theorem 6, and thus the Gaussian has a single block in its structure. By a similar contradiction argument, we can also argue that G has a large variance ($\Omega(\sigma^2)$) when projected in any direction. Since G ’s variance is at least $\Omega(\sigma^2)$ in any direction, while P is only supported over $\{0, \dots, o(\sigma)\}^k$, it can be shown that P ’s contribution to the distribution is negligible using Proposition 6, and thus we can remove it at low cost; i.e. $d_{\text{TV}}(G + P, G)$ is small. Since Theorem 6 implied that $d_{\text{TV}}(X, G + P)$ was small, by triangle inequality, we have shown that the original PMD X and the discretized Gaussian from the cover G are close in total variation distance.

Next, we bound the distance between the discretized Gaussian from the cover, G , and a discretized Gaussian with the same moments as the original PMD, G_X . At this point, we know that X and G are close in total variation distance. By projecting both distributions in some direction and considering true Gaussians with the same moments as X and G , it can be shown that the first two moments are similar in this direction – otherwise, the true Gaussians would be far from each other in the Kolmogorov metric. This implies that the first two moments of X and G are close in *every* direction, as guaranteed by Proposition 8. Applying Lemma 2 tells us that bona-fide Gaussians with moments which are close in every direction are therefore close in total variation distance. The proof is concluded by applying the Data Processing inequality, which shows that the corresponding discretized Gaussians G and G_X are close as well.

We state and prove many useful lemmas in Section 3.1, which we combine to complete the proof of Theorem 1 in Section 3.2.

3.1 Useful Lemmas

The following two propositions bound the Kolmogorov distance between a univariate Gaussian and the projection of a GMD or a discretized Gaussian, respectively.

Proposition 3. *Suppose that there exists an (n, k) -generalized multinomial random vector X , with mean vector μ and covariance matrix Σ . Then for any unit vector v ,*

$$d_K(v^T X, \mathcal{N}(v^T \mu, v^T \Sigma v)) \leq \frac{1}{\sigma},$$

where σ^2 is the minimum eigenvalue of Σ .

Proof. We apply the Berry-Esseen theorem (Proposition 1). Let $Y_i = X_i - E[X_i]$ to recenter the random variables, and we will now compare $Y = \sum_i Y_i$ with $\mathcal{N}(0, v^T \Sigma v)$. We note that $v^T Y_i \in [-\sqrt{2}, \sqrt{2}]$. Letting $\sigma_i^2 = \mathbf{Var}(v^T Y_i)$ and $\rho_i = E[|v^T Y_i|^3]$, this implies that $\rho_i \leq \sqrt{2}\sigma_i^2$, and thus the Berry-Esseen bound gives

$$d_K(v^T Y, \mathcal{N}(0, v^T \Sigma v)) \leq \frac{0.56 (\sum_i \rho_i)}{(\sum_i \sigma_i^2)^{3/2}} \leq \frac{(\sum_i \sigma_i^2)}{(\sum_i \sigma_i^2)^{3/2}} = \frac{1}{(\sum_i \sigma_i^2)^{1/2}} \leq \frac{1}{\sigma}.$$

□

Proposition 4. *Suppose there exists a random variable $X \sim \lfloor \mathcal{N}(\mu, \Sigma) \rfloor$. Then for any unit vector v ,*

$$d_K(v^T X, \mathcal{N}(v^T \mu, v^T \Sigma v)) \leq \frac{\sqrt{k}}{\sqrt{2\pi}\sigma},$$

where σ^2 is the minimum eigenvalue of Σ .

Proof. Let $Y \sim \mathcal{N}(\mu, \Sigma)$. We first show $|v^T(Y - \lfloor Y \rfloor)| \leq \frac{\sqrt{k}}{2}$, which holds by Cauchy-Schwarz: $\|v\|_2 = 1$ and $\|Y - \lfloor Y \rfloor\|_2 \leq \sqrt{k} \cdot \|Y - \lfloor Y \rfloor\|_\infty \leq \frac{\sqrt{k}}{2}$. Thus,

$$v^T Y - \frac{\sqrt{k}}{2} \leq v^T \lfloor Y \rfloor \leq v^T Y + \frac{\sqrt{k}}{2}.$$

Using F to denote the corresponding CDFs, this stochastic dominance condition implies that for any $y \in \mathbb{R}$,

$$F_{v^T Y - \frac{\sqrt{k}}{2}}(y) \leq F_{v^T \lfloor Y \rfloor}(y) \leq F_{v^T Y + \frac{\sqrt{k}}{2}}(y).$$

Furthermore,

$$F_{v^T Y - \frac{\sqrt{k}}{2}}(y) \leq F_{v^T Y}(y) \leq F_{v^T Y + \frac{\sqrt{k}}{2}}(y)$$

and

$$F_{v^T Y + \frac{\sqrt{k}}{2}}(y) - F_{v^T Y - \frac{\sqrt{k}}{2}}(y) \leq \sqrt{k} \cdot \frac{1}{\sqrt{2\pi}\sigma},$$

because the two distributions are univariate Gaussians with the same variance (which is at least σ^2) and means shifted by \sqrt{k} . This implies

$$|F_{v^T Y}(y) - F_{v^T \lfloor Y \rfloor}(y)| \leq \frac{\sqrt{k}}{\sqrt{2\pi}\sigma},$$

as desired. □

The following proposition compares a Gaussian X and an arbitrary distribution Y . It shows that if Y 's variance is much smaller than X 's, then they must be far in Kolmogorov distance.

Proposition 5. *Suppose there exists a univariate Gaussian X with variance σ_X^2 , and a distribution Y with variance $\sigma_Y^2 < \sigma_X^2$. Then the Kolmogorov distance between X and Y is at least $\frac{1}{2} - \left(\frac{\sigma_Y}{\sigma_X}\right)^{2/3}$.*

Proof. We consider the event that a sample falls in an interval of width $2k$ centered at $E[Y]$. As a certificate of a large Kolmogorov distance between X and Y , we show that the probability assigned to this interval is very different for X versus Y .

First, by Chebyshev's inequality, we know that

$$\Pr[|Y - E[Y]| \leq k] \geq 1 - \frac{\sigma_Y^2}{k^2}.$$

On the other hand, we know that

$$\Pr[|X - E[Y]| \leq k] \leq \Pr[|X - E[X]| \leq k] = \text{erf}\left(\frac{k}{\sqrt{2}\sigma_X}\right) \leq \frac{k}{\sqrt{2\pi}\sigma_X},$$

where the last inequality uses the Taylor expansion of the error function.

The difference in probability assigned to this interval is at least

$$1 - \frac{\sigma_Y^2}{k^2} - \frac{k}{\sqrt{2\pi}\sigma_X}.$$

Setting $k = \sigma_Y^{2/3} \sigma_X^{1/3}$ gives

$$d_K(X, Y) \geq \frac{1}{2} \left(1 - \left(\frac{\sigma_Y}{\sigma_X}\right)^{2/3} - \frac{1}{\sqrt{2\pi}} \left(\frac{\sigma_Y}{\sigma_X}\right)^{2/3} \right) \geq \frac{1}{2} - \left(\frac{\sigma_Y}{\sigma_X}\right)^{2/3},$$

as desired. \square

The following proposition tells us if we are considering the sum of two random variables, one being a Gaussian with a large variance and one being an arbitrary distribution with a small support, we can remove all contribution from the distribution with small support and not pay a large cost in total variation distance.

Proposition 6. *Suppose X and Y are independent random variables, where $X \sim [\mathcal{N}(\mu, \Sigma)] \in \mathbb{R}^k$ and Y is supported on $S = \{0, \dots, m\}^k$. Then $d_{\text{TV}}(X, X + Y) \leq \frac{m\sqrt{k}}{\sqrt{2\pi}\sigma}$, where σ is the minimum eigenvalue of Σ .*

Proof. We start by applying a law of total probability for total variation distance:

$$d_{\text{TV}}(X, X + Y) \leq \sum_{v \in S} \Pr(Y = v) d_{\text{TV}}(X, X + v) = \sum_{v \in S} \Pr(Y = v) d_{\text{TV}}([\mathcal{N}(\mu, \Sigma)], [\mathcal{N}(\mu + v, \Sigma)]).$$

Using the data processing inequality for total variation distance (Lemma 1):

$$d_{\text{TV}}([\mathcal{N}(\mu, \Sigma)], [\mathcal{N}(\mu + v, \Sigma)]) \leq d_{\text{TV}}(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\mu + v, \Sigma)) \leq \frac{\|v\|}{\sqrt{2\pi}\sigma} \leq \frac{m\sqrt{k}}{\sqrt{2\pi}\sigma},$$

where the second last inequality follows from Proposition 2. We conclude by observing that $d_{\text{TV}}(X, X + Y)$ is a convex combination of such terms. \square

The next proposition tells us that Kolmogorov closeness implies parameter closeness for univariate Gaussians.

Proposition 7. *Consider two univariate Gaussians $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ where $\sigma_1 \leq \sigma_2$. For any $\alpha \in (0, 1)$, if $d_K(X, Y) \leq \frac{\alpha}{10}$, then $|\mu_2 - \mu_1| \leq \alpha\sigma_1$ and $|\sigma_2^2 - \sigma_1^2| \leq 3\alpha\sigma_1^2$.*

Proof. We start by proving the following statement: For any $\alpha \in (0, 1)$, if $|\mu_2 - \mu_1| \geq \alpha\sigma_1$ or $|\sigma_2 - \sigma_1| \geq \alpha\sigma_1$, then $d_K(X, Y) \geq \frac{\alpha}{10}$. The proof follows by contraposition, and observing that multiplying both sides of $|\sigma_2 - \sigma_1| \leq \alpha\sigma_1$ by $(\sigma_2 + \sigma_1)$, bounding $\sigma_2 \leq (1 + \alpha)\sigma_1$, and $\alpha \leq 1$ imply $|\sigma_2^2 - \sigma_1^2| \leq 3\alpha\sigma_1^2$.

Without loss of generality, assume $\mu_1 \leq \mu_2$. We will first show the conclusion assuming the means are separated, and then assuming the variances are separated.

Suppose $|\mu_2 - \mu_1| \geq \alpha\sigma_1$. Consider the point $x = \mu_2$. At this point, the CDF of the second Gaussian is equal to $\frac{1}{2}$. The CDF of the first Gaussian is $\frac{1}{2} \left(1 + \operatorname{erf}\left(\frac{\mu_2 - \mu_1}{\sqrt{2}\sigma_1}\right)\right) \geq \frac{1}{2} \left(1 + \operatorname{erf}\left(\frac{\alpha}{\sqrt{2}}\right)\right)$. Therefore, $d_K(\mathcal{N}_1, \mathcal{N}_2) \geq \frac{1}{2} \operatorname{erf}\left(\frac{\alpha}{\sqrt{2}}\right) \geq \frac{\alpha}{10}$, where the last inequality holds for all $\alpha \in (0, 1)$.

Now, suppose $|\sigma_2 - \sigma_1| \geq \alpha\sigma_1$. Consider the point $x = \mu_1 + \sqrt{2}\sigma_1$. At this point, the CDF of the first Gaussian equal to $\frac{1}{2}(1 + \operatorname{erf}(1))$. Similarly, the CDF of the second Gaussian is at most $\frac{1}{2} \left(1 + \operatorname{erf}\left(\frac{\sigma_1}{\sigma_2}\right)\right) \leq \frac{1}{2} \left(1 + \operatorname{erf}\left(\frac{1}{1+\alpha}\right)\right)$. Therefore, $d_K(\mathcal{N}_1, \mathcal{N}_2) \geq \frac{\operatorname{erf}(1) - \operatorname{erf}(\frac{1}{1+\alpha})}{2} \geq \frac{\alpha}{10}$ where the last inequality holds for all $\alpha \in (0, 1)$. \square

Our final proposition in this section applies the previous proposition, showing that total variation closeness implies parameter closeness (in any projection) when considering a GMD and a discretized Gaussian.

Proposition 8. *Suppose X is an (n, k) -GMD, and Y is a k -dimensional discretized Gaussian such that $d_{TV}(X, Y) \leq \alpha$. Let μ_X and Σ_X be the mean vector and covariance matrix (respectively) of X , and define μ_Y and Σ_Y similarly for Y . For a unit vector v , let $\sigma_v^2 = \min\{v^T \Sigma_X v, v^T \Sigma_Y v\}$, and let $\sigma^2 = \min_v \sigma_v^2$. If $\alpha + \frac{2\sqrt{k}}{\sigma} \leq 1/10$, then for all unit vectors v*

$$|v^T(\mu_X - \mu_Y)| \leq 10 \left(\alpha + \frac{2\sqrt{k}}{\sigma} \right) \sigma_v;$$

$$|v^T(\Sigma_X - \Sigma_Y)v| \leq 30 \left(\alpha + \frac{2\sqrt{k}}{\sigma} \right) \sigma_v^2.$$

Proof. Consider the projections of X and Y onto v . By Propositions 3 and 4 and the triangle inequality, the Kolmogorov distance between the univariate Gaussians with the same mean and variance is at most $\alpha + \frac{2\sqrt{k}}{\sigma}$. Applying Proposition 7 implies the desired result. \square

3.2 Proof of Theorem 1

We will prove the statement for a sufficiently large constant C . Thus we only need examine the case

$$\frac{k^3}{\sigma^{1/10}} \leq \frac{1}{C}, \tag{5}$$

otherwise the conclusion of the theorem statement is vacuous since total variation distance is at most 1.

As a starting point, we convert from a GMD to the corresponding (n, k) -Poisson multinomial random vector X and apply Theorem 6 with $\varepsilon = \frac{k^3}{\sigma^{1/10}}$. This gives us that

$$d_{\text{TV}}(X, G + P) \leq \frac{k^3}{\sigma^{1/10}},$$

where G is a Gaussian with a structure preserving rounding and P is a (tk^2, k) -Poisson multinomial random vector. By the definition of t in Section 2.4, we have that $t \leq \frac{C'\sigma^{9/10}}{k^2}$ for some constant C' . Thus, P is a $(C'\sigma^{9/10}, k)$ -Poisson multinomial random vector.

First, we argue that the Gaussian component G only has a single block in its structure. We prove this by contradiction – suppose there exist multiple blocks in its structure. Let one of the pivots be the pivot coordinate for the GMD, and ignore this dimension. If there are multiple blocks, the rounding procedure implies that there exists a direction v in which the variance of the resulting covariance matrix of the Gaussian is 0. In direction v , the maximum possible value for the variance of P is $\frac{C'\sigma^{9/10}}{4}$, giving us an upper bound for the variance of $G + P$. However, we know that the variance of X in direction v is at least σ^2 , by the assumption in the theorem statement. By Proposition 3, projecting X in direction v and converting to a univariate Gaussian X_g with the same mean and variance incurs a cost of at most $\frac{1}{\sigma}$ in Kolmogorov distance. Also projecting $G + P$ in direction v , Proposition 5 tells us that $d_K(v^T X, v^T(G + P)) \geq d_K(X_g, v^T(G + P)) - d_K(v^T X, X_g) \geq \frac{1}{2} - \left(\frac{C'}{4\sigma^{11/10}}\right)^{1/3} - \frac{1}{\sigma}$. Because $\sigma \geq C^{10}$ (as assumed in (5)), we have that $d_K(v^T X, v^T(G + P)) > \frac{1}{3}$. Since we know $d_{\text{TV}}(X, G + P) \leq \frac{k^3}{\sigma^{1/10}}$, this implies that $d_K(v^T X, v^T(G + P)) \leq d_{\text{TV}}(v^T X, v^T(G + P)) \leq \frac{k^3}{\sigma^{1/10}}$ should also hold, which is a contradiction for large C , as $\frac{k^3}{\sigma^{1/10}} \leq \frac{1}{C} < \frac{1}{3}$. Therefore, the Gaussian component G only has a single block in its structure.

Since we have established that the Gaussian component G only has a single block, we will convert back to the original GMD domain for the remainder of the proof. Recall that the original GMD is M^ρ , and we let D be the discretized Gaussian and S be the $(C'\sigma^{9/10}, k)$ -Generalized multinomial random vector with the same pivot coordinate as M^ρ . Now, we wish to upper bound $d_{\text{TV}}(M^\rho, D)$, i.e., we want to eliminate the sparse GMD from our statement. First, we wish to argue that D has a large variance in every direction, and thus removing S will not have a large effect. This is done by the same method in the above paragraph. Let the minimum variance of D in any direction be ζ^2 . Then to avoid the same contradiction as above, we require that

$$\frac{1}{2} - \left(\frac{\frac{C'\sigma^{9/10}}{4} + \zeta^2}{\sigma^2}\right)^{1/3} - \frac{1}{\sigma} \leq \frac{1}{C}.$$

This can be manipulated to show that

$$\zeta^2 \geq \frac{1}{16}\sigma^2. \quad (6)$$

Now, applying Proposition 6 and the triangle inequality, we get

$$d_{\text{TV}}(M^\rho, D) \leq \frac{k^3}{\sigma^{1/10}} + \frac{4C'\sqrt{k}}{\sqrt{2\pi}\sigma^{1/10}}. \quad (7)$$

Finally, to conclude, we must compare D with a discretized Gaussian with the same moments as M^ρ , i.e., we wish to upper bound $d_{\text{TV}}(D, [\mathcal{N}(\mu, \Sigma)])$. Recall that μ and Σ are the mean and covariance of M^ρ , and let μ_D and Σ_D be the mean and covariance of D . Apply Proposition 8 to

M^ρ and D using the guarantees of Equations (6) and (7). This implies that their moments are close:

$$|v^T(\mu - \mu_D)| \leq 10 \left(\frac{k^3}{\sigma^{1/10}} + \frac{4C'\sqrt{k}}{\sqrt{2\pi}\sigma^{1/10}} + \frac{8\sqrt{k}}{\sigma} \right) \sigma_v;$$

$$|v^T(\Sigma - \Sigma_D)v| \leq 30 \left(\frac{k^3}{\sigma^{1/10}} + \frac{4C'\sqrt{k}}{\sqrt{2\pi}\sigma^{1/10}} + \frac{8\sqrt{k}}{\sigma} \right) \sigma_v^2,$$

where $\sigma_v^2 = \min\{v^T \Sigma v, v^T \Sigma_D v\}$.

We use the Data Processing Inequality (Lemma 1) followed by Lemma 2 with these guarantees to give:

$$d_{\text{TV}}(D, \lfloor \mathcal{N}(\mu, \Sigma) \rfloor) \leq d_{\text{TV}}(\mathcal{N}(\mu_D, \Sigma_D), \mathcal{N}(\mu, \Sigma)) \leq 60 \left(\frac{k^3}{\sigma^{1/10}} + \frac{4C'\sqrt{k}}{\sqrt{2\pi}\sigma^{1/10}} + \frac{8\sqrt{k}}{\sigma} \right) \sqrt{k}.$$

Finally, applying the triangle inequality with Equation (7) gives

$$d_{\text{TV}}(M^\rho, \lfloor \mathcal{N}(\mu, \Sigma) \rfloor) \leq d_{\text{TV}}(M^\rho, D) \leq 61 \left(\frac{k^{7/2}}{\sigma^{1/10}} + \frac{4C'k}{\sqrt{2\pi}\sigma^{1/10}} + \frac{8k}{\sigma} \right).$$

Choosing the constant C sufficiently large completes the proof.

4 A PTAS for Anonymous Games

Here, we overview the algorithm of Theorem 2. The algorithm starts with a guess of $X = \sum_i X_i$ at a Nash equilibrium $V_X = (X_1, \dots, X_n)$ of the game, where X_i represents the mixed strategy of player i . While there are infinitely many X 's to guess, our proper cover theorem (Theorem 3) implies that every X can be approximated by some $Y = \sum_i Y_i$, where $V_Y = (Y_1, \dots, Y_n) \in S_\varepsilon$, $d_{\text{TV}}(X, Y) \leq \varepsilon$ and $|S_\varepsilon|$ is of the order of at most (3). What we would like to claim is that if Y approximates X , then V_Y is an approximate Nash equilibrium of the game up to a permutation of the Y_i 's. This is unfortunately not necessarily true, but the following guarantees would suffice:

$$\forall i : \text{support}(Y_i) \subseteq \text{support}(X_i) \quad \wedge \quad d_{\text{TV}} \left(\sum_{j \neq i} X_j, \sum_{j \neq i} Y_j \right) \leq \varepsilon. \quad (8)$$

Indeed, if the above guarantee held, then the expected payoff of every player i from any pure strategy σ would not change by more than an additive $O(\varepsilon)$ if we changed the strategies of all other players from $(X_j)_{j \neq i}$ to $(Y_j)_{j \neq i}$. So, if V_X were a Nash equilibrium and $\text{support}(Y_i) \subseteq \text{support}(X_i)$, it would follow that Y_i is an approximate best response of player i to $(Y_j)_{j \neq i}$. So V_Y would be an approximate equilibrium.

Unfortunately, we do not know how to construct a proper ε -cover S_ε of all (n, k) -PMDs that has size of order (3) and such that for any V_X there exists some $V_Y \in S_\varepsilon$ satisfying Condition (8). Nevertheless, we can exploit our CLT and the structural result of [DKT15] (restated as Theorem 6 in this paper) to bypass this difficulty. Roughly speaking [DKT15] approximate a given $V_X = (X_1, \dots, X_n)$ by first discretizing the parameters of all X_i 's into fine enough accuracy (this is shown to only cost some $O(\varepsilon)$ in total variation distance), then partitioning the X_i 's into a small group \mathcal{L} of size $\text{poly}(k/\varepsilon)$ that are left intact, and a large group whose sum is approximated by a discretized multidimensional Gaussian (up to another cost of $O(\varepsilon)$ in total variation distance). It

is further shown that the distribution of the sum of variables in \mathcal{L} can be summarized through the vector \vec{m} of its first $O(\log 1/\varepsilon)$ moments (at a loss of an additional $O(\varepsilon)$ in total variation distance), while the discretized Gaussian through its first two moments (μ, Σ) . Moreover, it is shown that the Gaussian has at least $\text{poly}(k/\varepsilon)$ variance in all directions where it has non-zero variance. By enumerating over all possible summary statistics (\vec{m}, μ, Σ) , a non-proper cover of all (n, k) -PMDs can be obtained, whose size is of the order of (3).

Suppose now that $V_X = (X_1, \dots, X_n)$ is a Nash equilibrium whose approximating statistic in the non-proper cover is some (\vec{m}, μ, Σ) . Given a correct guess for this statistic, our goal is to uncover an approximate Nash equilibrium $V_Y = (Y_1, \dots, Y_n)$ of the game. By the construction of the cover, we know that every player i either contributed his discretized X_i to the discretized Gaussian with parameters (μ, Σ) , or to the small group of variables with moments \vec{m} . So, letting \mathcal{C} be the set of k -CRVs whose parameters have the discretization accuracy used in the construction of the cover, we need to assign some $Y_i \in \mathcal{C}$ to each player i such that:

- (a) There exists a $\text{poly}(k/\varepsilon)$ -size subset \mathcal{L} of players such that $\sum_{i \in \mathcal{L}} Y_i$ has vector of moments \vec{m} , while $\sum_{i \notin \mathcal{L}} Y_i$ has first two moments (μ, Σ) .
- (b) For all i , Y_i is a best response to $\sum_{j \neq i} Y_j$.

To find a good assignment, we first construct a compatibility graph between players and mixed strategies in \mathcal{C} . We add an edge between some i and some $Y_i \in \mathcal{C}$ iff at least one of the following two conditions is met. We also annotate the edge with all conditions that are met:

1. (Y_i is compatible with $i \in \mathcal{L}$): Y_i is an approximate best response to the “environment” i would observe if i contributed to \vec{m} . If i contributed to \vec{m} and Condition (a) were met, then we can deduce what PMD player i would see in his environment. Indeed, this would be within some $O(\varepsilon)$ in total variation distance to the sum of a Gaussian random vector with parameters (μ, Σ) and a PMD whose first $O(\log(1/\varepsilon))$ moments are the same as \vec{m} after removing the contribution of Y_i . The updated moment vector can be computed from \vec{m} and Y_i as moments are symmetric polynomials of the underlying parameters. Given the updated moment vector, the PMD is determined to within ε in total variation distance, so its sum with the discretized Gaussian is also determined, and we can also efficiently determine whether Y_i is an approximate best response of player i to that distribution.
2. (Y_i is compatible with $i \in \bar{\mathcal{L}}$): Y_i is an approximate best response to the “environment” i would observe if i contributed to the discretized Gaussian with parameters (μ, Σ) . First, for this to be the case Y_i must be “compatible” with Σ , i.e. not correlating uncorrelated pairs of dimensions/adding variance in zero-variance dimensions (or in other words, the block structure of Σ should be preserved). Moreover, since all non-zero eigenvalues of Σ are at least $\text{poly}(k/\varepsilon)$ -large, the discretized Gaussian with parameters (μ, Σ) and $(\mu - \mathbf{E}[Y_i], \Sigma - \mathbf{cov}(Y_i))$ are approximately the same (Proposition 2). At the same time, due to the largeness of the non-zero eigenvalues of Σ , if condition (a) were eventually true, then our CLT (Theorem 1) would imply that $\sum_{j \in \bar{\mathcal{L}} \setminus \{i\}} Y_j$ is well-approximated by the discretized Gaussian with parameters $(\mu - \mathbf{E}[Y_i], \Sigma - \mathbf{cov}(Y_i))$, and hence by that with parameters (μ, Σ) . So, if $i \in \bar{\mathcal{L}}$, i is assigned Y_i , and Condition (a) is eventually met, then the PMD that player i sees in his environment is pinned down to within $O(\varepsilon)$ in total variation distance: it is approximately the sum of the discretized Gaussian with parameters (μ, Σ) and a PMD with moments \vec{m} . We can therefore check if Y_i is an approximate best response to that distribution.

After constructing the compatibility graph as above, we need to see if there is an assignment of players to compatible mixed strategies from \mathcal{C} so that (a) is satisfied. This looks non-trivial, but

it can be done using dynamic programming. We sweep through the players, maintaining as state all possible leftover moments $(\vec{m}', \mu', \Sigma')$ that may arise from assignments of a prefix of players to compatible mixed strategies. Given the discretization of \mathcal{C} , the set of possible states is bounded by (3). Importantly, the compatibility graph has the property that player i is happy when given a compatible strategy as long as the overall assignment matches (\vec{m}, μ, Σ) .

4.1 Preliminaries for Anonymous Games

Definition 10. An anonymous game is a triple $G = (n, k, \{u_j^i\})$ where $[n] = \{1, \dots, n\}$, $n \geq 2$, is the set of players, $[k] = \{1, \dots, k\}$, $k \geq 2$, is the set of strategies, and u_j^i with $i \in [n]$ and $j \in [k]$ is the utility of player i when she plays strategy j , a function mapping the set of partitions $\Pi_{n-1}^k = \{(x_1, \dots, x_k) : x_i \in \mathbb{N}_0 \text{ for all } i \in [k], \sum_{i=1}^k x_i = n-1\}$ to the interval $[0, 1]$.

A mixed strategy profile ρ is a set of n distributions $\{\rho_i \in \Delta^k\}_{i \in [n]}$, where by Δ^k we denote the $(k-1)$ -dimensional simplex, or, equivalently, the set of distributions over $[k]$. A mixed strategy profile ρ is an ε -approximately well supported Nash equilibrium (or an ε -Nash equilibrium, for short) if, for all $i \in [n]$ and $j \in [k]$,

$$E_{x \sim \rho_{-i}}[u_j^i(x)] < \max_{j' \in [k]} E_{x \sim \rho_{-i}}[u_{j'}^i(x)] - \varepsilon \Rightarrow \rho_{ij} = 0,$$

where ρ_{-i} is the distribution over Π_{n-1}^k obtained by drawing $n-1$ random samples from $[k]$ independently according to the distributions $\rho_{i'}, i' \neq i$, and forming the induced partition. We note that this an ε -Nash equilibrium is stronger than the related concept of an ε -approximate Nash equilibrium (see, i.e., [DGP09] for further discussion of this distinction). Throughout this paper, we solely consider the harder problem of computing an ε -Nash equilibrium.

A 0-Nash equilibrium is simply called a *Nash equilibrium* and it is always guaranteed to exist by Nash's theorem.

4.2 An Algorithm for Anonymous Games

In a Nash equilibrium ρ of an anonymous game every player uses a mixed strategy ρ_i selecting strategy j with probability ρ_{ij} . The distribution of the number of players which select each of the strategies is an (n, k) -PMD. Using the fact that there exist small size ε -covers for PMDs, we can efficiently search over the space of all strategies and identify a mixed strategy profile that produces an ε -Nash equilibrium. We show that there exists an efficient polynomial time approximation scheme (EPTAS) for computing an ε -Nash equilibrium, thus proving Theorem 2.

The algorithm works by guessing an aggregate statistic (m, μ, Σ) that describes the overall behavior of all players. This statistic is based on the structural theorem shown in [DKT15], which shows that the overall PMD that describes the mixed strategy profile can be approximately written as sum of a discretized Gaussian and a sparse PMD with only $\text{poly}(k/\varepsilon)$ components. Moreover, for the sparse PMD knowledge of the $\log(1/\varepsilon)$ moments (which is equivalent to knowing the power-sums of all the summands up to $\text{poly}(1/\varepsilon)$), suffices to describe it within ε in total variation distance. Thus, the algorithm requires guessing the power-sums m of the sparse PMD and the mean μ and covariance Σ of the discretized Gaussian.

As we will show, knowledge of an individual's strategy together with the aggregate statistic (m, μ, Σ) for the overall mixed strategy profile, allows us to compute an approximate distribution D_i that describes the player's view about the aggregate strategy of everyone else. If we manage to assign strategies ρ_i to every player so that ρ_{-i} approximately matched D_i and additionally each player only chooses strategies that corresponds to approximate best responses with respect to his

view D_i we will obtain an ε -Nash equilibrium. The following lemma formalizes this intuition and is the main tool we use in the proof of Theorem 2.

Lemma 4. *Consider the anonymous game $G = (n, k, \{u_j^i\})$ and let D_1, D_2, \dots, D_n be arbitrary distributions over \mathbb{Z}^k . If there exists an (n, k) -PMD ρ such that:*

- *For all $i \in [n]$, $d_{\text{TV}}(\rho_{-i}, D_i) \leq \varepsilon_1$*
- *For all $i \in [n]$ and $j \in [k]$, $E_{x \sim D_i}[u_j^i(x)] < \max_{j' \in [k]} E_{x \sim D_i}[u_{j'}^i(x)] - \varepsilon_2 \Rightarrow \rho_{ij} = 0$,*

Then, ρ is an $(2\varepsilon_1 + \varepsilon_2)$ -Nash equilibrium for the game G .

Proof. For any $i \in [n]$ and $j \in [k]$, we have that $|E_{x \sim D_i}[u_j^i(x)] - E_{x \sim \rho_{-i}}[u_j^i(x)]| \leq \varepsilon_1$, since $i \in [n]$, $d_{\text{TV}}(\rho_{-i}, D_i) \leq \varepsilon_1$. Therefore,

$$\begin{aligned} \max_{j' \in [k]} E_{x \sim \rho_{-i}}[u_{j'}^i(x)] - E_{x \sim \rho_{-i}}[u_j^i(x)] &> \varepsilon_2 + 2\varepsilon_1 \Rightarrow \\ \max_{j' \in [k]} E_{x \sim D_i}[u_{j'}^i(x)] - E_{x \sim D_i}[u_j^i(x)] &> \varepsilon_2 \Rightarrow \rho_{ij} = 0 \end{aligned}$$

□

Proof of Theorem 2. Consider the game $G = (n, k, \{u_j^i\})$. By Nash's theorem there always exists a Nash equilibrium. Let ρ be such an equilibrium where every player uses a mixed strategy ρ_i selecting strategy j with probability ρ_{ij} . The distribution of vectors which give the number of players which select each of the strategies is an (n, k) -PMD.

To get an efficient algorithm, we need to search over a restricted set of strategies for each player. To be able to do that we must show that an ε -Nash equilibrium exists in a more restricted space. To argue that, we begin by a Nash equilibrium ρ and perform a series of operations that maintain the property that the resulting mixed strategy profile is an ε -equilibrium.

1. We first proceed by rounding the probabilities ρ_{ij} so that they are either 0 or at least c as done in Lemma 3. This gives a PMD $\rho^{(1)}$ that is $O(c^{1/2}k^{5/2}\log^{1/2}(1/ck))$ -close in total variation distance to ρ . Moreover, if we consider the PMD $\rho_{-i}^{(1)}$, which is the $(n-1, k)$ -PMD obtained by removing the i -th component from the rounded PMD $\rho^{(1)}$, this is also $O(c^{1/2}k^{5/2}\log^{1/2}(1/ck))$ -close in total variation to ρ_{-i} , i.e. the PMD obtained after removing the i -th component from the original PMD ρ . The proof of this statement is almost identical to the proof in [DKT15] and is omitted. That proof uses Poisson approximations to bound the total variation between the rounded and the unrounded PMDs and uses the fact that the means of the two PMDs can differ by at most c in each coordinate. The only difference is that here, the means of the two PMDs can differ by at most $2c$ in each coordinate which results in the same asymptotic bound for total variation distance. Moreover, note the rounding procedure doesn't change any probabilities that were originally 0, i.e. $\rho_{ij} = 0 \Rightarrow \rho_{ij}^{(1)} = 0$.
2. We now discretize all parameters $\rho_{ij}^{(1)}$ into multiples of $\lceil \frac{nk}{\varepsilon} \rceil^{-1}$ to get a new PMD $\rho^{(2)}$. This preserves the support of every CRV and makes sure that parameters that were at least c originally remain at least $c - \frac{\varepsilon}{nk}$. Moreover, since $|\hat{r}_{ij} - \bar{r}_{ij}| < \frac{\varepsilon}{nk}$, it holds that $d_{\text{TV}}(\rho_i^{(1)}, \rho_i^{(2)}) < \varepsilon/n$ which implies that $d_{\text{TV}}(\rho_{-i}^{(1)}, \rho_{-i}^{(2)}) < \varepsilon$. This means that overall, for all $i \in [n]$ and $j \in [k]$, $d_{\text{TV}}(\rho_{-i}, \rho_{-i}^{(2)}) < \varepsilon + O(c^{1/2}k^{5/2}\log^{1/2}(1/ck)) < 2\varepsilon$ and $\rho_{ij} = 0 \Rightarrow \rho_{ij}^{(2)} = 0$.
3. By the structural theorem of [DKT15], the components $\rho_i^{(2)}$ of the PMD $\rho^{(2)}$ can be partitioned into two PMDs:

- a sparse PMD of size tk^2 : As in step 2, we can discretize all its probabilities into multiples of $\lceil \frac{tk^3}{\varepsilon} \rceil^{-1}$ to obtain a PMD ρ^{sparse} that is ε -close in total variation distance.
- a large PMD of size $n - tk^2$: This PMD ρ^{large} is shown in [DKT15] to be approximable within ε in total variation distance by a discretized Gaussian g (with a structure preserving rounding) that has the same mean and covariance. The Gaussian consists of one or many blocks and has minimum non-zero eigenvalue at least $\frac{tc}{2k^4}$. Since all the probabilities of the PMD are discretized into multiples of $\lceil \frac{nk}{\varepsilon} \rceil^{-1}$, the entries of the mean vector of the Gaussian are also multiples of $\lceil \frac{nk}{\varepsilon} \rceil^{-1}$ and the entries of the covariance matrix are integer multiples of $\lceil \frac{nk}{\varepsilon} \rceil^{-2}$.

Note that the support of every CRV in the PMD $\rho^{\text{sparse}} * \rho^{\text{large}}$ is a subset of the support of the corresponding CRV in the PMD of the Nash equilibrium ρ . Moreover for every CRV i in ρ^{sparse} and i' in ρ^{large} , it holds that $d_{\text{TV}}(\rho_{-i}, \rho_{-i}^{\text{sparse}} * \rho^{\text{large}}) < 3\varepsilon$ and $d_{\text{TV}}(\rho_{-i'}, \rho^{\text{sparse}} * \rho_{-i'}^{\text{large}}) < 3\varepsilon$.

After performing the steps above, we have shown that an $O(\varepsilon)$ -Nash equilibrium can be found by searching over a limited set of parameters. In particular we require to search over ρ^{sparse} with accuracy $\lceil \frac{tk^3}{\varepsilon} \rceil^{-1}$ and on ρ^{large} with accuracy $\lceil \frac{nk}{\varepsilon} \rceil^{-1}$. The search space unfortunately is still very large since it requires searching over ρ^{large} with high accuracy. The main idea to reduce the search space for the problem is to note that the large PMD is approximable by a discretized Gaussian g (with a structure preserving rounding) that has large non-zero eigenvalues, i.e. $d_{\text{TV}}(\rho^{\text{large}}, g) < \varepsilon$.

For every player i in the sparse PMD, his view about the aggregate strategy of the others is approximately the same as if the large PMD was replaced by the Gaussian, i.e.

$$d_{\text{TV}}(\rho_{-i}^{\text{sparse}} * \rho^{\text{large}}, \rho^{\text{sparse}} * g) < \varepsilon$$

Moreover, for every player i that corresponds to a CRV in the large PMD, his view about the aggregate strategy of the others is approximately the same as if the rest of the components in the large PMD were replaced by a Gaussian g_{-i} with the same mean and covariance ρ_{-i}^{large} , i.e.

$$d_{\text{TV}}(\rho^{\text{sparse}} * \rho_{-i}^{\text{large}}, \rho^{\text{sparse}} * g_{-i}) < \varepsilon$$

At this point, the aggregate behavior of all players can be summarized by describing the probabilities of the sparse PMD and providing the mean and covariance of the Gaussian. However, as shown in Lemma 22 and Lemma 23 of [DKT15], it is possible to reduce the search space by only keeping track of the first $\log(1/\varepsilon)$ moments/power-sums of the sparse PMD. In particular, for a PMD π let $m_{\alpha_1, \dots, \alpha_k}(\pi)$ be the power sum $\sum_i \prod_{j=1}^k (\pi_{ij})^{\alpha_j}$. If a PMD π^A has the same power sums $m_{\alpha_1, \dots, \alpha_k}(\pi^A)$ as the PMD π^B for $\alpha_1, \dots, \alpha_k \in \mathbb{Z}_{\geq 0}$ such that $\sum_{j=1}^k \alpha_j \leq \log(1/\varepsilon)$ and additionally $|\pi_{ij}^A - \pi_{i'j}^B| \leq (4ek^3)^{-1}$ then $d_{\text{TV}}(\pi^A, \pi^B) < 2\varepsilon$. Using this fact, we can partition the CRVs of the sparse PMD into at most $(4ek^3)^k$ smaller components according to the value of the probability in each of the coordinates and replace all CRVs within every partition with a PMD that matches their corresponding power-sums without significant loss in total variation. So knowledge of the power-sums $m_{\alpha_1, \dots, \alpha_k}(\pi)$ for every sub-PMD in the partition is sufficient to approximately describe the distribution of the sparse PMD.

With those observations in hand, we proceed to give the algorithm for computing ε -equilibria for anonymous games. To do this, we first guess the mean μ and covariance Σ of the Gaussian component as well as all the power-sums m of the sparse PMD. We then try to construct CRVs for every player so that the overall mean and covariance as well as the power-sums match those that we guessed and moreover every player's CRV assigns positive probability mass only to approximately optimal strategies. If we are able to do so, Lemma 4 implies that this gives an approximate Nash equilibrium. In more detail, the algorithm performs the following steps:

1. Guess the mean and covariance of the Gaussian component and the power sums of the sparse PMD. For every guess, we repeat the next steps until a feasible solution is found.

We need to guess the powersums for $(4ek^3)^k$ different PMDs since CRVs are first clustered according to their value in every coordinate. Since the parameters of the sparse PMD are all multiples of $\lceil \frac{tk^3}{\varepsilon} \rceil^{-1}$, this results in at most $2^{k^{5k} \log^{k+2}(\frac{1}{\varepsilon})}$ distinct power-sum vectors in total⁶. For the gaussian component all entries of the mean and covariance are multiples of $\lceil \frac{nk}{\varepsilon} \rceil$ which requires $\lceil \frac{nk}{\varepsilon} \rceil^{O(k^2)}$ guesses in total.

2. For every player, we need to compute the contribution of his mixed strategy (CRV) to the overall distribution. If that player is to be assigned in the sparse component, its probabilities are all multiples of $\lceil \frac{tk^3}{\varepsilon} \rceil^{-1}$ and we can compute its contribution to the power-sums m . Similarly, if that player is to be assigned in the gaussian component its probabilities are all multiples of $\lceil \frac{nk}{\varepsilon} \rceil^{-1}$ and we can easily compute its contribution to the mean and covariance.

However, not all assignments are feasible. We need to consider only CRVs for that player that assign positive probability mass to coordinates that are approximately best responses to the strategy of other players. Even though we don't know the strategies of the others exactly, we can compute a good approximate description of the players view by subtracting from the power sums m the players contribution (if any) and computing any *PMD* that matches those power-sums. Similarly, if the player is mapped to the gaussian component we subtract the players mean and covariance from the overall mean μ and covariance Σ and compute a discretized Gaussian with the resulting mean and covariance instead. We say that an assignment of a player to a component (sparse or Gaussian) and a specific distribution over strategies is feasible if it approximately maximizes the player's utility u with respect to his approximate view about the strategies of others.

3. To find if there exists a set of feasible strategies that matches the guessed statistic (m, μ, Σ) , we use dynamic programming. The states of our dynamic program are the following: For any prefix of players, we keep track of the remaining power-sums, mean and covariance we need to account for. We iteratively process players one by one keeping track of which states are reachable. Our estimation is feasible if after processing all players we have accounted for all the power-sums, mean and covariance in our original guess. If we find such a solution, we output the assignment of players to mixed strategies that resulted in this solution.

This algorithm is always guaranteed to find a solution $\hat{\rho}$, since the PMD $\rho^{\text{sparse}} * \rho^{\text{large}}$ that we got by modifying a Nash equilibrium for the game, satisfies all the constraints we imposed. We now claim that the resulting PMD from this algorithm is an ε -Nash equilibrium. The main ingredient to showing this is applying the CLT we developed in Theorem 1 to show that the view $\hat{\rho}_{-i}$ for every player i is close to the view that was assumed when choosing feasible strategies for every player. Indeed, by the CLT all the CRVs that were mapped in the Gaussian component are approximable by a Gaussian with the same mean and covariance, while CRVs that were mapped in the sparse component have the same power-sums as those that we had guessed.

Applying Lemma 4 directly shows that this is indeed an $O(\varepsilon)$ -Nash equilibrium.

The total runtime of the algorithm is polynomial on the number of states of the above dynamic program. Since there are $\lceil \frac{nk}{\varepsilon} \rceil^{O(k^2)}$ Gaussian parameters in total as well as $2^{k^{5k} \log^{k+2}(\frac{1}{\varepsilon})}$ power sums in total, the overall runtime is $n^{O(k^2)} 2^{\text{poly}(k, \log(1/\varepsilon))^k}$ and the theorem follows. \square

⁶This upper bound was derived in [DKT15]

5 An $n^{O(k)}$ Non-Proper Cover for PMDs

On the road to getting the proper cover described by Theorem 3, we first show Theorem 7. This constructs a non-proper cover of the same size. The main theorem of this section is the following:

Theorem 7. *For all $n, k \in \mathbb{N}$, and $\varepsilon > 0$, there exists a (non-proper) ε -cover, in total variation distance, of the set of all (n, k) -PMDs whose size is*

$$n^{O(k)} \cdot \min \left\{ 2^{\text{poly}(k/\varepsilon)}, 2^{O(k^{5k} \log^{k+2}(1/\varepsilon))} \right\}.$$

Moreover, we can efficiently enumerate this cover in time polynomial in its size.

This theorem should be contrasted with Theorem 3, which provides a proper cover of similar size. It should also be contrasted to Theorem 2 of [DKT15], which provides a cover with a leading factor of n^{k^2} , so the cover presented here improves the exponent of n from quadratic to linear in the dimension. This is the correct order of exponential dependence on k , as simply counting the number of (n, k) -PMDs with deterministic summands gives a lower bound of $n^{\Omega(k)}$. We also show in Section 7 that the quasi-polynomial dependence on $1/\varepsilon$ with an exponent of $\Omega(k)$ cannot be avoided, as we provide an essentially matching lower bound on the cover size.

The starting point for our cover will be Theorem 6, stating that every (n, k) -PMD is ε -close to the sum of an appropriately discretized Gaussian and a $(\text{poly}(k/\varepsilon), k)$ -PMD. We generate an $\varepsilon/2$ -cover for each and combine them by triangle inequality.

Covering the sparse PMD. We cover the sparse PMD component using the same methods as in [DKT15]. The first, naive way of covering this component involves gridding over all $\text{poly}(k/\varepsilon)$ parameters with $\text{poly}(\varepsilon/k)$ granularity. This results in a cover size of $2^{\text{poly}(k/\varepsilon)}$.

The more sophisticated way of covering this component uses a “moment matching” technique. A result by Roos [Roo02] shows that the probability mass function can be written as the weighted sum of partial derivatives of a standard multinomial distribution. When analyzed carefully, his result implies that the lower order moments of the distribution are sufficient to characterize the PMD. In other words, any two PMDs with identical “moment profiles” (which describe these lower order moments) are close in total variation distance, and it suffices to keep only one representative for each moment profile. This method results in a cover of size $2^{O(k^{5k} \log^{k+2}(1/\varepsilon))}$. Combining this with the other approach gives a cover of size

$$\min \left\{ 2^{\text{poly}(k/\varepsilon)}, 2^{O(k^{5k} \log^{k+2}(1/\varepsilon))} \right\}.$$

For more details, see the proof of Theorem 2 of [DKT15].

Covering the discretized Gaussian. To cover the Gaussian component, [DKT15] grid over all $O(k^2)$ parameters of the Gaussian component, arguing the effectiveness of the gridding using Proposition 2. This gridding results in the leading factor of $n^{O(k^2)}$ in the size of the cover. In contrast, we use a spectral covering approach: instead of trying to grid over all parameters of the covariance matrix, we first sparsify it and then match the magnitude of its projection in every direction. In particular, we establish a cover of the following nature:

Lemma 5. *Let $\mathcal{G}_{n,k,\varepsilon}$ be the set of all Gaussians with structure preserving roundings which may arise as a consequence of Theorem 6 when applied to (n, k) -Poisson multinomial random vectors*

with parameter ε . Then there exists a set \mathcal{S} of Gaussians with structure preserving roundings of size at most $n^{O(k)} \cdot \left(\frac{k}{\varepsilon}\right)^{O(k^3)}$ with the following properties:

For any $G \in \mathcal{G}_{n,k,\varepsilon}$, there exists a $\hat{G} \in \mathcal{S}$, such that G and \hat{G} have the same block structure (i.e., the partition of coordinates), and within each block, have the same pivot coordinate and sum for the mean vector coordinates. Furthermore, for each block i , letting (μ_i, Σ_i) and $(\hat{\mu}_i, \hat{\Sigma}_i)$ be the mean and covariance for the block (excluding the pivot coordinate), we have that for all unit vectors v ,

- $|v^T(\mu_i - \hat{\mu}_i)| \leq \frac{\varepsilon \sigma_{iv}}{k}$;
- $|v^T(\Sigma_i - \hat{\Sigma}_i)v| \leq \frac{\varepsilon \sigma_{iv}^2}{2k^{3/2}}$;

where $\sigma_{iv}^2 = \max\{v^T \Sigma_i v, v^T \hat{\Sigma}_i v\}$.

This lemma statement is slightly technical due to the nature of the Gaussians with structure preserving roundings. It essentially says that we cover the set of Gaussians arising from the structural theorem by matching their block structure exactly, and within each block, matching the moments spectrally. Plugging these guarantees into Lemma 2 and applying the data processing inequality for total variation distance (Lemma 1) gives the desired closeness.

For simplicity of exposition, for the remainder of this overview section, we assume that the Gaussian's structure preserving rounding consists of a single block, an assumption we do not make in the full proof (described in Section 5.1). By the guarantees of the structural result, in this case, the minimum eigenvalue of the covariance matrix is at least some $\text{poly}(k/\varepsilon)$. So the goal of our exposition in this section is to produce a cover of Gaussians that may result from Theorem 6 and whose covariance matrices have minimum eigenvalue at least $\text{poly}(k/\varepsilon)$.

Since the mean vector only has k parameters, we can grid over the entries. Though we require a spectral guarantee, this naive gridding is sufficient. This gives a set of size $\left(\frac{nk}{\varepsilon}\right)^{O(k)}$, such that, for any Gaussian which may arise from Theorem 6, its mean vector is approximated by a mean vector in our set with the approximation guarantees required by Lemma 2.

Covering the covariance matrix takes more care. At a high level, our approach views PMDs through the lens of spectral graph theory and exploits the existence of spectral sparsifiers. Recall the definition of the Laplacian matrix of a graph:

Definition 11. Given an undirected weighted graph $G = (V, E, w)$ on n vertices, its Laplacian matrix is an $n \times n$ matrix L_G where

$$L_G(i, j) = \begin{cases} \sum_{k \neq i} w(i, k) & \text{if } i = j \\ -w(i, j) & \text{if } i \neq j \wedge (i, j) \in E, \\ 0 & \text{otherwise} \end{cases}$$

To see the connection to PMDs, we observe that the covariance matrix of a PMD is the Laplacian matrix of a graph defined by the parameters. For a single k -CRV X with parameter vector π , it can be shown that the variance of X_i is $\pi(i)(1 - \pi(i))$ and the covariance of X_i and X_j is $-\pi(i)\pi(j)$. Since $\sum_{i=1}^k \pi(i) = 1$, the covariance matrix is equal to the Laplacian matrix of a graph on k nodes with $w(i, j) = \pi(i)\pi(j)$. This can be extended to (n, k) -PMDs by observing that the sum of random variables has a covariance matrix equal to the sum of the individual covariance matrices, and a similar statement holds for graphs and the corresponding Laplacian matrices. We summarize this connection in the following observation:

Observation 1. *The covariance matrix of an (n, k) -Poisson Multinomial Distribution M^π corresponds to the Laplacian matrix of a graph $G = (V, E, w)$ on k nodes, where the $w(i, j) = \sum_{\ell=1}^n \pi(\ell, i)\pi(\ell, j)$.*

At the core of our approach, we use the following celebrated result of Batson, Spielman, and Srivastava [BSS12], which says that the Laplacian matrix of a graph on k vertices can be spectrally approximated by the Laplacian matrix of a graph with only $O(k)$ edges:

Theorem 8 (Theorem 1.1 in [BSS12]). *For every $\varepsilon \in (0, 1)$, every undirected weighted graph $G = (V, E, w)$ on n vertices contains a weighted subgraph $H = (V, F, \tilde{w})$ with $\lceil (n-1)/\varepsilon^2 \rceil$ edges which satisfies*

$$(1 - \varepsilon)^2 L_G \preceq L_H \preceq (1 + \varepsilon)^2 L_G,$$

where L_G is the Laplacian matrix of the graph G .

Using this tool, the approach will proceed as follows. This theorem implies that, for every true covariance matrix Σ , there exists a matrix M_1 with only $O(k)$ entries which preserves every projection up to a multiplicative factor of $1/5$. We can obtain a matrix M_2 with the same sparsity pattern as M_1 by guessing which subset of $O(k)$ entries is non-zero, requiring $\exp(k \cdot \log k)$ guesses. Furthermore, we can grid over the non-zero entries of M_2 to ensure that it approximates every projection of M_1 up to a multiplicative factor of $1/25$. Since M_1 has minimum eigenvalue $\text{poly}(k/\varepsilon)$ and maximum entry $O(n)$, gridding requires only $(\frac{n \cdot k}{\varepsilon})^{O(k)}$ guesses, and we get that M_2 gives a $1/4$ multiplicative spectral approximation to Σ . To make our approximation finer, we will $O(\varepsilon/\sqrt{k})$ -cover the set of PSD matrices within a $1/4$ -neighborhood of M_2 . We first recall the definition of a cover in this context:

Definition 12. *Let S be a set of symmetric $k \times k$ PSD matrices. An ε -cover of the set S , denoted by S_ε , is a set of PSD matrices such that for any matrix $A \in S$, there exists a matrix $B \in S_\varepsilon$ such that for all vectors y : $|y^T(A - B)y| \leq \varepsilon y^T A y$.*

Now, if we could $O(\varepsilon/\sqrt{k})$ -cover the set of all matrices $1/4$ -close to M_2 , we would obtain an $O(\varepsilon/\sqrt{k})$ -approximation to Σ . We do so using the following lemma, which provides a method to generate such a cover. A slight generalization of this statement appeared as Lemma 9 in [DKT15], but we give a slightly simpler proof in Section 5.2 for completeness.

Lemma 6 (Lemma 9 in [DKT15]). *Let A be a symmetric $k \times k$ PSD matrix with minimum eigenvalue at least 1 and let S be the set of all matrices B such that $|y^T(A - B)y| \leq \varepsilon_1 y^T A y$ for all vectors y , where $\varepsilon_1 \in [0, 1/4]$. Then, there exists an ε -cover S_ε of S that has size $|S_\varepsilon| \leq (\frac{k}{\varepsilon})^{O(k^2)}$.*

Combining the above, we obtain a set of covariance matrices of size $n^{O(k)} \cdot (\frac{1}{\varepsilon})^{\text{poly}(k)}$ such that, for any Gaussian which may arise in Theorem 6, its covariance matrix is approximated by a covariance matrix in our cover as required by Lemma 2.

Combining the guarantees obtained for the mean and the covariance matrix, we find that they satisfy both conditions of Lemma 2. Therefore, we have described a cover of size $n^{O(k)} \cdot (\frac{1}{\varepsilon})^{\text{poly}(k)}$ for all possible Gaussian components. The proof of Theorem 7 is completed by taking the Cartesian product of this Gaussian cover with the cover for the $(\text{poly}(k/\varepsilon), k)$ -PMD component.

For more details on covering the Gaussian component, see Section 5.1.

5.1 Details on Covering the Gaussian Component

Recall that the Gaussian component will have a structure preserving rounding. The first step in designing our cover will be to guess the partitioning into blocks. There are k dimensions, resulting in at most $k!$ different block structures. In what follows, we will describe how to cover a single block up to accuracy $O(\frac{\varepsilon}{k})$, taking the Cartesian product of the resulting sets will give an $O(\varepsilon)$ -cover of the entire Gaussian at the additional cost of k in the exponent.

For a single block which consists of dimensions S_i , we must first guess the size parameter n_i and which dimension is to be used as the pivot. The former is an integer between 0 and n , and guessing it comes at a cost of n in our cover size. Guessing the latter comes at a $|S_i|$ cost in our cover size.

Recall that our strategy will be to spectrally match the parameters of the true Gaussian. We will conclude the two distributions are close using the guarantees provided by Lemma 2. We describe how to obtain such guarantees for both the mean and covariance matrix separately.

5.1.1 Covering the Mean Vector of a Block

We know the mean of the block will be contained in the cube $[0, n_i]^{|S_i|}$. For some $\alpha(k, \varepsilon)$ (which for simplicity, we assume divides n_i), consider the lattice $\{0, \alpha, 2\alpha, \dots, n_i\}^{|S_i|}$, which has $(\frac{n_i}{\alpha} + 1)^{|S_i|}$ points. We note that the maximum ℓ_2 distance between the mean μ and the closest point of this lattice $\hat{\mu}$ is at most $\alpha\sqrt{k}$, and therefore, for any unit vector v , we have that $|v^T(\mu - \hat{\mu})| \leq \alpha\sqrt{k}$. We also know that the minimum variance of any projection the Gaussian is large, in particular, at least $\frac{tc}{2k^4}$, so the standard deviation in any direction v is $\sigma_v \geq \sqrt{\frac{tc}{2k^4}}$. Choosing $\alpha \leq k^5/\varepsilon \leq \varepsilon\sigma_v/k^{3/2}$ implies that $\alpha\sqrt{k} \leq \varepsilon\sigma_v/k$. This shows that the first condition of Lemma 2 is satisfied to approximate this block up to $\frac{\varepsilon}{k}$ accuracy. Substituting the value of α , we cover the mean with a set of size at most $(\frac{n_i\varepsilon}{k^5} + 1)^{|S_i|}$.

5.1.2 Covering the Covariance Matrix of a Block

We will use the characterization provided by Observation 1, which tells us that the covariance matrix of an (n, k) -PMD is the Laplacian matrix of a graph defined by the parameters of the distribution. Recall from the proof of Theorem 6 (which appears in [DKT15]), the covariance matrix of the Gaussian we are attempting to match is also the covariance matrix of an $(n_i, |S_i|)$ -Generalized Multinomial Distribution. For the remainder of this proof, we let G be the graph defined by this characterization for the covariance matrix of the corresponding $(n_i, |S_i|)$ -Poisson Multinomial Distribution.

As a starting point, we use Theorem 8, which shows the existence of spectral sparsifiers. In particular it implies that, if given G on $|S_i|$ nodes and we want a subgraph H such that $(1 - 1/5)L_G \preceq L_H \preceq (1 + 1/5)L_G$, there exists an H with at most $110|S_i|$ edges which gives this approximation. The first step in covering the covariance matrix is to guess which edges are present in the graph. Since there are $\binom{|S_i|}{2}$ possible edges in the graph, this requires at most

$$\binom{\binom{|S_i|}{2}}{110|S_i|} \leq k^{220k}$$

guesses.

Now that we know which edges are present in the graph, the goal is to guess the weights of these edges. Ideally, we would like to obtain a graph M with the guarantee that $(1 - 1/25)L_H \preceq L_M \preceq (1 + 1/25)L_H$. However, this is stronger than we can hope for, since recalling that L_H has a zero

eigenvalue, it would require that the diagonals of L_M and L_H are exactly equal. Instead, we recall that we have a pivot coordinate which will be left out of the Gaussian's covariance matrix, and we only have to match projections which are orthogonal to this direction. Without loss of generality, assume that the pivot coordinate is 1. For any unit vector $v \in \mathbb{R}^k$ orthogonal to e_1 , we will obtain an L_M such that

$$v^T(L_H - L_M)v \leq \frac{1}{25}v^T L_H v,$$

which will imply

$$v^T(L_G - L_M)v \leq \frac{1}{4}v^T L_G v.$$

Further, recall that our structural result implies that $\frac{1}{25}v^T L_H v \geq \frac{tc}{100k^4}$, so it suffices to obtain a graph M such that

$$v^T(L_H - L_M)v \leq \frac{tc}{100k^4}.$$

For a unit vector v and $|S_i| \times |S_i|$ PSD matrices A and B ,

$$v^T(A - B)v = \sum_{i,j} v_i v_j (A(i, j) - B(i, j)) \leq |S_i|^2 \max_{i,j} |A(i, j) - B(i, j)|.$$

Suppose we guess the edge weights of M such that they are at most $\frac{tc}{100k^7}$ away from those of H . This tells us $\max_{i \neq j} |L_H(i, j) - L_M(i, j)| \leq \frac{tc}{100k^7}$, and since the diagonal entries of L_M are the sums of the off-diagonal entries, $\max_i |L_H(i, i) - L_M(i, i)| \leq \frac{tc}{100k^6}$. This implies that it suffices to additively estimate the edge weights up to accuracy $\frac{tc}{100k^6}$. Since the maximum entry of L_G is at most n_i , the spectral guarantee implies that the maximum entry of L_H is at most $\frac{6n_i}{5}$, and similarly, the maximum edge weight. Therefore, gridding over all $110|S_i|$ non-zero edge weights, we define a set with at most

$$\left(\frac{6n_i/5}{tc/100k^7} \right)^{110|S_i|} \leq \left(\frac{3n_i \varepsilon^5}{250k^{11}} \right)^{110|S_i|}$$

candidates.

At this point, we have a PSD matrix L_M which, when projected onto the subspace orthogonal to e_1 , is $1/4$ -spectrally close to the target covariance matrix. We wish to $\frac{\varepsilon}{2k^{3/2}}$ -cover the space of all PSD matrices which are $1/4$ -spectrally close to this matrix. We will use Lemma 6, which we instantiate with parameter “ ε ” set to $\frac{\varepsilon}{2k^{3/2}}$, allowing us to generate a $\frac{\varepsilon}{2k^{3/2}}$ -cover of a $\frac{1}{4}$ -neighborhood of a given PSD matrix with $\left(\frac{k}{\varepsilon}\right)^{O(k^2)}$ candidates. Since we knew one of the previous candidates was $\frac{1}{4}$ -close to the target, this gives us a matrix which satisfies the second condition of Lemma 2 to approximate this block up to $\frac{\varepsilon}{k}$ accuracy. The size of this cover is at most

$$k^{220k} \cdot \left(\frac{3n_i \varepsilon^5}{250k^{11}} \right)^{110|S_i|} \cdot \left(\frac{k}{\varepsilon} \right)^{O(k^2)} = n^{O(|S_i|)} \left(\frac{k}{\varepsilon} \right)^{O(k^2)}.$$

5.1.3 Putting the Guarantees Together

At this point, to cover a single block up to accuracy $O(\varepsilon/k)$, we have a set of size at most

$$n \cdot |S_i| \cdot \left(\frac{n_i \varepsilon}{k^5} + 1 \right)^{|S_i|} \cdot n^{O(|S_i|)} \left(\frac{k}{\varepsilon} \right)^{O(k^2)} = n^{O(|S_i|)} \left(\frac{k}{\varepsilon} \right)^{O(k^2)}.$$

Taking the Cartesian product of sets and multiplying by the number of guesses for the block structure of the Gaussian, we get an overall cover of size

$$k! \cdot \prod_{S_i} \left(n^{O(|S_i|)} \left(\frac{k}{\varepsilon} \right)^{O(k^2)} \right) = n^{O(k)} \left(\frac{k}{\varepsilon} \right)^{O(k^3)}.$$

Combining with the cover for the $(\text{poly}(k/\varepsilon), k)$ -PMD component, we obtain an overall cover for (n, k) -PMDs of size

$$n^{O(k)} \cdot \min \left\{ 2^{\text{poly}(k/\varepsilon)}, 2^{O(k^{5k} \cdot \log^{k+2}(1/\varepsilon))} \right\},$$

as desired.

5.2 Proof of Lemma 6

To construct the cover, we will make use of the eigenvalues and eigenvectors of the matrix A . We first show that for any matrix $B \in S$, its eigenvalues are close to the eigenvalues of A .

Proposition 9. *Let A, B be two symmetric $k \times k$ PSD matrices such that for all vectors y with $\|y\| = 1$, $|y^T(A - B)y| \leq \varepsilon_1 y^T A y$ for some constant $\varepsilon_1 > 0$. Then for the eigenvalues $\lambda_1^A \leq \dots \leq \lambda_k^A$ of A , and the eigenvalues $\lambda_1^B \leq \dots \leq \lambda_k^B$ of B , it holds that:*

$$|\lambda_i^A - \lambda_i^B| \leq \varepsilon_1 \lambda_i^A$$

Proof. From Courant's minimax principle, we have that the i -th eigenvalue of A is equal to:

$$\lambda_i^A = \max_C \min_{\substack{\|x\|=1 \\ Cx=0}} x^T A x$$

where C is an $(i - 1) \times k$ matrix. For the matrix B , we have that

$$\lambda_i^B = \max_C \min_{\substack{\|x\|=1 \\ Cx=0}} x^T B x \leq \max_C \min_{\substack{\|x\|=1 \\ Cx=0}} (1 + \varepsilon_1) x^T A x = (1 + \varepsilon_1) \lambda_i^A$$

Similarly, we have that $\lambda_i^B \geq (1 - \varepsilon_1) \lambda_i^A$, so the result follows. \square

By computing the eigenvalues $\mu_1 \leq \dots \leq \mu_k$ of A , we have estimates of the eigenvalues $\lambda_1, \dots, \lambda_k$ of B within a multiplicative factor of $1 \pm 2\varepsilon_1$. We can improve our estimates to a better multiplicative factor $1 \pm \varepsilon$ by gridding multiplicatively around each eigenvalue. This requires another $\log_{1+\varepsilon} \left(\frac{1+2\varepsilon_1}{1-2\varepsilon_1} \right) = O(1/\varepsilon)$ guesses per eigenvalue. So in total, we require $(\frac{1}{\varepsilon})^{O(k)}$ guesses for obtaining accurate estimates $\lambda'_1, \dots, \lambda'_k$ of the eigenvalues of B .

Once we know (approximately) the eigenvalues of B , we will try to guess also its eigenvectors v_1, \dots, v_k . We will do this by performing a careful gridding around the eigenvectors of A which we can assume, without loss of generality (by rotating), to be the standard basis vectors e_1, e_2, \dots, e_k . So for each eigenvector v_z of B , we will try to approximate it by guessing its projections to the eigenvectors of A .

We now bound the projections of eigenvectors of A to eigenvectors of B . Since we know that $e_i^T B e_i \leq (1 + \varepsilon_1) e_i^T A e_i$, we get that $\sum_z \lambda_z (v_z e_i)^2 \leq (1 + \varepsilon_1) \mu_i$ which implies that $v_{z,i} \leq \sqrt{\frac{2\mu_i}{\lambda_z}}$. Moreover, since $\lambda_z \geq \max\{(1 - \varepsilon_1)\mu_z, 1\} \geq \max\{\frac{1}{2}\mu_z, 1\}$, we know that the projection of v_z to e_i will be smaller than $2\sqrt{\frac{\mu_i}{\max\{\mu_z, 1\}}}$. An additional bound for the projection of v_z to e_i can be obtained

by considering the variance of the matrices A and B in the direction v_z . Since we know that $v_z^T B v_z \geq (1 - \varepsilon_1) v_z^T A v_z$, we get that $\sum_i \mu_i (v_z e_i)^2 \leq \frac{\lambda_z}{1 - \varepsilon_1} \leq 2\lambda_z$ which implies that $v_{z,i} \leq \sqrt{\frac{2\lambda_z}{\mu_i}}$.

We now guess vectors v'_1, \dots, v'_k that approximate the eigenvectors of B by additively gridding over the projections to each eigenvector of A . In particular, our candidate guesses for $v'_z \cdot e_i = v'_{z,i}$ will be $\ell \varepsilon' \min \left\{ 2\sqrt{\frac{\mu_i}{\max\{\mu_z, 1\}}}, 1 \right\}$ with $\ell \in \{0, 1, \dots, 1/\varepsilon'\}$, for a small enough ε' that only depends on k and ε . This will give us an approximation v'_z for the eigenvector v_z , with the guarantee that $|v'_{z,i} - v_{z,i}| \leq \varepsilon' \min \left\{ 2\sqrt{\frac{\mu_i}{\max\{\mu_z, 1\}}}, 1 \right\}$. This requires $\frac{1}{\varepsilon'}$ guesses for each projection, and thus $(\frac{1}{\varepsilon'})^{k^2}$ guesses for all k^2 projections. The final covariance matrix we output is then $\hat{B} = \sum_z \lambda'_z v'_z (v'_z)^T$.

We will now show that the covariance matrix \hat{B} satisfies the property that it is close in all directions to B . To do this we will make use of the following lemma from [DKT15]. This roughly states that two PSD matrices spectrally approximate each other in $O(k^2)$ particular directions, then they spectrally approximate each other in every direction.

Lemma 7 (Lemma 25 from [DKT15]). *Let $\Sigma, \hat{\Sigma} \in \mathbb{R}^{k \times k}$ be two symmetric, positive semi-definite matrices, and let $(\lambda_1, v_1), \dots, (\lambda_k, v_k)$ be the eigenvalue-eigenvector pairs of Σ . Suppose that*

- For all $i \in [k]$, $\left| \left(\frac{v_i}{\sqrt{\lambda_i}} \right)^T (\hat{\Sigma} - \Sigma) \left(\frac{v_i}{\sqrt{\lambda_i}} \right) \right| \leq \varepsilon$,
- For all $i, j \in [k]$, $\left| \left(\frac{v_i}{\sqrt{\lambda_i}} + \frac{v_j}{\sqrt{\lambda_j}} \right)^T (\hat{\Sigma} - \Sigma) \left(\frac{v_i}{\sqrt{\lambda_i}} + \frac{v_j}{\sqrt{\lambda_j}} \right) \right| \leq 4\varepsilon$.

Then for all $y \in \mathbb{R}^k$, $\left| y^T (\hat{\Sigma} - \Sigma) y \right| \leq 3k\varepsilon y^T \Sigma y$.

We will only consider directions $y = \frac{v_z}{\sqrt{\lambda_z}}$ for $z \in [k]$ and $y = \frac{v_z}{\sqrt{\lambda_z}} + \frac{v_{z'}}{\sqrt{\lambda_{z'}}}$ for $z, z' \in [k]$.

We first consider direction $y = \frac{v_z}{\sqrt{\lambda_z}}$. We have that:

$$\frac{v_z^T}{\sqrt{\lambda_z}} \hat{B} \frac{v_z}{\sqrt{\lambda_z}} = \sum_i \frac{\lambda'_i}{\lambda_z} (v_z^T v'_i)^2 = \sum_i \frac{\lambda'_i}{\lambda_z} (v_z^T v_i + v_z^T (v'_i - v_i))^2 = \frac{\lambda'_z}{\lambda_z} (1 + v_z^T (v'_z - v_z))^2 + \sum_{i \neq z} \frac{\lambda'_i}{\lambda_z} (v_z^T (v'_i - v_i))^2$$

The first term is in the range $[(1 - \varepsilon)(1 - k\varepsilon')^2, (1 + \varepsilon)(1 + k\varepsilon')^2]$, which for $\varepsilon' \leq \varepsilon/k$, becomes $(1 \pm O(\varepsilon))$. The rest of the terms can be bounded as follows:

$$\begin{aligned}
\frac{\lambda'_i}{\lambda_z} (v_z(v'_i - v_i))^2 &\leq (1 + \varepsilon) \frac{\lambda_i}{\lambda_z} \left(\sum_j v_{z,j} (v'_{i,j} - v_{i,j}) \right)^2 \\
&\leq (1 + \varepsilon) \frac{\lambda_i}{\lambda_z} \left(\sum_j \sqrt{2 \frac{\lambda_z}{\mu_j}} \varepsilon' 2 \sqrt{\frac{\mu_j}{\max\{\mu_i, 1\}}} \right)^2 \\
&\leq (1 + \varepsilon) \frac{\lambda_i}{\lambda_z} \left(\sum_j 2 \varepsilon' \sqrt{2 \lambda_z} \sqrt{\frac{1}{\max\{\mu_i, 1\}}} \right)^2 \\
&\leq (1 + \varepsilon) \left(\sum_j 2 \varepsilon' \sqrt{\frac{2 \lambda_i}{\max\{\mu_i, 1\}}} \right)^2 \\
&\leq (1 + \varepsilon) \left(4 k \varepsilon' \sqrt{\frac{\mu_i}{\max\{\mu_i, 1\}}} \right)^2 \\
&\leq (1 + \varepsilon) (8 k \varepsilon')^2 \\
&\leq \frac{\varepsilon}{k}
\end{aligned}$$

for $\varepsilon' = O(\sqrt{\frac{\varepsilon}{k^3}})$. This means that $v_z^T \hat{B} v_z \in (1 - \varepsilon, 1 + \varepsilon) \lambda_z$. The proof is similar for directions $y = \frac{v_z}{\sqrt{\lambda_z}} + \frac{v_{z'}}{\sqrt{\lambda_{z'}}}$ for $z, z' \in [k]$.

Overall, we can get an estimate \hat{B} of any matrix $B \in S$ by making at most $(\frac{k}{\varepsilon})^{O(k^2)}$ guesses, which implies an ε -cover of this size.

6 A Proper Cover for PMDs

We show how to turn the non-proper cover of Section 5 into a proper one as described by Theorem 3, using Theorem 1. We note that a non-constructive proper cover follows immediately from Theorem 7, since for each element of an improper $\varepsilon/2$ -cover that lies within $\varepsilon/2$ of a PMD, we can match it with such a PMD. The resulting set of PMDs defines then a proper ε -cover. Our focus in this section is to provide an efficient construction of a proper cover.

Our approach will be to enumerate the improper cover of Theorem 7 and convert each distribution to a nearby (n, k) -PMD. This cover consists of distributions which are the sum of a Gaussian with a structure preserving rounding and a $(\text{poly}(k/\varepsilon), k)$ -PMD. Since the $(\text{poly}(k/\varepsilon), k)$ -PMD component is already a collection of k -CRVs, this part of the cover is already proper, and it suffices to convert the Gaussian component into a nearby $(n - \text{poly}(k/\varepsilon), k)$ -PMD.

The main technical lemma we prove is the following, which states that if a discretized Gaussian G is spectrally close to a GMD ρ , we can obtain a new GMD ρ' which is spectrally close to ρ :

Lemma 8. *Let $\lfloor \mathcal{N}(\mu, \Sigma) \rfloor$ be a discretized Gaussian and suppose there exists a (n, k) -GMD ρ with mean μ^ρ and covariance Σ^ρ such that for all vectors v it holds that $|v^T(\mu - \mu^\rho)| \leq \varepsilon_1 \sqrt{v^T \Sigma v}$ and $|v^T(\Sigma - \Sigma^\rho)v| \leq \varepsilon_2 v^T \Sigma v$.*

Then, it is possible to compute in time $n^{O(k)}$ a (n, k) -GMD ρ' with mean $\mu^{\rho'}$ and covariance $\Sigma^{\rho'}$ such that for all vectors v it holds that $|v^T(\mu - \mu^{\rho'})| \leq \varepsilon_1 \sqrt{v^T \Sigma v} + 3k^{2.5} \|v\|_2$ and $|v^T(\Sigma - \Sigma^{\rho'})v| \leq \varepsilon_2 v^T \Sigma v + 3k^3 \|v\|_2^2$.

We prove this lemma using the Shapley-Folkman lemma [Sta69], which states that the Minkowski sum of a large number of sets is approximately convex:

Lemma 9 (Shapley-Folkman lemma). *Let S_1, \dots, S_n be a collection of sets in \mathbb{R}^d , and let $S = \{\sum_{i=1}^n x_i \mid x_1 \in S_1, \dots, x_n \in S_n\}$ be their Minkowski sum. Then, letting $\text{conv}(X)$ denote the convex hull of X , every $x \in \text{conv}(S) = \sum_{i=1}^n x_i$ where $x_i \in \text{conv}(S_i)$ for $i = 1, \dots, n$ and $|\{i \mid x_i \notin S_i\}| \leq d$.*

With this lemma in hand, the proof of Lemma 8 proceeds as follows. Let \mathcal{M} be the set of all possible mean and covariances for a single CRV, and $\mathcal{M}^{\oplus n}$ be the Minkowski sum of n copies of \mathcal{M} . Given a discretized Gaussian with mean and covariance $(\mu, \Sigma) \in \mathcal{M}^{\oplus n}$, we would ideally like to find $\{x_1, \dots, x_n\}$ such that $\sum_{i=1}^n x_i = (\mu, \Sigma)$. However, since this set is not convex, this optimization problem is not obviously tractable. Instead, we convert (μ, Σ) to a spectrally close $(\hat{\mu}, \hat{\Sigma})$ which lies on the convex hull of $\mathcal{M}^{\oplus n}$, which can be done using a linear program. At this point, we exploit the “almost convex” characterization provided the Shapley-Folkman lemma, and we will iteratively “peel off” plausible CRVs. More specifically, noting that the moment profile is at most $k^2 + k$ dimensional and applying Lemma 9, we can use a linear program to find the parameters of a single CRV such that subtracting its moments gives a moment profile which lies on the convex hull of $\mathcal{M}^{\oplus n-1}$. We repeat $n - k^2 - k$ times until we are left with a point on the convex hull of $\mathcal{M}^{\oplus k^2+k}$, at which point we may pick the last $k^2 + k$ CRVs arbitrarily. The proof is completed by arguing that the resulting GMD satisfies the theorem conditions. For the full proof of Lemma 8, see Section 6.1.

We now prove Theorem 3. As mentioned before, for our starting point, we relate our original PMD π to the sum of a discretized Gaussian with a structure preserving rounding and a $(\text{poly}(k/\varepsilon'), k)$ -PMD using 7, for some ε' to be set later. This comes at a cost of ε' in total variation distance. The CRVs corresponding to the sparse PMD are already in the form desired for the proper cover, and we ignore them for the remainder of the proof. We also know that the discretized Gaussian’s mean and covariance matrix arose from the mean and covariance matrix of some PMD. This covariance matrix has a block structure, where each block has a minimum eigenvalue of at least $\frac{k^{15}}{2\varepsilon'^6}$. At this point, we wish to show that each block of the current PMD $\tilde{\pi}$ is ε/k -close to each block of the PMD after applying the method of Lemma 8, π' . This will be proven by relating a block of $\tilde{\pi}$ and π' to the corresponding discretized Gaussians using Theorem 1, and arguing that the discretized Gaussians are close using Lemma 2.

We focus on one block of $\tilde{\pi}$. The guarantee of our cover, summarized in Lemma 5, tells us that the corresponding block of π' will have a matching pivot and constituent number of CRV’s n_i . Therefore, it suffices to consider the corresponding GMDs which exclude the pivot coordinate, namely $\tilde{\rho}$ and ρ' . We know that the minimum eigenvalue of this block of $\tilde{\rho}$ ’s covariance matrix is at least $\frac{k^{15}}{2\varepsilon'^6}$. The guarantees of Lemma 5 give us an input to Lemma 8 with $\varepsilon_1 = \frac{\varepsilon}{k}$ and $\varepsilon_2 = \frac{\varepsilon}{2k^{3/2}}$. Since the minimum variance of this block of $\tilde{\rho}$ is sufficiently large, the output of Lemma 8 is a relative spectral approximation to the mean and covariances, with multiplicative $2\varepsilon_1$ and $2\varepsilon_2$ factors, respectively. We note that this implies that the minimum eigenvalue of this block of ρ' ’s covariance matrix is at least $\frac{k^{15}}{4\varepsilon'^6}$.

We convert this block of $\tilde{\rho}$ to the corresponding discretized Gaussian using our CLT, Theorem 1. Given the aforementioned minimum eigenvalue condition, the cost incurred is at most

$$O\left(\frac{k^{7/2}}{(k^{15}/\varepsilon'^6)^{1/20}}\right) = O(k^{11/4}\varepsilon'^{3/10}).$$

We convert the same block of ρ' to a discretized Gaussian in the same way, incurring the same cost. Finally, we relate the two discretized Gaussians in total variation distance. As mentioned in the previous paragraph, the means and covariances are spectrally close up to relative accuracy $\frac{2\varepsilon}{k}$

and $\frac{\varepsilon}{k^{3/2}}$. We plug this guarantee into Lemma 2 and apply the data processing inequality (Lemma 1) to conclude that the two distributions are $O(\varepsilon/k)$ -close. The proof is concluded by setting $\varepsilon' = \varepsilon^{10/3}/k^{25/2}$ and rescaling ε by a constant factor.

6.1 Proof of Lemma 8

We first argue that rounding all constituent probability vectors in the (n, k) -GMD ρ so that all their coordinates are integer multiples of $1/n$ to obtain a (n, k) -GMD $\hat{\rho}$ approximately preserves the spectral closeness guarantees with the discretized Gaussian. More specifically, for all vectors v it holds that:

$$|v^T(\mu - \mu^{\hat{\rho}})| \leq \varepsilon_1 \sqrt{v^T \Sigma v} + \sqrt{k} \|v\|_2 \quad \text{and} \quad |v^T(\Sigma - \Sigma^{\hat{\rho}})v| \leq \varepsilon_2 v^T \Sigma v + k \|v\|_2^2.$$

We know that $\|\mu^\rho - \mu^{\hat{\rho}}\|_\infty \leq 1$ and thus $\|\mu^\rho - \mu^{\hat{\rho}}\|_2 \leq \sqrt{k}$, so

$$\begin{aligned} |v^T(\mu - \mu^{\hat{\rho}})| &\leq |v^T(\mu - \mu^\rho)| + |v^T(\mu^{\hat{\rho}} - \mu^\rho)| \\ &\leq \varepsilon_1 \sqrt{v^T \Sigma v} + \|\mu^\rho - \mu^{\hat{\rho}}\|_2 \|v\|_2 \\ &= \varepsilon_1 \sqrt{v^T \Sigma v} + \sqrt{k} \|v\|_2 \end{aligned}$$

Similarly we have that $\|\Sigma^\rho - \Sigma^{\hat{\rho}}\|_{\max} \leq 1$ which implies that $|v^T(\Sigma^\rho - \Sigma^{\hat{\rho}})v| \leq k \|v\|^2$ for all vectors v . Thus,

$$|v^T(\Sigma - \Sigma^{\hat{\rho}})v| \leq |v^T(\Sigma - \Sigma^\rho)v| + |v^T(\Sigma^{\hat{\rho}} - \Sigma^\rho)v| \leq \varepsilon_2 v^T \Sigma v + k \|v\|_2^2.$$

At this point, we have shown that there exists a (n, k) -GMD with mean and covariance close to that of the discretized Gaussian such that all its constituent probability vectors have coordinates that are integer multiples of $1/n$. Now, for every probability vector \vec{p} with probabilities that are multiples of $1/n$, consider its moment profile $(\mu^{\vec{p}}, \Sigma^{\vec{p}})$, where $\mu^{\vec{p}} = \vec{p}$ and $\Sigma^{\vec{p}}$ are the mean and covariance of the k -CRV with probabilities \vec{p} . Let \mathcal{M} be the set of all possible moment profiles generated by such probability vectors \vec{p} . Since there are at most n^{k-1} probability vectors \vec{p} the set \mathcal{M} has size at most n^{k-1} . Moreover, it is easy to see that for the rounded GMD $\hat{\rho}$, it holds that $(\mu^{\hat{\rho}}, \Sigma^{\hat{\rho}}) \in \mathcal{M}^{\oplus n}$ where $\mathcal{M}^{\oplus n} = \{x \mid \exists x_1, \dots, x_n \in \mathcal{M}, x = \sum_i x_i\}$ denotes the Minkowski addition of \mathcal{M} with itself n times. This is because the mean and covariance of the GMD is equal to the sum of the means and covariances of its constituent CRVs, which are all in \mathcal{M} since each CRV has probabilities that are integer multiples of $1/n$.

Naively searching over $\mathcal{M}^{\oplus n}$ for a GMD that satisfies the guarantees of $\hat{\rho}$ is not easy since it would require time that is exponential in n . To get a computationally efficient algorithm, we search instead in the set $\text{conv}(\mathcal{M}^{\oplus n}) = \text{conv}(\mathcal{M})^{\oplus n}$ where, for a set A , $\text{conv}(A)$ denotes its convex closure, and the equality is a basic property of Minkowski sums. The reason this is easy is that it is solvable by a linear program as follows:

- For $m \in \mathcal{M}$ and $i \in \{1, \dots, n\}$, we assign the variables $x_{i,m} \geq 0$ that denote whether we want to pick the moment profile m for the i -th CRV.
- For all i , we need that $\sum_m x_{i,m} = 1$. This ensures that for all i , $\sum_m x_{i,m} m \in \text{conv}(\mathcal{M})$.
- We need that the aggregate moment profile $(\hat{\mu}, \hat{\Sigma}) = \sum_{i,m} x_{i,m} m$ satisfies the closeness constraints with (μ, Σ) . For all v we require that:

$$|v^T(\mu - \hat{\mu})| \leq \varepsilon_1 \sqrt{v^T \Sigma v} + \sqrt{k} \|v\|_2 \quad \text{and} \quad |v^T(\Sigma - \hat{\Sigma})v| \leq \varepsilon_2 v^T \Sigma v + k \|v\|_2^2.$$

These are all linear constraints so a solution $(\hat{\mu}, \hat{\Sigma}) = \sum_{i,m} x_{i,m} m$, can be computed by solving the linear program using the Ellipsoid method. Note that the constraints of the third bullet are infinitely many but can be verified efficiently using a separation oracle. To check the first set of constraints, we can check whether the optimization problem

$$\min_{\|v\| \leq 1} \varepsilon_1 \sqrt{v^T \Sigma v} + \sqrt{k} \|v\|_2 - v^T (\mu - \hat{\mu})$$

has a negative solution. This is a convex optimization problem which can be solved in polynomial time. To check the second set of constraints, we note that $\varepsilon_2 v^T \Sigma v + k \|v\|_2^2 = v^T (\varepsilon_2 \Sigma + kI) v$. By setting $A \triangleq (\varepsilon_2 \Sigma + kI)$ and $u \triangleq A^{1/2} v$, we can rewrite the constraints as:

$$\frac{|u^T (A^{-1/2})^T (\Sigma - \hat{\Sigma}) A^{-1/2} u|}{u^T u} \leq 1$$

This is equivalent to checking whether the maximum eigenvalue of the matrix $(A^{-1/2})^T (\Sigma - \hat{\Sigma}) A^{-1/2}$ is greater than 1.

At this point, we have efficiently computed a solution $(\hat{\mu}, \hat{\Sigma}) \in \text{conv}(\mathcal{M})^{\oplus n}$ that satisfies the closeness guarantees and we need to convert it to a solution in the set $\mathcal{M}^{\oplus n}$ that is also appropriately close to (μ, Σ) and obtain a GMD with the guarantees of the lemma. By the Shapley-Folkman theorem, it holds that $\text{conv}(\mathcal{M})^{\oplus n} = \mathcal{M}^{\oplus (n-k^2-k)} \oplus \text{conv}(\mathcal{M})^{\oplus (k^2+k)}$ since $\mathcal{M} \subset \mathbb{R}^{k^2+k}$. We can greedily construct such a solution by iteratively picking points $m_i \in \mathcal{M}$ for $i = 1, \dots, (n - k^2 - k)$ such that $((\hat{\mu}, \hat{\Sigma}) - \sum_{j=1}^i m_j) \in \text{conv}(\mathcal{M})^{\oplus (n-i)}$. The Shapley-Folkman theorem for the space $\text{conv}(\mathcal{M})^{\oplus (n-i)}$, guarantees that for all $i \leq (n - k^2 - k)$, a point m_i with the required property always exists. Since membership in $\text{conv}(\mathcal{M})^{\oplus (n-i)}$ can be checked efficiently by writing a linear program similar to the one above, we can efficiently run the above process to generate $(n - k^2 - k)$ CRVs. For the remaining $k^2 + k$ CRVs, we arbitrarily choose points $m_{n-k^2-k+1}, \dots, m_n \in \mathcal{M}$ to obtain a complete (n, k) -GMD ρ' . We argue next that this GMD satisfies the conditions required by the lemma.

For any $m, m' \in \text{conv}(\mathcal{M})$, it holds that $\|m - m'\|_\infty \leq 1$. Moreover, $(\mu^{\rho'}, \Sigma^{\rho'}) = \sum_{i=1}^n m_i$ and $(\hat{\mu}, \hat{\Sigma}) = \sum_{i=1}^{(n-k^2-k)} m_i + \sum_{i=1}^{k^2+k} m'_i$. This implies that $\|\mu^{\rho'} - \hat{\mu}\|_\infty \leq k^2 + k$ and $\|\Sigma^{\rho'} - \hat{\Sigma}\|_{\max} \leq k^2 + k$. We have that:

$$\begin{aligned} |v^T (\mu - \mu^{\rho'})| &\leq |v^T (\mu - \hat{\mu})| + |v^T (\mu^{\rho'} - \hat{\mu})| \\ &\leq \varepsilon_1 \sqrt{v^T \Sigma v} + \sqrt{k} \|v\|_2 + \|v\|_2 \|\mu^{\rho'} - \hat{\mu}\|_2 \\ &\leq \varepsilon_1 \sqrt{v^T \Sigma v} + \sqrt{k} \|v\|_2 + (k^2 + k) \sqrt{k} \|v\|_2 \\ &= \varepsilon_1 \sqrt{v^T \Sigma v} + 3k^{2.5} \|v\|_2 \end{aligned}$$

Similarly, $|v^T (\Sigma - \Sigma^{\rho'}) v| \leq |v^T (\Sigma - \hat{\Sigma}) v| + |v^T (\hat{\Sigma} - \Sigma^{\rho'}) v| \leq \varepsilon_2 v^T \Sigma v + (k^3 + k^2 + k^1) \|v\|_2^2$.

7 A Lower Bound for Covers of PMDs

In this section, we discuss Theorem 4, the lower bound on the size of any ε -cover of (n, k) PMDs. This theorem shows that it is not possible to get significant improvement on the cover size obtained in Theorem 3. In particular, the dependence of the size of the cover on $1/\varepsilon$ is tight up to a difference of 3 in the exponent of $\log(1/\varepsilon)$.

It turns out that it is easy to prove a dependence of $O(n^k)$ on the size of any ε -cover and most of the work is involved in showing a lower bound of $T(k, \varepsilon) = 2^{\log^{k-1}(1/\varepsilon)}$ on the cover size. Thus, in this overview we only focus on the machinery required to show the lower bound of $T(k, \varepsilon)$ on the ε -cover size. We remark that prior to our work, for $k = 2$ (i.e. PBDs), Diakonikolas, Kane, and Stewart obtained a lower bound of $2^{\log^2(1/\varepsilon)}$ [DKS16b].

Showing the lower bound on the cover size is equivalent to showing the existence of $T(k, \varepsilon)$ -many (n_0, k) -PMDs which are all ε -far from each other where $n_0 \leq n$. The usual difficulty in showing cover size lower bounds, is that even if the parameters specifying two PMDs are significantly different, it is not necessarily true that the resulting PMDs are far in total variation distance. In fact, directly arguing that two PMDs are far apart in total variation distance seems difficult. Instead, our strategy is to carefully pick a family of $T(k, \varepsilon)$ PMDs and show that for any two distinct PMDs in this set, there is at least one (k -dimensional) moment $\alpha \in \mathbb{Z}^{+k}$ of size $O(\log(1/\varepsilon))$ such that the α^{th} moment of the two PMDs are ε -far from each other (by size of the moment α , we mean $\|\alpha\|_1$). Usually, gap in moments for two distributions need not translate to significant gap in total variation distance. However, in our setting, we can choose $n_0 \approx \log^k(1/\varepsilon)$. Since n_0 is small, it is easy to show that if two PMDs differ by ε in one of their moments of size $O(\log(1/\varepsilon))$, then they are $\approx \varepsilon$ far in total variation distance (Claim 3).

Note that the α^{th} -moment of a PMD is a multisymmetric polynomial in the parameters of the PMD (i.e. invariant under permuting its summands). Next consider the multidimensional multisymmetric polynomial map where each coordinate in the range corresponds to a moment of size $O(\log(1/\varepsilon))$. Since there are roughly $\Theta_k(\log^k(1/\varepsilon))$ moments of size $O(\log(1/\varepsilon))$, the dimension of the map is $\Theta_k(\log^k(1/\varepsilon))$. The problem of showing lower bounds on the cover size is now equivalent to showing that the range of this map contains $T(k, \varepsilon)$ -many points which are ε -far from each other. In other words, we need a way to show a lower bound on the metric entropy of this polynomial map. Such problems are usually treated with tools of algebraic geometry and we adopt the same strategy. In particular, rather than directly working over the reals, we change the domain to a finite field \mathbb{F} of appropriate size and consider the corresponding polynomial map in \mathbb{F} . Once we are in \mathbb{F} , we apply an extension of Bézout's theorem due to Wooley [Woo96] (Theorem 10) to show that this map has a large number of points in its range when the underlying domain is \mathbb{F} (Lemma 13). Because of the special structure of the polynomials involved, it is possible to show that the presence of a large range in a finite field corresponds to an appropriate lower bound on the metric entropy of the map. We remark that the application of Bézout's theorem in our context is not straightforward. In particular, to apply the theorem, one needs to reason about the Jacobian of this polynomial map. Despite being a very natural family of maps, to the best of our knowledge, properties of the corresponding Jacobian have not been previously investigated.

7.1 Details

We provide the proof of Theorem 4. The proof will use algebraic geometric tools to argue this fact. In particular, the main theorem we will prove will be the following:

Theorem 9. *There are (m, k) -PMDs Z_1, \dots, Z_ℓ such that for all $1 \leq i < j \leq \ell$, $d_{TV}(Z_i, Z_j) \geq \varepsilon$ and $\ell = 2^{\tilde{\Omega}(\log^{k-1}(1/\varepsilon))}$ where $m = O(\log^{k-1}(1/\varepsilon))$.*

We will now prove Theorem 4 using Theorem 9.

Proof. Note that by assumption $n > 2m$. It is easy to observe that for any $\alpha = (\alpha_1, \dots, \alpha_k) \in \mathbb{Z}^k$ such that $\sum \alpha_i = n/2$, there are k CRVs $X_1, \dots, X_{n/2}$ such that $X_1 + \dots + X_{n/2}$ is supported on α . Now, consider any $\alpha, \beta \in \mathbb{Z}^k$ such that $\|\alpha - \beta\|_1 > m$. Then, for any (m, k) PMD Z_i, Z_j , the supports

of $Z_i + \alpha$ and $Z_j + \beta$ are disjoint. It is now easy to see that we can choose $L = (\frac{n}{2m})^{k-1}$ points $\alpha^{(1)}, \dots, \alpha^{(L)}$ such that for $1 \leq j \leq L$, $\sum_{i=1}^k \alpha_i^{(j)} = n/2$ and $\|\alpha^{(j)} - \alpha^{(\ell)}\|_1 \geq m$ whenever $j \neq \ell$. Now, let Z_{i_1} and Z_{i_2} be two (m, k) PMDs from Theorem 9. Then, both $Z_{i_1} + \alpha^{(j)}$ and $Z_{i_2} + \alpha^{(\ell)}$ are (n, k) PMDs and further, $d_{TV}(Z_{i_1} + \alpha^{(j)}, Z_{i_2} + \alpha^{(\ell)}) \geq \varepsilon$. This gives a set of $L \cdot 2^{\tilde{\Omega}(\log^{k-1}(1/\varepsilon))}$ (n, k) -PMDs which are ε -far from each other. \square

Thus, it remains to prove Theorem 9. The proof of this theorem shall involve a combination of ideas using combinatorics of multisymmetric polynomials and tools from algebraic geometry. In particular, instead of directly arguing about total variation distance of PMDs, we will argue about the moments of PMDs. We first observe that for any (m, k) PMD Z , we can associate a matrix $P_Z \in \mathbb{R}^{m \times (k-1)}$ where the entries of the matrix are non-negative such that the entries of any row sum to at most 1. The semantics of the matrix are that $Z = X_1 + \dots + X_n$ where each X_i is an independent CRV with $\Pr[X_i = \mathbf{e}_j] = P_Z[i, j]$ (if $1 \leq j < k$) and $\Pr[X_i = \mathbf{e}_k] = 1 - \sum_{j < k} P_Z[i, j]$. Clearly, the distribution of Z is invariant under permuting the rows of P_Z . Further, up to permutations of columns, the matrix P_Z associated with such a Z is unique.

If Z is a (m, k) PMD and $\alpha \in \mathbb{Z}^{+k}$, then $M_\alpha(Z) = \mathbf{E}[Z^\alpha]$ i.e. the α^{th} moment of Z . Here Z^α is an abbreviation of $\prod_{i=1}^k Z_i^{\alpha_i}$ where Z_i the i^{th} component of Z . Thus, $M_\alpha(Z)$ is the α^{th} moment of Z . The following will be a very useful observation.

Observation 2. $M_\alpha(Z)$ is a multisymmetric polynomial of degree $\|\alpha\|$ in the variables $\{P_Z[i, j]\}$ where $1 \leq i \leq m$ and $1 \leq j < k$. By multisymmetric, we mean the polynomial is invariant under permuting the rows of P_Z .

While the moment $M_\alpha(Z)$ is very natural to consider, for the purposes of proving Theorem 9, it will be useful to define two more families of multisymmetric polynomials (in the variables $\{P_Z[i, j]\}_{1 \leq i \leq m, 1 \leq j < k}$). These polynomials will be the elementary multisymmetric polynomials and the power sum multisymmetric polynomials. Also, from now, we will only restrict our attention to α where the last entry is 0. This is because, for any $\beta \in \mathbb{Z}^{+k}$, $M_\beta(Z)$ can be expressed as a polynomial combination of $M_\alpha(Z)$ where $\alpha_k = 0$.

Definition 13. Let $\alpha \in \mathbb{Z}^{+k}$. Let $\beta \in \mathbb{Z}^{|\alpha|}$ such that the entry i occurs exactly α_i times in β . While there are multiple choices for such a β , any canonical choice is good enough. Let $1 \leq i_1 < \dots < i_{|\alpha|} \leq m$. Then,

$$\mathbf{E}_\alpha(Z) = \sum_{\sigma: \{1, \dots, \|\alpha\|\} \rightarrow [m]} \prod_{j=1}^{\|\alpha\|} P_Z[\sigma(i_j), \beta_j],$$

where the sum is taken over all surjections σ from $\{1, \dots, \|\alpha\|\} \rightarrow [m]$ where for $i < j$, if $\beta_i = \beta_j$, then $\sigma(i) < \sigma(j)$.

The definition above might look a little involved, so to illustrate the concept, we will consider a simple example. For example, if $k = 3$ and $\alpha = (1, 2, 0)$, then

$$\mathbf{E}_\alpha(Z) = \sum_{j < k, i \neq j, k} P_Z[i, 1] P_Z[j, 2] P_Z[k, 2].$$

Likewise, if $k = 4$ and $\alpha = (1, 1, 1, 0)$, then

$$\mathbf{E}_\alpha(Z) = \sum_{i, j, k \text{ all are distinct}} P_Z[i, 1] P_Z[j, 2] P_Z[k, 3].$$

Another family of polynomials which will be very useful in our reasoning will be the family of power sum multisymmetric polynomials.

Definition 14. Let $\alpha \in \mathbb{Z}^{+k}$. Then,

$$\mathbf{P}_\alpha(Z) = \sum_{1 \leq i \leq m} \prod_{j=1}^m P_Z[i, j]^{\alpha_j}.$$

We remark that the polynomials we have defined so far i.e. $M_\alpha(\cdot)$, $\mathbf{E}_\alpha(\cdot)$ and $\mathbf{P}_\alpha(\cdot)$ are well-defined formal polynomials and make sense even if the matrix $P_Z[i, j]$ has entries from field \mathbb{F} whose characteristic is not zero. This shall be useful for us going forward. For the moment, we will consider the connection between the family \mathbf{E}_α and M_α . We will require the following definition. For two vectors, $\alpha, \beta \in \mathbb{Z}^{+k}$, we say $\alpha \preceq \beta$ if $\alpha_i \leq \beta_i$ for $1 \leq i \leq k$. The first is the following observation.

Observation 3. For any (m, k) -PMD Z and $\alpha \in \mathbb{Z}^{+k}$ (with $\alpha_k = 0$), $M_\alpha(Z)$ can be expressed as a linear combination of $\mathbf{E}_\beta(Z)$ where $\beta \preceq \alpha$.

Proof. As we have already observed, $M_\alpha(Z)$ is a multisymmetric polynomial in entries of the matrix P_Z at degree is at most $\|\alpha\|$. To prove that is a linear combination of \mathbf{E}_β for $\beta \preceq \alpha$, it suffices to make the following observation: Note that $M_\alpha(Z)$ is

$$M_\alpha(Z) = \mathbf{E} \left[\prod_{j=1}^{k-1} \left(\sum_{i=1}^m X_{i,j} \right)^{\alpha_j} \right].$$

Observe that any monomial where $X_{i,j}$ and $X_{i,\ell}$ appear together with $j \neq \ell$ vanishes under the expectation. Likewise, since $X_{i,j}$ is supported on $\{0, 1\}$, hence $X_{i,j}^\ell = X_{i,j}$ for any $\ell \geq 1$. These two observations coupled with each other imply that M_α is a linear combination of $\mathbf{E}_\beta(\cdot)$ for $\beta \preceq \alpha$. \square

The next lemma implies bounds on the coefficients of \mathbf{E}_β in expressing M_α . Let us now assume that $M_\alpha = \sum_{\beta \preceq \alpha} \gamma_\beta \cdot \mathbf{E}_\beta$. It is easy to see the following claim.

Claim 1. For α with $\alpha_k = 0$, we have $\gamma_\alpha = \prod_{j=1}^{k-1} \alpha_j!$.

The next claim is also fairly easy to prove.

Claim 2. $\sum_{\beta \preceq \alpha} |\gamma_\beta| = 2^{O(k)} \cdot \prod_{j=1}^{k-1} \alpha_j!$.

Proof. Let $c_1, \dots, c_{k-1} \in \mathbb{Z}^{+k}$ be defined as the following:

$$c_j = \alpha \cdot I_j.$$

In other words, c_j is obtained by a pointwise product of α and the indicator vector of the singleton set $\{j\}$. Now, assume that for $1 \leq j \leq k-1$,

$$M_{c_j}(Z) = \sum_{\beta_j \preceq c_j} \gamma_{\beta_j} \cdot \mathbf{E}_{\beta_j}.$$

Then, it is not difficult to see that

$$M_\alpha(Z) = \sum_{\beta_1 \preceq c_1} \dots \sum_{\beta_{k-1} \preceq c_{k-1}} \mathbf{E}_\beta \cdot \prod_{j=1}^{k-1} \gamma_{\beta_j},$$

where $\beta = \beta_1 + \dots + \beta_{k-1}$. Thus, to prove our claim, it suffices to show that for any particular $1 \leq j \leq k-1$,

$$\sum_{\beta_j \preceq c_j} |\gamma_{\beta_j}| = O(1) \cdot \alpha_j!.$$

Note that c_j is just α_j at the j^{th} position and zero everywhere else. We introduce the following notation: For any integer k , we let $\mathcal{P}(k)$ denote the set of its partitions i.e. a tuple of strictly positive integers summing to k ordered in decreasing sequence. For example, for $k = 5$, we have 6 distinct partitions $(5), (4, 1), (3, 2), (3, 1, 1), (2, 2, 1), (2, 1, 1, 1)$. For any partition $P \in \mathcal{P}(k)$, we use $s(P)$ to denote the number of summands in P . For example, for the partition $P = (3, 2)$, $s(P) = 2$. Further, if $P = (x_1, \dots, x_k)$ is a partition of n , then

$$\binom{n}{P} = \binom{n}{x_1 \dots x_k}.$$

With these notations in place, it is easy to see that

$$\gamma_{\beta_j} = \sum_{P \in \mathcal{P}(\alpha_j): s(P) = \beta_j} \binom{\alpha_j}{P}$$

Now, it is easy to see that for any integer $x > 1$, $(1.4)^x \leq x!$. If $1(P)$ denotes the number of 1 in the partition P . With this, we have

$$|\gamma_{\beta_j}| \leq \sum_{P \in \mathcal{P}(\alpha_j): s(P) = \beta_j} \frac{\alpha_j!}{(1.4)^{\alpha_j}} \cdot 1.4^{1(P)}.$$

Thus, implies that

$$\sum_{\beta_j \preceq c_j} |\gamma_{\beta_j}| \leq \sum_{\beta_j \preceq c_j} \sum_{P \in \mathcal{P}(\alpha_j): s(P) = \beta_j} \frac{\alpha_j!}{(1.4)^{\alpha_j}} \cdot 1.4^{1(P)}.$$

Now, note that the total number of partitions of α_j with t ones in it is upper bounded by $|\mathcal{P}(\alpha - t)|$. However, it is a well-known fact in number theory, that $|\mathcal{P}(\alpha - t)| \leq 2^{O(\sqrt{\alpha - t})}$. Thus,

$$\sum_{\beta_j \preceq c_j} |\gamma_{\beta_j}| \leq \sum_{x=0}^{\alpha_j} \frac{\alpha_j!}{(1.4)^{\alpha_j}} \cdot 1.4^x \cdot 2^{O(\sqrt{\alpha_j - x})} \leq \alpha_j! \cdot \int_{x=0}^{\alpha_j} 1.4^{O(\sqrt{x}) - x} dx = O(\alpha_j!).$$

This finishes the proof. \square

Thus, using the last two claims, we infer that there is a linear map which given any $\alpha \in \mathbb{Z}^{+k}$ (with $\alpha_k = 0$), maps the set $\{\mathbf{E}_{\beta}(Z)\}_{\beta \preceq \alpha}$ to the set $\{M_{\beta}(Z)\}_{\beta \preceq \alpha}$. The next lemma bounds the condition number of this map.

Lemma 10. *Let Z and Z' be two (m, k) -PMDs such that $|\mathbf{E}_{\beta}(Z) - \mathbf{E}_{\beta}(Z')| \geq \delta$. Then, there exists $\beta_0 \preceq \beta$ such that*

$$|M_{\beta_0}(Z) - M_{\beta_0}(Z')| \geq \delta \cdot c^{-d}.$$

where $c = 2^{O(k)}$ is the constant appearing in Claim 2.

Proof. Let $c = 2^{O(k)}$ be the constant appearing in Claim 2. Let $|\beta| = d$ and i be the smallest integer such that there exists a $\beta_0 \preceq \beta$ with $|\beta_0| = i$ and

$$|\mathbf{E}_{\beta_0}(Z) - \mathbf{E}_{\beta_0}(Z')| \geq \delta \cdot (2c)^{i-d}.$$

Note that by assumption, there exists such a β_0 . Next,

$$\begin{aligned} |M_{\beta_0}(Z) - M_{\beta_0}(Z')| &= \left| \sum_{\kappa \preceq \beta_0} \gamma_{\kappa} \cdot (\mathbf{E}_{\kappa}(Z) - \mathbf{E}_{\kappa}(Z')) \right| \\ &\geq \gamma_{\beta_0} \cdot |(\mathbf{E}_{\beta_0}(Z) - \mathbf{E}_{\beta_0}(Z'))| - \left| \sum_{\kappa \prec \beta_0} \gamma_{\kappa} \cdot (\mathbf{E}_{\kappa}(Z) - \mathbf{E}_{\kappa}(Z')) \right| \end{aligned}$$

Applying Claim 1 and Claim 2, we get that

$$|M_{\beta_0}(Z) - M_{\beta_0}(Z')| \geq \prod_{i=1}^{k-1} \beta_{0,i}! \left(|(\mathbf{E}_{\beta_0}(Z) - \mathbf{E}_{\beta_0}(Z'))| \right) - c \cdot \prod_{i=1}^{k-1} \beta_{0,i}! \max_{\kappa \preceq \beta} |(\mathbf{E}_{\kappa}(Z) - \mathbf{E}_{\kappa}(Z'))|$$

Again applying the hypothesis on β_0 , we have

$$\begin{aligned} |M_{\beta_0}(Z) - M_{\beta_0}(Z')| &\geq \prod_{i=1}^{k-1} \beta_{0,i}! \cdot \delta \cdot (2c)^{\|\beta_0\| - d} - c \cdot \prod_{i=1}^{k-1} \beta_{0,i}! \cdot \delta \cdot (2c)^{\|\beta_0\| - d - 1} \\ &\geq \prod_{i=1}^{k-1} \beta_{0,i}! \cdot \delta \cdot c^{\|\beta_0\| - d} \geq \delta \cdot c^{-d}. \end{aligned}$$

□

The strategy for the rest of the proof is as follows: Instead of showing Theorem 9, we will show the following lemma.

Lemma 11. *There are (m, k) -PMDs Z_1, \dots, Z_{ℓ} such that $\ell = 2^{\tilde{\Omega}(\log^{k-1}(1/\varepsilon))}$, $m = O(\log^{k-1}(1/\varepsilon))$ and for every $1 \leq i < j \leq \ell$, there exists some $\alpha \in \mathbb{Z}^{+k}$ such that $\alpha_k = 0$, $\|\alpha\| = \tilde{O}(\log(1/\varepsilon))$ and $|M_{\alpha}(Z_i) - M_{\alpha}(Z_j)| \geq \varepsilon$.*

To see why it suffices to prove Lemma 11, we have the following claim.

Claim 3. *Let Z_1, Z_2 be two (m, k) -PMDs and $\alpha \in \mathbb{Z}^k$ such that $|M_{\alpha}(Z_1) - M_{\alpha}(Z_2)| \geq \delta$. Then, $d_{TV}(Z_1, Z_2) \geq \delta \cdot m^{-\|\alpha\|_1}$.*

Proof. Assume towards a contradiction that $d_{TV}(Z_1, Z_2) < \delta \cdot m^{-\|\alpha\|_1}$. By definition, this means that there is a coupling (Z'_1, Z'_2) such that the marginal Z'_1 is distributed as Z_1 , the marginal Z'_2 is distributed as Z_2 and $\Pr[Z'_1 \neq Z'_2] < \delta \cdot m^{-\|\alpha\|_1}$. As the support of both Z_1 and Z_2 is confined in the box $[0, m]^k$, it easily follows that $|M_{\alpha}(Z_1) - M_{\alpha}(Z_2)| < \Pr[Z'_1 \neq Z'_2] \cdot m^{\|\alpha\|_1} < \delta$. This results in a contradiction, thus completing the proof. □

In light of Lemma 10, it instead suffices to prove the following lemma.

Lemma 12. *There are (m, k) -PMDs Z_1, \dots, Z_{ℓ} such that $\ell = 2^{\tilde{\Omega}(\log^{k-1}(1/\varepsilon))}$, $m = O(\log^{k-1}(1/\varepsilon))$ and for every $1 \leq i < j \leq \ell$, there exists some $\alpha \in \mathbb{Z}^{+k}$ such that $\alpha_k = 0$, $\|\alpha\| = \tilde{O}(\log(1/\varepsilon))$ and $|\mathbf{E}_{\alpha}(Z_i) - \mathbf{E}_{\alpha}(Z_j)| \geq \varepsilon$.*

The rest of the proof is towards proving Lemma 12. As we said before, the proof is going to involve use of algebraic geometry tools. In fact, to prove Lemma 12, instead of considering the matrices P_Z to be real-valued matrices, we will instead first show an equivalent version of Lemma 12 over a finite field \mathbb{F} of appropriate size. This change to finite fields will make it easier to apply tools of algebraic geometry. In particular, we will prove the following lemma.

Lemma 13. *For any integer $d \in \mathbb{N}$ and any finite field \mathbb{F} of size $2 \cdot d \cdot m(k-1)$, there are ℓ matrices A_1, \dots, A_ℓ in $\mathbb{F}^{m \times (k-1)}$ where $\ell = 2^{\tilde{\Omega}(\log^{k-1}(1/\varepsilon))}$, $m = O(\log^{k-1}(1/\varepsilon))$ and for $1 \leq i < j \leq \ell$, there exists $\alpha \in \mathbb{Z}^{+k}$ where $\|\alpha\| \leq d$, $\alpha_k = 0$ and*

$$\mathbf{E}_\alpha(A_i) \neq \mathbf{E}_\alpha(A_j).$$

Before, we prove Lemma 13, let us see why it implies Lemma 12. To get PMD matrices from the matrices A_1, \dots, A_ℓ , we use the following map.

$$A_i \mapsto P_{Z_i} \text{ where } P_{Z_i}[j, j'] = \begin{cases} \frac{A_i[j, j']}{2k|\mathbb{F}|} & \text{if } j' < k \\ 1 - \sum_{j'' < k} P_{Z_i}[j, j''] & \text{if } j' = k \end{cases}$$

It is easy to see that this operation defines legitimate (m, k) PMD matrices. Further, note that \mathbf{E}_α is a homogenous polynomial of degree $\|\alpha\|$. Thus, it is easy to see that if $\mathbf{E}_\alpha(A_i) \neq \mathbf{E}_\alpha(A_j)$, then

$$|\mathbf{E}_\alpha(Z_i) - \mathbf{E}_\alpha(Z_j)| \geq \frac{1}{|\mathbb{F}|^{\|\alpha\|}}.$$

By choosing $|\mathbb{F}|$ to be a field of size $O(k \cdot \log(1/\varepsilon))$, we immediately see that it implies the bounds in Lemmas 12. Thus, all that remains to be proven here is Lemma 13. To prove this, let us set $d = \tilde{O}(\log(1/\varepsilon))$ and let $\mathcal{S} = \{\alpha \in \mathbb{Z}^{+k} : \alpha_k = 0 \text{ and } \|\alpha\| \leq d\}$. We now define the map $\mathbf{E}_\mathcal{S} : \mathbb{F}^{m \times (k-1)} \rightarrow \mathbb{F}^\mathcal{S}$ which is a multidimensional map indexed by \mathcal{S} where the coordinate for $\alpha \in \mathbb{Z}^{+k}$ is $\mathbf{E}_\alpha(\cdot)$. Note that Lemma 13 amounts to showing a lower bound on the entropy of this map.

We will need the notion of Jacobian of a map which is defined next.

Definition 15. *Let \mathbb{F} be any field and let $M : \mathbb{F}^n \rightarrow \mathbb{F}^m$. Then, the Jacobian of M , denoted by J_M is the $m \times n$ matrix in $\mathbb{F}[x_1, \dots, x_n]$ where the $(i, j)^{\text{th}}$ entry is given by $\partial M_i(x_1, \dots, x_n) / \partial x_j$ where M_i denotes the i^{th} coordinate of the map M .*

For us, the utility of Jacobian will come from its role in the following theorem.

Theorem 10. *[Woo96] Let \mathbb{F} be a prime field of size p . Let k and d be integers. Let $M : \mathbb{F}^s \rightarrow \mathbb{F}^s$ be such that any coordinate is a polynomial map of degree at most d . For $a \in \mathbb{F}^s$, let*

$$N_a = |\{c \in \mathbb{F}^s : M(c) = a \text{ and } J_M(c) \neq 0\}|$$

Then, for every $a \in \mathbb{F}^s$, $N_a \leq d^s$.

As a consequence, we have the following corollary.

Corollary 2. *Let \mathbb{F} be a prime field of size p . Let s and d be integers. Let $M : \mathbb{F}^{s'} \rightarrow \mathbb{F}^s$ be such that any coordinate is a polynomial map of degree at most d . Let us assume that $\text{rank}(J_M) = s$. If $|\mathbb{F}| > 2 \cdot d \cdot s$, then, $|\text{Range}(M)| \geq |\mathbb{F}|^s / 2d^s$.*

Proof. Since $\text{rank}(J_M) = s$, it means that there is a submatrix of size $s \times s$ (call it J'_M) such that $\det J_{M'} \neq 0$ (here the determinant is evaluated over the field of rational functions over \mathbb{F}). Let the s columns correspond to the set of variables \mathcal{L} . Since the determinant is a low degree polynomial (of degree at most $d \cdot s$ in each variable), hence there is choice of the variables outside \mathcal{L} to some values in \mathbb{F} such that $\det J_{M'} \neq 0$.

For this setting of variables, let $M' : \mathbb{F}^{\mathcal{L}} \rightarrow \mathbb{F}$ be the map restricted to the variables in \mathcal{L} . Since $\det J_{M'} \neq 0$, hence $J_{M'}$ is a non-zero polynomial of degree at most $d \cdot s$. Since $|\mathbb{F}| > 2 \cdot d \cdot s$, hence by Schwartz-Zippel lemma, the set $C = \{c : \det J_M(c) \neq 0\}$ has size at least $|\mathbb{F}|^s/2$. Applying Theorem 10, we get the stated claim. \square

To show a lower bound on the entropy of M_S , we will apply Corollary 2. To apply this, we need to prove that $\det J_{M_S} \neq 0$. It is not clear how to show this, so we introduce an intermediate map.

Definition 16. For $\alpha \in \mathbb{Z}^k$ (with $\alpha_k = 0$), we define the map $\mathbf{P}_\alpha : \mathbb{F}^{m \cdot (k-1)} \rightarrow \mathbb{F}$ where

$$\mathbf{P}(A) \mapsto \sum_{i=1}^n \prod_{j=1}^{k-1} A[i, j]^{\alpha_j}.$$

The family $\{\mathbf{P}_\alpha\}$ is usually referred to as power-sum multisymmetric polynomials.

The idea here will be that we will relate the family \mathbf{P}_α and \mathbf{E}_α and then argue about the Jacobian of a map defined in terms of \mathbf{P}_α . The following relation between \mathbf{E}_α and \mathbf{P}_α was established in Dalbec [Dal99].

Proposition 10.

$$\|\alpha\| \cdot \mathbf{E}_\alpha + \sum_{\substack{\alpha = \beta + \gamma, \\ \beta, \gamma \neq 0}} (-1)^{\|\beta\|} \binom{\|\beta\|}{\beta} \mathbf{P}_\beta \cdot \mathbf{E}_\gamma + (-1)^{\|\alpha\|} \binom{\|\alpha\|}{\alpha} \mathbf{P}_\alpha = 0.$$

Using induction, the following lemma is immediate.

Lemma 14. For any α , if either the characteristic of \mathbb{F} is 0 or is more than $\|\alpha\|$, there exist \mathbf{Q}_α such that

$$\mathbf{P}_\alpha = \mathbf{Q}_\alpha \left(\left\{ \mathbf{E}_\beta \right\}_{\beta \preceq \alpha} \right)$$

Rather than considering the map \mathbf{E}_α , we will consider a restricted version of it. In particular, choose some matrix $A \in \mathbb{F}^{m \cdot (k-2)}$. The exact choice will be specified later. However, given $x \in \mathbb{F}^m$, we can consider a matrix $A_x \in \mathbb{F}^{m \cdot (k-1)}$ which is obtained by concatenating x with A where x is the last row of A_x whereas the first $(k-1)$ rows are formed by A . Thus, fixing this choice of A , for every $\alpha \in \mathbb{Z}^{+k}$ with $\alpha_k = 0$, we can define the map

$$\mathbf{E}'_\alpha : x \mapsto \mathbf{E}_\alpha(A_x).$$

Our aim will be to argue that the map $\mathbf{E}'_S : \mathbb{F}^m \rightarrow \mathbb{F}^S$ (defined analogously to \mathbf{E}_α) has full rank and thus Corollary 2 is applicable here. Note that, we can also define the map $\mathbf{P}'_\alpha : \mathbb{F}^m \rightarrow \mathbb{F}$ and $\mathbf{P}'_S : \mathbb{F}^m \rightarrow \mathbb{F}^S$ analogously. As a consequence of Lemma 14, we have that for $\mathbf{P}'_\alpha(z_1, \dots, z_m)$ and for $1 \leq j \leq m$,

$$\frac{\partial \mathbf{P}'_\alpha}{\partial z_j} = \sum_{\gamma \preceq \alpha} \frac{\partial \mathbf{Q}_\alpha \left(\left\{ \mathbf{E}'_\beta \right\}_{\beta \preceq \alpha} \right)}{\partial \mathbf{E}'_\gamma} \cdot \frac{\partial \mathbf{E}'_\gamma}{\partial z_j}.$$

Consider the field $\mathbb{K} = \mathbb{F}(z_1, \dots, z_m)$ (i.e. the field of rational functions over \mathbb{F} in the variables z_1, \dots, z_m). If $J_{\mathbf{P}'_S}$ and $J_{\mathbf{E}'_S}$ are the Jacobians of the maps \mathbf{P}'_S and \mathbf{E}'_S respectively, then this immediately implies that there exists a matrix $B \in \mathbb{K}^{S \times m}$ such that

$$J_{\mathbf{P}'_S} = B \cdot J_{\mathbf{E}'_S}.$$

Immediately, we have that $\text{rank}(J_{\mathbf{P}'_S}) \leq \text{rank}(J_{\mathbf{E}'_S})$. Thus, to show a lower bound on $\text{rank}(J_{\mathbf{E}'_S})$, it suffices to show a lower bound on $\text{rank}(J_{\mathbf{P}'_S})$. We next show the following claim.

Claim 4. *Over \mathbb{K} , $\text{rank}(J_{\mathbf{P}'_S}) \geq |\mathcal{S}|/k^k$ provided $|\mathbb{F}| > d/k$.*

Proof. Consider the row in $J_{\mathbf{P}'_S}$ corresponding to $\alpha \in \mathbb{Z}^{+k}$ where $\alpha_k = 0$. Let us denote it by $J_{\mathbf{P}'_\alpha}$. It is given by

$$J_{\mathbf{P}'_\alpha} = \alpha_k \cdot \left[z_1^{\alpha_{k-1}} \cdot \prod_{j=1}^{k-2} A[1, j]^{\alpha_j} \dots \dots z_m^{\alpha_{k-1}} \cdot \prod_{j=1}^{k-2} A[m, j]^{\alpha_j} \right]$$

Now, consider the m points in \mathbb{F}^{k-1} given by A_z where $z = (z_1, \dots, z_m)$. Call these points y_1, \dots, y_m . Then, up to the scaling factor α_k , $J_{\mathbf{P}'_\alpha}$ is simply the evaluation of the monomial $\mathbf{y}^{\alpha'}$ where $\alpha' = \alpha - \mathbf{e}_k$ at the points y_1, \dots, y_m . Thus, if we restrict our attention to those α such that $\alpha_k \neq 0$, then these rows constitute the multivariate interpolation matrix for the monomials given by such α 's at the points y_1, \dots, y_m . We would like to prove the non-singularity of this multivariate interpolation matrix. Let us look at the subset of \mathcal{S} such that $1 \leq \alpha_{k-1} \leq d/k$ and $0 \leq \alpha_k < d/k$. While this is a subset of \mathcal{S} , note that the size of this subset is at least $|\mathcal{S}|/k^k$. Further, now, let us assume our points y_1, \dots, y_m are obtained as follows: Choose some subset L of \mathbb{F} of size at least d/k and consider the $(d/k)^k$ obtained by taking a direct product of these points. The interpolation matrix is then a k -fold tensor product of the univariate interpolation matrix at L (of degree d/k). If all the points in L are distinct and non-zero, then the univariate interpolation matrix is the Vandermonde matrix which has a non-zero determinant. This will imply that its k -fold tensor product has a non-zero determinant concluding the proof. \square

This implies that $\text{rank}(J_{\mathbf{E}'_S}) \geq d^k/k^k$. This means that we can choose a square submatrix of size $(d/k)^k \times (d/k)^k$ of $J_{\mathbf{E}'_S}$ of full rank. This means there is a subset of \mathcal{S} of size d^k/k^k (call it \mathcal{S}') such that

$$\text{rank}(J_{\mathbf{E}'_{\mathcal{S}'}}) = |\mathcal{S}'|.$$

Now, applying Corollary 2 to the map $\mathbf{E}'_{\mathcal{S}'}$, we see that the range of the map has size $2^{\tilde{\Omega}(\log^{k-1}(1/\varepsilon))}$. This immediately proves Lemma 13.

8 A Fourier-Based Learning Algorithm for PMDs

In this section, we discuss Theorem 5, our learning result for PMDs. Our technique crucially uses Fourier analysis. We note that the recent work of Diakonikolas, Kane, and Stewart [DKS16b] also uses Fourier analysis to learn k -SIIRVs, i.e. sums of independent integer valued random variables taking values in $\{0, 1, \dots, k-1\}$. We note that our use of Fourier analysis is somewhat different from theirs. In particular, [DKS16b] use the Fourier transform over some discrete group \mathbb{Z}_m for an appropriately chosen m . In contrast, we do the usual Fourier analysis over \mathbb{Z}^k . It turns out doing Fourier analysis over \mathbb{Z}^k (rather than a finite group) avoids many problems and may be viewed as the natural domain for Fourier analysis for such problems.

We believe the application of Fourier analysis to learn such structured distributions is interesting in its own right and might have application in the future towards obtaining learning algorithms for

other interesting classes of distributions. In particular, the recent work on the population recovery problem [WY12, MS13, LZ15] may also be viewed as an example of use of Fourier analysis towards learning of structured distributions.

We now give a high level description of our learning algorithm. The (n, k) -PMD Z , that we are aiming to learn is supported on \mathbb{Z}^k and hence the Fourier transform \widehat{Z} is defined for every $\xi \in [-1, 1]^k$ as $\widehat{Z}(\xi) = \mathbf{E}[e^{i \cdot \pi \cdot \langle \xi, Z \rangle}]$. While our actual algorithm does not perform Fourier inversion explicitly, it resembles Fourier inversion fairly closely. For the moment, assume that we are performing Fourier inversion. It immediately becomes clear that a vanilla Fourier inversion will not work – this is because the Fourier transform is supported on $[-1, 1]^k$ which is an uncountable set and thus we cannot evaluate $\widehat{Z}(\cdot)$ at all points of the support. Rather what we show is that the Fourier transform of a PMD decays exponentially around any point of the form $\{-1, 0, 1\}^k$. In particular, if Σ is the covariance matrix of the PMD, then we show that for $\xi \in [-1/2, 1/2]^k$,

$$|\widehat{Z}(\xi)| = e^{-\Theta(1)\xi^T \Sigma \xi}.$$

Refer to Corollary 3 for the precise bounds. Similar exponential decay of Fourier transform is also true around the other points of the form $\{-1, 0, 1\}^k$. Let us use $V = \prod_{i=1}^k (1 + \sigma_i)$ where σ_i^2 are the eigenvalues of Σ . It is not difficult to show that all but an ε -fraction of the mass of Z falls on a set of size $V \cdot \log^k(1/\varepsilon)$ (Lemma 18). On the other hand, using the exponential decay of the Fourier transform, we have the following crucial claim: We identify a region $\mathcal{S} \subseteq [-1, 1]^k$ of volume $\log^k(1/\varepsilon)/V$ such that

$$\int_{\xi \notin \mathcal{S}} |\widehat{Z}(\xi)|^2 d\xi \leq \widetilde{O}_k\left(\frac{\varepsilon}{V}\right). \quad (9)$$

Refer to Claim 7 for the precise bounds. Also, in this informal description, we use \widetilde{O} to hide the dependence on k as well as the polylogarithmic factors of $1/\varepsilon$. This implies that if H is another function such that $|\widehat{H}(\xi) - \widehat{Z}(\xi)| \leq \varepsilon$ inside \mathcal{S} and 0 outside \mathcal{S} , then

$$\int_{\xi \in [-1, 1]^k} |\widehat{H}(\xi) - \widehat{Z}(\xi)|^2 d\xi \leq \widetilde{O}_k\left(\frac{\varepsilon}{V}\right). \quad (10)$$

By using Plancherel's identity and Cauchy-Schwarz, it immediately follows that $\sum_{z \in \mathbb{Z}^k} |H(z) - Z(z)| \leq \widetilde{O}(\varepsilon)$. In other words, if we perform Fourier inversion by estimating \widehat{Z} pointwise to error ε within \mathcal{S} and setting it to be 0 outside \mathcal{S} , then the ℓ_1 distance between our hypothesis and Z is $\widetilde{O}(\varepsilon)$. We remark that the factor $1/V$ that we get in (9) and (10) is crucial for our algorithm to succeed. The only detail we have not specified is how to approximate \widehat{Z} to error ε inside \mathcal{S} . Note that \mathcal{S} still has infinitely many points. However, what we show is that there is a carefully chosen grid $\mathcal{S}_{\text{grid}}$ of size $\widetilde{O}_k((1/\varepsilon)^k)$ such that estimating $\widehat{Z}(\xi)$ on $\mathcal{S}_{\text{grid}}$ to error ε suffices to estimate $\widehat{Z}(\xi)$ on \mathcal{S} (to error 2ε). This is done by assigning the estimate of \widehat{Z} of the nearest grid point. This uses the choice of the grid points in \mathcal{S} along with the Lipschitz property of the Fourier transform. Note that since we are evaluating the Fourier transform at $(1/\varepsilon)^k$ points to error ε , we need $\widetilde{O}_k(1/\varepsilon^2)$ samples.

One caveat that remains to be discussed is that we have not commented on the time complexity of the Fourier inversion algorithm. In the actual algorithm, we do not perform Fourier inversion out of concerns of time complexity and the fact that the resulting measure obtained from Fourier inversion while computable need not be samplable. Instead, we use the structural characterization of PMDs from [DKT15] to decompose $Z \approx G + S$ where G is a discretized Gaussian and S is a $(\text{poly}(k/\varepsilon), k)$ PMD (Theorem 6). Using samples from Z , we can spectrally approximate its covariance matrix, which then gives us a good handle on the covariance matrix of G , as S has small

size. In particular, we can construct a $(1/\varepsilon)^{O(k)}$ -size spectral cover for the covariance matrix of G using the covariance matrix of Z . So we can assume that G is essentially known, and the challenge is to uncover S , using samples from Z . Of course, Z is not actually equal to $G + S$, but if our overall algorithm uses $\ell = \tilde{O}_k(1/\varepsilon^2)$ samples, and we have approximate equality of Z and $G + S$ to within variation distance $O(1/\ell^2)$, say, then we can pretend that Z is actually equal to $G + S$ for the purposes of our analysis (Claim 10). So knowing G , and getting samples from $G + S$ we need to uncover S . We follow a linear programming approach to find the probability density of S . We enforce constraints on this density so that the Fourier transform of $G + S$ approximately matches the empirical Fourier transform of Z . Our choice of the error and points at which we evaluate \hat{Z} and enforce this constraint is informed by the discussion above. What is crucial here is that the Fourier transform of S is a linear function of its probability density and thus we are left to solve a system of linear constraints.

8.1 Fourier Properties of PMDs

The main idea behind learning PMDs is to look at the Fourier spectrum of PMDs. Specifically, we will prove two structural results about PMDs. One is that the Fourier spectrum of PMDs (roughly) has an exponential decay around the origin. The second result we will prove is the Fourier spectrum is a Lipschitz function and thus to estimate the Fourier spectrum in the entire domain, it suffices to compute it at a few points. Combining these two results along with standard statements on Fourier inversion show that if we construct a hypothesis distribution which approximates the Fourier spectrum of the target PMD at the chosen points and also exhibits a similar exponential decay in the Fourier spectrum, then the hypothesis distribution is close to the target PMD. While the condition on Fourier decay is not algorithmically easy to impose, we show that using some ideas from [DKT15], the problem of imposing these constraints reduces to linear programming. We will first quickly review the notion of Fourier spectrum of integer valued distributions.

Definition 17. For a random variable Z supported in \mathbb{Z}^k and $\xi \in [-1, 1]^k$, we define

$$\hat{Z}(\xi) = \mathbf{E}_{z \sim Z}[e^{i \cdot \pi \cdot \xi \cdot z}].$$

We note that the reason to restrict $\xi \in [-1, 1]^k$ is because the Fourier spectrum of distributions supported on \mathbb{Z}^k is periodic with the fundamental period being the box $[-1, 1]^k$.

Let us now recall the setting: $P = Z_1 + \dots + Z_n$ where Z_i are independent random variables supported on $\{\mathbf{e}_1, \dots, \mathbf{e}_k\}$. Also for $1 \leq i \leq n$ and $1 \leq j \leq k$, let $p_{ij} = \Pr[X_i = j]$. To specify the next lemma, for any $\xi \in [-1, 1]^k$, we will need to define an associated vector $\zeta \in [-1, 1]^k$. For any $\xi \in [-1, 1]$, define the associated ζ as follows:

$$\zeta = \begin{cases} \xi & \text{if } \xi \in [-1/2, 1/2] \\ 1 - \xi & \text{if } \xi \in [1/2, 1] \\ -1 - \xi & \text{if } \xi \in [-1, -1/2] \end{cases}$$

For $\xi \in [-1, 1]^k$, we obtain $\zeta \in [-1/2, 1/2]^k$, by doing the above operation coordinatewise. Let $\mathcal{B}_{\ell_p}(z, r)$ denote the ℓ_p ball of radius r around z . To put it succinctly, for $\xi \in \mathcal{B}_{\ell_\infty}(z, 1/2)$ where $z \in \{-1, 0, 1\}^k$, we obtain $\zeta = (\xi - z) \circ (-1)^z$. Here $(-1)^z$ denotes the vector in $\{-1, 0, 1\}^k$ where the i^{th} coordinate is $(-1)^{z_i}$ and \circ denotes the Hadamard (the coordinate-wise) product of two vectors. For every $\xi \in [-1, 1]$, we call it Type 1 if $\xi \in [-1/2, 1/2]$, Type 2 if $\xi \in [1/2, 1]$ and Type 3 otherwise.

Claim 5. Let X be a CRV with covariance matrix Σ . Then, $|\widehat{X}(\xi)|^2 \leq 1 - \frac{1}{5} \cdot \zeta^T \cdot \Sigma \cdot \zeta$.

Proof.

$$\widehat{X}(\xi) = \sum_{j=1}^k p_j \cdot e^{i \cdot \pi \cdot \xi_j} = \sum_{j=1}^k p_j \cdot \cos(\pi \cdot \xi_j) + i \cdot p_j \cdot \sin(\pi \cdot \xi_j).$$

This implies

$$\begin{aligned} |\widehat{X}(\xi)|^2 &= \sum_{j=1}^k p_j^2 + 2 \cdot \sum_{1 \leq i < j \leq k} p_i \cdot p_j \cdot \cos(\pi(\xi_i - \xi_j)). \\ &= \sum_{j=1}^k p_j^2 + 2 \cdot \sum_{1 \leq i < j \leq k} p_i \cdot p_j \cdot \left(1 - 2 \sin^2 \left(\frac{\pi(\xi_i - \xi_j)}{2} \right)\right) \\ &= 1 - 4 \sum_{1 \leq i < j \leq k} p_i \cdot p_j \cdot \sin^2 \left(\frac{\pi(\xi_i - \xi_j)}{2} \right). \end{aligned}$$

We will first show that for every i, j ,

$$\sin^2 \left(\frac{\pi(\xi_i - \xi_j)}{2} \right) \geq \frac{1}{5} \cdot (\zeta_i - \zeta_j)^2. \quad (11)$$

To prove this, we do a simple case analysis, and use the inequality $\sin^2(\pi x/2) \geq x^2/5$ for $|x| \leq 3/2$:

- If both ξ_i and ξ_j are of the same type, then note that $|\xi_i - \xi_j| = |\zeta_i - \zeta_j| \leq 1$ which gives the required inequality.
- If ξ_i and ξ_j are type 2 and 3, then note that $|\xi_i - \xi_j| = |2 - (\zeta_i - \zeta_j)|$. This implies that

$$\sin^2 \left(\frac{\pi(\xi_i - \xi_j)}{2} \right) = \sin^2 \left(\frac{\pi(\zeta_i - \zeta_j)}{2} \right).$$

Noting that $|\zeta_i - \zeta_j| \leq 1$ gives the required inequality.

- If ξ_i is of type 1 and ξ_j is of type 2, then note that the maximum value that $|\xi_i - \xi_j|$ can take is $3/2$. On the other hand, notice that $|\xi_i - \xi_j| \geq |\zeta_i - \zeta_j|$. These two facts immediately imply that

$$\sin^2 \left(\frac{\pi(\xi_i - \xi_j)}{2} \right) \geq \frac{1}{5} \cdot (\zeta_i - \zeta_j)^2.$$

The exact same situation holds if ξ_i is of type 1 and ξ_j is of type 3.

Having shown (11), see that this implies that

$$|\widehat{X}(\xi)|^2 \leq 1 - \frac{1}{5} \sum_{1 \leq i < j \leq k} p_i p_j (\zeta_i - \zeta_j)^2.$$

However,

$$\sum_{1 \leq i < j \leq k} p_i p_j (\zeta_i - \zeta_j)^2 = \zeta^T \cdot \Sigma \cdot \zeta.$$

This finishes the proof. □

As a corollary, we have the following.

Corollary 3. *For any (n, k) -PMD P with covariance matrix Σ , we have that for any $\xi \in [-1, -1]^k$:*

$$|\widehat{P}(\xi)|^2 \leq \exp\left(-\frac{1}{5}\zeta^T \cdot \Sigma \cdot \zeta\right).$$

Proof. This follows simply by noticing that for a PMD $P = X_1 + \dots + X_n$,

$$\widehat{P}(\xi) = \prod_{j=1}^n \widehat{X}_j(\xi).$$

Using Claim 5, we have

$$|\widehat{P}(\xi)|^2 \leq \prod_{j=1}^n \left(1 - \zeta^T \cdot \Sigma_j \cdot \zeta\right),$$

where Σ_i is the covariance matrix of X_i . Using the inequality, $1 - x^2 \leq e^{-x^2/2}$ (for $|x| \leq 1$), we have

$$|\widehat{P}(\xi)|^2 \leq \prod_{j=1}^n e^{-\frac{1}{5}\zeta^T \cdot \Sigma_j \cdot \zeta}.$$

□

Lemma 15. *Let X be a random variable supported on \mathbb{R}^k with mean μ and covariance matrix Σ . Then the Fourier transform \widehat{X} is Lipschitz in the following sense:*

$$|\widehat{X}(\xi) - \widehat{X}(\xi')| \leq \pi \cdot (\xi - \xi') \cdot (\Sigma + \mu^T \cdot \mu) \cdot (\xi - \xi').$$

Proof.

$$|\mathbf{E}[e^{i\pi \cdot \xi \cdot X}] - \mathbf{E}[e^{i\pi \cdot \xi' \cdot X}]| = |\mathbf{E}[e^{i\pi \cdot \xi \cdot X} \cdot (e^{i\pi \cdot (\xi - \xi') \cdot X} - 1)]| \leq \mathbf{E}[|e^{i\pi \cdot (\xi - \xi') \cdot X} - 1|]$$

It is easy to observe that for any $\theta \in \mathbb{R}$, $|e^{i\theta} - 1| \leq \theta^2$. Applying this to the above inequality, we have

$$|\mathbf{E}[e^{i\pi \cdot \xi \cdot X}] - \mathbf{E}[e^{i\pi \cdot \xi' \cdot X}]| \leq \mathbf{E}[|\pi \cdot (\xi - \xi') \cdot X|^2] = \pi \cdot (\xi - \xi') \cdot (\Sigma + \mu^T \cdot \mu) \cdot (\xi - \xi').$$

□

We also have the following variant of the above lemma which will be useful for us.

Lemma 16. *Let X and Y be two distributions in \mathbb{R}^k with the same mean μ and covariance Σ . If for a point $\xi \in \mathbb{R}^k$, $|\widehat{X}(\xi) - \widehat{Y}(\xi)| \leq \varepsilon$, then*

$$\left| \widehat{X}(\xi + \zeta) - \widehat{Y}(\xi + \zeta) \right| \leq \varepsilon + 2\zeta^T \cdot \Sigma \cdot \zeta.$$

Proof. To prove this, note that

$$\left| \widehat{X}(\xi + \zeta) - \widehat{Y}(\xi + \zeta) \right| = \left| \widehat{X'}(\xi + \zeta) - \widehat{Y'}(\xi + \zeta) \right|$$

where X' and Y' are the centered random variables obtained by centering X and Y . Likewise,

$$\left| \widehat{X}(\xi) - \widehat{Y}(\xi) \right| = \left| \widehat{X'}(\xi) - \widehat{Y'}(\xi) \right|$$

However, by Lemma 15, we have

$$\left| \widehat{X}'(\xi) - \widehat{X}'(\xi + \zeta) \right| \leq \zeta^T \cdot \Sigma \cdot \zeta.$$

Applying the same for the \widehat{Y}' and applying triangle inequality, we get the claim. \square

We now state the Plancherel identity in this setting. In particular, we have the following easy claim (which can be found in any standard text on Fourier analysis).

Claim 6. *Let $F : \mathbb{Z}^k \rightarrow \mathbb{R}$. Then,*

$$\int_{\xi_1 \in [-1,1]} \cdots \int_{\xi_k \in [-1,1]} |\widehat{F}(\xi)|^2 d\xi_1 d\xi_2 \cdots d\xi_k = \sum_{z \in \mathbb{Z}^k} |F(z)|^2.$$

In our setting, $F = G - H$ where G and H are probability measures supported on \mathbb{Z}^k . The following easy consequence of Claim 6 will be useful for us.

Corollary 4. *Let $F, H : \mathbb{Z}^k \rightarrow [0, 1]$ be a probability distributions such that for some $\mathcal{S} \subseteq \mathbb{Z}^k$, $\Pr[F \notin \mathcal{S}], \Pr[H \notin \mathcal{S}] \leq \varepsilon$. Then*

$$d_{\text{TV}}(F, H) \leq \varepsilon + \sqrt{|\mathcal{S}|} \cdot \sqrt{\left(\int_{\xi_1 \in [-1,1]} \cdots \int_{\xi_k \in [-1,1]} |\widehat{F - H}(\xi)|^2 d\xi_1 d\xi_2 \cdots d\xi_k \right)}$$

Proof. By Claim 6,

$$\sum_{z \in \mathcal{S}} |F(z) - H(z)|^2 \leq \int_{\xi_1 \in [-1,1]} \cdots \int_{\xi_k \in [-1,1]} |\widehat{F - H}(\xi)|^2 d\xi_1 d\xi_2 \cdots d\xi_k = \sum_{z \in \mathbb{Z}^k} |F(z) - H(z)|^2.$$

Applying Cauchy-Schwarz inequality,

$$\sum_{z \in \mathcal{S}} |F(z) - H(z)| \leq \sqrt{|\mathcal{S}|} \cdot \sqrt{\left(\int_{\xi_1 \in [-1,1]} \cdots \int_{\xi_k \in [-1,1]} |\widehat{F - H}(\xi)|^2 d\xi_1 d\xi_2 \cdots d\xi_k \right)}$$

\square

The above corollary demonstrates that to learn the PMD to error ε , it suffices to produce another distribution H whose Fourier spectrum is very close to the Fourier spectrum of F (the “very small” is quantified by the effective support of F).

Lemma 17 (Lemma 8 from [DKT15]). *Given sample access to a (n, k) -PMD X with mean μ and covariance matrix Σ , there exists an algorithm which can produce estimates $\hat{\mu}$ and $\hat{\Sigma}$ such that with probability at least $9/10$ for every vector y :*

$$|y^T(\hat{\mu} - \mu)| \leq \varepsilon \sqrt{y^T \Sigma y} \quad \text{and} \quad |y^T(\hat{\Sigma} - \Sigma)y| \leq \varepsilon y^T \Sigma y \sqrt{1 + \frac{y^T y}{y^T \Sigma y}}$$

The sample and time complexity are $O(k^4/\varepsilon^2)$.

The following is guaranteed by the multidimensional Chernoff bound.

Lemma 18. Let X be a (n, k) -PMD with mean μ and covariance matrix Σ . Let $L_r = \{z : (z - \mu) \cdot (z - \mu)^t \preceq r \cdot \Sigma\}$. For $r = O(\log(1/\varepsilon) + \log k)$,

$$\Pr[X \notin L_r] \leq \varepsilon.$$

This implies the following corollary.

Lemma 19. Let $\mu \in \mathbb{R}^k$ and Σ be a PSD matrix with eigenvalues $\sigma_1^2 \geq \dots \geq \sigma_k^2 \geq 0$. Let $L_r = \{z : (z - \mu) \cdot (z - \mu)^t \preceq r \cdot \Sigma\}$. The total number of points of \mathbb{Z}^k which lie in L_r is bounded by $\prod_{i=1}^k (2\sigma_i \sqrt{kr} + 1)$.

Proof. Let the eigenvectors of Σ be v_1, \dots, v_k with the corresponding eigenvalues $\sigma_1^2, \dots, \sigma_k^2$. Consider any two distinct $x, y \in L_r$. Since x and y are distinct, hence there must be some $1 \leq i \leq k$ such that the projection of x and y along v_i is separated by $k^{-1/2}$.

Let us denote the projection of x along v_i by x_i . Then the condition of lying in L_r implies that $|x_i| \leq r \cdot \sigma_i$. It is then easy to see that if the number of integer points in L_r is more than $\prod_{i=1}^k (2\sigma_i \sqrt{kr} + 1)$, then there must be 2 points x and y and some $1 \leq i \leq k$, $|x_i - y_i| \leq k^{-1/2}$. \square

Given a (n, k) PMD Z , let $\hat{\mu}$ and $\hat{\Sigma}$ be the empirical mean and covariance matrices obtained from Lemma 17. For technical reasons, instead of working with $\hat{\Sigma}$, we create a new PSD matrix $\tilde{\Sigma}$ which is obtained as follows: $\hat{\Sigma}$ and $\tilde{\Sigma}$ have the same eigenvectors. If $\hat{\sigma}_i^2$ is the eigenvalue of $\hat{\Sigma}$ corresponding to v_i , then the corresponding eigenvalue $\tilde{\sigma}_i^2$ of $\tilde{\Sigma}$ is $(1 - 3\varepsilon) \cdot \hat{\sigma}_i^2$. Further, after this operation, if a particular eigenvalue of $\tilde{\Sigma}$ is smaller than ε , we modify that singular value to make it 0. Doing this operation ensures that

$$|y^T (\tilde{\Sigma} - \Sigma)y| \leq \varepsilon y^T \Sigma y \sqrt{1 + \frac{y^T y}{y^T \Sigma y}} \text{ and } \tilde{\Sigma} \preceq \Sigma$$

which implies that for all eigenvalues $\tilde{\sigma}_i^2 \leq \sigma_i^2$. Note that to learn the PMD, one possible strategy is to evaluate the Fourier transform of a (n, k) -PMD in the region $(\xi_1, \dots, \xi_k) \in [-1, 1]^k$ and then perform a Fourier inversion. Unfortunately, this is too expensive for us. Instead, we show that the Fourier transform only needs to be evaluated in a very small region.

Definition 18. For a point $z \in \{-1, 0, 1\}^k$, define $C_{z,r}$ as

$$C_{z,r} = \{y : \sum \tilde{\sigma}_i^2 (\tilde{v}_i \cdot ((-1)^z \circ (y - z)))^2 \leq r\}$$

and R_z as

$$R_z = \mathcal{B}_{\ell_\infty}(z, 1/2) \cap [-1, 1]^k.$$

Note that $[-1, 1]^k$ can be partitioned into the regions R_z (for $z \in \{-1, 0, 1\}^k$). In other words,

$$[-1, 1]^k = \cup_{z \in \{-1, 0, 1\}^k} R_z.$$

Claim 7. Let $S_r = \cup_{z \in \{-1, 0, 1\}^k} (R_z \cap C_{z,r})$ and let $\bar{S}_r = [-1, 1]^k \setminus S_r$. Then,

$$\int_{(\xi_1, \dots, \xi_k) \in \bar{S}_r} |\hat{Z}(\xi)|^2 d\xi_1 \dots d\xi_k \leq e^{-r/10} \prod_{i=1}^k \frac{1}{\max \left\{ \tilde{\sigma}_i, \frac{1}{k} \right\}}$$

Proof.

$$\int_{(\xi_1, \dots, \xi_k) \in \bar{S}_r} |\widehat{Z}(\xi)|^2 d\xi_1 \dots d\xi_k = \sum_{z \in \{-1, 0, 1\}^k} \int_{(\xi_1, \dots, \xi_k) \in R_z \setminus C_{z,r}} |\widehat{Z}(\xi)|^2 d\xi_1 \dots d\xi_k$$

We now individually bound each of the summands. Fix any particular z . Using Corollary 3

$$\int_{(\xi_1, \dots, \xi_k) \in R_z \setminus C_{z,r}} |\widehat{Z}(\xi)|^2 d\xi_1 \dots d\xi_k \leq \int_{(\xi_1, \dots, \xi_k) \in R_z \setminus C_{z,r}} e^{-\frac{1}{5} \zeta^T \cdot \Sigma \cdot \zeta} d\xi_1 \dots d\xi_k$$

where $\zeta = (-1)^z \circ (\xi - z)$. To bound this, note that since $\tilde{\Sigma} \preceq \Sigma$ and $R_z \subset B_{\ell_2(z, \sqrt{k}/2)}$, we get

$$\int_{(\xi_1, \dots, \xi_k) \in R_z \setminus C_{z,r}} e^{-\frac{1}{5} \zeta^T \cdot \Sigma \cdot \zeta} d\xi_1 \dots d\xi_k \leq \int_{(\xi_1, \dots, \xi_k) \in B_{\ell_2(z, \sqrt{k}/2)} \setminus C_{z,r}} e^{-\frac{1}{5} \zeta^T \cdot \tilde{\Sigma} \cdot \zeta} d\xi_1 \dots d\xi_k$$

Using the fact that ℓ_2 balls are invariant under rotation, the right hand integral becomes

$$\int_{\sum_i \tilde{\sigma}_i^2 w_i^2 > r; (w_1, \dots, w_k) \in B_{\ell_2(0, \sqrt{k}/2)}} e^{-\frac{1}{5} \cdot \Sigma \tilde{\sigma}_i^2 w_i^2} dw_1 \dots dw_k$$

Since, $B_{\ell_2(0, \sqrt{k}/2)} \subset B_{\ell_\infty(0, \sqrt{k}/2)}$, this is upper bounded by

$$\int_{\sum \tilde{\sigma}_i^2 w_i^2 > r; |w_i| \leq \sqrt{k}/2} e^{-\frac{1}{5} \cdot \Sigma \tilde{\sigma}_i^2 w_i^2} dw_1 \dots dw_k$$

To upper bound this integral, let $Y_1 = \{j : \tilde{\sigma}_j \leq 1/k\}$. Then,

$$\begin{aligned} \int_{\sum \tilde{\sigma}_i^2 w_i^2 > r; |w_i| \leq \sqrt{k}/2} e^{-\frac{1}{5} \cdot \Sigma \tilde{\sigma}_i^2 w_i^2} dw_1 \dots dw_k &\leq \int_{\sum \tilde{\sigma}_i^2 w_i^2 > r; |w_i| \leq \sqrt{k}/2} e^{-\frac{1}{5} \cdot \sum_{i \notin Y_1} \tilde{\sigma}_i^2 w_i^2} dw_1 \dots dw_k \\ &\leq \int_{\sum_{i \notin Y_1} \tilde{\sigma}_i^2 w_i^2 > r/2; |w_i| \leq \sqrt{k}/2} e^{-\frac{1}{5} \cdot \sum_{i \notin Y_1} \tilde{\sigma}_i^2 w_i^2} dw_1 \dots dw_k \end{aligned}$$

The last inequality uses that $r > 2$. This integral is now easily seen to be bounded by

$$\prod_{i \in Y_1} k \cdot e^{-r/10} \cdot \prod_{i \notin Y_1} \frac{1}{\tilde{\sigma}_i}.$$

This is exactly the same bound as stated in the claim. \square

Claim 8. Let $S_r = \cup_{z \in \{-1, 0, 1\}^k} (R_z \cap C_{z,r})$. Then,

$$\int_{(\xi_1, \dots, \xi_k) \in S_r} d\xi_1 \dots d\xi_k \leq 3^k \prod_{i=1}^k \min \left\{ \sqrt{k}, \frac{2\sqrt{r}}{\tilde{\sigma}_i} \right\}$$

Proof. Doing the exact same calculation as in the proof of Claim 7,

$$\int_{(\xi_1, \dots, \xi_k) \in S_r} d\xi_1 \dots d\xi_k \leq 3^k \cdot \int_{\sum_i \tilde{\sigma}_i^2 w_i^2 \leq r; |w_i| \leq \sqrt{k}/2} dw_1 \dots dw_k$$

By using the same manipulation as before, we can upper bound this integral by $3^k \prod_{i=1}^k \min \left\{ \sqrt{k}, \frac{2\sqrt{r}}{\tilde{\sigma}_i} \right\}$. \square

8.2 Learning algorithm for PMDs

Theorem 6, the structure theorem from [DKT15], allows us to assume that the PMD Z is essentially a discretized Gaussian G convolved with a sparse PMD S where the sparse PMD is supported on only $\text{poly}(k/\varepsilon)$ summands.

By setting ‘ ε ’ from the Theorem statement to be ε^{10} , we get that $d_{TV}(Z, G + S) \leq \varepsilon^{10}$. Because our subsequent learning algorithm will take $\ll O(\varepsilon^{-10})$ samples, we assume that we are getting samples from $G + S$ instead of Z and that $Z = G + S$. Furthermore, using the following claim from [DKT15], we can get a spectral estimate with accuracy ε^{10} of the mean and covariance of the Gaussian G by guessing the partition of coordinates in the covariance matrix of the Gaussian and going through all elements of the spectral cover of PSD matrices around a fine estimate \hat{S} for Σ obtained using k/ε^2 samples from Lemma 17.

Claim 9 (Lemma 9 from [DKT15]). *Let A be a symmetric $k \times k$ PSD matrix with minimum eigenvalue 1 and let \mathcal{S} be the set of all matrices B such that $|y^T \cdot (A - B) \cdot y| \leq \varepsilon_1 y^T \cdot A \cdot y + \varepsilon_2 y^T y$ where $\varepsilon_1 \in [0, 1/4)$ and $\varepsilon_2 \in [0, \infty)$. Then, there exists a cover \mathcal{S}_ε of size $(k \cdot (1 + \varepsilon_2)/\varepsilon)^{k^2}$ such that any $B \in \mathcal{S}$ is ε -spectrally close to some element in the cover.*

The spectral closeness translates to closeness ε^{10} in total variation distance between Gaussians (Lemma 2) and again since we will be taking $\ll O(\varepsilon^{-10})$ samples in the learning algorithm, we can assume that the gaussian G has exactly the mean μ_G and covariance Σ_G we guessed.

Similarly, we can assume that the sparse-PMD has known mean and covariance μ_S and Σ_S . This is because any PMD with n' summands is ε^{10} -close in total variation to a PMD where all the probabilities are rounded to multiples of $\lceil n'k/\varepsilon^{10} \rceil^{-1}$. This fact follows from union-bounding all the errors of the individual summands. Since $n' = \text{poly}(k/\varepsilon)$ for the sparse PMD, all coordinates are multiples of $\text{poly}(\varepsilon/k)$, which implies that the mean and covariance coordinates are also multiples of $\text{poly}(\varepsilon/k)$ and we can guess them exactly using $\text{poly}(k/\varepsilon)^{k^2}$ guesses. Again, since this sparse PMD is ε^{10} close and we will be getting much fewer samples, we can assume that the sparse PMD has exactly the mean and covariance we guessed.

At this point, we have argued the following:

Claim 10. *The PMD Z is equal to the sum of a discretized Gaussian G and a sparse PMD S with $\text{poly}(k/\varepsilon)$ summands. The mean and covariance of the Gaussian (μ_G, Σ_G) and of the sparse PMD (μ_S, Σ_S) are known, which implies that the mean and covariance of the overall PMD Z is equal to $(\mu, \Sigma) = (\mu_S, \Sigma_S) + (\mu_G, \Sigma_G)$.*

Our learning algorithm attempts to recover the sparse PMD in order to learn the overall distribution Z . However, imposing the condition that the distribution we are trying to estimate is a sparse PMD will involve solving non linear equations making the computation intractable. Rather, we will seek to learn a sparse distribution S' supported on $[0, T]^k$ where $T = \text{poly}(k/\varepsilon)$.

To learn this distribution, we will attempt to estimate its Fourier Transform. We will be mostly interested in points on the grid:

$$\mathcal{V} = \left\{ \alpha_1 \cdot \frac{\varepsilon}{k^{2k} \cdot 6^k} \cdot \frac{\vec{v}_1}{\max\{1, \sigma_1\}} + \cdots + \alpha_k \cdot \frac{\varepsilon}{k^{2k} \cdot 6^k} \cdot \frac{\vec{v}_k}{\max\{1, \sigma_k\}} : \alpha_i \in \mathbb{Z} \right\}$$

where (\vec{v}_i, σ_i^2) are the eigenvector, eigenvalue pairs of the matrix Σ . From Corollary 3, we know that the Fourier transform decays exponentially as we move away from $\{-1, 0, 1\}^k$, and in particular Claim 7 bounds the total mass contained at a distance at least r from all the points. For our purposes, we set $r = O(k \log k + k \log(1/\varepsilon))$ and perform the following steps to learn the sparse distribution S' .

1. Create variables p_α for every $\alpha \in [0, T]^k$ with the constraints $0 \leq p_\alpha \leq 1$ and $\sum_{\alpha \in T^k} p_\alpha = 1$.
2. Let $\mathcal{A}_1 = \cup_{z \in \{-1, 0, 1\}^k} \{\xi : \sum \sigma_i^2 (v_i \cdot ((-1)^z \circ (\xi - z)))^2 \leq r\}$. Let \mathcal{V}_1 be the points of the grid \mathcal{V} that lie in \mathcal{A}_1 . For each of those points, get an estimate \hat{Z}_{est} of \hat{Z} such that $|\hat{Z}_{est} - \hat{Z}| < \frac{\varepsilon}{6^k \cdot k^{2k}}$ and then impose linear constraints on $\{p_\alpha\}$ so that $|Re[\hat{S}'(\xi) \cdot \hat{G}(\xi) - \hat{Z}_{est}(\xi)]| \leq \frac{\varepsilon}{6^k \cdot k^{2k}}$ and $|Im[\hat{S}'(\xi) \cdot \hat{G}(\xi) - \hat{Z}_{est}(\xi)]| \leq \frac{\varepsilon}{6^k \cdot k^{2k}}$.
3. Let $\sigma_{G,i}^2, \vec{v}_{G,i}$ be the eigenvalues and eigenvectors of Σ_G^7 and consider the set:

$$\mathcal{A}_2 = \cup_{z \in \{-1, 0, 1\}^k} \{\xi : \sum \sigma_{G,i}^2 (\vec{v}_{G,i} \cdot ((-1)^z \circ (\xi - z)))^2 \leq \frac{r}{2} \wedge \sum \sigma_i^2 (\vec{v}_i \cdot ((-1)^z \circ (\xi - z)))^2 > r\}$$

Construct a grid of points in $[-1, 1]^k$ with a spacing of $\frac{\varepsilon^{2k}}{k^{2k} \cdot 6^k}$ in every direction. Let \mathcal{V}_2 be the subset of these points which fall in \mathcal{A}_2 . For all these points impose the conditionss that $|Re[\hat{S}'(\xi)]| \leq e^{-\zeta^T \Sigma_S \zeta}$ and $|Im[\hat{S}'(\xi)]| \leq e^{-\zeta^T \Sigma_S \zeta}$ that follow from Corollary 3.

4. Finally, add the constraints $\sum_\alpha p_\alpha \alpha = \mu_S$ and $\sum_\alpha p_\alpha (\alpha - \mu_S)(\alpha - \mu_S)^T = \Sigma_S$

Note that in Step 2, \mathcal{V}_1 has size at most $\left(\frac{\sqrt{T} \cdot k^{2k} \cdot 6^k}{\varepsilon}\right)^k$. If we naively estimated every Fourier coefficient in \mathcal{V}_1 the number of samples would be too high because every Fourier coefficient requires $\log(1/\delta)/\varepsilon^2$ samples to learn with accuracy ε and probability of failure $1 - \delta$. However, we can instead take $O(k \log(r/\varepsilon)/\varepsilon^2)$ samples and reuse the same samples to compute all the required Fourier coefficients. Since the probability of error is very small a simple union bound among all of the coefficients, shows that with at least constant probability all of them can be estimated within ε .

To complete the learning algorithm, we repeat the steps above for each of the guessed mean and covariance matrices $(\mu_G, \Sigma_G), (\mu_S, \Sigma_S)$. We then perform a hypothesis selection algorithm to choose a distribution within $O(\varepsilon)$ from each of the distributions we obtain. We made $O(\text{poly}(k/\varepsilon)^{k^2})$ guesses, and thus obtained $O(\text{poly}(k/\varepsilon)^{k^2})$ candidate hypotheses. Applying the following tournament theorem for hypothesis selection from [DK14], we can select a good estimate in $O\left(\left(\frac{k}{\varepsilon}\right)^2 \log(k/\varepsilon)\right)$ samples in $O(\text{poly}(k/\varepsilon)^{k^2})$ runtime.

Theorem 11 (Theorem 19 of [DK14]). *There is an algorithm **FastTournament** $(X, \mathcal{H}, \varepsilon, \delta)$, which is given sample access to some distribution X and a collection of distributions $\mathcal{H} = \{H_1, \dots, H_N\}$ over some set \mathcal{D} , access to a PDF comparator for every pair of distributions $H_i, H_j \in \mathcal{H}$, an accuracy parameter $\varepsilon > 0$, and a confidence parameter $\delta > 0$. The algorithm makes $O\left(\frac{\log 1/\delta}{\varepsilon^2} \cdot \log N\right)$ draws from each of X, H_1, \dots, H_N and returns some $H \in \mathcal{H}$ or declares “failure.” If there is some $H^* \in \mathcal{H}$ such that $d_{TV}(H^*, X) \leq \varepsilon$ then with probability at least $1 - \delta$ the distribution H that **FastTournament** returns satisfies $d_{TV}(H, X) \leq 512\varepsilon$. The total number of operations of the algorithm is $O\left(\frac{\log 1/\delta}{\varepsilon^2} (N \log N + \log^2 \frac{1}{\delta})\right)$. Furthermore, the expected number of operations of the algorithm is $O\left(\frac{N \log N / \delta}{\varepsilon^2}\right)$.*

Proof of correctness:

⁷ We note that, since we may have eigenvalues which are both large and small in magnitude, a naive eigendecomposition algorithm would incur a cost which depends on n . However, as we only require the eigenvalues and eigenvectors approximately, this cost can be avoided by applying an appropriate power-iteration method. The cost in terms of k and $1/\varepsilon$ is dominated by the other steps in our algorithm.

We first show that there is a solution to $\{p_\alpha\}$ which satisfies all the constraints. Indeed, if we set the sparse distribution S' to be equal to the distribution S of the sparse PMD we defined above, we get:

1. $\sum_{\alpha \in T^k} p_\alpha = 1$ since S is a probability distribution supported on $[0, T]^k$.
2. The constraint $|Re[\widehat{S}'(\xi) \cdot \widehat{G}(\xi) - \widehat{Z}_{est}(\xi)]| \leq \frac{\varepsilon}{6^k \cdot k^{2k}}$ is satisfied since for $S' = S$,

$$|Re[\widehat{S}(\xi) \cdot \widehat{G}(\xi) - \widehat{Z}_{est}(\xi)]| = |Re[\widehat{Z}(\xi) - \widehat{Z}_{est}(\xi)]| \leq |\widehat{Z}(\xi) - \widehat{Z}_{est}(\xi)| \leq \frac{\varepsilon}{6^k \cdot k^{2k}}.$$

The derivation for the constraint on the imaginary part is identical.

3. From Corollary 3, the sparse PMD satisfies $|\widehat{S}(\xi)| \leq e^{-(1/5) \cdot \zeta^T \Sigma_S \zeta}$ everywhere in $[-1, 1]^k$. This condition implies the imposed constraints which are only evaluated in few points.
4. The distribution S has mean μ_S and covariance Σ_S , so the last constraint is satisfied.

We now prove that any feasible solution $\{p_\alpha\}$ to the above system of constraints defines a distribution S' such that $d_{TV}(S + G, S' + G) \leq \varepsilon$. To show this, we divide the space $[-1, 1]^k$ into three parts: \mathcal{A}_1 , \mathcal{A}_2 and $\mathcal{A}_3 = [-1, 1]^k \setminus (\mathcal{A}_1 \cup \mathcal{A}_2)$.

Claim 11.

$$\int_{\xi \in \mathcal{A}_1} |\widehat{S + G}(\xi) - \widehat{S' + G}(\xi)|^2 d\xi = O\left(\frac{\varepsilon^2 \cdot r^{k/2}}{k^{3k} \cdot \prod_{i=1}^k \max\{\sigma_i, 1\}}\right)$$

Proof. Consider any point ξ in \mathcal{A}_1 . Then, note that there is some $\xi' \in \mathcal{V}$ such that for $1 \leq i \leq k$, $\langle \xi - \xi', \vec{v}_i \rangle \leq \frac{\varepsilon}{k^{2k} \cdot 6^k \cdot \max\{1, \sigma_i\}}$. Applying Lemma 16, we get that

$$|\widehat{S + G}(\xi) - \widehat{S' + G}(\xi)| \leq \frac{\varepsilon \cdot \sqrt{k}}{6^k \cdot k^{2k}} + |\widehat{S + G}(\xi') - \widehat{S' + G}(\xi')| \leq \frac{\varepsilon \cdot 2\sqrt{k}}{6^k \cdot k^{2k}}.$$

Applying Claim 8, we have

$$\int_{\xi \in \mathcal{A}_1} |\widehat{S + G}(\xi) - \widehat{S' + G}(\xi)|^2 d\xi \leq \max_{\xi \in \mathcal{A}_1} |\widehat{S + G}(\xi) - \widehat{S' + G}(\xi)|^2 \cdot \int_{\xi \in \mathcal{A}_1} d\xi = O\left(\frac{\varepsilon^2 \cdot r^{k/2}}{k^{3k} \cdot \prod_{i=1}^k \max\{\sigma_i, 1\}}\right).$$

This finishes the proof. \square

Claim 12.

$$\int_{\xi \in \mathcal{A}_2} |\widehat{S}(\xi) - \widehat{S'}(\xi)|^2 d\xi = O\left(\frac{\varepsilon^2 \cdot r^{k/2}}{k^{3k} \cdot \prod_{i=1}^k \max\{\sigma_i, 1\}}\right)$$

Proof. Note that \mathcal{A}_2 is a subset of the set

$$B_2 = \cup_{z \in \{-1, 0, 1\}^k} \{\xi : \sum \sigma_{G,i}^2 (\vec{v}_{G,i} \cdot ((-1)^z \circ (\xi - z)))^2 \leq \frac{r}{2}\}.$$

We bound the volume of the set B_2 . To do this, we again apply Claim 8, and get that

$$\int_{\xi \in B_2} d\xi = 3^k \cdot \frac{r^{k/2} \cdot k^{k/2}}{\prod_{i=1}^k \max\{\sigma_{G,i}, 1\}}.$$

Note that for any point $\xi \in A_2$, we there is a point ξ' such that $\|\xi - \xi'\|_2 \leq \frac{\varepsilon^{2k}}{k^{2k} \cdot 6^k}$ and that $|\widehat{S}'(\xi')| \leq e^{-(1/5) \cdot \zeta^T \Sigma_S \cdot \zeta'}$. Since the variance of Σ_S is at most $\text{poly}(k/\varepsilon)$ in every direction, we get that

$$|\widehat{S}'(\xi)| \leq e^{-(1/5) \cdot \zeta^T \Sigma_S \cdot \zeta} + \frac{\varepsilon^{2k}}{k^{2k} \cdot 6^k}.$$

This implies that

$$\int_{\xi \in A_2} |\widehat{S}(\xi) - \widehat{S}'(\xi)|^2 \leq \int_{\xi \in A_2} 2 \cdot |\widehat{S}(\xi)|^2 + 2 \cdot |\widehat{S}'(\xi)|^2 d\xi$$

By applying Claim 8 to bound the volume of the set $A_2 \subseteq B_2$ and using the fact that $|\widehat{S}(\xi)|^2$ is at most $e^{-r/20}$, we get that the first integral is at most

$$\begin{aligned} \int_{\xi \in A_2} 2 \cdot |\widehat{S}(\xi)|^2 &\leq e^{-r/20} \cdot \prod_{i=1}^k \frac{1}{\max\{\sigma_{G,i}, 1/k\}} \\ &\leq e^{-r/20} \cdot \text{poly}(k/\varepsilon)^k \cdot \prod_{i=1}^k \frac{1}{\max\{\sigma_i, 1/k\}} \end{aligned}$$

The last inequality uses the fact that whenever $\sigma_{G,i} \leq \sigma_i$, it must imply that all the variance comes from S and thus $\sigma_i \leq \text{poly}(k/\varepsilon)$. By plugging the value of r , we get that

$$\int_{\xi \in A_2} 2 \cdot |\widehat{S}(\xi)|^2 \leq \varepsilon^k \cdot \prod_{i=1}^k \frac{1}{\max\{\sigma_i, 1\}}.$$

The calculation for the second integral is similar.

$$\begin{aligned} \int_{\xi \in A_2} 2 \cdot |\widehat{S}'(\xi)|^2 d\xi &\leq \int_{\xi \in A_2} e^{-(1/5) \cdot \zeta^T \Sigma_S \cdot \zeta} + \int_{\xi \in A_2} \frac{\varepsilon^{2k}}{k^{2k} \cdot 6^k} d\xi \\ &\leq \varepsilon^k \cdot \prod_{i=1}^k \frac{1}{\max\{\sigma_i, 1\}} + \int_{\xi \in A_2} \frac{\varepsilon^{2k}}{k^{2k} \cdot 6^k} d\xi \\ &\leq \varepsilon^k \cdot \prod_{i=1}^k \frac{1}{\max\{\sigma_i, 1\}} + \int_{\xi \in B_2} \frac{\varepsilon^{2k}}{k^{2k} \cdot 6^k} d\xi \end{aligned}$$

Here the first inequality follows by exactly the same calculation we did for the first integral whereas the second inequality uses that $A_2 \subseteq B_2$. Now, that we had derived that

$$\int_{\xi \in B_2} d\xi = 3^k \cdot \frac{r^{k/2} \cdot k^{k/2}}{\prod_{i=1}^k \max\{\sigma_{G,i}, 1\}}.$$

However, $\max\{\sigma_{G,i}, 1\} \geq \varepsilon^{\Theta(1)} \cdot \max\{\sigma_i, 1\}$ (because the variance of S is at most $\text{poly}(1/\varepsilon)$ in any direction. This implies that

$$\int_{\xi \in B_2} d\xi = \left(\frac{3}{\varepsilon}\right)^k \cdot \frac{r^{k/2} \cdot k^{k/2}}{\prod_{i=1}^k \max\{\sigma_i, 1\}}.$$

This implies that

$$\int_{\xi \in A_2} 2 \cdot |\widehat{S}'(\xi)|^2 \leq \varepsilon^k \cdot \prod_{i=1}^k \frac{1}{\max\{\sigma_i, 1\}}.$$

□

Claim 13.

$$\int_{\xi \in A_3} |\widehat{S+G}(\xi) - \widehat{S'+G}(\xi)|^2 d\xi = O\left(\frac{\varepsilon^2 \cdot r^{k/2}}{k^{3k} \cdot \prod_{i=1}^k \max\{\sigma_i, 1\}}\right)$$

Proof. Note that $\widehat{S'+G}(\xi) = \widehat{G}(\xi) \cdot \widehat{S'}(\xi)$. Thus, $|\widehat{S'+G}(\xi)|^2 \leq |\widehat{G}(\xi)|^2$. Applying Claim 7 and noting that

$$A_3 \subseteq \cup_{z \in \{-1, 0, 1\}^k} \{\xi : \sum \sigma_{G,i}^2 (\vec{v}_{G,i} \cdot ((-1)^z \circ (\xi - z)))^2 > \frac{r}{2}\}$$

we obtain that

$$\int_{\xi \in A_3} |\widehat{G}(\xi)|^2 d\xi = e^{-r/10} \cdot k^k \cdot \prod_{i=1}^k \frac{1}{\max\{1, \sigma_{G,i}\}}$$

Again using the fact that the variance of S in any direction is at most $\text{poly}(k/\varepsilon)$,

$$\int_{\xi \in A_3} |\widehat{G}(\xi)|^2 d\xi \leq e^{-r/10} \cdot \text{poly}(k/\varepsilon)^k \cdot \prod_{i=1}^k \frac{k}{\max\{1, \sigma_i\}}$$

Plugging in the value of r , we get that

$$\int_{\xi \in A_3} |\widehat{G}(\xi)|^2 d\xi \leq \varepsilon^k \cdot \prod_{i=1}^k \frac{1}{\max\{1, \sigma_i\}}$$

This immediately implies the claim. □

Combining Claim 11, Claim 12 and Claim 13, we get that

$$\int_{\xi \in [-1, 1]^k} |\widehat{S+G}(\xi) - \widehat{S'+G}(\xi)|^2 d\xi = \varepsilon^2 \cdot (k \log(1/\varepsilon))^{O(k)} \cdot \prod_{i=1}^k \frac{1}{\max\{\sigma_i, 1\}}.$$

We now apply Corollary 4 to derive that

$$d_{TV}(S+G, S'+G) \leq \varepsilon \cdot (k \log(1/\varepsilon))^{O(k)} \cdot \sqrt{\prod_{i=1}^k \frac{1}{\max\{\sigma_i, 1\}} \cdot \prod_{i=1}^k (2\sigma_i \sqrt{kr} + 1)}.$$

This is at most $d_{TV}(S+G, S'+G) \leq \varepsilon \cdot (k \log(1/\varepsilon))^{O(k)}$. Setting ε to be $\frac{\varepsilon'}{\text{poly}(k, \log(1/\varepsilon'))^k}$, we complete the proof of Theorem 5.

9 Open Problems

A number of interesting questions regarding Poisson Multinomial distributions are left open by this work and [DKS16a]. We outline a few of them here.

1. **The complexity of learning Poisson Multinomials.** This work and [DKS16a] both give algorithms for learning PMDs. The sample and time complexities are polynomial in $1/\varepsilon$ and exponential in k . Meanwhile, [DKT15] gives an algorithm with a sample complexity polynomial in both parameters, but the time complexity is exponential in k and $1/\varepsilon$. Is there an algorithm for learning PMDs with sample and time complexities both polynomial in k and $1/\varepsilon$?

2. **Exploring the connection between Poisson Multinomials and Laplacian matrices.** In this work, we described a cover for the set of (n, k) -PMDs of size $O_{k,\varepsilon}(n^{O(k)})$. Our construction relied crucially on Observation 1 (which states that the covariance matrix of a PMD is Laplacian) and spectral sparsification results for Laplacian matrices. With this connection in hand, can one derive other results for PMDs using the wealth of literature on Laplacian matrices?
3. **A tighter central limit theorem.** [VV11] proves a central limit theorem between an (n, k) -GMD and a discretized Gaussian with the same mean and covariance, upper bounding their total variation distance by $O(k^{4/3}\sigma^{-1/3}\log^{2/3}n)$, where σ^2 is the smallest eigenvector of the covariance matrix of the GMD. Both this paper and [DKS16a] qualitatively improve this bound by removing the dependence on n , while keeping the dependence on k and $1/\sigma$ still polynomial. How well can a GMD be approximated by a discretized Gaussian? In one dimension, the answer is $\Theta(1/\sigma)$ [CGS10], which implies a the answer for multiple dimensions is at least $\Omega(\sqrt{k}/\sigma)$. [DKS16a] achieves this dependence on $1/\sigma$ (up to log factors), but the optimal dependence on k is currently unknown.
4. **Sums of independent integer random vectors.** Poisson Multinomial distributions are the natural multivariate generalization of Poisson Binomial distributions, which have now been explored in this paper and other recent works [DKT15, DKS16a]. However, we currently have minimal understanding of any multivariate analogue of sums of independent integer random variables (i.e., SIIRVs, the object of study in [BĆ02, DDO⁺13, DKS16b]), which we will denote as *vector SIIRVs* (VSIIRVs). The natural definition of such an object is not immediately clear; one potential definition of an (n, k, d) -VSIIRV may be as the sum of n independent random vectors in \mathbb{N}^d , where each is a distribution over all positive lattice points at ℓ_1 distance at most k from the origin. We note that an $(n, 1, d)$ -VSIIRV is an (n, d) -PMD, so these objects generalize PMDs at well. An interesting line of study would be to obtain structural, covering, and learning results for VSIIRVs.

References

- [Bar88] Andrew D. Barbour. Stein’s method and Poisson process convergence. *Journal of Applied Probability*, 25:175–184, 1988.
- [BĆ02] Andrew D. Barbour and Ćekanavičius. Total variation asymptotics for sums of independent integer random variables. *The Annals of Probability*, 30(2):509–545, 2002.
- [Ben05] Vidmantas Bentkus. A Lyapunov-type bound in \mathbb{R}^d . *Theory of Probability & Its Applications*, 49(2):311–323, 2005.
- [Ber41] Andrew C. Berry. The accuracy of the Gaussian approximation to the sum of independent variates. *Transactions of the American Mathematical Society*, 49(1):122–136, 1941.
- [Blo99] Matthias Blonski. Anonymous games with binary actions. *Games and Economic Behavior*, 28(2):171–180, 1999.
- [Blo05] Matthias Blonski. The women of Cairo: Equilibria in large anonymous games. *Journal of Mathematical Economics*, 41(3):253–264, 2005.

- [BSS12] Joshua D. Batson, Daniel A. Spielman, and Nikhil Srivastava. Twice-Ramanujan sparsifiers. *SIAM Journal on Computing*, 41(6):1704–1721, 2012.
- [BSST13] Joshua D. Batson, Daniel A. Spielman, Nikhil Srivastava, and Shang-Hua Teng. Spectral sparsification of graphs: Theory and algorithms. *Communications of the ACM*, 56(8):87–94, 2013.
- [CDO15] Xi Chen, David Durfee, and Anthi Orfanou. On the complexity of Nash equilibria in anonymous games. In *Proceedings of the 47th Annual ACM Symposium on the Theory of Computing*, STOC ’15, pages 381–390, New York, NY, USA, 2015. ACM.
- [CDT09] Xi Chen, Xiaotie Deng, and Shang-Hua Teng. Settling the complexity of computing two-player Nash equilibria. *Journal of the ACM*, 56(3):14:1–14:57, 2009.
- [CGS10] Louis H.Y. Chen, Larry Goldstein, and Qi-Man Shao. *Normal approximation by Steins method*. Springer, 2010.
- [CST14] Xi Chen, Rocco A. Servedio, and Li Yang Tan. New algorithms and lower bounds for monotonicity testing. In *Proceedings of the 55th Annual IEEE Symposium on Foundations of Computer Science*, FOCS ’14, pages 286–295, Washington, DC, USA, 2014. IEEE Computer Society.
- [Dal99] John Dalbec. Multisymmetric functions. *Beiträge zur Algebra und Geometrie*, 40(1):27–51, 1999.
- [DDO⁺13] Constantinos Daskalakis, Ilias Diakonikolas, Ryan O’Donnell, Rocco A. Servedio, and Li Yang Tan. Learning sums of independent integer random variables. In *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science*, FOCS ’13, pages 217–226, Washington, DC, USA, 2013. IEEE Computer Society.
- [DGP09] Constantinos Daskalakis, Paul W. Goldberg, and Christos H. Papadimitriou. The complexity of computing a Nash equilibrium. *SIAM Journal on Computing*, 39(1):195–259, 2009.
- [DK14] Constantinos Daskalakis and Gautam Kamath. Faster and sample near-optimal algorithms for proper learning mixtures of Gaussians. In *Proceedings of the 27th Annual Conference on Learning Theory*, COLT ’14, pages 1183–1213, 2014.
- [DKS16a] Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. The Fourier transform of Poisson multinomial distributions and its algorithmic applications. In *Proceedings of the 48th Annual ACM Symposium on the Theory of Computing*, STOC ’16, New York, NY, USA, 2016. ACM.
- [DKS16b] Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Optimal learning via the Fourier transform for sums of independent integer random variables. In *Proceedings of the 29th Annual Conference on Learning Theory*, COLT ’16, 2016.
- [DKT15] Constantinos Daskalakis, Gautam Kamath, and Christos Tzamos. On the structure, covering, and learning of Poisson multinomial distributions. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*, FOCS ’15, Washington, DC, USA, 2015. IEEE Computer Society.

- [DP88] Paul Deheuvels and Dietmar Pfeifer. Poisson approximations of multinomial distributions and point processes. *Journal of multivariate analysis*, 25(1):65–89, 1988.
- [DP07] Constantinos Daskalakis and Christos H. Papadimitriou. Computing equilibria in anonymous games. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '07, pages 83–93, Washington, DC, USA, 2007. IEEE Computer Society.
- [DP08] Constantinos Daskalakis and Christos H. Papadimitriou. Discretized multinomial distributions and Nash equilibria in anonymous games. In *Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '08, pages 25–34, Washington, DC, USA, 2008. IEEE Computer Society.
- [DP09] Constantinos Daskalakis and Christos H. Papadimitriou. On oblivious PTAS's for Nash equilibrium. In *Proceedings of the 41st Annual ACM Symposium on the Theory of Computing*, STOC '09, pages 75–84, New York, NY, USA, 2009. ACM.
- [DP15] Constantinos Daskalakis and Christos H. Papadimitriou. Approximate Nash equilibria in anonymous games. *Journal of Economic Theory*, 156:207–245, 2015.
- [Ess42] Carl-Gustaf Esseen. On the Liapounoff limit of error in the theory of probability. *Arkiv för matematik, astronomi och fysik*, 28A(2):1–19, 1942.
- [Kal05] Ehud Kalai. Partially-specified large games. In *Proceedings of the 1st International Workshop on Internet and Network Economics*, WINE '05, pages 3–13, Berlin, Heidelberg, 2005. Springer.
- [Loh92] Wei-Liem Loh. Stein's method and multinomial approximation. *The Annals of Applied Probability*, 2(3):536–554, 08 1992.
- [LZ15] Shachar Lovett and Jiapeng Zhang. Improved noisy population recovery, and reverse Bonami-Beckner inequality for sparse functions. In *Proceedings of the 47th Annual ACM Symposium on the Theory of Computing*, STOC '15, pages 137–142, New York, NY, USA, 2015. ACM.
- [Mil96] Igal Milchtaich. Congestion games with player-specific payoff functions. *Games and Economic Behavior*, 13(1):111–124, 1996.
- [MS13] Ankur Moitra and Michael Saks. A polynomial time algorithm for lossy population recovery. In *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '13, pages 110–116, Washington, DC, USA, 2013. IEEE Computer Society.
- [Roo02] Bero Roos. Multinomial and Krawtchouk approximations to the generalized multinomial distribution. *Theory of Probability & Its Applications*, 46(1):103–117, 2002.
- [She10] I.G. Shevtsova. An improvement of convergence rate estimates in the Lyapunov theorem. *Doklady Mathematics*, 82(3):862–864, 2010.
- [SS11] Daniel A. Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. *SIAM Journal on Computing*, 40(6):1913–1926, 2011.

- [ST11] Daniel A. Spielman and Shang-Hua Teng. Spectral sparsification of graphs. *SIAM Journal on Computing*, 40(4):981–1025, 2011.
- [Sta69] Ross M Starr. Quasi-equilibria in markets with non-convex preferences. *Econometrica*, 37(1):25–38, 1969.
- [VdV00] A. W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge University Press, 2000.
- [VV10] Gregory Valiant and Paul Valiant. A CLT and tight lower bounds for estimating entropy. *Electronic Colloquium on Computational Complexity (ECCC)*, 17(179), 2010.
- [VV11] Gregory Valiant and Paul Valiant. Estimating the unseen: An $n/\log n$ -sample estimator for entropy and support size, shown optimal via new CLTs. In *Proceedings of the 43rd Annual ACM Symposium on the Theory of Computing*, STOC '11, pages 685–694, New York, NY, USA, 2011. ACM.
- [Woo96] Trevor D. Wooley. A note on simultaneous congruences. *Journal of Number Theory*, 58(2):288–297, 1996.
- [WY12] Avi Wigderson and Amir Yehudayoff. Population recovery and partial identification. In *Proceedings of the 53rd Annual IEEE Symposium on Foundations of Computer Science*, FOCS '12, pages 390–399, Washington, DC, USA, 2012. IEEE Computer Society.

A Proof of Lemma 2

We instead prove that $d_{TV}(X, Y) \leq \varepsilon\sqrt{k}$ when $|v^T(\mu_1 - \mu_2)| \leq \varepsilon\sqrt{k}s_v$ and $|v^T(\Sigma_1 - \Sigma_2)v| \leq \frac{\varepsilon s_v^2}{2}$, which we can see is equivalent to the lemma statement by a rescaling.

Without loss of generality, assume that Σ_1 and Σ_2 are full rank. If not, the guarantees in the statement ensure that their nullspace is identical, and we can project to a lower dimension such that the resulting matrices are full rank.

First, we note that the assumptions in the lemma statement can be converted to be in terms of the minimum of the two variances, instead of the maximum. Define $\sigma_v^2 = \min\{v^T\Sigma_1 v, v^T\Sigma_2 v\}$. The second assumption can be rearranged to see that $(1 - \frac{\varepsilon}{2})s_v^2 \leq \sigma_v^2$. Plugging this back into the second assumption gives that

$$|v^T(\Sigma_1 - \Sigma_2)v| \leq \frac{\varepsilon s_v^2}{2} \leq \frac{\varepsilon \sigma_v^2}{2(1 - \frac{\varepsilon}{2})} \leq \varepsilon \sigma_v^2,$$

where the last inequality holds for $\varepsilon \leq 1$ (otherwise, the lemma's conclusion is trivial). Similarly, the second assumption also implies $\sqrt{1 - \frac{\varepsilon}{2}}s_v \leq \sigma_v$, when plugged into the first assumption gives

$$|v^T(\mu_1 - \mu_2)| \leq \varepsilon\sqrt{k}s_v \leq \frac{\varepsilon}{\sqrt{1 - \frac{\varepsilon}{2}}}\sqrt{k}\sigma_v \leq \sqrt{2}\varepsilon\sqrt{k}\sigma_v.$$

For the remainder of the proof, we will use these guarantees instead of the ones in the lemma statement.

We recall the standard formula for KL-divergence between two Gaussian distributions. Let $\{\lambda_i\}$ be the eigenvalues of $\Sigma_2^{-1/2}\Sigma_1\Sigma_2^{-1/2}$.

$$\begin{aligned} d_{\text{KL}}(X||Y) &= \frac{1}{2} \left((\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) + \text{Tr}(\Sigma_2^{-1/2} \Sigma_1 \Sigma_2^{-1/2}) - \ln \left(\det \left(\Sigma_2^{-1/2} \Sigma_1 \Sigma_2^{-1/2} \right) \right) - k \right) \\ &= \frac{1}{2} \left((\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) + \sum_{i=1}^k (\lambda_i - \ln \lambda_i - 1) \right) \end{aligned}$$

We bound the divergence induced by differences in the means and covariances separately. We start with the means. Note that

$$|v^T(\mu_2 - \mu_1)| \leq \sqrt{2}\varepsilon\sqrt{k}\sigma_v \Rightarrow \frac{|v^T(\mu_2 - \mu_1)|}{\sqrt{v^T \Sigma_2 v}} \leq \sqrt{2}\varepsilon\sqrt{k}.$$

Substituting $u = \Sigma_2 v$ gives

$$\frac{|u^T \Sigma_2^{-1}(\mu_2 - \mu_1)|}{\sqrt{u^T \Sigma_2^{-1} u}} \leq \sqrt{2}\varepsilon\sqrt{k}.$$

We let $u = \mu_2 - \mu_1$, giving

$$\sqrt{(\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1)} \leq \sqrt{2}\varepsilon\sqrt{k},$$

which implies

$$(\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) \leq 2\varepsilon^2 k.$$

Now we bound the divergence induced by differences in the covariances. We bound the eigenvalues of $\Sigma_2^{-1/2}\Sigma_1\Sigma_2^{-1/2}$. Note that

$$|v^T(\Sigma_1 - \Sigma_2)v| \leq \varepsilon\sigma_v^2 \Rightarrow \frac{1}{1 + \varepsilon} \leq \frac{v^T \Sigma_1 v}{v^T \Sigma_2 v} \leq 1 + \varepsilon.$$

Substituting $u = \Sigma_2^{1/2}v$ makes the latter condition equivalent to

$$\frac{1}{1 + \varepsilon} \leq \frac{u^T \Sigma_2^{-1/2} \Sigma_1 \Sigma_2^{-1/2} u}{u^T u} \leq 1 + \varepsilon.$$

The Courant-Fischer Theorem implies that $\frac{1}{1+\varepsilon} \leq \lambda_i \leq 1 + \varepsilon$ for all i .

At this point, we note that $x - \ln x - 1 \leq (1 - x)^2$ for all $x \geq 1$. This implies

$$\sum_{i=1}^k (\lambda_i - \ln \lambda_i - 1) \leq \sum_{i=1}^k (1 - \lambda_i)^2 \leq \varepsilon^2 k.$$

Thus, $d_{\text{KL}}(X||Y) \leq 2\varepsilon^2 k$. Applying Pinsker's inequality gives $d_{\text{TV}}(X, Y) \leq \varepsilon\sqrt{k}$, as desired.