2015 IEEE 56th Annual Symposium on Foundations of Computer Science

On the Structure, Covering, and Learning of Poisson Multinomial Distributions

Constantinos Daskalakis*, Gautam Kamath† and Christos Tzamos‡
Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA, USA
*Email: *costis@mit.edu, †g@csail.mit.edu, ‡tzamos@mit.edu,*

Abstract

An (n, k)-*Poisson Multinomial Distribution* (PMD) is the distribution of the sum of n independent random vectors supported on the set $\mathcal{B}_k = \{e_1, \ldots, e_k\}$ of standard basis vectors in $\mathbb{R}^k$. We prove a structural characterization of these distributions, showing that, for all $\varepsilon > 0$, any (n, k)-Poisson multinomial random vector is ε-close, in total variation distance, to the sum of a discretized multidimensional Gaussian and an independent $(\text{poly}(k/\varepsilon), k)$-Poisson multinomial random vector. Our structural characterization extends the multi-dimensional CLT of [2], by simultaneously applying to all approximation requirements ε. In particular, it overcomes factors depending on $\log n$ and, importantly, the minimum eigenvalue of the PMD's covariance matrix.

We use our structural characterization to obtain an ε-cover, in total variation distance, of the set of all (n, k)-PMDs, significantly improving the cover size of [3], [4], and obtaining the same qualitative dependence of the cover size on n and ε as the $k = 2$ cover of [5], [6]. We further exploit this structure to show that (n, k)-PMDs can be learned to within ε in total variation distance from $\tilde{O}_k(1/\varepsilon^2)$ samples, which is near-optimal in terms of dependence on ε and independent of n. In particular, our result generalizes the single-dimensional result of [7] for Poisson binomials to arbitrary dimension. Finally, as a corollary of our results on PMDs, we give a $\tilde{O}_k(1/\varepsilon^2)$ sample algorithm for learning (n, k)-sums of independent integer random variables (SIIRVs), which is near-optimal for constant k.

Keywords

Structure; Learning; Applied probability; Gaussian distribution; Multivariate statistics; Limit theorem

I. Introduction

Poisson Multinomial Distributions (PMDs) are one the most basic nonparametric multidimensional families of distributions. They express the distribution of how many out of n thrown balls will fall into k bins, when the balls (perhaps because of weight or other characteristics) have different biases towards falling into the different bins. Mathematically, a (n, k)-PMD is the distribution of the sum $\sum_{i=1}^{n} X_i$ of n independent random vectors X_i supported on the set $\mathcal{B}_k = \{e_1, \ldots, e_k\}$ of standard basis vectors in $\mathbb{R}^k$. In particular, a (n, k)-PMD requires for its description $n \cdot (k-1)$ probabilities, specifying the distribution of each summand random vector.

In this paper, we advance our understanding of the structure and learnability of this fundamental family of distributions by studying the following questions:

1) Can we approximate PMDs via simpler distributions such as multi-dimensional Gaussians or Poissons? Do they always "behave as" discretized multi-dimensional Gaussians or Poissons? If not, what is the range of possible "behaviors" that PMDs may exhibit?
2) Given n, k and ε, is there a small set of distributions that ε-cover, in total variation distance, the set of all (n, k)-PMDs? And, how does the size of the cover scale with n, k and ε?
3) How many samples from a (n, k)-PMD do we need to learn its density to within ε in total variation distance? What is the dependence of the learning complexity on the size $n^{O(k)}$ of their support?

The full version of this paper can be found at [1].

0272-5428/15 $31.00 © 2015 IEEE
DOI 10.1109/FOCS.2015.77

Structure of PMDs: It is hard to do justice to the probability literature studying Question 1. The multi-dimensional CLT informs us that the limiting behavior of (n, k)-PMDs, as $n \to +\infty$, is Gaussian, under conditions on the eigenvalues of the summands' covariance matrices; see, e.g., [8].[1] The CLT is quantified for finite n by the multi-dimensional Berry-Esseen theorem, which bounds the difference between the probability masses assigned to convex (or a bit more general) subsets of $\mathbb{R}^k$ by a (n, k)-PMD and the multi-dimensional Gaussian distribution with the same mean vector and covariance matrix, with the bound's quality typically degrading as the PMD's covariance matrix tends to singularity; see, e.g., [9]. More recently, Valiant and Valiant [2] provide a bound in total variation distance, between a (n, k)-PMD and the corresponding *discretized* multi-dimensional Gaussian, whose quality degrades mildly with n and worse with the minimum eigenvalue of the PMD's covariance matrix (see Theorem 6).[2] Finally, older results using Stein's method bound the total variation distance between a (n, k)-PMD and a multivariate Poisson [10], [11], or a (bona fide) multinomial distribution [12].

In summary, known bounds show that a (n, k)-PMD can be approximated by simpler, poly(k)-parameter, distributions, but the quality of their approximation depends on the first few moments of the PMD or its summands. Our goal instead is to provide universal approximation theorems showing how to approximate a given (n, k)-PMD by simpler distributions for *any desired approximation* ε and *without assumptions about the moments of the PMD or its summands*. Our main structural theorem is the following.

Theorem 1 (PMD Structure)**.** *For all $n, k \in \mathbb{N}$, and all $\varepsilon > 0$, a (n, k)-Poisson multinomial random vector is ε-close, in total variation distance, to the sum of a discretized multidimensional Gaussian and an independent $(\mathrm{poly}(k/\varepsilon), k)$-Poisson multinomial random vector.*

By introducing the independent $(\mathrm{poly}(k/\varepsilon), k)$-PMD, our structural result side-steps the degradation of the CLT bound of [2] with $\log n$ and the smallest eigenvalue of the PMD's covariance matrix, correcting it to any desired approximation ε. Interestingly, there may be directions where the variance of the discretized Gaussian used in our result may be arbitrarily far from that of the approximated PMD. The sparse PMD added to the Gaussian serves to correct the variance in those directions, but does so in a correlated manner across several directions. Moreover, while [2] discretize their approximating multidimensional Gaussian to the closest lattice point, our discretization is more faithful to the structure of its covariance matrix; see Definition 6. We provide more intuition about our structural result in Section I-A, where we also outline its proof. A more detailed proof of Theorem 1 appears in Section III and a more detailed statement is given as Theorem 5.

Covers for PMDs: Building covers for (n, k)-PMDs was pursued in [3], [4] as a means to develop approximation algorithms for Nash equilibria in anonymous games. These are games where n players share the same action set, say $\{1, \ldots, k\}$, and each player's utility depends on their own choice of action as well as the distribution of how many of the other players choose each of the available actions, but players' utility functions may otherwise be different. It was shown that proper ε-covers, in total variation distance, of (n, k)-PMDs[3] imply approximation algorithms for Nash equilibria in these games, whose complexity scales with the size of the cover. Intuitively, this is because switching from a mixed Nash equilibrium to a mixed strategy profile with the same distribution of how many players choose each action does not affect players' payoffs by more than ε.

The covers for (n, k)-PMDs obtained in the anonymous games papers cited above have size:

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

Such covers are of theoretical interest, their interesting feature being that the size is polynomial in n. Indeed, the standard discretization of the parameters of a PMD's constituent vectors results in covers of size exponential in

[1] When we approximate some (n, k)-PMD or refer to the eigenvalues of its covariance matrix, we typically project the PMD onto a $(k-1)$-dimensional space, e.g. by excluding one of its coordinates, as otherwise the covariance matrix always has a 0 eigenvalue and the distribution does not have full-dimensional support.

[2] Notice that bounds on total variation distance are stronger than bounds on the probabilities of all events defined by convex sets in $\mathbb{R}^k$ that Berry-Esseen-type theorems establish.

[3] An ε-cover $\mathcal{F}_\varepsilon$ of a set of distributions $\mathcal{F}$ is called *proper* iff $\mathcal{F}_\varepsilon \subseteq \mathcal{F}$.

n, so a more delicate "global" discretization is needed to obtain covers whose size is polynomial in n.

Besides providing an asymptotically smaller search space for Nash equilibria in anonymous games, or any other optimization problem over PMDs, the polynomial rather than exponential dependence of the cover size on n has direct consequences to the learnability of these distributions; see Theorem 7 (from [13]) and [14] for a similar result, which improve a long line of similar results in the probability literature [15]. In particular, a cover of polynomial size implies directly that these distributions can be learned from a number of samples logarithmic in n, despite their support being polynomial in n. Motivated by such applications of covers to algorithms and learning we use our structural result to obtain an improved cover theorem.

Theorem 2 (PMD Covers)**.** *For all $n, k \in \mathbb{N}$, and $\varepsilon > 0$, there exists an ε-cover, in total variation distance, of the set of all (n,k)-PMDs whose size is*

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

We make a few remarks about our cover. First, the cover is non-proper, containing distributions that are of the form specified in Theorem 1, i.e. are convolutions of a discretized Gaussian and a PMD. Moreover, it is straightforward to see that any cover has size at least $n^{\Omega(k)}$ and at least $(1/\varepsilon)^{\Omega(k)}$. For the first lower bound, count the number of (n,k)-PMDs whose summands are deterministic. For the second, count the number of $(1,k)$-PMDs whose probabilities are integer multiples of ε. So, for fixed k, our bound has the right qualitative dependence on n (namely polynomial), and a near-right dependence on $1/\varepsilon$ (namely quasi-polynomial rather than polynomial). Moreover, it obtains the same qualitative dependence on n and ε as the $k = 2$ cover of [5], [6], namely polynomial in n and quasi-polynomial in $1/\varepsilon$.

Learning PMDs: In view of tools for hypothesis selection from a cover (see, i.e., Theorem 7), our cover theorem directly implies that (n,k)-PMDs can be learned from $O(k^{5k} \cdot \log n \cdot \log^{k+2}(1/\varepsilon)/\varepsilon^2)$ samples. These are near-optimal in terms of ε, as $\Omega(k/\varepsilon^2)$ samples are necessary even for learning a $(1,k)$-PMD. We show that the dependence on n can be completely removed from the learner, generalizing the results on Poisson Binomial Distributions [7].

Theorem 3 (PMD Learning)**.** *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*

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

samples from X, runs in time[4]

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

and with probability at least $9/10$ outputs a (succinct description of a) random vector $\tilde{X}$ such that $d_{\mathrm{TV}}(X, \tilde{X}) \leq \varepsilon$.

Additional Results: Learning k-SIIRVs: A (n,k)-SIIRV is the sum of n independent (single-dimensional) random variables supported on $\{0, \ldots, k-1\}$. SIIRVs generalize Poisson Binomial distributions, which correspond to the case $k = 2$. At the same time, SIIRVs can be viewed as projections of PMDs onto the vector $(0, 1, \ldots, k-1)$. In particular, if X is a (n,k)-SIIRV, there exists a (n,k)-Poisson multinomial random vector Y, such that $X = (0, 1, \ldots, k-1)^{\mathrm{T}} \cdot Y$.

Recent work has established that (n,k)-SIIRVs can be learned from $\mathrm{poly}(k/\varepsilon)$ samples, independent of n, when even learning a $(1,k)$-SIIRV already requires $O(k/\varepsilon^2)$ samples [16]. A question arising from this work is

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

finding the optimal dependence of the sample complexity on ε. Exploiting the connection between SIIRVs and PMDs, we show that the optimal dependence is actually $\tilde{O}_k(1/\varepsilon^2)$.

Theorem 4 (SIIRV Learning)**.** *For all $n, k \in \mathbb{N}$ and $\varepsilon > 0$, there is a learning algorithm for (n,k)-SIIRVs with the following properties: Let $X = \sum_{i=1}^{n} X_i$ be any (n,k)-SIIRV. The algorithm uses $k^{5k} \cdot O(\log^{k+2}(1/\varepsilon)/\varepsilon^2)$ samples from X, runs in time $2^{O(k^{5k}\cdot\log^{k+2}(1/\varepsilon))}$, and with probability at least $9/10$ outputs a random vector $\tilde{X}$ such that $d_{\mathrm{TV}}(X,\tilde{X}) \leq \varepsilon$.*

A. Approach

Structure: The multi-dimensional nature of PMDs poses challenges in understanding their structure. The projection of a (n,k)-Poisson multinomial random vector onto each standard basis vector is a n-Poisson Binomial random variable, i.e. distributed as the sum of n independent indicators. Depending on our choice of ε, the latter may be ε-close (in total variation distance) to a discretized Normal distribution ("heavy projection") or a distribution whose essential support is a length $O(1/\varepsilon^3)$ subinterval of $\{0,\ldots,n\}$ ("light projection") [6]. Intuitively, one would like to aggregate all heavy projections into a discretized multi-dimensional Gaussian and all light projections into a distribution of small support, independent of n. However, projections onto different standard basis vectors may be correlated, and they cannot be disentangled this simply.

In fact, even if all projections of a PMD onto the standard basis vectors are heavy—even if they have variance super-polynomial in k/ε, it is still unclear whether the PMD can always be well approximated by a discretized multi-dimensional Gaussian. In particular, the multi-dimensional CLT of Valiant and Valiant [2] (Theorem 6) does pay a penalty that scales with $\log n$.

Finally, projections onto non-standard basis vectors may behave more erratically. As we pointed out earlier, the projection of a (n,k)-PMD onto the vector $\vec{v} = (0,1,\ldots,k-1)$ is a (n,k)-SIIRV, which need not be log-concave or even unimodal, and could even exhibit "mod-structure" and be n-modal; think of the distribution of $Y + 2\cdot Z$ where Z is sampled from a Binomial$(n,0.5)$ and Y is a Bernoulli$(1/3)$. Whichever simpler distribution we identify to approximate a given (n,k)-PMD thus needs to respect the potential mod-structure that the PMD's projection onto $\vec{v}$, its permutations or other integral vectors may exhibit.

Our analysis sidesteps the difficulties identified above by showing that, for all ε, n, k, a (n,k)-Poisson multinomial random vector is ε-close to the sum of a discretized Gaussian and an independent $(\mathrm{poly}(k/\varepsilon),k)$-Poisson multinomial random vector. Roughly speaking, the Gaussian absorbs the variance in the heavy dimensions, and explains the correlation between light and heavy dimensions, while the sparse PMD explains the remaining variance in the light dimensions. Of course, what dimensions are "light" and "heavy" in the above discussion depends on our desired approximation ε.

At the heart of our proof lies the aforecited CLT by Valiant and Valiant [2], approximating a Poisson Multinomial by a discretized Gaussian. There are several issues with its application here: the accuracy of the approximation cannot be made an arbitrary ε, but worse, it deteriorates (logarithmically) as we increase n or decrease the minimum eigenvalue of the covariance matrix of the PMD. The main intuition behind our structural theorem and the main technical roadblock for its proof lies in avoiding paying these two penalties.

To mitigate the latter cost (corresponding to the smallest eigenvalue), we use a stripped down version of the trickle-down sampling procedure from [3] to round the parameters of our given PMD. This allows us to shift the parameters of the PMD's constituent random vectors such that they are either equal to 0 or 1, or sufficiently far from 0 or 1. A coordinated "rounding" of these parameters combined with a coupling argument and single-dimensional Poisson approximations allow us to argue that the effect of the rounding is small in the total variation distance of the resulting PMD compared to the original PMD. Each constituent random vector in the resulting PMD now has decent variance in every axis direction where it has non-zero variance. Partitioning the PMD's constituent vectors into sets based on the axis directions where they have non-zero variance, we get that the

minimum eigenvalue of each resulting sub-PMD is large in the span of these directions.[5]

To avoid paying the logarithmic cost in the value of n (the number of summands) which appears in the CLT, we repeatedly partition and sort the random vectors into bins. The sub-PMD corresponding to each bin will have the property that the logarithm of the number of summands is negligible compared to the minimum eigenvalue of its covariance matrix, so that we can apply the central limit theorem from [2]. We note that there will be a small number of random vectors which do not fall into a bin that has this property – these leftover vectors result in the sparse Poisson Multinomial component in our structural result.

The above approximations result in a distribution comprising several discretized Gaussians and a sparse Poisson multinomial. We subsequently merge all component discretized Gaussians into a single distribution. It is well-known that the sum of two Gaussians is another Gaussian whose parameters are equal to the sum of the parameters of its two components. The same is not true for discretized Gaussians, and we must quantify the error induced by this merging operation.

Our structural results are described further in Section III.

Cover: We provide two covers for (n, k)-PMDs, which are advantageous for different regimes of k and ε. The first cover follows directly from Theorem 5, which gives a structural characterization of a PMD as the sum of an appropriately discretized Gaussian and a $(\mathrm{poly}(k/\varepsilon), k)$-PMD. We simply take an additive grid over all the parameters of this characterization to achieve a cover size which is polynomial in n and exponential in k and $1/\varepsilon$.

Similar to [6], we can reduce the dependence of the cover size to pseudo-polynomial in $1/\varepsilon$, albeit at an increased cost in k. This is done using a generalization of the moment matching techniques known for Poisson Binomial distributions. At a high level, this avoids the naive gridding over all $(\mathrm{poly}(k/\varepsilon), k)$-PMDs by filtering out the ones with unique "moment profiles," which describe the first several moments of the distribution. We prove that any two distributions with matching moment profiles will have small total variation distance by leveraging results by Roos on Krawtchouk approximations to PMDs [17].

A further description of our cover results is provided in Section IV.

Learning: Our cover theorem (Theorem 2) directly implies (using Theorem 7) that (n, k)-PMDs can be learned from $O(\log N/\varepsilon^2)$ samples, where N is the size of our cover. Given that N is polynomial in n, the resulting sample complexity is logarithmic in n. To remove the dependence on n from our sample complexity, we need to exploit not just the size but also the structure of the cover.

In particular, we know from our structural characterization (Theorem 1) that any (n, k)-Poisson Multinomial random vector is ε-close to the sum of a discretized multi-dimensional Gaussian and an independent $(\mathrm{poly}(k/\varepsilon), k)$-PMD. The dependence of the cover size on n is due to enumerating over a cover of discretized multi-dimensional Gaussians, as enumerating over $(\mathrm{poly}(k/\varepsilon), k)$-PMDs has no dependence on n. The challenge is this: given sample access to an unknown (n, k)-PMD can we zoom in to a smaller set of candidate discretized multi-dimensional Gaussians whose size is independent of n and which suffice for the purposes of guaranteeing an approximation to the unknown PMD?

Let us start with an easier task. Suppose that our structural theorem decides that a (n, k)-PMD is ε-close in total variation distance to a discretized multi-dimensional Gaussian. In this case, is it possible to recover the Gaussian from $\mathrm{poly}(k/\varepsilon)$ samples from the PMD? Intuitively the answer should be "yes," as learning a multi-dimensional Gaussian to within ε in total variation distance is feasible from $O(k/\varepsilon^2)$ samples. Only there are two complications. First, we are seeking to actually learn a discretized multi-dimensional Gaussian and, most importantly, we do not have sample access to the Gaussian, but a distribution that is ε-close to it in total variation distance. The first complication becomes an issue when the covariance matrix of the Gaussian has minimum eigenvalue that does not scale with some $\mathrm{poly}(k/\varepsilon)$, which may very well be the case. The second is more severe as it necessitates robust estimators for the moments of a (discretized) multi-dimensional Gaussian that

[5]Again, as pointed out earlier, when we refer to the eigenvalues of the covariance matrix of a PMD spanning a certain subspace, we always project the PMD onto a subspace of one dimension less, as otherwise the covariance matrix always has a 0 eigenvalue since the distribution does not have full-dimensional support.

are resilient to an arbitrary movement of ε probability mass. We are not aware of such estimators even for a (continuous) multi-dimensional Gaussian.

Despite these apparent issues, even in the simple case we are considering, the saving grace comes from a closer examination of the proof of our structural result. When our structural theorem deems a (n, k)-PMD approximable by a discretized multi-dimensional Gaussian, we can argue that the covariance matrices Σ of the former and Σ_G of the latter are spectrally close, satisfying $|x^{\mathrm{T}}\Sigma x - x^{\mathrm{T}}\Sigma_G x| \leq \varepsilon \cdot x^{\mathrm{T}}\Sigma x$, for all x. So it suffices to learn the covariance matrix of the PMD to which we have direct sample access, thereby obviating the need for a robust estimator. Learning the covariance matrix of a PMD is feasible from $\mathrm{poly}(k/\varepsilon)$ samples by bounding the kurtosis of any projection of the PMD (Lemma 8).

The bigger challenge is generalizing the approach to when our structural theorem deems a (n, k)-Poisson Multinomial random vector X approximable by the sum of a discretized multi-dimensional Gaussian G and a $(\mathrm{poly}(k/\varepsilon), k)$-Poisson Multinomial random vector Y. We can enumerate over the latter, but enumerating over the former is too expensive (i.e. will incur a dependence on n). So we have to learn it with sample access to X. Unfortunately, our spectral approximation is now much weaker. The covariance matrices Σ of X and Σ_G of G are now related as follows, for all x: $|x^{\mathrm{T}}\Sigma x - x^{\mathrm{T}}\Sigma_G x| \leq \varepsilon \cdot x^{\mathrm{T}}\Sigma x + \mathrm{poly}(k/\epsilon)$. Hence, for directions x where the variance $x^{\mathrm{T}}\Sigma x$ of X is small, this approximation is quite loose to just approximate Σ_G with Σ.

Our approach is instead to use samples from X to get a handle on the spectrum of Σ_G. As before, by bounding the kurtosis of any projection of the PMD, we can produce an estimate $\hat{\Sigma}$ that approximates Σ spectrally: for all x, $|x^{\mathrm{T}}\Sigma x - x^{\mathrm{T}}\hat{\Sigma}x| \leq \varepsilon \cdot x^{\mathrm{T}}\Sigma x$ (Lemma 8). Then, using Courant minimax principle through the proof of our structural result, we can argue that the i-th eigenvalue λ_i^G of Σ_G and $\hat{\lambda}_i$ of $\hat{\Sigma}$ are related as follows: $|\lambda_i^G - \hat{\lambda}_i| \leq O(\varepsilon)\hat{\lambda}_i + \mathrm{poly}(k/\epsilon)$. So, using the eigenvalues of our learned $\hat{\Sigma}$, we can produce a small cover for the eigenvalues of Σ_G. Unfortunately, the corresponding eigenvectors of Σ_G and $\hat{\Sigma}$ need not be as closely related, and it is not clear how to grid over those as the ratio of the smallest to the largest eigenvalue may be polynomial in n. We show how to use the knowledge of the eigenvalues and the spectral relation between $\hat{\Sigma}$ and Σ_G to produce a small cover over matrices $\hat{\Sigma}_G$ (and not eigenvectors) such that at least one matrix in the cover spectrally approximates our target Σ_G. The details are provided in the appendix of the full version. At this point, we have a small cover over possible distributions Y and a small cover over possible discretized multi-dimensional Gaussians. So we can select among these hypotheses using Theorem 7.

Our learning algorithm is described in Section V.

II. Preliminaries

A. Parameters

Throughout this paper, we will repeatedly refer to three key parameters, $c = c(\varepsilon, k) = \mathrm{poly}(\varepsilon/k)$, $t = t(\varepsilon, k) = \mathrm{poly}(k/\varepsilon)$, and $\gamma = O(1)$. We set

$$c = \left(\frac{\varepsilon^2}{k^5}\right)^{1+\delta_c}, \qquad t = \left(\frac{k^{19}}{c\varepsilon^6}\right)^{1+\delta_t}, \qquad \gamma = 6 + \delta_\gamma,$$

for constants $\delta_c, \delta_t, \delta_\gamma > 0$.

B. Definitions

We start by defining several of the distribution classes we will consider. First, and most importantly, we start with a formal definition of Poisson Multinomial Distributions.

Definition 1. *A* k-Categorical Random Variable *(k-CRV) is a random variable that takes values in* $\{e_1, \ldots, e_k\}$ *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 [2].

Definition 3. *A* Truncated k-Categorical Random Variable *is a random variable that takes values in $\{0, e_1, \ldots, e_{k-1}\}$ 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 $\lfloor x \rceil$ 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 $\lfloor \mathcal{N}(\mu, \Sigma) \rceil$, is the distribution with support $\mathbb{Z}^k$ obtained by picking a sample according to the k-dimensional Gaussian $\mathcal{N}(\mu, \Sigma)$, 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.*

III. Structure of PMDs

In this section, we show a structural result, stating that any (n, k)-PMD is close to the sum of an appropriately discretized Gaussian and a $(\mathrm{poly}(k/\varepsilon), k)$-PMD:

Theorem 5. *For parameters c and t as described in Section II-A, 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 $\frac{tc}{2k^4}$.*

There are three main steps in the proof of this theorem.

Step 1 First, we replace our (n, k)-PMD with one where all parameters are sufficiently far from 0 and 1, while still being close to the original in total variation distance. To motivate this operation, we introduce one of our main tools in our approach, the central limit theorem of Valiant and Valiant [2], which approximates an (n, k)-GMD by a discretized multivariate Gaussian.

Theorem 6 (Theorem 4 from [19])**.** *Given a generalized multinomial distribution G^ρ, with k dimensions and n rows, let μ denote its mean and Σ denote its covariance matrix, then*

$$d_{\mathrm{TV}}\left(G^\rho, \lfloor \mathcal{N}(\mu, \Sigma) \rceil\right) \leq \frac{k^{4/3}}{\sigma^{1/3}} \cdot 2.2 \cdot (3.1 + 0.83 \log n)^{2/3}$$

where σ^2 is the minimum eigenvalue of Σ.

We note that this has an error term which depends on the minimum eigenvalue of the covariance matrix of the GMD. If we perform this rounding procedure and ignore any zero coordinates, then we are given the guarantee that the minimum eigenvalue will be sufficiently large.

Recall that in Section II-A we have set $c = \mathrm{poly}(\varepsilon/k)$. This lemma summarizes the result of the rounding procedure:

Lemma 1. *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*

$$d_{\mathrm{TV}}\left(M^\rho, M^{\hat{\rho}}\right) < O\left(c^{1/2} k^{5/2} \log^{1/2}\left(\frac{1}{ck}\right)\right).$$

The procedure starts by fixing two coordinates i and j, and considers all CRVs with a parameter in i which is close to 0, and has maximum parameter in coordinate j. We move some of the weight in this "heavy" coordinate either to or from the "light" coordinate, while approximately preserving the overall mean vector of the set of CRVs.

The analysis of this process uses a stripped-down version of the "trickle-down" process in [3]. This gives an approximate way to sample from a PMD, resulting in a distribution which is very close in total variation distance. While we postpone technical details to the appendix of the full version, roughly speaking, it works as follows. First, take a sample from the PMD but disregard the values for its light coordinate i and heavy coordinate j. Instead, sample a new value for coordinate i according to a Poisson distribution with parameter μ_i, the mean value for coordinate i. Finally, set coordinate j to ensure that all coordinates of the sample sum to n. As mentioned before, the rounding process approximately preserves the value of μ_i, and thus this alternate sampling procedure is closely coupled for the rounded and original PMD. Thus, by triangle inequality, the rounded and original PMDs are close in total variation distance.

We repeat this rounding procedure for each i and j, eventually leading to all parameters either being equal to or far from 0 and 1. A full description and analysis of the rounding procedure are in the appendix of the full version.

Step 2 Now, we have a "massaged" (n, k)-PMD $M^{\hat{\rho}}$, with no parameters lying in the intervals $(0, c)$ or $(1 - c, 1)$. Next, we will show how to relate the massaged (n, k)-Poisson multinomial random vector to a sum of k Gaussians with a structure preserving rounding plus a "sparse" $(\mathrm{poly}(k/\varepsilon), k)$-PMD. The general roadmap

is as follows. We start by partitioning the constituent k-CRVs into k sets, $S_1, \ldots, S_k$, based on which basis vector we are most likely to observe. We work seperately for each set S_i by considering the GMD formed by leaving out the coordinate i. Our goal is to use the CLT of Theorem 6 to bound the total variation distance between the corresponding GMD and a discretized Gaussian with the same mean and covariance matrix. We must be careful when applying Theorem 6, since the bound depends on the size of the GMD. Instead of applying the theorem directly, to get a useful bound, we further partition the set S_i into smaller subsets and apply the theorem to each of the resulting subsets. We can then "merge" the resulting discretized Gaussians together using the following lemma whose proof is given in the appendix of the full version:

Lemma 2. *Let $X_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $X_2 \sim \mathcal{N}(\mu_2, \Sigma_2)$ be k-dimensional Gaussian random variables, and let $\sigma = \min_j \max_i \sigma_{i,j}$ where $\sigma_{i,j}$ is the standard deviation of X_i in the direction parallel to the jth coordinate axis. Then*

$$d_{\mathrm{TV}}\left(\lfloor X_1 + X_2 \rceil, \lfloor X_1 \rceil + \lfloor X_2 \rceil\right) \leq \frac{k}{2\sigma}.$$

In more detail, we partition each set S_i into 2^{k-1} subsets, grouping together CRVs according to the dimensions they are non-zero in, i.e. set $S_i^{\mathcal{I}}$ contains all CRVs that are non zero in the coordinates given by set $\mathcal{I} \subseteq [k] \setminus \{i\}$. We then group these sets into buckets, where a set is assigned to a bucket depending on its cardinality; bucket B^l gets all sets $S_i^{\mathcal{I}}$ with $|S_i^{\mathcal{I}}| \in [l^\gamma t, (l+1)^\gamma t)$, with $\gamma = O(1)$ and $t = \mathrm{poly}(k/\varepsilon)$ as defined in Section II-A. This bounds the ratio between the size and the minimum eigenvalue of the covariance of the GMD within every bucket other than B^0. This allows us to apply Theorem 6 and replace the CRVs within each bucket B^l for $l \geq 1$ with a discretized Gaussian, leaving us with a $(\mathrm{poly}(2^k/\varepsilon), k)$-GMD consisting of all the CRVs of bucket B^0. To reduce the number of remaining CRVs to polynomial in k, we show that by removing only $\mathrm{poly}(k/\varepsilon)$ of these CRVs, we can apply Theorem 6 again to the rest and obtain another discretized Gaussian. In particular, in the appendix of the full version, we prove the following lemma:

Lemma 3. *Let $G^{\hat{\rho}_k^0}$ be the $(|B^0|, k)$-GMD induced by the truncated CRVs in bucket B^0. Given $\hat{\rho}_k^0$, we can efficiently compute a partition of B^0 into S and $\bar{S}$, where $|\bar{S}| \leq kt$. Letting μ_S and Σ_S be the mean and covariance matrix of the $(|S|, k)$-GMD induced by S, and $G^{\hat{\rho}_k^{\bar{S}}}$ be the $(|\bar{S}|, k)$-GMD induced by $\bar{S}$,*

$$d_{\mathrm{TV}}\left(G^{\hat{\rho}_k^0}, \lfloor \mathcal{N}(\mu_S, \Sigma_S) \rceil * G^{\hat{\rho}_k^{\bar{S}}}\right) \leq \frac{8.646 k^{3/2} \log^{2/3}(2^k t)}{t^{1/6} c^{1/6}}.$$

Furthermore, the minimum non-zero eigenvalue of Σ_S is at least $\frac{tc}{k}$.

After merging together all discretized Gaussians (at most one coming from each bucket B^l for all $l \geq 0$) by iteratively applying Lemma 2, we are able to approximate each original set of CRVs S_i as the sum of a single discretized Gaussian and a $(\mathrm{poly}(k/\varepsilon), k)$-PMD. Combining the result from each of the sets S_i of the initial partition, we obtain the sum of k discretized Gaussians and a $(\mathrm{poly}(k/\varepsilon), k)$-PMD. The details of this step are described in the appendix of the full version.

Step 3 The final step is to show that the k discretized Gaussians can be merged into a single Gaussian with a structure preserving rounding. We note that we cannot apply Lemma 2 here, since each discretized Gaussian has a different pivot coordinate that has been left out. (Recall that by construction, the CRVs in set S_i are approximated by a discretized Gaussian that leaves out coordinate i). We thus need a new tool to enable us to merge Gaussians defined in different dimensions. The main idea is that if two Gaussians with a structure preserving rounding overlap in some dimension, we can use the common dimension as the pivot. We then add the mean vectors and covariance matrices to merge the distributions. Iteratively repeating this process will merge all distributions which overlap in some coordinate. This leaves us with one or many discretized Gaussians that lie in completely disjoint coordinates which we can describe as a single Gaussian with a structure preserving rounding (defining blocks according to the coordinates spanned by each Gaussian). If these were (continuous) Gaussians, the swapping and merging operations would have no cost, but some care

is required when dealing with discretized Gaussians. There are two costs which we must bound here. First, we must show that swapping the pivot of a PMD is inexpensive, and second, we need to bound the cost of repeatedly merging Gaussians.

We bound the cost of swapping the pivot by proving the following lemma:

Lemma 4 (Total Variation Swap Lemma)**.** *For $\mu \in \mathbb{R}^k$, positive semidefinite $\Sigma \in \mathbb{R}^{k\times k}$, $n \in \mathbb{Z}$, let*

- *X_i be the distribution $\mathcal{N}(\mu_{-i}, \Sigma_{-i})$, where $\mu_{-i} \in \mathbb{R}^{k-1}$ is μ with the ith coordinate removed, and $\Sigma_{-i} \in \mathbb{R}^{(k-1)\times(k-1)}$ is Σ with the ith row and column removed;*
- *Y_i be the distribution in which we draw a sample $(x_1, \ldots, x_{k-1}) \sim X_i$ and return*

$$(\lfloor x_1 \rceil, \ldots, \lfloor x_{i-1} \rceil, (n - \sum_{j=1}^{k-1} \lfloor x_j \rceil), \lfloor x_i \rceil, \ldots, \lfloor x_{k-1} \rceil).$$

Then $d_{\mathrm{TV}}(Y_i, Y_j) \le \frac{k}{2\sigma}$ for any $i, j \in [k]$, where $\sigma^2 = \max(\sigma_{-i}^2, \sigma_{-j}^2)$ and σ_{-i}^2 is the smallest eigenvalue of Σ_{-i}.

By applying Lemma 4, we can make two discretized Gaussians have the same left out coordinate and then merge them using Lemma 2 if at least one of them has large variance in every direction. While each of the k discretized Gaussians starts with this property (for the dimensions in which it is non-deterministic), it is not clear whether this is true after a sequence of pivot swaps and merges.

In many cases, swapping the pivot decreases the minimum eigenvalue of the distribution's covariance matrix by a factor of $\mathrm{poly}(k)$. This is acceptable if we only perform a single swap, but naively applying this bound for a sequence of k swaps and merges results in the minimum eigenvalue dropping by a factor of $k^{O(k)}$. We show that such a bad situation cannot occur, no matter how one performs the sequence of swaps and merges, by proving the following lemma:

Lemma 5 (Variance Swap Lemma)**.** *Let $\Sigma^{(1)}, \ldots, \Sigma^{(m)} \in \mathbb{R}^{k\times k}$ be a sequence of symmetric positive-semidefinite matrices, and define $S^{(i)} = \{j \mid e_j^T \Sigma^{(i)} e_j \neq 0\}$ to be the set of coordinates in which $\Sigma^{(i)}$ is non-zero. Furthermore, let $\Sigma = \sum_i \Sigma^{(i)}$ and $S = \cup_i S^{(i)}$. Suppose the following hold for all i:*

1) *$\Sigma^{(i)}$ has eigenvalue 0 with corresponding eigenvector $\vec{1}$*
2) *There exists coordinate $j^* \in S^{(i)}$ such that $\Sigma^{(i)}_{S^{(i)}\setminus\{j^*\}}$ has minimum eigenvalue at least λ*
3) *$\left(\cup_{\ell<i} S^{(\ell)}\right) \cap S^{(i)} \neq \emptyset$*

Then, for all $j \in S$, the minimum eigenvalue of $\Sigma_{S\setminus\{j\}}$ is at least $\frac{\lambda}{2k^3}$.

The details of this step, the proofs of Lemma 4 and Lemma 5 as well as the proof of Theorem 5 are described in the appendix of the full version.

IV. Covers for PMDs

In this section, we describe a pair of covers for (n, k)-PMDs.

The first cover follows directly from Theorem 5, which gives a structural characterization of a (n, k)-Poisson multinomial random vector as the sum of an appropriately discretized Gaussian and an (tk^2, k)-Poisson multinomial random vector. We grid over all possible mean vectors and covariance matrices for the Gaussian component, and all possible parameter values for the (tk^2, k)-PMD. These are covered by sets of size $(n \cdot \mathrm{poly}(k/\varepsilon))^{k^2}$ and $2^{\mathrm{poly}(k/\varepsilon)}$ respectively, resulting in an overall cover of size $n^{k^2} \cdot 2^{\mathrm{poly}(k/\varepsilon)}$.

Lemma 6. *For all $n, k \in \mathbb{N}$, and all $\varepsilon > 0$, there exists an ε-cover of the set of all (n, k)-PMDs whose size is*

$$n^{k^2} \cdot 2^{\mathrm{poly}(k/\varepsilon)}.$$

The proof of this lemma is presented in the appendix of the full version.

The second cover further sparsifies the cover for the (tk^2, k)-PMD component, by using a multivariate generalization of the moment matching technique described in [6]. This reduces the cover size for this component

to $2^{O(k^{5k}\log^{k+2}(1/\varepsilon))}$. In [17], Roos shows that a PMD can be written as the weighted sum of partial derivatives of a regular multinomial distribution. He goes on to show that dropping the higher order derivatives in this sum results in a total variation approximation, where the quality of the approximation depends on the parameters of the PMD and the point at which we evaluate the derivatives. We take advantage of this tool to obtain an ε-approximation, through a careful partitioning of the CRVs and choice of point at which to evaluate the derivatives of the multinomial distributions. This implies that any two distributions which have matching "moment profiles" (which roughly describe the lower order derivatives of the distribution) are ε-close to each other, and thus only one representative element must be kept from each such equivalence class. The size of the cover follows by a counting argument on the number of moment profiles.

Lemma 7. *For all $n, k \in \mathbb{N}$, and all $\varepsilon > 0$, there exists an ε-cover of the set of all (n,k)-PMDs whose size is*

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

The proof of this lemma is given in the appendix of the full version. We note that this cover can be efficiently enumerated over, using a dynamic program similar to that of [6].

By combining these two lemmas, we obtain Theorem 2.

V. Learning PMDs

As mentioned before, Theorem 2 combined with Theorem 7 below (taken from [13]) immediately implies that (n,k)-PMDs can be learned from $O(\log N/\varepsilon^2)$ samples, where N is the size of our cover.

Theorem 7 (Theorem 19 of [13])**.** *There is an algorithm* `FastTournament`$(X, \mathcal{H}, \varepsilon, \delta)$*, which is given sample access to some distribution X and a collection of distributions $\mathcal{H} = \{H_1, \ldots, H_N\}$ 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 $X, H_1, \ldots, H_N$ and returns some $H \in \mathcal{H}$ or declares "failure." If there is some $H^* \in \mathcal{H}$ such that $d_{\mathrm{TV}}(H^*, X) \le \varepsilon$ then with probability at least $1-\delta$ the distribution H that* `FastTournament` *returns satisfies $d_{\mathrm{TV}}(H, X) \le 512\varepsilon$. The total number of operations of the algorithm is $O\left(\frac{\log 1/\delta}{\varepsilon^2}\left(N \log N + \log^2 \frac{1}{\delta}\right)\right)$. Furthermore, the expected number of operations of the algorithm is $O\left(\frac{N \log N/\delta}{\varepsilon^2}\right)$.*

Theorem 7 is using a tournament-style algorithm for hypothesis selection, which takes a set of candidate distributions and outputs one which is $O(\varepsilon)$-close to the unknown distribution (if such a distribution exists)[6]. Given that N is polynomial in n, the resulting sample complexity is logarithmic in n. To remove the dependence on n from our sample complexity, we need to exploit not just the size but also the Gaussian structure of the cover. Instead of trying all possible Gaussians that the cover could describe, we instead estimate the moments of the Gaussian directly.

Our strategy will not be to generate an ε-cover for *all* (n,k)-PMDs, but instead we take samples and select only distributions from our cover which are consistent with the data. Similar to before, we will apply Theorem 7 to do hypothesis selection but instead of applying it to the complete cover resulting from Theorem 2, we will apply it to a much smaller set of hypothesis that we obtain after making several "guesses" for the parameters of our distribution. At least one set of these parameters will be sufficiently accurate to obtain an ε total variation distance guarantee and we will be able to determine a good candidate using Theorem 7.

The first step of our learning algorithm is to guess the block-diagonal structure of the Gaussian component of our distribution by guessing the partition of the coordinates and choosing an arbitrary pivot within each block. This requires at most k^k guesses. Note that any choice of pivot in the partition is acceptable (as shown in Lemma 4 above).

[6]We note that this tournament additionally requires a "PDF comparator," which we describe for our setting in the appendix of the full version.

1) Guess the block structure/partition of the coordinates.
2) Estimate (using a single sample) the number of CRVs in each block.
3) For each Gaussian in the block structure, use $\mathrm{poly}(k)/\varepsilon^2$ samples to find its mean vector and covariance matrix, as follows:
   a) With $\mathrm{poly}(k)/\varepsilon^2$ samples, estimate the mean vector and covariance matrix of the PMD.
   b) Convert these estimates to the mean and covariance of the Gaussian by searching over a spectral cover of positive semidefinite matrices.
4) Guess the sparse component by enumerating over elements in either of the two covers.
5) Run a tournament on the set of guessed distributions to identify one which is ε-close.

Figure 1. Steps of the learning algorithm

The next step is to guess the sum of the means for the Gaussian component within each block. We need this to know how to fill in the pivot coordinate once we sampled the rest of the coordinates in the block. This will be the number of CRVs which result in this block of the Gaussian component, and thus an integer between 0 and n. Since the total variation distance between the sampled distribution and the distribution from the cover is at most ε, with probability at least $1-\varepsilon$, the sample has non-zero probability to be generated by the distribution from the cover. In this case, the sum of the sample's values within each block will be equal to the sum of the means from the Gaussian component, plus the contribution from the sparse (tk^2, k)-PMD component. Therefore, for each block, we can guess the sum of the means via the following procedure: Take a single sample $X \in \mathbb{R}^k$, and for each block $\mathcal{B}$, guess the sum of the means to be $\sum_{i\in\mathcal{B}} X_i - \ell$, for all $\ell \in \{0, 1, \ldots, tk^2\}$. Since there are at most k blocks, this requires $(tk^2+1)^k$ guesses.

Next, we estimate the mean and covariance of the Gaussian component for each block. We need to estimate them accurately enough in order to learn each block of the discretized Gaussians to within $O(\varepsilon/k)$ in total variation distance. A useful tool for showing this is the following proposition:

Proposition 1. *Let $\mu, \mu' \in \mathbb{R}^k$ and $\Sigma, \Sigma' \in \mathbb{R}^{k\times k}$, such that for all $y \in \mathbb{R}^k$*

$$|y^T(\mu'-\mu)| \le \varepsilon\sqrt{y^T\Sigma y} \quad \textit{and} \quad |y^T(\Sigma'-\Sigma)y| \le \varepsilon y^T\Sigma y.$$

Then

$$d_{\mathrm{TV}}(\mathcal{N}(\mu,\Sigma), \mathcal{N}(\mu',\Sigma')) \le 2\varepsilon k.$$

Proposition 1 implies that, in order to achieve the required bound in total variation distance, it suffices to get an estimate that approximately matches the mean and variance of the Gaussian component in every direction. In the appendix of the full version, we prove Lemma 8 which shows that using $\mathrm{poly}(k)/\varepsilon^2$ samples from the PMD, we can get an estimate of the mean and covariance matrix that achieves this guarantee in every direction. However, this estimate is with respect to the PMD we are sampling from and *not* with respect to the Gaussian component, which is the guarantee we desire.

Lemma 8. *Given sample access to a (n,k)-PMD X with mean μ and covariance matrix Σ (with minimum eigenvalue at least 1), there exists an algorithm which can produce estimates $\hat{\mu}$ and $\hat{\Sigma}$ such that with probability at least $9/10$:*

$$|y^T(\hat{\mu}-\mu)| \le \varepsilon\sqrt{y^T\Sigma y} \quad \textit{and} \quad |y^T(\hat{\Sigma}-\Sigma)y| \le \varepsilon y^T\Sigma y$$

for all vectors y.

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

In order to obtain a guarantee for the Gaussian component, we observe that there are two possible sources of errors in our estimation:

- The first source of error comes from the rounding step. In proving our structural result, the real PMD had to be rounded so that no CRV has any probability that is in the range $(0, c)$, which affected the mean and covariance. In the appendix of the full version, we show that this only affects the mean and variance in each direction up to a small multiplicative factor.
- The second source of error is due to the existence of the sparse component creates an additional additive error in each direction. This error might be very significant in some directions as the variance of the Gaussian component can be very small compared to the number of sparse CRVs.

Understanding that our estimation is off by an additive error and a multiplicative error, we show how to efficiently correct this estimation by searching around it for the underlying covariance matrice of the Gaussian distribution. In particular, we obtain a cover of positive semidefinite matrices that are close to the estimated covariance matrix and which contains a good approximation to the covariance matrix of the underlying Gaussian. This is challenging because the above two sources of error might affect the spectrum of the covariance matrix significantly. However, we are able to tackle this issue by carefully guessing appropriate corrections to the eigenvectors and eigenvalues of the matrix. We prove Lemma 9 which states that this cover has cardinality at most $(k/\varepsilon)^{O(k^2)}$, and thus we can get a very accurate estimate for the underlying Gaussian distribution by guessing different points in the cover.

Lemma 9. *Let A be a symmetric $k \times k$ PSD matrix with minimum eigenvalue 1 and let S be the set of all matrices B such that $|y^T(A-B)y| \leq \varepsilon_1 y^T A y + \varepsilon_2 y^T y$ for all vectors y, where $\varepsilon_1 \in [0, 1/2)$ and $\varepsilon_2 \in [0, \infty)$. Then, there exists an ε-cover S_ε of S that has size $|S_\varepsilon| \leq \left(\frac{k\varepsilon_2}{\varepsilon_1 \varepsilon}\right)^{O(k^2)}$.*

At this point, we have a collection of distributions such that at least one is close to the Gaussian component. We do the same for the sparse PMD component by simply enumerating over all the elements in the cover. By reading the corresponding term from the statement of Theorem 2, this requires $\min\{2^{\mathrm{poly}(k/\varepsilon)}, 2^{O(k^{5k}\cdot \log^{k+2}(1/\varepsilon))}\}$ guesses.

In conclusion, using $\mathrm{poly}(k)/\varepsilon^2$ samples, we have generated a set $\mathcal{S}$ of size

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

which contains a distribution which is ε-close to the true distribution with constant probability. In order to choose a "good" distribution from this set, we apply the hypothesis selection algorithm of Theorem 7 to obtain a distribution which is $O(\varepsilon)$-close to the unknown distribution with constant probability, which concludes the proof of Theorem 3. More details about the learning steps and complete proofs can be found in the full version.

VI. Learning k-SIIRVs

We demonstrate the expressive power of PMDs by demonstrating their applicability to learning (n, k)-SIIRVs. In particular, we leverage our cover results to give a $\tilde{O}_k(1/\varepsilon^2)$ sample algorithm for this problem.

The proof uses the structural result of [16], which says that any (n, k)-SIIRV is close to either a low variance distribution with limited support, or a high variance distribution which enjoys certain Gaussian structural properties.

Lemma 10 (Corollary 4.8 of [16])**.** *Let $S = X_1 + \cdots + X_n$ be a (n, k)-SIIRV for some positive integer k. Let μ and σ^2 be respectively the mean and variance of S. Then for all $\varepsilon > 0$, the distribution of S is $O(\varepsilon)$-close in total variation distance to one of the following:*

1) *a random variable supported on $\frac{k^9}{\varepsilon^4}$ consecutive integers; or*
2) *the sum of two independent random variables $S_1 + cS_2$, where c is some positive integer $1 \leq c \leq k-1$, S_2 is distributed according to $\lfloor \mathcal{N}(\mu, \sigma^2) \rceil$, and S_1 is a c-IRV; in this case, $\sigma^2 = \Omega\left(\frac{k^{18}}{\varepsilon^6}\log^2(1/\varepsilon)\right)$.*

As we did for PMDs, we will use the tournament based approach, in which we generate a set of probability distributions $\mathcal{S}$, containing at least one distribution which is ε-close to S. We then use Theorem 7 to select a distribution which is $O(\varepsilon)$-close to S, using $\tilde{O}(|\mathcal{S}|/\varepsilon^2)$ samples.

To cover the former case, we use the PMD cover of Theorem 2. In this setting, the SIIRV has a variance upper bounded by $\mathrm{poly}(k/\varepsilon)$. By applying a rounding procedure, it can be shown that this can be approximated by an offset $(\mathrm{poly}(k/\varepsilon), k)$-SIIRV. Recalling that any (n, k)-SIIRV can be expressed as the projection of an (n, k)-PMD onto the vector $(0, 1, \ldots, k-1)$ and applying our quasi-polynomial cover result in Theorem 2 covers this case with $2^{O(k^{5k} \cdot \log^{k+2}(1/\varepsilon))}$ candidates.

To cover the latter case, we first perform $k-1$ guesses for the value of $c \in [k-1]$. For each guess, we learn the two distributions S_1 and S_2 separately. To learn S_1, we use the same approach as [16], which uses the empirical distribution obtained after mapping the samples onto $\{0, 1, \ldots, c-1\}$ using their residue mod c. Our method for learning S_2 is novel – we first round the value of each sample down to the next multiple of c, and examine the distribution on this support, which will be close in total variation distance to S_2. We estimate the moments of this distribution using robust statistical tools, as in [16]. The empirical median is used to estimate the mean, and a rescaling of the interquartile range is used to estimate the standard deviation. Thus, we cover this case using only $k-1$ candidates, one for each guess of c.

Full details are provided in the appendix of the full version.

REFERENCES

[1] C. Daskalakis, G. Kamath, and C. Tzamos, "On the structure, covering, and learning of Poisson multinomial distributions," *ArXiV*, vol. abs/1504.08363, Apr. 2015.

[2] G. Valiant and P. Valiant, "Estimating the unseen: An $n/\log n$-sample estimator for entropy and support size, shown optimal via new CLTs," in *Proceedings of the 43rd Annual ACM Symposium on the Theory of Computing*, ser. STOC '11. New York, NY, USA: ACM, 2011, pp. 685–694.

[3] C. Daskalakis and C. H. Papadimitriou, "Discretized multinomial distributions and Nash equilibria in anonymous games," in *Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science*, ser. FOCS '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 25–34.

[4] ——, "Approximate Nash equilibria in anonymous games," *Journal of Economic Theory*, vol. 156, pp. 207–245, 2015.

[5] ——, "On oblivious PTAS's for Nash equilibrium," in *Proceedings of the 41st Annual ACM Symposium on the Theory of Computing*, ser. STOC '09. New York, NY, USA: ACM, 2009, pp. 75–84.

[6] ——, "Sparse covers for sums of indicators," *Probability Theory and Related Fields*, 2014.

[7] C. Daskalakis, I. Diakonikolas, and R. A. Servedio, "Learning Poisson binomial distributions," in *Proceedings of the 44th Annual ACM Symposium on the Theory of Computing*, ser. STOC '12. New York, NY, USA: ACM, 2012, pp. 709–728.

[8] A. W. Van der Vaart, *Asymptotic statistics*. Cambridge University Press, 2000, vol. 3.

[9] V. Bentkus, "A Lyapunov-type bound in Rd," *Theory of Probability & Its Applications*, vol. 49, no. 2, pp. 311–323, 2005.

[10] A. D. Barbour, "Stein's method and Poisson process convergence," *Journal of Applied Probability*, vol. 25, pp. 175–184, 1988.

[11] P. Deheuvels and D. Pfeifer, "Poisson approximations of multinomial distributions and point processes," *Journal of multivariate analysis*, vol. 25, no. 1, pp. 65–89, 1988.

[12] W.-L. Loh, "Stein's method and multinomial approximation," *The Annals of Applied Probability*, vol. 2, no. 3, pp. 536–554, 08 1992.

[13] C. Daskalakis and G. Kamath, "Faster and sample near-optimal algorithms for proper learning mixtures of Gaussians," in *Proceedings of the 27th Annual Conference on Learning Theory*, ser. COLT '14, 2014, pp. 1183–1213.

[14] J. Acharya, A. Jafarpour, A. Orlitsky, and A. T. Suresh, "Sorting with adversarial comparators and application to density estimation," in *Proceedings of the 2014 IEEE International Symposium on Information Theory*, ser. ISIT '14. Washington, DC, USA: IEEE Computer Society, 2014, pp. 1682–1686.

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

[16] C. Daskalakis, I. Diakonikolas, R. O'Donnell, R. A. Servedio, and L. Y. Tan, "Learning sums of independent integer random variables," in *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science*, ser. FOCS '13. Washington, DC, USA: IEEE Computer Society, 2013, pp. 217–226.

[17] B. Roos, "Multinomial and Krawtchouk approximations to the generalized multinomial distribution," *Theory of Probability & Its Applications*, vol. 46, no. 1, pp. 103–117, 2002.

[18] J. Acharya and C. Daskalakis, "Testing Poisson binomial distributions," in *Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA '15. Philadelphia, PA, USA: SIAM, 2015, pp. 1829–1840.

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