Properly Learning Poisson Binomial Distributions in Almost Polynomial Time

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

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

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

November 12, 2015

Abstract

We give an algorithm for properly learning Poisson binomial distributions. A Poisson binomial distribution (PBD) of order $n \in \mathbb{Z}_+$ is the discrete probability distribution of the sum of n mutually independent Bernoulli random variables. Given $\widetilde{O}(1/\epsilon^2)$ samples from an unknown PBD $\mathbf{P}$, our algorithm runs in time $(1/\epsilon)^{O(\log\log(1/\epsilon))}$, and outputs a hypothesis PBD that is ϵ-close to $\mathbf{P}$ in total variation distance. The sample complexity of our algorithm is known to be nearly-optimal, up to logarithmic factors, as established in previous work [DDS12]. However, the previously best known running time for properly learning PBDs [DDS12, DKS15b] was $(1/\epsilon)^{O(\log(1/\epsilon))}$, and was essentially obtained by enumeration over an appropriate ϵ-cover. We remark that the running time of this cover-based approach cannot be improved, as any ϵ-cover for the space of PBDs has size $(1/\epsilon)^{\Omega(\log(1/\epsilon))}$ [DKS15b].

As one of our main contributions, we provide a novel structural characterization of PBDs, showing that any PBD $\mathbf{P}$ is ϵ-close to another PBD $\mathbf{Q}$ with $O(\log(1/\epsilon))$ distinct parameters. More precisely, we prove that, for all $\epsilon > 0$, there exists an explicit collection $\mathcal{M}$ of $(1/\epsilon)^{O(\log\log(1/\epsilon))}$ vectors of multiplicities, such that for any PBD $\mathbf{P}$ there exists a PBD $\mathbf{Q}$ with $O(\log(1/\epsilon))$ distinct parameters whose multiplicities are given by some element of $\mathcal{M}$, such that $\mathbf{Q}$ is ϵ-close to $\mathbf{P}$. Our proof combines tools from Fourier analysis and algebraic geometry.

Our approach to the proper learning problem is as follows: Starting with an accurate non-proper hypothesis, we fit a PBD to this hypothesis. More specifically, we essentially start with the hypothesis computed by the computationally efficient non-proper learning algorithm in our recent work [DKS15b]. Our aforementioned structural characterization allows us to reduce the corresponding fitting problem to a collection of $(1/\epsilon)^{O(\log\log(1/\epsilon))}$ systems of low-degree polynomial inequalities. We show that each such system can be solved in time $(1/\epsilon)^{O(\log\log(1/\epsilon))}$, which yields the overall running time of our algorithm.

*Supported by EPSRC grant EP/L021749/1 and a Marie Curie Career Integration grant.
†Some of this work was performed while visiting the University of Edinburgh.
‡Supported by EPSRC grant EP/L021749/1.

1 Introduction

The Poisson binomial distribution (PBD) is the discrete probability distribution of a sum of mutually independent Bernoulli random variables. PBDs comprise one of the most fundamental nonparametric families of discrete distributions. They have been extensively studied in probability and statistics [Poi37, Che52, Hoe63, DP09b], and are ubiquitous in various applications (see, e.g., [CL97] and references therein). Recent years have witnessed a flurry of research activity on PBDs and generalizations from several perspectives of theoretical computer science, including learning [DDS12, DDO$^+$13, DKS15b, DKT15, DKS15a], pseudorandomness and derandomization [GMRZ11, BDS12, De15, GKM15], property testing [AD15, CDGR15], and computational game theory [DP07, DP09a, DP14a, DP14b, GT14].

Despite their seeming simplicity, PBDs have surprisingly rich structure, and basic questions about them can be unexpectedly challenging to answer. We cannot do justice to the probability literature studying the following question: Under what conditions can we approximate PBDs by simpler distributions? See Section 1.2 of [DDS15] for a summary. In recent years, a number of works in theoretical computer science [DP07, DP09a, DDS12, DP14a, DKS15b] have studied, and essentially resolved, the following questions: Is there a small set of distributions that approximately cover the set of all PBDs? What is the number of samples required to learn an unknown PBD?

We study the following natural computational question: Given independent samples from an unknown PBD $\mathbf{P}$, can we efficiently find a hypothesis PBD $\mathbf{Q}$ that is close to $\mathbf{P}$, in total variation distance? That is, we are interested in *properly learning* PBDs, a problem that has resisted recent efforts [DDS12, DKS15b] at designing efficient algorithms. In this work, we propose a new approach to this problem that leads to a significantly faster algorithm than was previously known. At a high-level, we establish an interesting connection of this problem to algebraic geometry and polynomial optimization. By building on this connection, we provide a new structural characterization of the space of PBDs, on which our algorithm relies, that we believe is of independent interest. In the following, we motivate and describe our results in detail, and elaborate on our ideas and techniques.

Distribution Learning. We recall the standard definition of learning an unknown probability distribution from samples [KMR$^+$94, DL01]: Given access to independent samples drawn from an unknown distribution $\mathbf{P}$ in a given family $\mathcal{C}$, and an error parameter $\epsilon > 0$, a learning algorithm for $\mathcal{C}$ must output a hypothesis $\mathbf{H}$ such that, with probability at least $9/10$, the total variation distance between $\mathbf{H}$ and $\mathbf{P}$ is at most ϵ. The performance of a learning algorithm is measured by its *sample complexity* (the number of samples drawn from $\mathbf{P}$) and its *computational complexity*.

In *non-proper* learning (density estimation), the goal is to output an approximation to the target distribution without any constraints on its representation. In *proper* learning, we require in addition that the hypothesis $\mathbf{H}$ is a member of the family $\mathcal{C}$. Note that these two notions of learning are essentially equivalent in terms of sample complexity (given any accurate hypothesis, we can do a brute-force search to find its closest distribution in $\mathcal{C}$), but not necessarily equivalent in terms of computational complexity. A typically more demanding notion of learning is that of *parameter estimation*. The goal here is to identify the parameters of the unknown model, e.g., the means of the individual Bernoulli components for the case of PBDs, up to a desired accuracy ϵ.

Discussion. In many learning situations, it is desirable to compute a proper hypothesis, i.e., one that belongs to the underlying distribution family $\mathcal{C}$. A proper hypothesis is typically preferable due to its interpretability. In the context of distribution learning, a practitioner may not want to use a density estimate, unless it is proper. For example, one may want the estimate to have the properties of the underlying family, either because this reflects some physical understanding of the inference problem, or because one might only be using the density estimate as the first stage of

a more involved procedure. While parameter estimation may arguably provide a more desirable guarantee than proper learning in some cases, its sample complexity is typically prohibitively large.

For the class of PBDs, we show (Proposition 14, Appendix A) that parameter estimation requires $2^{\Omega(1/\epsilon)}$ samples, for PBDs with $n = \Omega(1/\epsilon)$ Bernoulli components, where $\epsilon > 0$ is the accuracy parameter. In contrast, the sample complexity of (non-)proper learning is known to be $\widetilde{O}(1/\epsilon^2)$ [DDS12]. Hence, proper learning serves as an attractive middle ground between non-proper learning and parameter estimation. Ideally, one could obtain a proper learner for a given family whose running time matches that of the best non-proper algorithm.

Recent work by the authors [DKS15b] has characterized the computational complexity of non-properly learning PBDs, which was shown to be $\widetilde{O}(1/\epsilon^2)$, i.e., nearly-linear in the sample complexity of the problem. Motivated by this progress, a natural research direction is to obtain a computationally efficient proper learning algorithm, i.e., one that runs in time $\mathrm{poly}(1/\epsilon)$ and outputs a PBD as its hypothesis. Besides practical applications, we feel that this is an interesting algorithmic problem, with intriguing connections to algebraic geometry and polynomial optimization (as we point out in this work). We remark that several natural approaches fall short of yielding a polynomial–time algorithm. More specifically, the proper learning of PBDs can be phrased in a number of ways as a structured non-convex optimization problem, albeit it is unclear whether any such formulation may lead to a polynomial–time algorithm.

This work is part of a broader agenda of systematically investigating the computational complexity of proper distribution learning. We believe that this is a fundamental goal that warrants study for its own sake. The complexity of proper learning has been extensively investigated in the supervised setting of PAC learning Boolean functions [KV94, Fel15], with several algorithmic and computational intractability results obtained in the past couple of decades. In sharp contrast, very little is known about the complexity of proper learning in the unsupervised setting of learning probability distributions.

1.1 Preliminaries.

For $n, m \in \mathbb{Z}_+$ with $m \leq n$, we will denote $[n] \stackrel{\text{def}}{=} \{0, 1, \ldots, n\}$ and $[m, n] \stackrel{\text{def}}{=} \{m, m+1, \ldots, n\}$. For a distribution $\mathbf{P}$ supported on $[m]$, $m \in \mathbb{Z}_+$, we write $\mathbf{P}(i)$ to denote the value $\Pr_{X \sim \mathbf{P}}[X = i]$ of the probability mass function (pmf) at point i. The *total variation distance* between two distributions $\mathbf{P}$ and $\mathbf{Q}$ supported on a finite domain A is $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) \stackrel{\text{def}}{=} \max_{S \subseteq A} |\mathbf{P}(S) - \mathbf{Q}(S)| = (1/2) \cdot \|\mathbf{P} - \mathbf{Q}\|_1$. If X and Y are random variables, their total variation distance $d_{\mathrm{TV}}(X, Y)$ is defined as the total variation distance between their distributions.

Poisson Binomial Distribution. A *Poisson binomial distribution of order* $n \in \mathbb{Z}_+$ or *n-PBD* is the discrete probability distribution of the sum $\sum_{i=1}^n X_i$ of n mutually independent Bernoulli random variables $X_1, \ldots, X_n$. An n-PBD $\mathbf{P}$ can be represented uniquely as the vector of its n parameters $p_1, \ldots, p_n$, i.e., as $(p_i)_{i=1}^n$, where we can assume that $0 \leq p_1 \leq p_2 \leq \ldots \leq p_n \leq 1$. To go from $\mathbf{P}$ to its corresponding vector, we find a collection $X_1, \ldots, X_n$ of mutually independent Bernoullis such that $\sum_{i=1}^n X_i$ is distributed according to $\mathbf{P}$ with $\mathbb{E}[X_1] \leq \ldots \leq \mathbb{E}[X_n]$, and we set $p_i = \mathbb{E}[X_i]$ for all i. An equivalent unique representation of an n-PBD with parameter vector $(p_i)_{i=1}^n$ is via the vector of its *distinct parameters* $p'_1, \ldots, p'_k$, where $1 \leq k \leq n$, and $p'_i \neq p'_j$ for $i \neq j$, together with their corresponding integer multiplicities $m_1, \ldots, m_k$. Note that $m_i \geq 1$, $1 \leq i \leq k$, and $\sum_{i=1}^k m_i = n$. This representation will be crucial for the results and techniques of this paper.

Discrete Fourier Transform. For $x \in \mathbb{R}$ we will denote $e(x) \stackrel{\text{def}}{=} \exp(2\pi i x)$. The *Discrete Fourier Transform (DFT) modulo M* of a function $F : [n] \to \mathbb{C}$ is the function $\widehat{F} : [M-1] \to \mathbb{C}$ defined as $\widehat{F}(\xi) = \sum_{j=0}^n e(-\xi j/M) F(j)$, for integers $\xi \in [M-1]$. The DFT modulo M, $\widehat{\mathbf{P}}$, of a distribution $\mathbf{P}$ is the DFT modulo M of its probability mass function. The *inverse DFT modulo M* onto the range $[m, m+M-1]$ of $\widehat{F} : [M-1] \to \mathbb{C}$, is the function $F : [m, m+M-1] \cap \mathbb{Z} \to \mathbb{C}$ defined by

$F(j) = \frac{1}{M}\sum_{\xi=0}^{M-1} e(\xi j/M)\widehat{F}(\xi)$, for $j \in [m, m+M-1] \cap \mathbb{Z}$. The L_2 norm of the DFT is defined as $\|\widehat{F}\|_2 = \sqrt{\frac{1}{M}\sum_{\xi=0}^{M-1} |\widehat{F}(\xi)|^2}$.

1.2 Our Results and Comparison to Prior Work. We are ready to formally describe the main contributions of this paper. As our main algorithmic result, we obtain a near-sample optimal and almost polynomial-time algorithm for properly learning PBDs:

Theorem 1 (Proper Learning of PBDs)**.** *For all $n \in \mathbb{Z}_+$ and $\epsilon > 0$, there is a proper learning algorithm for n-PBDs with the following performance guarantee: Let $\mathbf{P}$ be an unknown n-PBD. The algorithm uses $\widetilde{O}(1/\epsilon^2)$ samples from $\mathbf{P}$, runs in time $(1/\epsilon)^{O(\log\log(1/\epsilon))}$[1], and outputs (a succinct description of) an n-PBD $\mathbf{Q}$ such that with probability at least $9/10$ it holds that $d_{\mathrm{TV}}(\mathbf{Q}, \mathbf{P}) \leq \epsilon$.*

We now provide a comparison of Theorem 1 to previous work. The problem of learning PBDs was first explicitly considered by Daskalakis *et al.* [DDS12], who gave two main results: (i) a non-proper learning algorithm with sample complexity and running time $\widetilde{O}(1/\epsilon^3)$, and (ii) a proper learning algorithm with sample complexity $\widetilde{O}(1/\epsilon^2)$ and running time $(1/\epsilon)^{\mathrm{polylog}(1/\epsilon)}$. In recent work [DKS15b], the authors of the current paper obtained a near-optimal sample and time algorithm to non-properly learn a more general family of discrete distributions (containing PBDs). For the special case of PBDs, the aforementioned work [DKS15b] yields the following implications: (i) a non-proper learning algorithm with sample and time complexity $\widetilde{O}(1/\epsilon^2)$, and (ii) a proper learning algorithm with sample complexity $\widetilde{O}(1/\epsilon^2)$ and running time $(1/\epsilon)^{\Theta(\log(1/\epsilon))}$. Prior to this paper, this was the fastest algorithm for properly learning PBDs. Hence, Theorem 1 represents a super-polynomial improvement in the running time, while still using a near-optimal sample size.

In addition to obtaining a significantly more efficient algorithm, the proof of Theorem 1 offers a novel approach to the problem of properly learning PBDs. The proper algorithms of [DDS12, DKS15b] exploit the cover structure of the space of PBDs, and (essentially) proceed by running an appropriate tournament procedure over an ϵ-cover (see, e.g., Lemma 10 in [DDS15])[2]. This cover-based approach, when applied to an ϵ-covering set of size N, clearly has runtime $\Omega(N)$, and can be easily implemented in time $O(N^2/\epsilon^2)$. [DDS12] applies the cover-based approach to the ϵ-cover construction of [DP14a], which has size $(1/\epsilon)^{O(\log^2(1/\epsilon))}$, while [DKS15b] proves and uses a new cover construction of size $(1/\epsilon)^{O(\log(1/\epsilon))}$. Observe that if there existed an explicit ϵ-cover of size $\mathrm{poly}(1/\epsilon)$, the aforementioned cover-based approach would immediately yield a $\mathrm{poly}(1/\epsilon)$ time proper learning algorithm. Perhaps surprisingly, it was shown in [DKS15b] that *any* ϵ-cover for n-PBDs with $n = \Omega(\log(1/\epsilon))$ Bernoulli coordinates has size $(1/\epsilon)^{\Omega(\log(1/\epsilon))}$. In conclusion, the cover-based approach for properly learning PBDs inherently leads to runtime of $(1/\epsilon)^{\Omega(\log(1/\epsilon))}$.

In this work, we circumvent the $(1/\epsilon)^{\Omega(\log(1/\epsilon))}$ cover size lower bound by establishing a new structural characterization of the space of PBDs. Very roughly speaking, our structural result allows us to reduce the proper learning problem to the case that the underlying PBD has $O(\log(1/\epsilon))$ *distinct* parameters. Indeed, as a simple corollary of our main structural result (Theorem 4 in Section 2), we obtain the following:

Theorem 2 (A "Few" Distinct Parameters Suffice)**.** *For all $n \in \mathbb{Z}_+$ and $\epsilon > 0$ the following holds: For any n-PBD $\mathbf{P}$, there exists an n-PBD $\mathbf{Q}$ with $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) \leq \epsilon$ such that $\mathbf{Q}$ has $O(\log(1/\epsilon))$ distinct parameters.*

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

[2] Note that any ϵ-cover for the space of n-PBDs has size $\Omega(n)$. However, for the task of properly learning PBDs, by a simple (known) reduction, one can assume without loss of generality that $n = \mathrm{poly}(1/\epsilon)$. Hence, the tournament-based algorithm only needs to consider ϵ-covers over PBDs with $\mathrm{poly}(1/\epsilon)$ Bernoulli components.

We note that in subsequent work [DKS15a] the authors generalize the above theorem to Poisson multinomial distributions.

Remark. We remark that Theorem 2 is quantitatively tight, i.e., $O(\log(1/\epsilon))$ distinct parameters are in general necessary to ϵ-approximate PBDs. This follows directly from the explicit cover lower bound construction of [DKS15b].

We view Theorem 2 as a natural structural result for PBDs. Alas, its statement does not quite suffice for our algorithmic application. While Theorem 2 guarantees that $O(\log(1/\epsilon))$ distinct parameters are enough to consider for an ϵ-approximation, it gives no information on the multiplicities these parameters may have. In particular, the upper bound on the number of different combinations of multiplicities one can derive from it is $(1/\epsilon)^{O(\log(1/\epsilon))}$, which is not strong enough for our purposes. The following stronger structural result (see Theorem 4 and Lemma 5 for detailed statements) is critical for our improved proper algorithm:

Theorem 3 (A "Few" Multiplicities and Distinct Parameters Suffice)**.** *For all $n \in \mathbb{Z}_+$ and $\epsilon > 0$ the following holds: For any $\widetilde{\sigma} > 0$, there exists an explicit collection $\mathcal{M}$ of $(1/\epsilon)^{O(\log\log(1/\epsilon))}$ vectors of multiplicities computable in $\mathrm{poly}(|\mathcal{M}|)$ time, so that for any n-PBD $\mathbf{P}$ with variance $\Theta(\widetilde{\sigma}^2)$ there exists a PBD $\mathbf{Q}$ with $O(\log(1/\epsilon))$ distinct parameters whose multiplicities are given by some element of $\mathcal{M}$, such that $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) \leq \epsilon$.*

Now suppose we would like to properly learn an unknown PBD with $O(\log(1/\epsilon))$ *distinct* parameters and known multiplicities for each parameter. Even for this very restricted subset of PBDs, the construction of [DKS15b] implies a cover lower bound of $(1/\epsilon)^{\Omega(\log(1/\epsilon))}$. To handle such PBDs, we combine ingredients from Fourier analysis and algebraic geometry with careful Taylor series approximations, to construct an appropriate system of low-degree polynomial inequalities whose solution approximately recovers the unknown distinct parameters.

In the following subsection, we provide a detailed intuitive explanation of our techniques.

1.3 Techniques. The starting point of this work lies in the non-proper learning algorithm from our recent work [DKS15b]. Roughly speaking, our new proper algorithm can be viewed as a two-step process: We first compute an accurate non-proper hypothesis $\mathbf{H}$ using the algorithm in [DKS15b], and we then post-process $\mathbf{H}$ to find a PBD $\mathbf{Q}$ that is close to $\mathbf{H}$. We note that the non-proper hypothesis $\mathbf{H}$ output by [DKS15b] is represented succinctly via its Discrete Fourier Transform; this property is crucial for the computational complexity of our proper algorithm. (We note that the description of our proper algorithm and its analysis, presented in Section 3, are entirely self-contained. The above description is for the sake of the intuition.)

We now proceed to explain the connection in detail. The crucial fact, established in [DKS15b] for a more general setting, is that the Fourier transform of a PBD has small effective support (and in particular the effective support of the Fourier transform has size roughly inverse to the effective support of the PBD itself). Hence, in order to learn an unknown PBD $\mathbf{P}$, it suffices to find another PBD, $\mathbf{Q}$, with similar mean and standard deviation to $\mathbf{P}$, so that the Fourier transform of $\mathbf{Q}$ approximates the Fourier transform of $\mathbf{P}$ on this small region. (The non-proper algorithm of [DKS15b] for PBDs essentially outputs the empirical DFT of $\mathbf{P}$ over its effective support.)

Note that the Fourier transform of a PBD is the product of the Fourier transforms of its individual component variables. By Taylor expanding the logarithm of the Fourier transform, we can write the log Fourier transform of a PBD as a Taylor series whose coefficients are related to the moments of the parameters of $\mathbf{P}$ (see Equation (2)). We show that for our purposes it suffices to find a PBD $\mathbf{Q}$ so that the first $O(\log(1/\epsilon))$ moments of its parameters approximate the corresponding moments for $\mathbf{P}$. Unfortunately, we do not actually know the moments for $\mathbf{P}$, but since we can easily

approximate the Fourier transform of $\mathbf{P}$ from samples, we can derive conditions that are sufficient for the moments of $\mathbf{Q}$ to satisfy. This step essentially gives us a system of polynomial inequalities in the moments of the parameters of $\mathbf{Q}$ that we need to satisfy.

A standard way to solve such a polynomial system is by appealing to Renegar's algorithm [Ren92b, Ren92a], which allows us to solve a system of degree-d polynomial inequalities in k real variables in time roughly d^k. In our case, the degree d will be at most poly-logarithmic in $1/\epsilon$, but the number of variables k corresponds to the number of parameters of $\mathbf{Q}$, which is $k = \mathrm{poly}(1/\epsilon)$. Hence, this approach is insufficient to obtain a faster proper algorithm.

To circumvent this obstacle, we show that it actually suffices to consider only PBDs with $O(\log(1/\epsilon))$ many distinct parameters (Theorem 2). To prove this statement, we use a recent result from algebraic geometry due to Riener [Rie11] (Theorem 6), that can be used to relate the number of distinct parameters of a solution of a polynomial system to the degree of the polynomials involved. Note that the problem of matching $O(\log(1/\epsilon))$ moments can be expressed as a system of polynomial equations, where each polynomial has degree $O(\log(1/\epsilon))$. We can thus find a PBD $\mathbf{Q}$, which has the same first $O(\log(1/\epsilon))$ moments as $\mathbf{P}$, with $O(\log(1/\epsilon))$ distinct parameters such that $d_{\mathrm{TV}}(\mathbf{Q}, \mathbf{P}) \leq \epsilon$. For PBDs with $O(\log(1/\epsilon))$ distinct parameters and *known* multiplicities for these parameters, we can reduce the runtime of solving the polynomial system to $O(\log(1/\epsilon))^{O(\log(1/\epsilon))} = (1/\epsilon)^{O(\log\log(1/\epsilon))}$.

Unfortunately, the above structural result is not strong enough, as in order to set up an appropriate system of polynomial inequalities for the parameters of $\mathbf{Q}$, we must first guess the multiplicities to which the distinct parameters appear. A simple counting argument shows that there are roughly $k^{\log(1/\epsilon)}$ ways to choose these multiplicities. To overcome this second obstacle, we need the following refinement of our structural result on distinct parameters: We divide the parameters of $\mathbf{Q}$ into categories based on how close they are to 0 or 1. We show that there is a tradeoff between the number of parameters in a given category and the number of distinct parameters in that category (see Theorem 4). With this more refined result in hand, we show that there are only $(1/\epsilon)^{O(\log\log(1/\epsilon))}$ many possible collections of multiplicities that need to be considered (see Lemma 5]).

Given this stronger structural characterization, our proper learning algorithm is fairly simple. We enumerate over the set of possible collections of multiplicities as described above. For each such collection, we set up a system of polynomial equations in the distinct parameters of $\mathbf{Q}$, so that solutions to the system will correspond to PBDs whose distinct parameters have the specified multiplicities which are also ϵ-close to $\mathbf{P}$. For each system, we attempt to solve it using Renegar's algorithm. Since there exists at least one PBD $\mathbf{Q}$ close to $\mathbf{P}$ with such a set of multiplicities, we are guaranteed to find a solution, which in turn must describe a PBD close to $\mathbf{P}$.

One technical issue that arises in the above program occurs when $\mathrm{Var}[\mathbf{P}] \ll \log(1/\epsilon)$. In this case, the effective support of the Fourier transform of $\mathbf{P}$ cannot be restricted to a small subset. This causes problems with the convergence of our Taylor expansion of the log Fourier transform for parameters near $1/2$. However, then only $O(\log(1/\epsilon))$ parameters are not close to 0 and 1, and we can deal with such parameters separately.

1.4 Related Work. Distribution learning is a classical problem in statistics with a rich history and extensive literature (see e.g., [BBBB72, DG85, Sil86, Sco92, DL01]). During the past couple of decades, a body of work in theoretical computer science has been studying these questions from a computational complexity perspective; see e.g., [KMR+94, FM99, AK01, CGG02, VW02, FOS05, BS10, KMV10, MV10, DDS12, DDO+13, CDSS14a, CDSS14b, ADLS15].

We remark that the majority of the literature has focused either on non-proper learning (density estimation) or on parameter estimation. Regarding proper learning, a number of recent works in the statistics community have given proper learners for structured distribution families, by using a

maximum likelihood approach. See e.g., [DR09, GW09, Wal09, DW13, CS13, KS14, BD14] for the case of continuous log-concave densities. Alas, the computational complexity of these approaches has not been analyzed. Two recent works [ADK15, CDGR15] yield computationally efficient proper learners for discrete log-concave distributions, by using an appropriate convex formulation. Proper learning has also been recently studied in the context of mixture models [FOS05, DK14, SOAJ14, LS15]. Here, the underlying optimization problems are non-convex, and efficient algorithms are known only when the number of mixture components is small.

1.5 Organization. In Section 2, we prove our main structural result, and in Section 3, we describe our algorithm and prove its correctness. In Section 4, we conclude with some directions for future research.

2 Main Structural Result

In this section, we prove our main structural results thereby establishing Theorems 2 and 3. Our proofs rely on an analysis of the Fourier transform of PBDs combined with recent results from algebraic geometry on the solution structure of systems of symmetric polynomial equations. We show the following:

Theorem 4. *Given any n-PBD $\mathbf{P}$ with $\mathrm{Var}[\mathbf{P}] = \mathrm{poly}(1/\epsilon)$, there is an n-PBD $\mathbf{Q}$ with $d_{\mathrm{TV}}(\mathbf{P},\mathbf{Q}) \leq \epsilon$ such that $\mathbb{E}[\mathbf{Q}] = \mathbb{E}[\mathbf{P}]$ and $\mathrm{Var}[\mathbf{P}] - \epsilon^3 \leq \mathrm{Var}[\mathbf{Q}] \leq \mathrm{Var}[\mathbf{P}]$, satisfying the following properties:*

Let $R \stackrel{\text{def}}{=} \min\{1/4, \sqrt{\ln(1/\epsilon)/\mathrm{Var}[\mathbf{P}]}\}$. Let $B_i \stackrel{\text{def}}{=} R^{2^i}$, for the integers $0 \leq i \leq \ell$, where $\ell = O(\log\log(1/\epsilon))$ is selected such that $B_\ell = \mathrm{poly}(\epsilon)$. Consider the partition $\mathcal{I} = \{I_i, J_i\}_{i=0}^{\ell+1}$ of $[0,1]$ into the following set of intervals: $I_0 = [B_0, 1/2]$, $I_{i+1} = [B_{i+1}, B_i)$, $0 \leq i \leq \ell-1$, $I_{\ell+1} = (0, B_\ell)$; and $J_0 = (1/2, 1-B_0]$, $J_{i+1} = (1-B_i, 1-B_{i+1}]$, $0 \leq i \leq \ell-1$, $J_{\ell+1} = (1-B_\ell, 1]$. Then we have the following:

(i) For each $0 \leq i \leq \ell$, each of the intervals I_i and J_i contains at most $O(\log(1/\epsilon)/\log(1/B_i))$ distinct parameters of $\mathbf{Q}$.

(ii) $\mathbf{Q}$ has at most one parameter in each of the intervals $I_{\ell+1}$ and $J_{\ell+1} \setminus \{1\}$.

(iii) The number of parameters of $\mathbf{Q}$ equal to 1 is within an additive $\mathrm{poly}(1/\epsilon)$ of $\mathbb{E}[\mathbf{P}]$.

(iv) For each $0 \leq i \leq \ell$, each of the intervals I_i and J_i contains at most $2\mathrm{Var}[\mathbf{P}]/B_i$ parameters of $\mathbf{Q}$.

Theorem 4 implies that one needs to only consider $(1/\epsilon)^{O(\log\log(1/\epsilon))}$ different combinations of multiplicities:

Lemma 5. *For every $\mathbf{P}$ as in Theorem 4, there exists an explicit set $\mathcal{M}$ of multisets of triples $(m_i, a_i, b_i)_{1\leq i\leq k}$ so that*

(i) For each element of $\mathcal{M}$ and each i, $[a_i, b_i]$ is either one of the intervals I_i or J_i as in Theorem 4 or $[0,0]$ or $[1,1]$.

(ii) For each element of $\mathcal{M}$, $k = O(\log(1/\epsilon))$.

(iii) There exist an element of $\mathcal{M}$ and a PBD $\mathbf{Q}$ as in the statement of Theorem 4 with $d_{\mathrm{TV}}(\mathbf{P},\mathbf{Q}) < \epsilon^2$ so that $\mathbf{Q}$ has a parameter of multiplicity m_i between a_i and b_i for each $1 \leq i \leq k$ and no other parameters.

(iv) $\mathcal{M}$ has size $\left(\frac{1}{\epsilon}\right)^{O(\log\log(1/\epsilon))}$ and can be enumerated in $\mathrm{poly}(|\mathcal{M}|)$ time.

This is proved in Appendix B.1 by a simple counting argument. We multiply the number of multiplicities for each interval, which is at most the maximum number of parameters to the power of the maximum number of distinct parameters in that interval, giving $(1/\epsilon)^{O(\log\log(1/\epsilon))}$ possibilities. We now proceed to prove Theorem 4. We will require the following result from algebraic geometry:

Theorem 6 (Part of Theorem 4.2 from [Rie11])**.** *Given $m+1$ symmetric polynomials in n variables $F_j(x)$, $0 \le j \le m$, $x \in \mathbb{R}^n$, let $K = \{x \in \mathbb{R}^n \mid F_j(x) \ge 0, \text{for all } 1 \le j \le m\}$. Let $k = \max\{2, \lceil \deg(F_0)/2 \rceil, \deg(F_1), \deg(F_2), \ldots, \deg(F_m)\}$. Then, the minimum value of F_0 on K is achieved by a point with at most k distinct co-ordinates.*

As an immediate corollary, we obtain the following:

Corollary 7. *If a set of multivariate polynomial equations $F_i(x) = 0$, $x \in \mathbb{R}^n$, $1 \le i \le m$, with the degree of each $F_i(x)$ being at most d has a solution $x \in [a,b]^n$, then it has a solution $y \in [a,b]^n$ with at most d distinct values of the variables in y.*

The following lemma will be crucial:

Lemma 8. *Let $\epsilon > 0$. Let $\mathbf{P}$ and $\mathbf{Q}$ be n-PBDs with $\mathbf{P}$ having parameters $p_1, \ldots, p_k \le 1/2$ and $p'_1, \ldots, p'_m > 1/2$ and $\mathbf{Q}$ having parameters $q_1, \ldots, q_k \le 1/2$ and $q'_1, \ldots, q'_m > 1/2$. Suppose furthermore that $\mathrm{Var}[\mathbf{P}] = \mathrm{Var}[\mathbf{Q}] = V$ and let $C > 0$ be a sufficiently large constant. Suppose furthermore that for $A = \min(3, C\sqrt{\log(1/\epsilon)/V})$ and for all positive integers ℓ it holds*

$$A^\ell \left(\left| \sum_{i=1}^k p_i^\ell - \sum_{i=1}^k q_i^\ell \right| + \left| \sum_{i=1}^m (1-p'_i)^\ell - \sum_{i=1}^m (1-q'_i)^\ell \right| \right) < \epsilon / C \log(1/\epsilon). \tag{1}$$

Then $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) < \epsilon$.

In practice, we shall only need to deal with a finite number of ℓ's, since we will be considering the case where all p_i, q_i or $1-p'_i, 1-q'_i$ that do not appear in pairs will have size less than $1/(2A)$. Therefore, the size of the sum in question will be sufficiently small automatically for ℓ larger than $\Omega(\log((k+m)/\epsilon))$.

The basic idea of the proof will be to show that the Fourier transforms of $\mathbf{P}$ and $\mathbf{Q}$ are close to each other. In particular, we will need to make use of the following intermediate lemma:

Lemma 9. *Let $\mathbf{P}$, $\mathbf{Q}$ be PBDs with $|\mathbb{E}[\mathbf{P}] - \mathbb{E}[\mathbf{Q}]| = O(\mathrm{Var}[\mathbf{P}]^{1/2})$ and $\mathrm{Var}[\mathbf{P}] + 1 = \Theta(\mathrm{Var}[\mathbf{Q}] + 1)$. Let $M = \Theta(\log(1/\epsilon) + \sqrt{\mathrm{Var}[\mathbf{P}]\log(1/\epsilon)})$ and $\ell = \Theta(\log(1/\epsilon))$ be positive integers with the implied constants sufficiently large. If $\sum_{-\ell \le \xi \le \ell} |\widehat{\mathbf{P}}(\xi) - \widehat{\mathbf{Q}}(\xi)|^2 \le \epsilon^2/16$, then $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) \le \epsilon$.*

The proof of this lemma, which is given in Appendix B.2, is similar to (part of) the correctness analysis of the non-proper learning algorithm in [DKS15b].

Proof of Lemma 8. We proceed by means of Lemma 9. We need only show that for all ξ with $|\xi| = O(\log(1/\epsilon))$ that $|\widehat{\mathbf{P}}(\xi) - \widehat{\mathbf{Q}}(\xi)| \ll \epsilon/\sqrt{\log(1/\epsilon)}$. For this we note that

$$\begin{aligned}
\widehat{\mathbf{P}}(\xi) &= \prod_{i=1}^k ((1-p_i) + p_i e(\xi/M)) \prod_{i=1}^m ((1-p'_i) + p'_i e(\xi/M)) \\
&= e(m\xi/M) \prod_{i=1}^k (1 + p_i(e(\xi/M) - 1)) \prod_{i=1}^m (1 + (1-p'_i)(e(-\xi/M) - 1)).
\end{aligned}$$

Taking a logarithm and Taylor expanding, we find that

$$\log(\widehat{\mathbf{P}}(\xi)) = 2\pi i m\xi/M + \sum_{\ell=1}^{\infty} \frac{(-1)^{1+\ell}}{\ell} \left((e(\xi/M)-1)^{\ell} \sum_{i=1}^{k} p_i^{\ell} + (e(-\xi/M)-1)^{\ell} \sum_{i=1}^{m} (1-p_i')^{\ell} \right). \quad (2)$$

A similar formula holds for $\log(\widehat{\mathbf{Q}}(\xi))$. Therefore, we have that

$$|\widehat{\mathbf{P}}(\xi) - \widehat{\mathbf{Q}}(\xi)| \leq |\log(\widehat{\mathbf{P}}(\xi)) - \log(\widehat{\mathbf{Q}}(\xi))|,$$

which is at most

$$\begin{aligned}
&\sum_{\ell=1}^{\infty} |e(\xi/M)-1|^{\ell} \left(\left| \sum_{i=1}^{k} p_i^{\ell} - \sum_{i=1}^{k} q_i^{\ell} \right| + \left| \sum_{i=1}^{m} (1-p_i')^{\ell} - \sum_{i=1}^{m} (1-q_i')^{\ell} \right| \right) \\
\leq &\sum_{\ell=1}^{\infty} (2A/3)^{\ell} \left(\left| \sum_{i=1}^{k} p_i^{\ell} - \sum_{i=1}^{k} q_i^{\ell} \right| + \left| \sum_{i=1}^{m} (1-p_i')^{\ell} - \sum_{i=1}^{m} (1-q_i')^{\ell} \right| \right) \\
\leq &\sum_{\ell=1}^{\infty} (2/3)^{\ell} \epsilon / C \log(1/\epsilon) \\
\ll &\epsilon / C \log(1/\epsilon).
\end{aligned}$$

An application of Lemma 9 completes the proof. □

Proof of Theorem 4. The basic idea of the proof is as follows. First, we will show that it is possible to modify $\mathbf{P}$ in order to satisfy (ii) without changing its mean, increasing its variance (or decreasing it by too much), or changing it substantially in total variation distance. Next, for each of the other intervals I_i or J_i, we will show that it is possible to modify the parameters that $\mathbf{P}$ has in this interval to have the appropriate number of distinct parameters, without substantially changing the distribution in variation distance. Once this holds for each i, conditions (iii) and (iv) will follow automatically.

To begin with, we modify $\mathbf{P}$ to have at most one parameter in $I_{\ell+1}$ in the following way. We repeat the following procedure. So long as $\mathbf{P}$ has two parameters, p and p' in $I_{\ell+1}$, we replace those parameters by 0 and $p+p'$. We note that this operation has the following properties:

- The expectation of $\mathbf{P}$ remains unchanged.
- The total variation distance between the old and new distributions is $O(pp')$, as is the change in variances between the distributions.
- The variance of $\mathbf{P}$ is decreased.
- The number of parameters in $I_{\ell+1}$ is decreased by 1.

All of these properties are straightforward to verify by considering the effect of just the sum of the two changed variables. By repeating this procedure, we eventually obtain a new PBD, $\mathbf{P}'$ with the same mean as $\mathbf{P}$, smaller variance, and at most one parameter in $I_{\ell+1}$. We also claim that $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{P}')$ is small. To show this, we note that in each replacement, the error in variation distance is at most a constant times the increase in the sum of the squares of the parameters of the relevant PBD. Therefore, letting p_i be the parameters of $\mathbf{P}$ and letting p_i' be the parameters of $\mathbf{P}'$, we have that $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{P}') = O(\sum (p_i')^2 - p_i^2)$. We note that this difference is entirely due to the parameters that were modified by this procedure. Therefore, it is at most $(2B_{\ell})^2$ times the number

of non-zero parameters created. Note that all but one of these parameters contributes at least $B_\ell/2$ to the variance of $\mathbf{P}'$. Therefore, this number is at most $2\mathrm{Var}[\mathbf{P}]/B_\ell + 1$. Hence, the total variation distance between $\mathbf{P}$ and $\mathbf{P}'$ is at most $O(B_\ell^2)(\mathrm{Var}[\mathbf{P}]/B_\ell + 1) \leq \epsilon^3$. Similarly, the variance of our distribution is decreased by at most this much. This implies that it suffices to consider $\mathbf{P}$ that have at most one parameter in $I_{\ell+1}$. Symmetrically, we can also remove all but one of the parameters in $J_{\ell+1}$, and thus it suffices to consider $\mathbf{P}$ that satisfy condition (ii).

Next, we show that for any such $\mathbf{P}$ that it is possible to modify the parameters that $\mathbf{P}$ has in I_i or J_i, for any i, so that we leave the expectation and variance unchanged, introduce at most ϵ^2 error in variation distance, and leave only $O(\log(1/\epsilon)/\log(1/B_i))$ distinct parameters in this range. The basic idea of this is as follows. By Lemma 8, it suffices to keep $\sum p_i^\ell$ or $\sum(1-p_i)^\ell$ constant for parameters p_i in that range for some range of values of ℓ. On the other hand, Theorem 6 implies that this can be done while producing only a small number of distinct parameters.

Without loss of generality assume that we are dealing with the interval I_i. Note that if $i = 0$ and $\mathrm{Var}[\mathbf{P}] \ll \log(1/\epsilon)$, then $B_0 = 1/4$, and there can be at most $O(\log(1/\epsilon))$ parameters in I_0 to begin with. Hence, in this case there is nothing to show. Thus, assume that either $i \geq 0$ or that $\mathrm{Var}[\mathbf{P}] \gg \log(1/\epsilon)$ with a sufficiently large constant. Let $p_1, \ldots, p_m$ be the parameters of p that lie in I_i. Consider replacing them with parameters $q_1, \ldots, q_m$ also in I_i to obtain $\mathbf{Q}$. By Lemma 8, we have that $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) < \epsilon^2$ so long as the first two moments of $\mathbf{P}$ and $\mathbf{Q}$ agree and

$$\min(3, C\sqrt{\log(1/\epsilon)/\mathrm{Var}[\mathbf{P}]})^\ell \left| \sum_{j=1}^m p_j^\ell - \sum_{j=1}^m q_j^\ell \right| < \epsilon^3 \,, \tag{3}$$

for all ℓ (the terms in the sum in Equation (1) coming from the parameters not being changed cancel out). Note that $\min(3, C\sqrt{\log(1/\epsilon)/\mathrm{Var}[\mathbf{P}]}) \max(p_j, q_j) \leq B_i^{O(1)}$. This is because by assumption either $i > 0$ and $\max(p_j, q_j) \leq \sqrt{B_i} \leq 1/4$ or $i = 0$ and $B_i = \sqrt{\log(1/\epsilon)/\mathrm{Var}[\mathbf{P}]} \ll 1$. Furthermore, note that $\mathrm{Var}[\mathbf{P}] \geq mB_{i+1}$. Therefore, $m \leq \mathrm{poly}(1/\epsilon)$. Combining the above, we find that Equation (3) is automatically satisfied for any $q_j \in I_i$ so long as ℓ is larger than a sufficiently large multiple of $\log(1/\epsilon)/\log(1/B_i)$. On the other hand, Theorem 6 implies that there is some choice of $q_j \in I_i$ taking on only $O(\log(1/\epsilon)/\log(1/B_i))$ distinct values, so that $\sum_{j=1}^m q_j^\ell$ is exactly $\sum_{j=1}^m p_j^\ell$ for all ℓ in this range. Thus, replacing the p_j's in this range by these q_j's, we only change the total variation distance by ϵ^2, leave the expectation and variance the same (as we have fixed the first two moments), and have changed our distribution in variation distance by at most ϵ^2.

Repeating the above procedure for each interval I_i or J_i in turn, we replace $\mathbf{P}$ by a new PBD, $\mathbf{Q}$ with the same expectation and smaller variance and $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) < \epsilon$, so that $\mathbf{Q}$ satisfies conditions (i) and (ii). We claim that (iii) and (iv) are necessarily satisfied. Condition (iii) follows from noting that the number of parameters not 0 or 1 is at most $2 + 2\mathrm{Var}[\mathbf{P}]/B_\ell$, which is $\mathrm{poly}(1/\epsilon)$. Therefore, the expectation of $\mathbf{Q}$ is the number of parameters equal to $1 + \mathrm{poly}(1/\epsilon)$. Condition (iv) follows upon noting that $\mathrm{Var}[\mathbf{Q}] \leq \mathrm{Var}[\mathbf{P}]$ is at least the number of parameters in I_i or J_i times $B_i/2$ (as each contributes at least $B_i/2$ to the variance). This completes the proof of Theorem 4. □

3 Proper Learning Algorithm

Given samples from an unknown PBD $\mathbf{P}$, and given a collection of intervals and multiplicities as described in Theorem 4, we wish to find a PBD $\mathbf{Q}$ with those multiplicities that approximates $\mathbf{P}$. By Lemma 8, it is sufficient to find such a $\mathbf{Q}$ so that $\widehat{\mathbf{Q}}(\xi)$ is close to $\widehat{\mathbf{P}}(\xi)$ for all small ξ. On the other hand, by Equation (2) the logarithm of the Taylor series of $\widehat{\mathbf{Q}}$ is given by an appropriate expansion in the parameters. Note that if $|\xi|$ is small, due to the $(e(\xi/M) - 1)^\ell$ term, the terms of our sum with $\ell \gg \log(1/\epsilon)$ will automatically be small. By truncating the Taylor series, we

get a polynomial in the parameters that gives us an approximation to $\log(\widehat{\mathbf{Q}}(\xi))$. By applying a truncated Taylor series for the exponential function, we obtain a polynomial in the parameters of $\mathbf{Q}$ which approximates its Fourier coefficients. This procedure yields a system of polynomial equations whose solution gives the parameters of a PBD that approximates $\mathbf{P}$. Our main technique will be to solve this system of equations to obtain our output distribution using the following result:

Theorem 10 ([Ren92b, Ren92a])**.** *Let $P_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, m$, be m polynomials over the reals each of maximum degree at most d. Let $K = \{x \in \mathbb{R}^n : P_i(x) \geq 0, \text{ for all } i = 1, \ldots, m\}$. If the coefficients of the P_i's are rational numbers with bit complexity at most L, there is an algorithm that runs in time $\mathrm{poly}(L, (d \cdot m)^n)$ and decides if K is empty or not. Further, if K is non-empty, the algorithm runs in time $\mathrm{poly}(L, (d \cdot m)^n, \log(1/\delta))$ and outputs a point in K up to an L_2 error δ.*

In order to set up the necessary system of polynomial equations, we have the following theorem:

Theorem 11. *Consider a PBD $\mathbf{P}$ with $\mathrm{Var}[\mathbf{P}] < \mathrm{poly}(1/\epsilon)$, and real numbers $\tilde{\sigma} \in [\sqrt{\mathrm{Var}[\mathbf{P}]}/2, 2\sqrt{\mathrm{Var}[\mathbf{P}]}+1]$ and $\tilde{\mu}$ with $|\mathbb{E}[\mathbf{P}]-\tilde{\mu}| \leq \tilde{\sigma}$. Let M be as above and let ℓ be a sufficiently large multiple of $\log(1/\epsilon)$. Let h_ξ be complex numbers for each integer ξ with $|\xi| \leq \ell$ so that $\sum_{|\xi| \leq \ell} |h_\xi - \widehat{\mathbf{P}}(\xi)|^2 < \epsilon^2/16$.*

Consider another PBD with parameters q_i of multiplicity m_i contained in intervals $[a_i, b_i]$ as described in Theorem 4. There exists an explicit system $\mathcal{P}$ of $O(\log(1/\epsilon))$ real polynomial inequalities each of degree $O(\log(1/\epsilon))$ in the q_i so that:

(i) If there exists such a PBD of the form of $\mathbf{Q}$ with $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) < \epsilon/\ell$, $\mathbb{E}[\mathbf{Q}] = \mathbb{E}[\mathbf{P}]$, and $\mathrm{Var}[\mathbf{P}] \geq \mathrm{Var}[\mathbf{Q}] \geq \mathrm{Var}[\mathbf{P}]/2$, then its parameters q_i yield a solution to $\mathcal{P}$.

(ii) Any solution $\{q_i\}$ to $\mathcal{P}$ corresponds to a PBD $\mathbf{Q}$ with $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) < \epsilon/2$.

Furthermore, such a system can be found with rational coefficients of encoding size $O(\log^2(1/\epsilon))$ bits.

Proof. For technical reasons, we begin by considering the case that $\mathrm{Var}[\mathbf{P}]$ is larger than a sufficiently large multiple of $\log(1/\epsilon)$, as we will need to make use of slightly different techniques in the other case. In this case, we construct our system $\mathcal{P}$ in the following manner. We begin by putting appropriate constraints on the mean and variance of $\mathbf{Q}$ and requiring that the q_i's lie in appropriate intervals.

$$\widetilde{\mu} - 2\widetilde{\sigma} \leq \sum_{j=1}^{k} m_j p_j \leq \widetilde{\mu} + 2\widetilde{\sigma} \tag{4}$$

$$\widetilde{\sigma}^2/2 - 1 \leq \sum_{j=1}^{k} m_j p_j (1 - p_j) \leq 2\widetilde{\sigma}^2 \tag{5}$$

$$a_j \leq p_j \leq b_j, \tag{6}$$

Next, we need a low-degree polynomial to express the condition that Fourier coefficients of $\mathbf{Q}$ are approximately correct. To do this, we let S denote the set of indices i so that $[a_i, b_i] \subset [0, 1/2]$ and T the set so that $[a_i, b_i] \subset [1/2, 1]$ and let $m = \sum_{i \in T} m_i$. We let

$$g_\xi = 2\pi i \xi m/M + \sum_{k=1}^{\ell} \frac{(-1)^{k+1}}{k} \left((e(\xi/M) - 1)^k \sum_{i \in S} m_i q_i^k + (e(-\xi/M) - 1)^k \sum_{i \in T} m_i (1 - q_i)^k \right) \tag{7}$$

be an approximation to the logarithm of $\widehat{\mathbf{Q}}(\xi)$. We next define $\exp'$ to be a Taylor approximation to the exponential function

$$\exp'(z) := \sum_{k=0}^{\ell} \frac{z^k}{k!}.$$

By Taylor's theorem, we have that

$$|\exp'(z) - \exp(z)| \leq \frac{z^{\ell+1} \exp(z)}{(\ell+1)!},$$

and in particular that if $|z| < \ell/3$ that $|\exp'(z) - \exp(z)| = \exp(-\Omega(\ell))$.

We would ideally like to use $\exp'(g_\xi)$ as an approximation to $\widehat{\mathbf{Q}}(\xi)$. Unfortunately, g_ξ may have a large imaginary part. To overcome this issue, we let o_ξ, defined as the nearest integer to $\tilde{\mu}\xi/M$, be an approximation to the imaginary part, and we set

$$q_\xi = \exp'(g_\xi + 2\pi i o_\xi)\ . \tag{8}$$

We complete our system $\mathcal{P}$ with the final inequality:

$$\sum_{-\ell \leq \xi \leq \ell} |q_\xi - h_\xi|^2 \leq \epsilon^2/8. \tag{9}$$

In order for our analysis to work, we will need for q_ξ to approximate $\widehat{\mathbf{Q}}(\xi)$. Thus, we make the following claim:

Claim 12. *If Equations (4), (5), (6), (7), and (8) hold, then $|q_\xi - \widehat{\mathbf{Q}}(\xi)| < \epsilon^3$ for all $|\xi| \leq \ell$.*

This is proved in Appendix C by showing that g_ξ is close to a branch of the logarithm of $\widehat{\mathbf{Q}}(\xi)$ and that $|g_\xi + 2\pi i o_\xi| \leq O(\log(1/\epsilon))$, so $\exp'$ is a good enough approximation to the exponential.

Hence, our system $\mathcal{P}$ is defined as follows:

Variables:

- q_i for each distinct parameter i of $\mathbf{Q}$.
- g_ξ for each $|\xi| \leq \ell$.
- q_ξ for each $|\xi| \leq \ell$.

Equations: Equations (4), (5), (6), (7), (8), and (9).

To prove (i), we note that such a $\mathbf{Q}$ will satisfy (4) and (5), because of the bounds on its mean and variance, and will satisfy Equation (6) by assumption. Therefore, by Claim 12, q_ξ is approximately $\widehat{\mathbf{Q}}(\xi)$ for all ξ. On the other hand, since $d_{TV}(\mathbf{P}, \mathbf{Q}) < \epsilon/\ell$, we have that $|\widehat{\mathbf{P}}(\xi) - \widehat{\mathbf{Q}}(\xi)| < \epsilon/\ell$ for all ξ. Therefore, setting g_ξ and q_ξ as specified, Equation (9) follows. To prove (ii), we note that a $\mathbf{Q}$ whose parameters satisfy $\mathbf{P}$ will by Claim 12 satisfy the hypotheses of Lemma 9. Therefore, $d_{TV}(\mathbf{P}, \mathbf{Q}) \leq \epsilon/2$.

As we have defined it so far, the system $\mathcal{P}$ does not have rational coefficients. Equation (7) makes use of $e(\pm\xi/M)$ and π, as does Equation (8). To fix this issue, we note that if we approximate the appropriate powers of $(\pm 1 \pm e(\pm\xi/M))$ and $q\pi i$ each to accuracy $(\epsilon/\sum_{i \in S} m_i))^{10}$, this produces an error of size at most ϵ^4 in the value g_ξ, and therefore an error of size at most ϵ^3 for q_ξ, and this leaves the above argument unchanged.

Also, as defined above, the system $\mathcal{P}$ has complex constants and variables and many of the equations equate complex quantities. The system can be expressed as a set of real inequalities

by doubling the number of equations and variables to deal with the real and imaginary parts separately. Doing so introduces binomial coefficients into the coefficients, which are no bigger than $2^{O(\log(1/\epsilon))} = \mathrm{poly}(1/\epsilon)$ in magnitude. To express $\exp'$, we need denominators with a factor of $\ell! = \log(1/\epsilon)^{\Theta(\log(1/\epsilon))}$. All other constants can be expressed as rationals with numerator and denominator bounded by $\mathrm{poly}(1/\epsilon)$. So, the encoding size of any of the rationals that appear in the system is $\log(\log(1/\epsilon)^{O(\log(1/\epsilon))}) = O(\log^2(1/\epsilon))$.

One slightly more difficult problem is that the proof of Claim 12 depended upon the fact that $\mathrm{Var}[\mathbf{P}] \gg \log(1/\epsilon)$. If this is not the case, we will in fact need to slightly modify our system of equations. In particular, we redefine S to be the set of indices, i, so that $b_i \leq 1/4$ (rather than $\leq 1/2$), and let T be the set of indices i so that $a_i \geq 3/4$. Finally, we let R be the set of indices for which $[a_i, b_i] \subset [1/4, 3/4]$. We note that, since each $i \in R$ contributes at least $m_i/8$ to $\sum_i m_i q_i(1-q_i)$, if Equations (6) and (5) both hold, we must have $|R| = O(\mathrm{Var}[\mathbf{P}]) = O(\log(1/\epsilon))$.

We then slightly modify Equation (8), replacing it by

$$q_\xi = \exp'(g_\xi) \prod_{i \in R} (q_i e(\xi/M) + (1-q_i))^{m_i}. \tag{10}$$

Note that by our bound on $\sum_{i \in R} m_i$, this is of degree $O(\log(1/\epsilon))$.

We now need only prove the analogue of Claim 12 in order for the rest of our analysis to follow.

Claim 13. *If Equations (4), (5), (6), (7), and (10) hold, then $|q_\xi - \widehat{\mathbf{Q}}(\xi)| < \epsilon^3$ for all $|\xi| \leq \ell$.*

We prove this in Appendix C, by proving similar bounds to those needed for Claim 12. This completes the proof of our theorem in the second case. □

Our algorithm for properly learning PBDs is given in pseudocode below:

Algorithm `Proper-Learn-PBD`
Input: sample access to a PBD $\mathbf{P}$ and $\epsilon > 0$.
Output: A hypothesis PBD that is ϵ-close to $\mathbf{P}$ with probability at least 9/10.

Let C be a sufficiently large universal constant.

1. Draw $O(1)$ samples from $\mathbf{P}$ and with confidence probability 19/20 compute: (a) $\widetilde{\sigma}^2$, a factor 2 approximation to $\mathrm{Var}_{X \sim \mathbf{P}}[X] + 1$, and (b) $\widetilde{\mu}$, an approximation to $\mathbb{E}_{X \sim \mathbf{P}}[X]$ to within one standard deviation. Set $M \stackrel{\text{def}}{=} \lceil C(\log(1/\epsilon) + \widetilde{\sigma}\sqrt{\log(1/\epsilon)}) \rceil$. Let $\ell \stackrel{\text{def}}{=} \lceil C^2 \log(1/\epsilon) \rceil$.

2. If $\widetilde{\sigma} > \Omega(1/\epsilon^3)$, then we draw $O(1/\epsilon^2)$ samples and use them to learn a shifted binomial distribution, using algorithms `Learn-Poisson` and `Locate-Binomial` from [DDS15]. Otherwise, we proceed as follows:

3. Draw $N = C^3(1/\epsilon^2)\ln^2(1/\epsilon)$ samples $s_1, \ldots, s_N$ from $\mathbf{P}$. For integers ξ with $|\xi| \leq \ell$, set h_ξ to be the empirical DFT modulo M. Namely, $h_\xi := \frac{1}{N}\sum_{i=1}^N e(-\xi s_i/M)$.

4. Let $\mathcal{M}$ be the set of multisets of multiplicities described in Lemma 5. For each element $m \in \mathcal{M}$, let $\mathcal{P}_m$ be the corresponding system of polynomial equations as described in Theorem 11.

5. For each such system, use the algorithm from Theorem 10 to find a solution to precision $\epsilon/(2k)$, where k is the sum of the multiplicities not corresponding to 0 or 1, if such a solution exists. Once such a solution is found, return the PBD $\mathbf{Q}$ with parameters q_i to multiplicity m_i, where m_i are the terms from m and q_i in the approximate solution to $\mathcal{P}_m$.

Proof of Theorem 1. We first note that the algorithm succeeds in the case that $\mathrm{Var}_{X\sim\mathbf{P}}[X] = \Omega(1/\epsilon^6)$: [DDS15] describes procedures `Learn-Poisson` and `Locate-Binomial` that draw $\mathrm{O}(1/\epsilon^2)$ samples, and return a shifted binomial ϵ-close to a PBD $\mathbf{P}$, provided $\mathbf{P}$ is not close to a PBD in "sparse form" in their terminology. This hods for any PBD with effective support $\Omega(1/\epsilon^3)$, since by definition a PBD in "sparse form" has support of size $O(1/\epsilon^3)$.

It is clear that the sample complexity of our algorithm is $O(\epsilon^{-2}\log^2(1/\epsilon))$. The runtime of the algorithm is dominated by Step 5. We note that by Lemma 5, $|\mathcal{M}| = (1/\epsilon)^{O(\log\log(1/\epsilon))}$. Furthermore, by Theorems 10 and 11, the runtime for solving the system $\mathcal{P}_m$ is $O(\log(1/\epsilon))^{O(\log(1/\epsilon))} = (1/\epsilon)^{O(\log\log(1/\epsilon))}$. Therefore, the total runtime is $(1/\epsilon)^{O(\log\log(1/\epsilon))}$.

It remains to show correctness. We first note that each h_ξ is an average of independent random variables $e(-\xi p_i/M)$, with expectation $\widehat{\mathbf{P}}(\xi)$. Therefore, by standard Chernoff bounds, with high probability we have that $|h_\xi - \widehat{\mathbf{P}}(\xi)| = O(\sqrt{\log(\ell)}/\sqrt{N}) \ll \epsilon/\sqrt{\ell}$ for all ξ, and therefore we have that

$$\sum_{|\xi|\leq\ell} |h_\xi - \widehat{\mathbf{P}}(\xi)|^2 < \epsilon^2/8.$$

Now, by Lemma 5, for some $m \in \mathcal{M}$ there will exist a PBD $\mathbf{Q}$ whose distinct parameters come in multiplicities given by m and lie in the corresponding intervals so that $d_{\mathrm{TV}}(\mathbf{P},\mathbf{Q}) \leq \epsilon^2$. Therefore, by Theorem 11, the system $\mathcal{P}_m$ will have a solution. Therefore, at least one $\mathcal{P}_m$ will have a solution and our algorithm will necessarily return *some* PBD $\mathbf{Q}$.

On the other hand, any $\mathbf{Q}$ returned by our algorithm will correspond to an approximation of some solution of $\mathcal{P}_m$, for some $m \in \mathcal{M}$. By Theorem 11, any solution to any $\mathcal{P}_m$ will give a PBD $\mathbf{Q}$ with $d_{\mathrm{TV}}(\mathbf{P},\mathbf{Q}) \leq \epsilon/2$. Therefore, the actual output of our algorithm is a PBD $\mathbf{Q}'$, whose parameters approximate those of such a $\mathbf{Q}$ to within $\epsilon/(2k)$. On the other hand, from this it is clear that $d_{\mathrm{TV}}(\mathbf{Q},\mathbf{Q}') \leq \epsilon/2$, and therefore, $d_{\mathrm{TV}}(\mathbf{P},\mathbf{Q}') \leq \epsilon$. In conclusion, our algorithm will always return a PBD that is within ϵ total variation distance of $\mathbf{P}$. □

4 Conclusions and Open Problems

In this work, we gave a nearly-sample optimal algorithm for properly learning PBDs that runs in almost polynomial time. We also provided a structural characterization for PBDs that may be of independent interest. The obvious open problem is to obtain a polynomial-time proper learning algorithm. We conjecture that such an algorithm is possible, and our mildly super-polynomial runtime may be viewed as an indication of the plausibility of this conjecture. Currently, we do not know of a $\mathrm{poly}(1/\epsilon)$ time algorithm even for the special case of an n-PBD with $n = O(\log(1/\epsilon))$.

A related open question concerns obtaining faster proper algorithms for learning more general families of discrete distributions that are amenable to similar techniques, e.g., sums of independent integer-valued random variables [DDO+13, DKS15b], and Poisson multinomial distributions [DKT15, DKS15a]. Here, we believe that progress is attainable via a generalization of our techniques.

The recently obtained cover size lower bound for PBDs [DKS15b] is a bottleneck for other non-convex optimization problems as well, e.g., the problem of computing approximate Nash equilibria in anonymous games [DP14b]. The fastest known algorithms for these problems proceed by enumerating over an ϵ-cover. Can we obtain faster algorithms in such settings, by avoiding enumeration over a cover?

References

[AD15] J. Acharya and C. Daskalakis. Testing poisson binomial distributions. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA*, pages 1829–1840, 2015.

[ADK15] J. Acharya, C. Daskalakis, and G. Kamath. Optimal testing for properties of distributions. *CoRR*, abs/1507.05952, 2015.

[ADLS15] J. Acharya, I. Diakonikolas, J. Li, and L. Schmidt. Sample-optimal density estimation in nearly-linear time. *CoRR*, abs/1506.00671, 2015.

[AK01] S. Arora and R. Kannan. Learning mixtures of arbitrary Gaussians. In *Proceedings of the 33rd Symposium on Theory of Computing*, pages 247–257, 2001.

[BBBB72] R.E. Barlow, D.J. Bartholomew, J.M. Bremner, and H.D. Brunk. *Statistical Inference under Order Restrictions*. Wiley, New York, 1972.

[BD14] F. Balabdaoui and C. R. Doss. Inference for a Mixture of Symmetric Distributions under Log-Concavity. Available at http://arxiv.org/abs/1411.4708, 2014.

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

[BS10] M. Belkin and K. Sinha. Polynomial learning of distribution families. In *FOCS*, pages 103–112, 2010.

[CDGR15] C. Canonne, I. Diakonikolas, T. Gouleakis, and R. Rubinfeld. Testing shape restrictions of discrete distributions. *CoRR*, abs/1507.03558, 2015.

[CDSS14a] S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Efficient density estimation via piecewise polynomial approximation. In *STOC*, pages 604–613, 2014.

[CDSS14b] S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Near-optimal density estimation in near-linear time using variable-width histograms. In *NIPS*, pages 1844–1852, 2014.

[CGG02] M. Cryan, L. Goldberg, and P. Goldberg. Evolutionary trees can be learned in polynomial time in the two state general Markov model. *SIAM Journal on Computing*, 31(2):375–397, 2002.

[Che52] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statist.*, 23:493–507, 1952.

[CL97] S.X. Chen and J.S. Liu. Statistical applications of the Poisson-Binomial and Conditional Bernoulli Distributions. *Statistica Sinica*, 7:875–892, 1997.

[CS13] Y. Chen and R. J. Samworth. Smoothed log-concave maximum likelihood estimation with applications. *Statist. Sinica*, 23:1373–1398, 2013.

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

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

[DDS15] C. Daskalakis, I. Diakonikolas, and R. A. Servedio. Learning poisson binomial distributions. *Algorithmica*, 72(1):316–357, 2015.

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

[DG85] L. Devroye and L. Györfi. *Nonparametric Density Estimation: The L_1 View.* John Wiley & Sons, 1985.

[DK14] C. Daskalakis and G. Kamath. Faster and sample near-optimal algorithms for proper learning mixtures of gaussians. In *Proceedings of The 27th Conference on Learning Theory, COLT 2014*, pages 1183–1213, 2014.

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

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

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

[DL01] L. Devroye and G. Lugosi. *Combinatorial methods in density estimation.* Springer Series in Statistics, Springer, 2001.

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

[DP09a] C. Daskalakis and C. Papadimitriou. On Oblivious PTAS's for Nash Equilibrium. In *STOC*, pages 75–84, 2009.

[DP09b] D. Dubhashi and A. Panconesi. *Concentration of measure for the analysis of randomized algorithms.* Cambridge University Press, Cambridge, 2009.

[DP14a] C. Daskalakis and C. Papadimitriou. Sparse covers for sums of indicators. *Probability Theory and Related Fields*, pages 1–27, 2014.

[DP14b] C. Daskalakis and C. H. Papadimitriou. Approximate nash equilibria in anonymous games. *Journal of Economic Theory*, 2014.

[DR09] L. D umbgen and K. Rufibach. Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency. *Bernoulli*, 15(1):40–68, 2009.

[DW13] C. R. Doss and J. A. Wellner. Global Rates of Convergence of the MLEs of Log-concave and s-concave Densities. Available at http://arxiv.org/abs/1306.1438, 2013.

[Fel15] V. Feldman. Hardness of proper learning (1988; pitt, valiant). In *Encyclopedia of Algorithms.* 2015.

[FM99] Y. Freund and Y. Mansour. Estimating a mixture of two product distributions. In *Proceedings of the 12th Annual COLT*, pages 183–192, 1999.

[FOS05] J. Feldman, R. O'Donnell, and R. Servedio. Learning mixtures of product distributions over discrete domains. In *Proc. 46th IEEE FOCS*, pages 501–510, 2005.

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

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

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

[GW09] F. Gao and J. A. Wellner. On the rate of convergence of the maximum likelihood estimator of a k-monotone density. *Science in China Series A: Mathematics*, 52:1525–1538, 2009.

[Hoe63] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[KMR+94] M. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R. Schapire, and L. Sellie. On the learnability of discrete distributions. In *Proc. 26th STOC*, pages 273–282, 1994.

[KMV10] A. T. Kalai, A. Moitra, and G. Valiant. Efficiently learning mixtures of two Gaussians. In *STOC*, pages 553–562, 2010.

[KS14] A. K. H. Kim and R. J. Samworth. Global rates of convergence in log-concave density estimation. Available at http://arxiv.org/abs/1404.2298, 2014.

[KV94] M. Kearns and U. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, Cambridge, MA, 1994.

[LS15] J. Li and L. Schmidt. A nearly optimal and agnostic algorithm for properly learning a mixture of k gaussians, for any constant k. *CoRR*, abs/1506.01367, 2015.

[MV10] A. Moitra and G. Valiant. Settling the polynomial learnability of mixtures of Gaussians. In *FOCS*, pages 93–102, 2010.

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

[Ren92a] J. Renegar. On the computational complexity and geometry of the first-order theory of the reals. *J. Symb. Comput.*, 13(3):255–352, 1992.

[Ren92b] J. Renegar. On the computational complexity of approximating solutions for real algebraic formulae. *SIAM J. Comput.*, 21(6):1008–1025, 1992.

[Rie11] C. Riener. *Symmetries in Semidefinite and Polynomial Optimization*. PhD thesis, Johann Wolfgang Goethe-Universitat, 2011.

[Sco92] D.W. Scott. *Multivariate Density Estimation: Theory, Practice and Visualization*. Wiley, New York, 1992.

[Sil86] B. W. Silverman. *Density Estimation*. Chapman and Hall, London, 1986.

[SOAJ14] A. T. Suresh, A. Orlitsky, J. Acharya, and A. Jafarpour. Near-optimal-sample estimators for spherical gaussian mixtures. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1395–1403, 2014.

[VW02] S. Vempala and G. Wang. A spectral algorithm for learning mixtures of distributions. In *Proceedings of the 43rd Annual Symposium on Foundations of Computer Science*, pages 113–122, 2002.

[Wal09] G. Walther. Inference and modeling with log-concave distributions. *Statistical Science*, 24(3):319–327, 2009.

Appendix

A Sample Complexity Lower Bound for Parameter Estimation

Proposition 14. *Suppose that $n \geq 1/\epsilon$. Any learning algorithm that takes N samples from an n-PBD and returns estimates of these parameters to additive error at most ϵ with probability at least $2/3$ must have $N \geq 2^{\Omega(1/\epsilon)}$.*

Proof. We may assume that $n = \Theta(1/\epsilon)$ (as we could always make the remaining parameters all 0) and demonstrate a pair of PBDs whose parameters differ by $\Omega(\epsilon)$, and yet have variation distance $2^{-\Omega(1/\epsilon)}$. Therefore, if such an algorithm is given one of these two PBDs, it will be unable to distinguish which one it is given, and therefore unable to learn the parameters to ϵ accuracy with at least $2^{\Omega(1/\epsilon)}$ samples.

In order to make this construction work, we take $\mathbf{P}$ to have parameters $p_j := (1+\cos\left(\frac{2\pi j}{n}\right))/8$, and let $\mathbf{Q}$ have parameters $q_j := (1+\cos\left(\frac{2\pi j+\pi}{n}\right))/8$. Suppose that $j = n/4 + O(1)$. We claim that none of the q_i are closer to p_j that $\Omega(1/n)$. This is because for all i we have that $\left(\frac{2\pi i+\pi}{n}\right)$ is at least $\Omega(1/n)$ from $\left(\frac{2\pi j}{n}\right)$ and $\left(\frac{2\pi(n-j)}{n}\right)$.

On the other hand, it is easy to see that the p_j are roots of the polynomial $(T_n(8x-1)-1)$, and q_j are the roots of $(T_n(8x-1)+1)$, where T_n is the n^{th} Chebyshev polynomial. Since these polynomials have the same leading term and identical coefficients other than their constant terms, it follows that the elementary symmetric polynomials in p_j of degree less than n equal the corresponding polynomials in the q_j. From this, by the Newton-Girard formulae, we have that $\sum_{i=1}^n p_i^l = \sum_{i=1}^n q_i^l$ for $1 \leq l \leq n-1$. For any $l \geq n$, we have that $3^l(\sum_{i=1}^n (p_i^l - q_i^l)) \leq n(3/4)^n$, and so by Lemma 8, we have that $d_{\mathrm{TV}}(\mathbf{P},\mathbf{Q}) = 2^{-\Omega(n)}$. This completes our proof. □

B Omitted Proofs from Section 2

B.1 Proof of Lemma 5.

For completeness, we restate the lemma below.

Lemma 5. *For every $\mathbf{P}$ as in Theorem 4, there exists an explicit set $\mathcal{M}$ of multisets of triples $(m_i, a_i, b_i)_{1\leq i\leq k}$ so that*

(i) For each element of $\mathcal{M}$ and each i, $[a_i, b_i]$ is either one of the intervals I_i or J_i as in Theorem 4 or $[0,0]$ or $[1,1]$.

(ii) For each element of $\mathcal{M}$, $k = O(\log(1/\epsilon))$.

(iii) There exist an element of $\mathcal{M}$ and a PBD $\mathbf{Q}$ as in the statement of Theorem 4 with $d_{\mathrm{TV}}(\mathbf{P},\mathbf{Q}) < \epsilon^2$ so that $\mathbf{Q}$ has a parameter of multiplicity m_i between a_i and b_i for each $1 \leq i \leq k$ and no other parameters.

(iv) $\mathcal{M}$ has size $\left(\frac{1}{\epsilon}\right)^{O(\log\log(1/\epsilon))}$ and can be enumerated in $\mathrm{poly}(|\mathcal{M}|)$ time.

Proof of Lemma 5 assuming Theorem 4. Replacing ϵ in Theorem 4 by ϵ^2, we take $\mathcal{M}$ to be the set of all possible ways to have at most $O(\log(1/\epsilon)/\log(1/B_i))$ terms with $[a_i, b_i]$ equal to I_i or J_i and having the sum of the corresponding m's at most $4\mathrm{Var}[\mathbf{P}]/B_i$, having one term with $a_i = b_i = 1$ and $m_i = \mathbb{E}[\mathbf{P}] + \mathrm{poly}(1/\epsilon)$, and one term with $a_i = b_i = 0$ and m_i such that the sum of all of the m_i's equals n.

For this choice of $\mathcal{M}$, (i) is automatically satisfied, and (iii) follows immediately from Theorem 4. To see (ii), we note that the total number of term in an element of $\mathcal{M}$ is at most

$$O(1) + \sum_{i=1}^{\ell} O(\log(1/\epsilon)/\log(1/B_i)) = O(1) + \sum_{i=1}^{\ell} O(\log(1/\epsilon)2^{-i}) = O(\log(1/\epsilon)).$$

To see (iv), we need a slightly more complicated counting argument. To enumerate $\mathcal{M}$, we merely need to enumerate each integer of size $\mathbb{E}[\mathbf{P}]+\mathrm{poly}(1/\epsilon)$ for the number of 1's, and enumerate for each $0 \le i \le \ell$ all possible multi-sets of m_i of size at most $O(\log(1/\epsilon)/\log(1/B_i))$ with sum at most $2\mathrm{Var}[\mathbf{P}]/B_i$ to correspond to the terms with $[a_i, b_i] = I_i$, and again for the terms with $[a_i, b_i] = J_i$. This is clearly enumerable in $\mathrm{poly}(|\mathcal{M}|)$ time, and the total number of possible multi-sets is at most

$$\mathrm{poly}(1/\epsilon)\prod_{i=0}^{\ell}(2\mathrm{Var}[\mathbf{P}]/B_i)^{O(\log(1/\epsilon)/\log(1/B_i))}.$$

Therefore, we have that

$$\begin{aligned}
|\mathcal{M}| &\le \mathrm{poly}(1/\epsilon)\prod_{i=0}^{\ell}(2\mathrm{Var}[\mathbf{P}]/B_i)^{O(\log(1/\epsilon)/\log(1/B_i))} \\
&= \mathrm{poly}(1/\epsilon)\prod_{i=0}^{\ell} B_i^{-O(\log_{1/B_i}(1/\epsilon))}\prod_{i=0}^{\ell} O(\mathrm{Var}[\mathbf{P}])^{O(\log(1/\epsilon)/(2^i\log(1/B_0)))} \\
&= \mathrm{poly}(1/\epsilon)\prod_{i=0}^{\ell} \mathrm{poly}(1/\epsilon) O(\mathrm{Var}[\mathbf{P}])^{O(\log(1/\epsilon)/\log(1/B_0))} \\
&= (1/\epsilon)^{O(\log\log(1/\epsilon))} O(\mathrm{Var}[\mathbf{P}])^{O(\log(1/\epsilon)/\log(1/B_0))} \\
&= (1/\epsilon)^{O(\log\log(1/\epsilon))}.
\end{aligned}$$

The last equality above requires some explanation. If $\mathrm{Var}[\mathbf{P}] < \log^2(1/\epsilon)$, then

$$O(\mathrm{Var}[\mathbf{P}])^{O(\log(1/\epsilon)/\log(1/B_0))} \le \log(1/\epsilon)^{O(\log(1/\epsilon))} = (1/\epsilon)^{O(\log\log(1/\epsilon))}.$$

Otherwise, if $\mathrm{Var}[\mathbf{P}] \ge \log^2(1/\epsilon)$, $\log(1/B_0) \gg \log(\mathrm{Var}[\mathbf{P}])$, and thus

$$O(\mathrm{Var}[\mathbf{P}])^{O(\log(1/\epsilon)/\log(1/B_0))} \le \mathrm{poly}(1/\epsilon).$$

This completes our proof. □

B.2 Proof of Lemma 9. For completeness, we restate the lemma below.

Lemma 9. *Let $\mathbf{P}$, $\mathbf{Q}$ be PBDs with $|\mathbb{E}[\mathbf{P}] - \mathbb{E}[\mathbf{Q}]| = O(\mathrm{Var}[\mathbf{P}]^{1/2})$ and $\mathrm{Var}[\mathbf{P}] = \Theta(\mathrm{Var}[\mathbf{Q}])$. Let $M = \Theta(\log(1/\epsilon) + \sqrt{\mathrm{Var}[\mathbf{P}]\log(1/\epsilon)})$ and $\ell = \Theta(\log(1/\epsilon))$ be positive integers with the implied constants sufficiently large. If $\sum_{-\ell\le\xi\le\ell} |\widehat{\mathbf{P}}(\xi) - \widehat{\mathbf{Q}}(\xi)|^2 \le \epsilon^2/16$, then $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) \le \epsilon$.*

Proof. The proof of this lemma is similar to the analysis of correctness of the non-proper learning algorithm in [DKS15b].

The basic idea of the proof is as follows. By Bernstein's inequality, $\mathbf{P}$ and $\mathbf{Q}$ both have nearly all of their probability mass supported in the same interval of length M. This means that is suffices to show that the distributions $\mathbf{P} \pmod M$ and $\mathbf{Q} \pmod M$ are close. By Plancherel's Theorem, it suffices to show that the DFTs $\widehat{\mathbf{P}}$ and $\widehat{\mathbf{Q}}$ are close. However, it follows by Lemma 6 of [DKS15b] that these DFTs are small in magnitude outside of $-\ell \leq \xi \leq \ell$.

Let m be the nearest integer to the expected value of $\mathbf{P}$. By Bernstein's inequality, it follows that both $\mathbf{P}$ and $\mathbf{Q}$ have $1-\epsilon/10$ of their probability mass in the interval $I = [m - M/2, m + M/2)$. We note that any given probability distribution X over $\mathbb{Z}/M\mathbb{Z}$ has a unique lift to a distribution taking values in I. We claim that $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) \leq \epsilon/5 + d_{\mathrm{TV}}(\mathbf{P} \pmod M, \mathbf{Q} \pmod M)$. This is because after throwing away the at most $\epsilon/5$ probability mass where $\mathbf{P}$ or $\mathbf{Q}$ take values outside of I, there is a one-to-one mapping between values in I taken by $\mathbf{P}$ or $\mathbf{Q}$ and the values taken by $\mathbf{P} \pmod M$ or $\mathbf{Q} \pmod M$. Thus, it suffices to show that $d_{\mathrm{TV}}(\mathbf{P} \pmod M, \mathbf{Q} \pmod M) \leq 4\epsilon/5$.

By Cauchy-Schwarz, we have that

$$d_{\mathrm{TV}}(\mathbf{P} \pmod M, \mathbf{Q} \pmod M) \leq \sqrt{M} \|\mathbf{P} \pmod M - \mathbf{Q} \pmod M\|_2 .$$

By Plancherel's Theorem, the RHS above is

$$\sqrt{\sum_{\xi \pmod M} |\widehat{\mathbf{P}}(\xi) - \widehat{\mathbf{Q}}(\xi)|^2}. \tag{11}$$

By assumption, the sum of the above over all $|\xi| \leq \ell$ is at most $\epsilon^2/16$. However, applying Lemma 6 of [DKS15b] with $k = 2$, we find that for any $|\xi| \leq M/2$ that each of $|\widehat{\mathbf{P}}(\xi)|, |\widehat{\mathbf{Q}}(\xi)|$ is $\exp(-\Omega(\xi^2 \mathrm{Var}[\mathbf{P}]/M^2)) = \exp(-\Omega(\xi^2/\log(1/\epsilon)))$. Therefore, the sum above over ξ not within ℓ of some multiple of M is at most

$$\begin{aligned}\sum_{n>\ell} \exp(-\Omega(n^2/\log(1/\epsilon))) &\leq \sum_{n>\ell} \exp(-\Omega((\ell^2 + (n-\ell)\ell)/\log(1/\epsilon))) \\ &\leq \sum_{n>\ell} \exp(-(n-\ell)) \exp(-\Omega(\ell^2/\log(1/\epsilon))) \leq \epsilon^2/16\end{aligned}$$

assuming that the constant defining ℓ is large enough. Therefore, the sum in (11) is at most $\epsilon^2/8$. This completes the proof. □

C Omitted Proofs from Section 3

In this section, we prove Claims 12 and 13 which we restate here.

Claim 12. *If Equations (4), (5), (6), (7), and (8) hold, then $|q_\xi - \widehat{\mathbf{Q}}(\xi)| < \epsilon^3$ for all $|\xi| \leq \ell$.*

Proof. First we begin by showing that g_ξ approximates $\log(\widehat{\mathbf{Q}}(\xi))$. By Equation (2), we would have equality if the sum over k were extended to all positive integers. Therefore, the error between g_ξ and $\log(\widehat{\mathbf{Q}}(\xi))$ is equal to the sum over all $k > \ell$. Since $\tilde{\sigma} \gg \log(1/\epsilon)$, we have that $M \gg \ell$ and therefore, $|1 - e(\xi/m)|$ and $|e(-\xi/M) - 1|$ are both less than $1/2$. Therefore, the term for a particular value of k is at most $2^{-k}\left(\sum_{i \in S} m_i q_i + \sum_{i \in T} m_i (1 - q_i)\right) \gg 2^{-k}\tilde{\sigma}$. Summing over $k > \ell$, we find that

$$|g_\xi - \log(\widehat{\mathbf{Q}}(\xi))| < \epsilon^4.$$

We have left to prove that $\exp'(g_\xi - 2\pi i o_\xi)$ is approximately $\exp(g_\xi) = \exp(g_\xi - 2\pi i o_\xi)$. By the above, it suffices to prove that $|g_\xi - 2\pi i o_\xi| < \ell/3$. We note that

$$\begin{aligned}
g_\xi &= 2\pi i\xi m/M + \sum_{k=1}^{\ell} \frac{(-1)^{k+1}}{k} \left((e(\xi/M)-1)^k \sum_{i\in S} m_i q_i^k + (e(-\xi/M)-1)^k \sum_{i\in T} m_i (1-q_i)^k \right) \\
&= 2\pi i\xi m/M + (e(\xi/M)-1)\sum_{i\in S} m_i q_i + (e(-\xi/M)-1)\sum_{i\in T} m_i(1-q_i) + \\
&+ O\left(\sum_{k=2}^{\ell} |\xi|^2/M^2 2^{-k} \left(\sum_i m_i q_i (1-q_i) \right) \right) \\
&= 2\pi i\xi m/M + 2\pi i\xi/M \left(\sum_{i\in S} m_i q_i - \sum_{i\in T} m_i(1-q_i) \right) + O(|\xi|^2/M^2\tilde{\sigma}^2) \\
&= 2\pi i\xi/M \sum_i m_i q_i + O(|\xi|^2/M^2\tilde{\sigma}^2) \\
&= 2\pi i\xi/M\tilde{\mu} + O(|\xi|/M\tilde{\sigma}) + O(|\xi|^2/M^2\tilde{\sigma}^2) \\
&= 2\pi i o_\xi + O(\log(1/\epsilon)).
\end{aligned}$$

This completes the proof. □

Claim 13. *If Equations (4), (5), (6), (7), and (10) hold, then $|q_\xi - \widehat{\mathbf{Q}}(\xi)| < \epsilon^3$ for all $|\xi| \leq \ell$.*

Proof. Let $\mathbf{Q}'$ be the PBD obtained from $\mathbf{Q}$ upon removing all parameters corresponding to elements of R. We note that

$$\widehat{\mathbf{Q}}(\xi) = \widehat{\mathbf{Q}'}(\xi) \prod_{i\in R} (q_i e(\xi/M) + (1-q_i))^{m_i}.$$

Therefore, it suffices to prove our claim when $R = \emptyset$.

Once again it suffices to show that g_ξ is within ϵ^4 of $\log(\widehat{\mathbf{Q}}(\xi))$ and that $|g_\xi| < \ell/3$. For the former claim, we again note that, by Equation (2), we would have equality if the sum over k were extended to all integers, and therefore only need to bound the sum over all $k > \ell$. On the other hand, we note that $q_i \leq 1/4$ for $i \in S$ and $(1-q_i) \leq 1/4$ for $i \in T$. Therefore, the k^{th} term in the sum would have absolute value at most

$$O\left(2^{-k} \left(\sum_{i\in S} m_i q_i + \sum_{i\in T} m_i(1-q_i) \right) \right) = O(2^{-k}\tilde{\sigma}_i).$$

Summing over $k > \ell$, proves the appropriate bound on the error. Furthermore, summing this bound over $1 \leq k \leq \ell$ proves that $|g_\xi| < \ell/3$, as required. Combining these results with the bounds on the Taylor error for $\exp'$ completes the proof. □