Learning Sums of Independent Random Variables with Sparse Collective Support

Anindya De*
University of Pennsylvania
anindyad@cis.upenn.edu

Philip M. Long
Google
plong@google.com

Rocco A. Servedio†
Columbia University
rocco@cs.columbia.edu

November 9, 2020

Abstract

We study the learnability of sums of independent integer random variables given a bound on the size of the union of their supports. For $\mathcal{A} \subset \mathbb{Z}_+$, a *sum of independent random variables with collective support* $\mathcal{A}$ (called an $\mathcal{A}$-sum in this paper) is a distribution $\mathbf{S} = \mathbf{X}_1 + \cdots + \mathbf{X}_N$ where the $\mathbf{X}_i$'s are mutually independent (but not necessarily identically distributed) integer random variables with $\cup_i \mathrm{supp}(\mathbf{X}_i) \subseteq \mathcal{A}$.

We give two main algorithmic results for learning such distributions:

1. For the case $|\mathcal{A}| = 3$, we give an algorithm for learning $\mathcal{A}$-sums to accuracy ε that uses $\mathrm{poly}(1/\varepsilon)$ samples and runs in time $\mathrm{poly}(1/\varepsilon)$, independent of N and of the elements of $\mathcal{A}$.
2. For an arbitrary constant $k \geq 4$, if $\mathcal{A} = \{a_1, ..., a_k\}$ with $0 \leq a_1 < ... < a_k$, we give an algorithm that uses $\mathrm{poly}(1/\varepsilon) \cdot \log\log a_k$ samples (independent of N) and runs in time $\mathrm{poly}(1/\varepsilon, \log a_k)$.

We prove an essentially matching lower bound: if $|\mathcal{A}| = 4$, then any algorithm must use

$$\Omega(\log\log a_4)$$

samples even for learning to constant accuracy. We also give similar-in-spirit (but quantitatively very different) algorithmic results, and essentially matching lower bounds, for the case in which $\mathcal{A}$ is not known to the learner.

Our algorithms and lower bounds together settle the question of how the sample complexity of learning sums of independent integer random variables scales with the elements in the union of their supports, both in the known-support and unknown-support settings. Finally, all our algorithms easily extend to the "semi-agnostic" learning model, in which training data is generated from a distribution that is only $c\varepsilon$-close to some $\mathcal{A}$-sum for a constant $c > 0$.

Keywords– Central limit theorem; sample complexity; sums of independent random variables; equidistribution; semi-agnostic learning

*Corresponding author. Supported by a start-up grant from Northwestern University and NSF CCF-1814706.

†Supported by NSF grants CCF-1420349 and CCF-1563155.

1 Introduction

The theory of sums of independent random variables forms a rich strand of research in probability. Indeed, many of the best-known and most influential results in probability theory are about such sums; prominent examples include the weak and strong law of large numbers, a host of central limit theorems, and (the starting point of) the theory of large deviations. Within computer science, the well-known "Chernoff-Hoeffding" bounds — i.e., large deviation bounds for sums of independent random variables — are a ubiquitous tool of great utility in many contexts. Not surprisingly, there are several books [GK54, Pet75, Pet95, PS00, Kle14, BB85] devoted to the study of sums of independent random variables.

Given the central importance of sums of independent random variables both within probability theory and for a host of applications, it is surprising that even very basic questions about *learning* these distributions were not rigorously investigated until very recently. The problem of learning probability distributions from independent samples has attracted a great deal of attention in theoretical computer science for almost two decades (see [KMR+94, Das99, AK01, VW02, KMV10, MV10, BS10] and a host of more recent papers), but most of this work has focused on other types of distributions such as mixtures of Gaussians, hidden Markov models, etc. While sums of independent random variables may seem to be a very simple type of distribution, as we shall see below the problem of learning such distributions turns out to be surprisingly tricky.

Before proceeding further, let us recall the standard PAC-style model for learning distributions that was essentially introduced in [KMR+94] and that we use in this work. In this model the unknown target distribution $\mathbf{X}$ is assumed to belong to some class $\mathcal{C}$ of distributions. A learning algorithm has access to i.i.d. samples from $\mathbf{X}$, and must produce an efficiently samplable description of a hypothesis distribution $\mathbf{H}$ such that with probability at least (say) $9/10$, the total variation distance $d_{\mathrm{TV}}(\mathbf{X}, \mathbf{H})$ between $\mathbf{X}$ and $\mathbf{H}$ is at most ε. (In the language of statistics, this task is usually referred to as *density estimation*, as opposed to *parametric estimation* in which one seeks to approximately identify the parameters of the unknown distribution $\mathbf{X}$ when $\mathcal{C}$ is a parametric class like Gaussians or mixtures of Gaussians.) In fact, all our positive results hold for the more challenging *semi-agnostic* variant of this model, which is as above except that the assumption that $\mathbf{X} \in \mathcal{C}$ is weakened to the requirement $d_{\mathrm{TV}}(\mathbf{X}, \mathbf{X}^*) \leq c\varepsilon$ for some constant c and some $\mathbf{X}^* \in \mathcal{C}$.

Learning sums of independent random variables: Formulating the problem. To motivate our choice of learning problem it is useful to recall some relevant context. Recent years have witnessed many research works in theoretical computer science studying the learnability and testability of discrete probability distributions (see e.g. [DDS12a, DDS12b, DDO+13, RSS14, ADK15, AD15, Can15, LRSS15, CDGR16, Can16, DKS16a, DKS16c, DDKT16]); our paper belongs to this line of research. A folklore result in this area is that a simple brute-force algorithm can learn *any* distribution over an M-element set using $\Theta(M/\varepsilon^2)$ samples, and that this is best possible if the distribution may be arbitrary. Thus it is of particular interest to learn classes of distributions over M elements for which a sample complexity dramatically better than this "trivial bound" (ideally scaling as $\log M$, or even independent of M altogether) can be achieved.

This perspective on learning, along with a simple result which we now describe, strongly motivates considering sums of random variables which have small *collective support*. Consider the following very simple learning problem: Let $\{\mathbf{X}_i\}_{i=1}^n$ be independent random variables where $\mathbf{X}_i$ is promised to be supported on the two-element set $\{0, i\}$ but $\mathbf{Pr}[\mathbf{X}_i = i]$ is unknown: what is the sample complexity of learning $\mathbf{X} = \mathbf{X}_1 + \cdots + \mathbf{X}_N$? Even though each random variable $\mathbf{X}_i$ is "as simple as a non-trivial random variable can be" — supported on just two values, one of which is zero — a straightforward lower bound given in [DDS12b] shows that any algorithm for learning $\mathbf{X}$ even to constant accuracy must use $\Omega(N)$ sam-

ples, which is not much better than the trivial brute-force algorithm based on support size. (We note that this learning problem is the problem of learning a weighted sum of independent Bernoulli random variables in which the i-th Bernoulli random variable has weight equal to i, and hence the collective support of $\mathbf{X}_1, \ldots, \mathbf{X}_N$ is $|\{0, 1, \ldots, N\}| = N + 1$.)

Given this lower bound, it is natural to restrict the learning problem by requiring the random variables $\mathbf{X}_1, \ldots, \mathbf{X}_N$ to have *small* collective support, i.e. the union $\mathrm{supp}(\mathbf{X}_1) \cup \cdots \cup \mathrm{supp}(\mathbf{X}_N)$ of their support sets is small. Inspired by this, Daskalakis *et al.* [DDS12b] studied the simplest non-trivial version of this learning problem, in which each $\mathbf{X}_i$ is a Bernoulli random variable (so the union of all supports is simply $\{0, 1\}$; note, though, that the $\mathbf{X}_i$'s may have distinct and arbitrary biases). The main result of [DDS12b] is that this class (known as *Poisson Binomial Distributions*) can be learned to error ε with $\mathsf{poly}(1/\varepsilon)$ samples — so, perhaps unexpectedly, the complexity of learning this class is completely independent of N, the number of summands. The proof in [DDS12b] relies on several sophisticated results from probability theory, including a discrete central limit theorem from [CGS11] (proved using Stein's method) and a "moment matching" result due to Roos [Roo00]. (A subsequent sharpening of the [DDS12b] result in [DKS16b], giving improved time and sample complexities, also employed sophisticated tools, namely Fourier analysis and algebraic geometry.)

Motivated by this first success, there has been a surge of recent work which studies the learnability of sums of richer classes of random variables. In particular, Daskalakis et al. [DDO+13] considered a generalization of [DDS12b] in which each $\mathbf{X}_i$ is supported on the set $\{0, 1, \ldots, k-1\}$, and Daskalakis et al. [DKT15] considered a vector-valued generalization in which each $\mathbf{X}_i$ is supported on the set $\{e_1, \ldots, e_k\}$, the standard basis unit vectors in $\mathbb{R}^k$. We will elaborate on these results shortly, but here we first highlight a crucial feature shared by all these results; in all of [DDS12b, DDO+13, DKT15] the collective support of the individual summands forms a "nice and simple" set (either $\{0, 1\}$, $\{0, 1, \ldots, k-1\}$, or $\{e_1, \ldots, e_k\}$). Indeed, the technical workhorses of all these results are various central limit theorems which crucially exploit the simple structure of these collective support sets. (These central limit theorems have since found applications in other settings, such as the design of algorithms for approximating equilibrium [DDKT16, DKT15, DKS16c, CDS17] as well as stochastic optimization [De18].)

In this paper we go beyond the setting in which the collective support of $\mathbf{X}_1, \ldots, \mathbf{X}_N$ is a "nice" set, by studying the learnability of $\mathbf{X}_1 + \cdots + \mathbf{X}_N$ where the collective support may be an *arbitrary* set of non-negative integers. Two questions immediately suggest themselves:

1. How (if at all) does the sample complexity depend on the elements in the common support?

2. Does knowing the common support set help the learning algorithm — how does the complexity vary depending on whether or not the learning algorithm knows the common support?

In this paper we give essentially complete answers to these questions. Intriguingly, the answers to these questions emerge from the interface of probability theory and number theory: our algorithms rely on new central limit theorems for sums of independent random variables which we establish, while our matching lower bounds exploit delicate properties of continued fractions and sophisticated equidistribution results from analytic number theory. The authors find it quite surprising that these two disparate sets of techniques "meet up" to provide matching upper and lower bounds on sample complexity.

We now formalize the problem that we consider.

Our learning problem. Let $\mathbf{X}_1, \ldots, \mathbf{X}_N$ be independent (but not necessarily identically distributed) random variables. Let $\mathcal{A} = \cup_i \mathrm{supp}(\mathbf{X}_i)$ be the union of their supports and assume w.l.o.g. that $\mathcal{A} = \{a_1, ..., a_k\}$ for $a_1 < a_2 < \cdots < a_k \in \mathbb{Z}_{\geq 0}$. Let $\mathbf{S}$ be the sum of these independent random variables, $\mathbf{S} = \mathbf{X}_1 + \cdots + \mathbf{X}_N$. We refer to such a random variable $\mathbf{S}$ as an *$\mathcal{A}$-sum*.

We study the problem of learning a unknown $\mathcal{A}$-sum $\mathbf{S}$, given access to i.i.d. draws from $\mathbf{S}$. $\mathcal{A}$-sums generalize several classes of distributions which have recently been intensively studied in unsupervised learning [DDS12b, DDO+13, DKS16a], namely Poisson Binomial Distributions and "k-SIIRVs," and are closely related to other such distributions [DKS16c, DDKT16] (k-Poisson Multinomial Distributions). These previously studied classes of distributions have all been shown to have learning algorithms with sample complexity $\mathrm{poly}(1/\varepsilon)$ for all constant k.

In contrast, in this paper we show that the picture is more varied for the sample complexity of learning when $\mathcal{A}$ can be any finite set. Roughly speaking (we will give more details soon), two of our main results are as follows:

- *Any* $\mathcal{A}$-sum with $|\mathcal{A}| = 3$ is learnable from $\mathrm{poly}(1/\varepsilon)$ samples independent of N and of the elements of $\mathcal{A}$. This is a significant (and perhaps unexpected) generalization of the efficient learnability of Poisson Binomial Distributions, which corresponds to the case $|\mathcal{A}| = 2$.

- No such guarantee is possible for $|\mathcal{A}| = 4$: if N is large enough, there are infinitely many sets $\mathcal{A} = \{a_1, a_2, a_3, a_4\}$ with $0 \leq a_1 < ... < a_4$ such that $\Omega(\log \log a_4)$ examples are needed even to learn to constant accuracy (for a small absolute constant).

Before presenting our results in more detail, to provide context we recall relevant previous work on learning related distributions.

1.1 Previous work

A *Poisson Binomial Distribution of order* N, or PBD_N, is a sum of N independent (not necessarily identical) Bernoulli random variables, i.e. an $\mathcal{A}$-sum for $\mathcal{A} = \{0, 1\}$. Efficient algorithms for learning PBD_N distributions were given in [DDS12c, DKS16b], which gave learning algorithms using $\mathrm{poly}(1/\varepsilon)$ samples and $\mathrm{poly}(1/\varepsilon)$ runtime, independent of N.

Generalizing a PBD_N distribution, a k-SIIRV_N *(Sum of Independent Integer Random Variables)* is a $\mathcal{A}$-sum for $\mathcal{A} = \{0, ..., k-1\}$. Daskalakis et al. [DDO+13] (see also [DKS16a]) gave $\mathrm{poly}(k, 1/\varepsilon)$-time and sample algorithms for learning any k-SIIRV_N distribution to accuracy ε, independent of N.

Finally, a different generalization of PBDs is provided by the class of (N, k)*-Poisson Multinomial Distributions*, or k-PMD_N distributions. Such a distribution is $\mathbf{S} = \mathbf{X}_1 + \cdots + \mathbf{X}_N$ where the $\mathbf{X}_i$'s are independent (not necessarily identical) k-dimensional vector-valued random variables each supported on $\{e_1, \ldots, e_k\}$, the standard basis unit vectors in $\mathbb{R}^k$. Daskalakis et al. [DKT15] gave an algorithm that learns any unknown k-PMD_N using $\mathrm{poly}(k/\varepsilon)$ samples and running in time $\min\{2^{O(k^{O(k)}) \cdot \log^{O(k)}(1/\varepsilon)}, 2^{\mathrm{poly}(k/\varepsilon)}\}$; this result was subsequently sharpened in [DKS16c, DDKT16].

Any $\mathcal{A}$-sum with $|\mathcal{A}| = k$ has an associated underlying k-PMD_N distribution: if $\mathcal{A} = \{a_1, ..., a_k\}$, then writing $\bar{a}$ for the vector $(a_1, \ldots, a_k) \in \mathbb{Z}^k$, an $\mathcal{A}$-sum $\mathbf{S}'$ is equivalent to $\bar{a} \cdot \mathbf{S}$ where $\mathbf{S}$ is an k-PMD_N, as making a draw from $\mathbf{S}'$ is equivalent to making a draw from $\mathbf{S}$ and outputting its inner product with the vector $\bar{a}$. However, this does *not* mean that the [DKT15] learning result for k-PMD_N distributions implies a corresponding learning result for $\{a_1, ..., a_k\}$-sums. If an $\mathcal{A}$-sum learning algorithm *were given draws from the underlying* k*-PMD*$_N$, then of course it would be straightforward to run the [DKT15] algorithm, construct a high-accuracy hypothesis distribution $\mathbf{H}$ over $\mathbb{R}^k$, and output $\bar{a} \cdot \mathbf{H}$ as the hypothesis distribution for the unknown $\mathcal{A}$-sum. But when learning $\mathbf{S}'$, the algorithm does not receive draws from the underlying k-PMD_N $\mathbf{S}$; instead it only receives draws from $\bar{a} \cdot \mathbf{S}$. In fact, as we discuss below, this more limited access causes a crucial *qualitative* difference in learnability, namely an inherent dependence on the a_i's in the necessary sample complexity once $k \geq 4$. (The challenge to the learner arising from the blending of the contributions

to a $\mathcal{A}$-sum is roughly analogous to the challenge that arises in learning a DNF formula; if each positive example in a DNF learning problem were annotated with an identifier for a term that it satisfies, learning would be trivial.)

1.2 The questions we consider and our algorithmic results.

As detailed above, previous work has extensively studied the learnability of PBDs, k-SIIRVs, and k-PMDs; however, we believe that the current work is the first to study the learnability of general $\mathcal{A}$-sums. A first simple observation is that since any $\mathcal{A}$-sum with $|\mathcal{A}| = 2$ is a scaled and translated PBD, the results on learning PBDs mentioned above easily imply that the sample complexity of learning any $\{a_1, a_2\}$-sum is $\mathrm{poly}(1/\varepsilon)$, independent of the number of summands N and the values a_1, a_2. A second simple observation is that any $\{a_1, ..., a_k\}$-sum with $0 \le a_1 < ... < a_k$ can be learned using $\mathrm{poly}(a_k, 1/\varepsilon)$ samples, simply by viewing it as an a_k-SIIRV$_N$. But this bound is in general quite unsatisfying – indeed, for large a_k it could be even larger than the trivial $O(N^k/\varepsilon^2)$ upper bound that holds since any $\mathcal{A}$-sum with $|\mathcal{A}| = k$ is supported on a set of size $O(N^k)$.

Once $k \ge 3$ there can be non-trivial additive structure present in the set of values $a_1, \ldots, a_k$. This raises a natural question: is $k = 2$ the only value for which $\mathcal{A}$-sums are learnable from a number of samples that is independent of the domain elements $a_1, \ldots, a_k$? Perhaps surprisingly, our first main result is an efficient algorithm which gives a negative answer. We show that for $k = 3$, the values of the a_i's don't matter; we do this by giving an efficient learning algorithm (even a semi-agnostic one) for learning $\{a_1, a_2, a_3\}$-sums, whose running time and sample complexity are completely independent of a_1, a_2 and a_3:

Theorem 1 (Learning $\mathcal{A}$-sums with $|\mathcal{A}| = 3$, known support)**.** *There is an algorithm and a positive constant c with the following properties: The algorithm is given N, an accuracy parameter $\varepsilon > 0$, distinct values $a_1 < a_2 < a_3 \in \mathbb{Z}_{\ge 0}$, and access to i.i.d. draws from an unknown distribution $\mathbf{S}^*$ that has total variation distance at most $c\varepsilon$ from an $\{a_1, a_2, a_3\}$-sum. The algorithm uses $\mathrm{poly}(1/\varepsilon)$ draws from $\mathbf{S}^*$, runs in $poly(1/\varepsilon)$ time[1], and with probability at least $9/10$ outputs a concise representation of a hypothesis distribution $\mathbf{H}$ such that $d_{\mathrm{TV}}(\mathbf{H}, \mathbf{S}^*) \le \varepsilon$.*

We also give an algorithm for $k \ge 4$. More precisely, we show:

Theorem 2 (Learning $\mathcal{A}$-sums, known support)**.** *For any $k \ge 4$, there is an algorithm and a constant $c > 0$ with the following properties: it is given N, an accuracy parameter $\varepsilon > 0$, distinct values $a_1 < \cdots < a_k \in \mathbb{Z}_{\ge 0}$, and access to i.i.d. draws from an unknown distribution $\mathbf{S}^*$ that has total variation distance at most $c\varepsilon$ from some $\{a_1, \ldots, a_k\}$-sum. The algorithm runs in time $(1/\varepsilon)^{2^{O(k^2)}} \cdot (\log a_k)^{\mathrm{poly}(k)}$, uses $(1/\varepsilon)^{2^{O(k^2)}} \cdot \log\log a_k$ samples, and with probability at least $9/10$ outputs a concise representation of a hypothesis distribution $\mathbf{H}$ such that $d_{\mathrm{TV}}(\mathbf{H}, \mathbf{S}^*) \le \varepsilon$.*

In contrast with $k = 3$, our algorithm for general $k \ge 4$ has a sample complexity which depends (albeit doubly logarithmically) on a_k. This is a doubly exponential improvement over the naive $\mathrm{poly}(a_k)$ bound which follows from previous a_k-SIIRV learning algorithms [DDO+13, DKS16a].

Secondary algorithmic results: Learning with unknown support. We also give algorithms for a more challenging *unknown-support* variant of the learning problem. In this variant the values $a_1, \ldots, a_k$ are not provided to the learning algorithm, but instead only an upper bound $a_{\max} \ge a_k$ is given. Interestingly, it turns out that the unknown-support problem is significantly different from the known-support problem: as

[1] Here and throughout we assume a unit-cost model for arithmetic operations $+, \times, \div$.

explained below, in the unknown-support variant the dependence on $a_{\max}$ kicks in at a smaller value of k than in the known-support variant, and this dependence is exponentially more severe than in the known-support variant.

Using well-known results from hypothesis selection, it is straightforward to show that upper bounds for the known-support case yield upper bounds in the unknown-support case, essentially at the cost of an additional additive $O(k \log a_{\max})/\varepsilon^2$ term in the sample complexity. This immediately yields the following:

Theorem 3 (Learning with unknown support of size k)**.** *For any $k \geq 3$, there is an algorithm and a positive constant c with the following properties: The algorithm is given N, the value k, an accuracy parameter $\varepsilon > 0$, an upper bound $a_{\max} \in \mathbb{Z}_{\geq 0}$, and access to i.i.d. draws from an unknown distribution $\mathbf{S}^*$ that has total variation distance at most $c\varepsilon$ from an $\mathcal{A}$-sum for $\mathcal{A} = \{a_1, \ldots, a_k\} \subset \mathbb{Z}_{\geq 0}$ where $\max_i a_i \leq a_{\max}$. The algorithm uses $O(k \log a_{\max})/\varepsilon^2 + (1/\varepsilon)^{2^{O(k^2)}} \cdot \log\log a_{\max}$ draws from $\mathbf{S}^*$, runs in $\mathrm{poly}((a_{\max})^k) \cdot (1/\varepsilon)^{2^{O(k^2)}} \cdot (\log a_{\max})^{\mathrm{poly}(k)}$ time, and with probability at least $9/10$ outputs a concise representation of a hypothesis distribution $\mathbf{H}$ such that $d_{\mathrm{TV}}(\mathbf{H}, \mathbf{S}^*) \leq \varepsilon$.*

Recall that a $\{a_1, a_2\}$-sum is simply a rescaled and translated PBD_N distribution. Using known results for learning PBDs, it is not hard to show that the $k = 2$ case is easy even with unknown support:

Theorem 4 (Learning with unknown support of size 2)**.** *There is an algorithm and a positive constant c with the following properties: The algorithm is given N, an accuracy parameter $\varepsilon > 0$, an upper bound $a_{\max} \in \mathbb{Z}_+$, and access to i.i.d. draws from an unknown distribution $\mathbf{S}^*$ that has total variation distance at most $c\varepsilon$ from an $\{a_1, a_2\}$-sum where $0 \leq a_1 < a_2 \leq a_{\max}$. The algorithm uses $\mathrm{poly}(1/\varepsilon)$ draws from $\mathbf{S}^*$, runs in $\mathrm{poly}(1/\varepsilon)$ time, and with probability at least $9/10$ outputs a concise representation of a hypothesis distribution $\mathbf{H}$ such that $d_{\mathrm{TV}}(\mathbf{H}, \mathbf{S}^*) \leq \varepsilon$.*

1.3 Our lower bounds.

We establish sample complexity lower bounds for learning $\mathcal{A}$-sums that essentially match the above algorithmic results.

Known support. Our first lower bound deals with the known support setting. We give an $\Omega(\log\log a_4)$-sample lower bound for the problem of learning an $\{a_1, ..., a_4\}$-sum for $0 \leq a_1 < a_2 < a_3 < a_4$. This matches the dependence on a_k of our $\mathrm{poly}(1/\varepsilon) \cdot \log\log a_k$ upper bound. More precisely, we show:

Theorem 5 (Lower Bound for Learning $\{a_1, ..., a_4\}$-sums, known support)**.** *Let A be any algorithm with the following properties: algorithm A is given N, an accuracy parameter $\varepsilon > 0$, distinct values $0 \leq a_1 < a_2 < a_3 < a_4 \in \mathbb{Z}$, and access to i.i.d. draws from an unknown $\{a_1, ..., a_4\}$-sum $\mathbf{S}^*$; and with probability at least $9/10$ algorithm A outputs a hypothesis distribution $\tilde{\mathbf{S}}$ such that $d_{\mathrm{TV}}(\tilde{\mathbf{S}}, \mathbf{S}^*) \leq \varepsilon$. Then there are infinitely many quadruples (a_1, a_2, a_3, a_4) such that for sufficiently large N, A must use $\Omega(\log\log a_4)$ samples even when run with ε set to a (suitably small) positive absolute constant.*

This lower bound holds even though the target is exactly an $\{a_1, ..., a_4\}$-sum (i.e. it holds even in the easier non-agnostic setting).

Since Theorem 1 gives a $\mathrm{poly}(1/\varepsilon)$ sample and runtime algorithm independent of the size of the a_i's for $k = 3$, the lower bound of Theorem 5 establishes a phase transition between $k = 3$ and $k = 4$ for the sample complexity of learning $\mathcal{A}$-sums: when $k = 3$ the sample complexity is always independent of the actual set $\{a_1, a_2, a_3\}$, but for $k = 4$ it can grow as $\Omega(\log\log a_4)$ (but no faster).

Unknown support. Our second lower bound deals with the unknown support setting. We give an $\Omega(\log a_{\max})$-sample lower bound for the problem of learning an $\{a_1, a_2, a_3\}$-sum with unknown support $0 \leq a_1 < a_2 < a_3 \leq a_{\max}$, matching the dependence on $a_{\max}$ of our algorithm from Theorem 3. More precisely, we prove:

Theorem 6 (Lower Bound for Learning $\{a_1, a_2, a_3\}$-sums, unknown support)**.** *Let A be any algorithm with the following properties: algorithm A is given N, an accuracy parameter $\varepsilon > 0$, a value $0 < a_{\max} \in \mathbb{Z}$, and access to i.i.d. draws from an unknown $\{a_1, a_2, a_3\}$-sum $\mathbf{S}^*$ where $0 \leq a_1 < a_2 < a_3 \leq a_{\max}$; and A outputs a hypothesis distribution $\tilde{\mathbf{S}}$ which with probability at least $9/10$ satisfies $d_{\mathrm{TV}}(\tilde{\mathbf{S}}, \mathbf{S}^*) \leq \varepsilon$. Then for sufficiently large N, A must use $\Omega(\log a_{\max})$ samples even when run with ε set to a (suitably small) positive absolute constant.*

Taken together with our algorithm from Theorem 4 for the case $k = 2$, Theorem 6 establishes another phase transition, but now between $k = 2$ and $k = 3$, for the sample complexity of learning $\mathcal{A}$-sums when $\mathcal{A}$ is unknown. When $|\mathcal{A}| = 2$ the sample complexity is always independent of the actual set, but for $|\mathcal{A}| = 3$ and $0 \leq a_1 < ... < a_3$ it can grow as $\Omega(\log a_3)$ (but no faster).

In summary, taken together the algorithms and lower bounds of this paper essentially settle the question of how the sample complexity of learning sums of independent integer random variables with sparse collective support scales with the elements in the collective support, both in the known-support and unknown-support settings.

Discussion. As described above, for an arbitrary set $\{a_1, \ldots, a_k\}$, the sample complexity undergoes a significant phase transition between $k = 3$ and $k = 4$ in the known-support case and between 2 and 3 in the unknown-support case. In each setting the phase transition is a result of "number-theoretic phenomena" (we explain this more later) which can only occur for the larger number and cannot occur for the smaller number of support elements. We find it somewhat surprising that the sample complexities of these learning problems are determined by number-theoretic properties of the support sets.

Organization. In the next section we give some of the key ideas that underlie our algorithms. See Section 3 for an overview of the ideas behind our lower bounds. Full proofs are given starting in Section 4.

2 Techniques for our algorithms

In this section we give an intuitive explanation of some of the ideas that underlie our algorithms and their analysis. While our learning results are for the semi-agnostic model, for simplicity's sake, we focus on the case in which the target distribution $\mathbf{S}$ is actually an $\mathcal{A}$-sum.

A first question, which must be addressed before studying the algorithmic (running time) complexity of learning $\mathcal{A}$-sums, is to understand the sample complexity of learning them. In fact, in a number of recent works on learning various kinds of "structured" distributions, just understanding the sample complexity of the learning problem is a major goal that requires significant work [DDS12c, WY12, DDO+13, DDS14, DKT15].

In many of the above-mentioned papers, an upper bound on both sample complexity and algorithmic complexity is obtained via a structural characterization of the distributions to be learned; our work follows a similar conceptual paradigm. To give a sense of the kind of structural characterization that can be helpful for learning, we recall the characterization of SIIRV$_N$ distributions that was obtained in [DDO+13] (which is the one most closely related to our work). The main result of [DDO+13] shows that if $\mathbf{S}$ is any k-SIIRV$_N$ distribution, then at least one of the following holds:

1. $\mathbf{S}$ is ε-close to being supported on $\mathrm{poly}(k/\varepsilon)$ many integers;

2. $\mathbf{S}$ is ε-close to a distribution $c \cdot \mathbf{Z} + \mathbf{Y}$, where $1 \leq c \leq k-1$, $\mathbf{Z}$ is a discretized Gaussian, $\mathbf{Y}$ is a distribution supported on $\{0, \ldots, c-1\}$, and $\mathbf{Y}, \mathbf{Z}$ are mutually independent.

In other words, [DDO+13] shows that a k-SIIRV$_N$ distribution is either close to sparse (supported on $\mathrm{poly}(k/\varepsilon)$ integers), or close to a c-scaled discretized Gaussian convolved with a sparse component supported on $\{0, \ldots, c-1\}$. This leads naturally to an efficient learning algorithm that handles Case (1) above "by brute-force" and handles Case (2) by learning $\mathbf{Y}$ and $\mathbf{Z}$ separately (handling $\mathbf{Y}$ "by brute force" and handling $\mathbf{Z}$ by estimating its mean and variance).

In a similar spirit, in this work we seek a more general characterization of $\mathcal{A}$-sums. It turns out, though, that even when $|\mathcal{A}| = 3$, $\mathcal{A}$-sums can behave in significantly more complicated ways than the k-SIIRV$_N$ distributions discussed above.

To be more concrete, let $\mathbf{S}$ be a $\{a_1, a_2, a_3\}$-sum with $0 \leq a_1 < a_2 < a_3$. By considering a few simple examples it is easy to see that there are at least four distinct possibilities for "what $\mathbf{S}$ is like" at a coarse level:

- **Example #1:** One possibility is that $\mathbf{S}$ is essentially sparse, with almost all of its probability mass concentrated on a small number of outcomes (we say that such an $\mathbf{S}$ has "small essential support").

- **Example #2:** Another possibility is that $\mathbf{S}$ "looks like" a discretized Gaussian scaled by $|a_i - a_j|$ for some $1 \leq i < j \leq 3$ (this would be the case, for example, if $\mathbf{S} = \sum_{i=1}^{N} \mathbf{X}_i$ where each $\mathbf{X}_i$ is uniform over $\{a_1, a_2\}$).

- **Example #3:** A third possibility is that $\mathbf{S}$ "looks like" a discretized Gaussian with no scaling (the analysis of [DDO+13] shows that this is what happens if, for example, N is large and each $\mathbf{X}_i$ is uniform over $\{a_1 = 6, a_2 = 10, a_3 = 15\}$, since $\gcd(6, 10, 15) = 1$).

- **Example #4:** Finally, yet another possibility arises if, say, a_3 is very large (say $a_3 \approx N^2$) while a_2, a_1 are very small (say $O(1)$), and $\mathbf{X}_1, \ldots, \mathbf{X}_{N/2}$ are each uniform over $\{a_1, a_3\}$ while $\mathbf{X}_{N/2+1}, \ldots, \mathbf{X}_N$ are each supported on $\{a_1, a_2\}$ and $\sum_{i=N/2+1}^{N} \mathbf{X}_i$ has very small essential support. In this case, for large N, $\mathbf{S}$ would (at a coarse scale) "look like" a discretized Gaussian scaled by $a_3 - a_1 \approx N^2$, but zooming in, locally each "point" in the support of this discretized Gaussian would actually be a copy of the small-essential-support distribution $\sum_{i=N/2+1}^{N} \mathbf{X}_i$.

Given these possibilities for how $\mathbf{S}$ might behave, it should not be surprising that our actual analysis for the case $|\mathcal{A}| = 3$ (given in Section 9) involves four cases (and the above four examples land in the four distinct cases). The overall learning algorithm "guesses" which case the target distribution belongs to and runs a different algorithm for each one; the guessing step is ultimately eliminated using the standard tool of hypothesis testing from statistics. We stress that while the algorithms for the various cases differ in some details, there are many common elements across their analyses, and the well known *kernel method* for density estimation provides the key underlying core learning routine that is used in all the different cases.

In the following intuitive explanation we first consider the case of $\mathcal{A}$-sums for general finite $|\mathcal{A}|$, and later explain how we sharpen the algorithm and analysis in the case $|\mathcal{A}| = 3$ to obtain our stronger results for that case. Our discussion below highlights a new structural result (roughly speaking, a new limit theorem that exploits both "long-range" and "short-range" shift-invariance) that plays a crucial role in our algorithms.

2.1 Learning $\mathcal{A}$-sums with $|\mathcal{A}| = k$

For clarity of exposition in this intuitive overview we make some simplifying assumptions. First, we make the assumption that the $\mathcal{A}$-sum $\mathbf{S}$ that is to be learned has 0 as one value in its k-element support, i.e. we assume that $\mathbf{S} = \mathbf{X}_1 + \ldots + \mathbf{X}_N$ where the support of each $\mathbf{X}_i$ is contained in the set $\{0, a_1, \ldots, a_{k-1}\}$. In fact, we additionally assume that each $\mathbf{X}_i$ is 0-*moded*, meaning that $\mathbf{Pr}[\mathbf{X}_i = 0] \geq \mathbf{Pr}[\mathbf{X}_i = a_j]$ for all $i \in [N]$ and all $j \in [k-1]$. (Getting rid of this assumption in our actual analysis requires us to work with zero-moded variants of the $\mathbf{X}_i$ distributions that we denote $\mathbf{X}'_i$, supported on $O(k^2)$ values that can be positive or negative, but we ignore this for the sake of our intuitive explanation here.) For $j \in [k-1]$ we define

$$\gamma_j := \sum_{i=1}^{N} \mathbf{Pr}[\mathbf{X}_i = a_j],$$

which can be thought of as the "weight" that $\mathbf{X}_1, \ldots, \mathbf{X}_N$ collectively put on the outcome a_j.

A useful tool: hypothesis testing. To explain our approach it is helpful to recall the notion of hypothesis testing in the context of distribution learning [DL01]. Informally, given T candidate hypothesis distributions, one of which is ε-close to the target distribution $\mathbf{S}$, a hypothesis testing algorithm uses $O(\varepsilon^{-2} \cdot \log T)$ draws from $\mathbf{S}$, runs in $\mathrm{poly}(T, 1/\varepsilon)$ time, and with high probability identifies a candidate distribution which is $O(\varepsilon)$-close to $\mathbf{S}$. We use this tool in a few different ways. Sometimes we will consider algorithms that "guess" certain parameters from a "small" (size-T) space of possibilities; hypothesis testing allows us to assume that such algorithms guess the right parameters, at the cost of increasing the sample complexity and running time by only small factors. In other settings we will show via a case analysis that one of several different learning algorithms will succeed; hypothesis testing yields a combined algorithm that learns no matter which case the target distribution falls into. (This tool has been used in many recent works on distribution learning, see e.g. [DDS12c, DDS15, DDO+13].)

Our analysis. Let $t_1 = O_{k,\varepsilon}(1) \ll t_2 = O_{k,\varepsilon}(1) \ll \cdots \ll t_{k-1} = O_{k,\varepsilon}(1)$ be fixed values (the exact values are not important here). Let us reorder $a_1, \ldots, a_{k-1}$ so that the weights $\gamma_1 \leq \cdots \leq \gamma_{k-1}$ are sorted in non-decreasing order. An easy special case for us (corresponding to Section 8.1) is when each $\gamma_j \leq t_j$. If this is the case, then $\mathbf{S}$ has small "essential support": in a draw from $\mathbf{S} = \mathbf{X}_1 + \cdots + \mathbf{X}_N$, with very high probability for each $j \in [k-1]$ the number of $\mathbf{X}_i$ that take value a_j is at most $\mathrm{poly}(t_{k-1})$, so w.v.h.p. a draw from $\mathbf{S}$ takes one of at most $\mathrm{poly}(t_{k-1})^k$ values. In such a case it is not difficult to learn $\mathbf{S}$ using $\mathrm{poly}((t_{k-1})^k, 1/\varepsilon) = O_{k,\varepsilon}(1)$ samples (see Fact 24). We henceforth may assume that some $\gamma_j > t_j$.

For ease of understanding it is helpful to first suppose that *every* $j \in [k-1]$ has $\gamma_j > t_j$, and to base our understanding of the general case (that some $j \in [k-1]$ has $\gamma_j > t_j$) off of how this case is handled; we note that this special case is the setting for the structural results of Section 7. (It should be noted, though, that our actual analysis of the main learning algorithm given in Section 8.2 does not distinguish this special case.) So let us suppose that for all $j \in [k-1]$ we have $\gamma_j > t_j$. To analyze the target distribution $\mathbf{S}$ in this case, we consider a multinomial distribution $\mathbf{M} = \mathbf{Y}_1 + \cdots + \mathbf{Y}_N$ defined by independent vector-valued random variables $\mathbf{Y}_i$, supported on $0, \boldsymbol{e}_1, \ldots, \boldsymbol{e}_{k-1} \in \mathbb{Z}^{k-1}$, such that for each $i \in [N]$ and $j \in [k-1]$ we have $\mathbf{Pr}[\mathbf{Y}_i = e_j] = \mathbf{Pr}[\mathbf{X}_i = a_j]$. Note that for the multinomial distribution $\mathbf{M}$ defined in this way we have $(a_1, \ldots, a_{k-1}) \cdot \mathbf{M} = \mathbf{S}$.

Using the fact that each γ_j is "large" (at least t_j), recent results from [DDKT16] imply that the multinomial distribution $\mathbf{M}$ is close to a $(k-1)$-dimensional discretized Gaussian whose covariance matrix has all eigenvalues large (working with zero-moded distributions is crucial to obtain this intermediate result). In turn, such a discretized multidimensional Gaussian can be shown to be close to a vector-valued random

variable in which each marginal (coordinate) is a (± 1)-weighted sum of *independent* large-variance Poisson Binomial Distributions. It follows that $\mathbf{S} = (a_1, \ldots, a_{k-1}) \cdot \mathbf{M}$ is close to a a weighted sum of $k-1$ signed PBDs. [2] A distribution $\tilde{\mathbf{S}}$ is a weighted sum of $k-1$ signed PBDs if $\tilde{\mathbf{S}} = a_1 \cdot \tilde{\mathbf{S}}_1 + \cdots + a_{k-1} \cdot \tilde{\mathbf{S}}_{k-1}$ where $\tilde{\mathbf{S}}_1, \ldots, \tilde{\mathbf{S}}_{k-1}$ are *independent* signed PBDs; in turn, a signed PBD is a sum of independent random variables each of which is either supported on $\{0, 1\}$ or on $\{0, -1\}$. The $\tilde{\mathbf{S}}$ that $\mathbf{S}$ is close to further has the property that each $\tilde{\mathbf{S}}_i$ has "large" variance (large compared with $1/\varepsilon$).

Given the above analysis, to complete the argument in this case that each $\gamma_j > t_j$ we need a way to learn a weighted sum of signed PBDs $\tilde{\mathbf{S}} = a_1 \cdot \tilde{\mathbf{S}}_1 + \cdots + a_{k-1} \cdot \tilde{\mathbf{S}}_{k-1}$ where each $\tilde{\mathbf{S}}_j$ has large variance. This is done with the aid of a new limit theorem, Lemma 41, that we establish for distributions of this form. We discuss (a simplified version of) this limit theorem in Section 2.3; here, omitting many details, let us explain what this new limit theorem says in our setting and how it is useful for learning. Suppose w.l.o.g. that $\mathbf{Var}[a_{k-1} \cdot \tilde{\mathbf{S}}_{k-1}]$ contributes at least a $\frac{1}{k-1}$ fraction of the total variance of $\tilde{\mathbf{S}}$. Let MIX denote the set of those $j \in \{1, \ldots, k-2\}$ such that $\mathbf{Var}[\tilde{\mathbf{S}}_j]$ is large compared with a_{k-1}, and let $\mathrm{MIX}' = \mathrm{MIX} \cup \{k-1\}$. The new limit theorem implies that the sum $\sum_{j \in \mathrm{MIX}'} a_j \cdot \tilde{\mathbf{S}}_j$ "mixes," meaning that it is very close (in d_{TV}) to a *single* scaled PBD $a_{\mathrm{MIX}'} \cdot \tilde{\mathbf{S}}_{\mathrm{MIX}'}$ where $a_{\mathrm{MIX}'} = \gcd\{a_j : j \in \mathrm{MIX}'\}$. (The proof of the limit theorem involves a generalization of the notion of shift-invariance from probability theory [BX99] and a coupling-based method. We elaborate on the ideas behind the limit theorem in Section 2.3.)

Given this structural result, it is enough to be able to learn a distribution of the form

$$\mathbf{T} := a_1 \cdot \tilde{\mathbf{S}}_1 + \cdots + a_\ell \cdot \tilde{\mathbf{S}}_\ell + a_{\mathrm{MIX}'} \cdot \tilde{\mathbf{S}}_{\mathrm{MIX}'}$$

for which we now know that $a_{\mathrm{MIX}'} \cdot \tilde{\mathbf{S}}_{\mathrm{MIX}'}$ has at least $\frac{1}{\ell+1}$ of the total variance, and each $\tilde{\mathbf{S}}_j$ for $j \in [\ell]$ has $\mathbf{Var}[\tilde{\mathbf{S}}_j]$ which is "not too large" compared with a_{k-1} (but large compared with $1/\varepsilon$). We show how to learn such a distribution using $O_{k,\varepsilon}(1) \cdot \log\log a_{k-1}$ samples (this is where the $\log\log$ dependence in our overall algorithm comes from). This is done, intuitively, by guessing various parameters that essentially define $\mathbf{T}$, specifically the variances $\mathbf{Var}[\tilde{\mathbf{S}}_1], \ldots, \mathbf{Var}[\tilde{\mathbf{S}}_\ell]$. Since each of these variances is roughly at most a_{k-1} (crucially, the limit theorem allowed us to get rid of the $\tilde{\mathbf{S}}_j$'s that had larger variance), via multiplicative gridding there are $O_{\varepsilon,k}(1) \cdot \log a_{k-1}$ possible values for each candidate variance, and via our hypothesis testing procedure this leads to an $O_{\varepsilon,k}(1) \cdot \log\log a_{k-1}$ number of samples that are used to learn.

We now turn to the general case, that some $j \in [k-1]$ has $\gamma_j > t_j$. Suppose w.l.o.g. that $\gamma_1 \leq t_1, \ldots \gamma_{\ell-1} \leq t_{\ell-1}$ and $\gamma_\ell > t_\ell$ (intuitively, think of $\gamma_1, \ldots, \gamma_{\ell-1}$ as "small" and $\gamma_\ell, \ldots, \gamma_{k-1}$ as "large"). Via an analysis (see Lemma 46) akin to the "Light-Heavy Experiment" analysis of [DDO+13], we show that in this case the distribution $\mathbf{S}$ is close to a distribution $\tilde{\mathbf{S}}$ with the following structure: $\tilde{\mathbf{S}}$ is a mixture of at most $\mathrm{poly}(t_{\ell-1})^{k-1}$ many distributions each of which is a different shift of a *single* distribution, call it $\mathbf{S}_{\mathrm{heavy}}$, that falls into the special case analyzed above: all of the relevant parameters $\gamma_\ell, \ldots, \gamma_{k-1}$ are large (at least t_ℓ). Intuitively, having at most $\mathrm{poly}(t_{\ell-1})^{k-1}$ many components in the mixture corresponds to having $\gamma_1, \ldots, \gamma_{\ell-1} < t_{\ell-1}$ and $\ell \leq k-1$, and having each component be a shift of the same distribution $\mathbf{S}_{\mathrm{heavy}}$ follows from the fact that there is a "large gap" between $\gamma_{\ell-1}$ and γ_ℓ.

Thus in this general case, the learning task essentially boils down to learning a distribution that is (close to) a mixture of translated copies of a distribution of the form $\mathbf{T}$ given above. Learning such a mixture of translates is a problem that is well suited to the "kernel method" for density estimation. This method has been well studied in classical density estimation, especially for continuous probability densities (see e.g. [DL01]), but results of the exact type that we need did not seem to previously be present in the literature.

[2] This is a simplification of what the actual analysis establishes, but it gets across the key ideas.

(We believe that ours is the first work that applies kernel methods to learn sums of independent random variables.)

In Section 5 we develop tools for multidimensional kernel based learning that suit our context. At its core, the kernel method approach that we develop allows us to do the following: Given a mixture of r translates of $\mathbf{T}$ and constant-factor approximations to $\gamma_\ell, \ldots, \gamma_{k-1}$, the kernel method allows us to learn this mixture to error $O(\varepsilon)$ using only $\mathrm{poly}(1/\varepsilon^\ell, r)$ samples. Further, this algorithm is robust in the sense that the same guarantee holds even if the target distribution is only $O(\varepsilon)$ close to having this structure (this is crucial for us). Theorem 49 in Section 8 combines this tool with the ideas described above for learning a $\mathbf{T}$-type distribution, and thereby establishes our general learning result for $\mathcal{A}$-sums with $|\mathcal{A}| \geq 4$.

2.2 The case $|\mathcal{A}| = 3$

In this subsection we build on the discussion in the previous subsection, specializing to $k = |\mathcal{A}| = 3$, and explain the high-level ideas of how we are able to learn with sample complexity $\mathrm{poly}(1/\varepsilon)$ independent of a_1, a_2, a_3.

For technical reasons (related to zero-moded distributions) there are three relevant parameters $t_1 \ll t_2 \ll t_3 = O_\varepsilon(1)$ in the $k = 3$ case. The easy special case that each $\gamma_j \leq t_j$ is handled as discussed earlier (small essential support). As in the previous subsection, let $\ell \in [3]$ be the least value such that $\gamma_\ell > t_\ell$.

In all the cases $\ell = 1, 2, 3$ the analysis proceeds by considering the Light-Heavy-Experiment as discussed in the preceding subsection, i.e. by approximating the target distribution $\mathbf{S}$ by a mixture $\tilde{\mathbf{S}}$ of shifts of the *same* distribution $\mathbf{S}_{\mathrm{heavy}}$. When $\ell = 3$, the "heavy" component $\mathbf{S}_{\mathrm{heavy}}$ is simply a distribution of the form $q_3 \cdot \mathbf{S}_3$ where $\mathbf{S}_3$ is a signed PBD. Crucially, while learning the distribution $\mathbf{T}$ in the previous subsection involved guessing certain variances (which could be as large as a_k, leading to $\log a_k$ many possible outcomes of guesses and $\log \log a_k$ sample complexity), in the current setting the extremely simple structure of $\mathbf{S}_{\mathrm{heavy}} = q_3 \cdot \mathbf{S}_3$ obviates the need to make $\log a_3$ many guesses. Instead, as we discuss in Section 9.2, its variance can be approximated in a simple direct way by sampling just two points from $\mathbf{T}$ and taking their difference; this easily gives a constant-factor approximation to the variance of $\mathbf{S}_3$ with non-negligible probability. This success probability can be boosted by repeating this experiment several times (but the number of times does not depend on the a_i values.) We thus can use the kernel-based learning approach in a sample-efficient way, without any dependence on a_1, a_2, a_3 in the sample complexity.

For clarity of exposition, in the remaining intuitive discussion (of the $\ell = 1, 2$ cases) we only consider a special case: we assume that $\mathbf{S} = a_1 \cdot \mathbf{S}_1 + a_2 \cdot \mathbf{S}_2$ where both $\mathbf{S}_1$ and $\mathbf{S}_2$ are large-variance PBDs (so each random variable $\mathbf{X}_i$ is either supported on $\{0, a_1\}$ or on $\{0, a_2\}$, but not on all three values $0, a_1, a_2$). We further assume, clearly without loss of generality, that $\gcd(a_1, a_2) = 1$. (Indeed, our analysis essentially proceeds by reducing the $\ell = 1, 2$ case to this significantly simpler scenario, so this is a fairly accurate rendition of the true case.) Writing $\mathbf{S}_1 = \mathbf{X}_1 + \ldots + \mathbf{X}_{N_1}$ and $\mathbf{S}_2 = \mathbf{Y}_1 + \ldots + \mathbf{Y}_{N_2}$, by zero-modedness we have that $\mathbf{Pr}[\mathbf{X}_i = 0] \geq \frac{1}{2}$ and $\mathbf{Pr}[\mathbf{Y}_i = 0] \geq \frac{1}{2}$ for all i, so $\mathbf{Var}[\mathbf{S}_j] = \Theta(1) \cdot \gamma_j$ for $j = 1, 2$. We assume w.l.o.g. in what follows that $a_1^2 \cdot \gamma_1 \geq a_2^2 \cdot \gamma_2$, so $\mathbf{Var}[\mathbf{S}]$, which we henceforth denote σ^2, is $\Theta(1) \cdot a_1^2 \cdot \gamma_1$.

We now branch into three separate possibilities depending on the relative sizes of γ_2 and a_1^2. Before detailing these possibilities we observe that using the fact that γ_1 and γ_2 are both large, it can be shown that if we sample two points $s^{(1)}$ and $s^{(2)}$ from $\mathbf{S}$, then with constant probability the value $\frac{|s^{(1)} - s^{(2)}|}{a_1}$ provides a constant-factor approximation to γ_1.

First possibility: $\gamma_2 < \varepsilon^2 \cdot a_1^2$. The algorithm samples two more points $s^{(3)}$ and $s^{(4)}$ from the distribution $\mathbf{S}$. The crucial idea is that with constant probability these two points can be used to obtain a constant-factor

approximation to γ_2; we now explain how this is done. For $j \in \{3,4\}$, let $s^{(j)} = a_1 \cdot s_1^{(j)} + a_2 \cdot s_2^{(j)}$ where $s_1^{(j)} \sim \mathbf{S}_1$ and $s_2^{(j)} \sim \mathbf{S}_2$, and consider the quantity $s^{(3)} - s^{(4)}$. Since γ_2 is so small relative to a_1, the "sampling noise" from $a_1 \cdot s_1^{(3)} - a_1 \cdot s_1^{(4)}$ is likely to overwhelm the difference $a_2 \cdot s_2^{(3)} - a_2 \cdot s_2^{(4)}$ at a "macroscopic" level. The key idea to deal with this is to *analyze the outcomes modulo* a_1. In the modular setting, because $\mathbf{Var}[\mathbf{S}_2] = \Theta(1) \cdot \gamma_2 \ll a_1^2$, one can show that with constant probability $|(a_2^{-1} \cdot (s_2^{(3)} - s_2^{(4)})) \bmod a_1|$ is a constant-factor approximation to γ_2. (Note that as a_1 and a_2 are coprime, the operation a_2^{-1} is well defined modulo a_1.) A constant-factor approximation to γ_2 can be used together with the constant-factor approximation to γ_1 to employ the aforementioned "kernel method" based algorithm to learn the target distribution $\mathbf{S}$. The fact that here we can use only two samples (as opposed to $\log \log a_1$ samples) to estimate γ_2 is really the crux of why for the $k = 3$ case, the sample complexity is independent of a_1. (Indeed, we remark that our analysis of the lower bound given by Theorem 5 takes place in the modular setting and this "mod a_1" perspective is crucial for constructing the lower bound examples in that proof.)

Second possibility: $a_1^2/\varepsilon^2 > \gamma_2 > \varepsilon^2 \cdot a_1^2$. Here, by multiplicative gridding we can create a list of $O(\log(1/\varepsilon))$ guesses such that at least one of them is a constant-factor approximation to γ_2. Again, we use the kernel method and the approximations to γ_1 and γ_2 to learn $\mathbf{S}$.

Third possibility: The last possibility is that $\gamma_2 \geq a_1^2/\varepsilon^2$. In this case, we show that $\mathbf{S}$ is in fact ε-close to the discretized Gaussian (with no scaling; recall that $\gcd(a_1, a_2) = 1$) that has the appropriate mean and variance. Given this structural fact, it is easy to learn $\mathbf{S}$ by just estimating the mean and the variance and outputting the corresponding discretized Gaussian. This structural fact follows from our new limit theorem, Lemma 41, mentioned earlier; we conclude this section with a discussion of this new limit theorem.

2.3 Lemma 41 and limit theorems.

Here is a simplified version of our new limit theorem, Lemma 41, specialized to the case in which its "D" parameter is set to 2:

Simplified version of Lemma 41. *Let* $\mathbf{S} = r_1 \cdot \mathbf{S}_1 + r_2 \cdot \mathbf{S}_2$ *where* $\mathbf{S}_1, \mathbf{S}_2$ *are independent signed PBDs and* r_1, r_2 *are nonzero integers such that* $\gcd(r_1, r_2) = 1$, $\mathbf{Var}[r_1 \cdot \mathbf{S}_1] \geq \mathbf{Var}[r_2 \cdot \mathbf{S}_2]$, *and* $\mathbf{Var}[\mathbf{S}_2] \geq \max\{\frac{1}{\varepsilon^8}, \frac{r_1}{\varepsilon}\}$. *Then* $\mathbf{S}$ *is* $O(\varepsilon)$*-close in total variation distance to a signed PBD* $\mathbf{S}'$ *(and hence to a signed discretized Gaussian) with* $\mathbf{Var}[\mathbf{S}'] = \mathbf{Var}[\mathbf{S}]$.

If a distribution $\mathbf{S}$ is close to a discretized Gaussian in Kolmogorov distance and is $1/\sigma$-shift invariant (i.e. $d_{\mathrm{TV}}(\mathbf{S}, \mathbf{S}+1) \leq 1/\sigma$), then $\mathbf{S}$ is close to a discretized Gaussian in total variation distance [RÖ7, Bar15]. Gopalan et al. [GMRZ11] used a coupling based argument to establish a similar central limit theorem to obtain pseudorandom generators for certain space bounded branching programs. Unfortunately, in the setting of the lemma stated above, it is not immediately clear why $\mathbf{S}$ should have $1/\sigma$-shift invariance. To deal with this, we give a novel analysis exploiting *shift-invariance at multiple different scales*. Roughly speaking, because of the $r_1 \cdot \mathbf{S}_1$ component of $\mathbf{S}$, it can be shown that $d_{\mathrm{TV}}(\mathbf{S}, \mathbf{S} + r_1) = 1/\sqrt{\mathbf{Var}[\mathbf{S}_1]}$, i.e. $\mathbf{S}$ has good "shift-invariance at the scale of r_1"; by the triangle inequality $\mathbf{S}$ is also not affected much if we shift by a small integer multiple of r_1. The same is true for a few shifts by r_2, and hence also for a few shifts by *both* r_1 and r_2. If $\mathbf{S}$ is approximated well by a discretized Gaussian, though, then it is also not affected by small shifts, including shifts by 1, and in fact we need such a guarantee to prove approximation by a discretized Gaussian through coupling. However, since $\gcd(r_1, r_2) = 1$, basic number theory implies that we can achieve any small integer shift via a small number of shifts by r_1 and r_2, and therefore $\mathbf{S}$ has the required "fine-grained" shift-invariance (at scale 1) as well. Intuitively, for this to work we need samples from $r_2 \cdot \mathbf{S}_2$ to "fill in the gaps" between successive values of $r_1 \cdot \mathbf{S}_1$ – this is why we need $\mathbf{Var}[\mathbf{S}_2] \gg r_1$.

Based on our discussion with researchers in this area [Bar15] the idea of exploiting both long-range and short-range shift invariance is new to the best of our knowledge and seems likely to be of use in proving new central limit theorems.

3 Lower bound techniques

In this section we give an overview of the ideas behind our lower bounds. Both of our lower bounds actually work by considering restricted $\mathcal{A}$-sums: our lower bounds can be proved using only distributions $\mathbf{S}$ of the form $\mathbf{S} = \sum_{i=1}^{k} a_i \cdot \mathbf{S}_i$, where $\mathbf{S}_1, \ldots, \mathbf{S}_k$ are independent PBDs; equivalently, $\mathbf{S} = \sum_{i=1}^{N} \mathbf{X}_i$ where each $\mathbf{X}_i$ is supported on one of $\{0, a_1\}, \ldots, \{0, a_k\}$.

A useful reduction. The problem of learning a distribution modulo an integer plays a key role in both of our lower bound arguments. More precisely, both lower bounds use a reduction, which we establish, showing that an efficient algorithm for learning weighted PBDs with weights $0 < a_1 < ... < a_k$ implies an efficient algorithm for learning with weights $a_1, ..., a_{k-1}$ modulo a_k. This problem is specified as follows: Consider an algorithm which is given access to i.i.d. draws from the distribution $(\mathbf{S} \mod a_k)$ (note that this distribution is supported over $\{0, 1, \ldots, a_k - 1\}$) where $\mathbf{S}$ is of the form $a_1 \cdot \mathbf{S}_1 + ... + a_{k-1} \cdot \mathbf{S}_{k-1}$ and $\mathbf{S}_1, ..., \mathbf{S}_{k-1}$ are PBDs. The algorithm should produce a high-accuracy hypothesis distribution for $(\mathbf{S} \mod a_k)$. We stress that the example points provided to the learning algorithm all lie in $\{0, \ldots, a_k - 1\}$ (so certainly any reasonable hypothesis distribution should also be supported on $\{0, \ldots, a_k - 1\}$). Such a reduction is useful for our lower bounds because it enables us to prove a lower bound for learning $\sum_{i=1}^{k} a_i \cdot \mathbf{S}_i$ by proving a lower bound for learning $\sum_{i=1}^{k-1} a_i \cdot \mathbf{S}_i \mod a_k$.

The high level idea of this reduction is fairly simple so we sketch it here. Let $\mathbf{S} = a_1 \cdot \mathbf{S}_1 + \cdots + a_{k-1} \cdot \mathbf{S}_{k-1}$ be a weighted sum of PBDs such that $(\mathbf{S} \mod a_k)$ is the target distribution to be learned and let N be the total number of summands in all of the PBDs. Let $\mathbf{S}_k$ be an independent PBD with mean and variance $\Omega(N^\star)$. The key insight is that by taking $N^\star$ sufficiently large relative to N, the distribution of $(\mathbf{S} \mod a_k) + a_k \cdot \mathbf{S}_k$ (which can easily be simulated by the learner given access to draws from $(\mathbf{S} \mod a_k)$ since it can generate samples from $a_k \cdot \mathbf{S}_k$ by itself) can be shown to be statistically very close to that of $\mathbf{S}' := \mathbf{S} + a_k \cdot \mathbf{S}_k$. Here is an intuitive justification: We can think of the different possible outcomes of $a_k \cdot \mathbf{S}_k$ as dividing the support of $\mathbf{S}'$ into bins of width a_k. Sampling from $\mathbf{S}'$ can be performed by picking a bin boundary (a draw from $a_k \cdot \mathbf{S}_k$) and an offset $\mathbf{S}$. While adding $\mathbf{S}$ may take the sample across multiple bin boundaries, if $\mathbf{Var}[\mathbf{S}_k]$ is sufficiently large, then adding $\mathbf{S}$ typically takes $a_k \cdot \mathbf{S}_k + \mathbf{S}$ across a small fraction of the bin boundaries. Thus, the conditional distribution given membership in a bin is similar between bins that have high probability under $\mathbf{S}'$, which means that all of these conditional distributions are similar to the distribution of $\mathbf{S}' \mod a_k$ (which is a mixture of them). Finally, $(\mathbf{S}' \mod a_k)$ has the same distribution as $(\mathbf{S} \mod a_k)$. Thus, given samples from $(\mathbf{S} \mod a_k)$, the learner can essentially simulate samples from $\mathbf{S}'$. However, $\mathbf{S}'$ is is a weighted sum of k PBDs, which by the assumption of our reduction theorem can be learned efficiently. Now, assuming the learner has a hypothesis $\mathbf{H}$ such that $d_{\mathrm{TV}}(\mathbf{H}, \mathbf{S}') \leq \varepsilon$, it immediately follows that $d_{\mathrm{TV}}((\mathbf{H} \mod a_k), (\mathbf{S}' \mod a_k)) \leq d_{\mathrm{TV}}(\mathbf{H}, \mathbf{S}') \leq \varepsilon$ as desired.

Proof overview of Theorem 5. At this point we have the task of proving a lower bound for learning weighted PBDs over $\{0, a_1, a_2\}$ mod a_3. We establish such a lower bound using Fano's inequality (stated precisely as Theorem 28 in Section 4). To get a sample complexity lower bound of $\Omega(\log \log a_3)$ from Fano's inequality, we must construct $T = \log^{\Omega(1)} a_3$ distributions $\mathbf{S}_1, \ldots, \mathbf{S}_T$, where each $\mathbf{S}_i$ is a weighted PBD on $\{0, a_1, a_2\}$ modulo a_3, meeting the following requirements: $d_{\mathrm{TV}}(\mathbf{S}_i, \mathbf{S}_j) = \Omega(1)$ if $i \neq j$, and $D_{KL}(\mathbf{S}_i || \mathbf{S}_j) = O(1)$ for all $i, j \in T$. In other words, applying Fano's inequality requires us to exhibit

a large number of distributions (belonging to the family for which we are proving the lower bound) such that any two distinct distributions in the family are *far* in total variation distance but *close* in terms of KL-divergence. The intuitive reason for these two competing requirements is that if $\mathbf{S}_i$ and $\mathbf{S}_j$ are 2ε-far in total variation distance, then a successful algorithm for learning to error at most ε must be able to distinguish $\mathbf{S}_i$ and $\mathbf{S}_j$. On the other hand, if $\mathbf{S}_i$ and $\mathbf{S}_j$ are close in KL divergence, then it is difficult for any learning algorithm to distinguish between $\mathbf{S}_i$ and $\mathbf{S}_j$.

Now we present the high-level idea of how we may construct distributions $\mathbf{S}_1, \mathbf{S}_2, \ldots$ with the properties described above to establish Theorem 5. The intuitive description of $\mathbf{S}_i$ that we give below does not align perfectly with our actual construction, but this simplified description is hopefully helpful in getting across the main idea.

For the construction we fix $a_1 = 1$, $a_2 = p$ and $a_3 = q$. (We discuss how p and q are selected later; this is a crucial aspect of our construction.) The i-th distribution $\mathbf{S}_i$ is $\mathbf{S}_i = \mathbf{U}_i + p\mathbf{V}_i \bmod q$; we describe the distribution $\mathbf{S}_i = \mathbf{U}_i + p\mathbf{V}_i \mod q$ in two stages, first by describing each $\mathbf{V}_i$, and then by describing the corresponding $\mathbf{U}_i$. In the actual construction $\mathbf{U}_i$ and $\mathbf{V}_i$ will be shifted binomial distributions. Since a binomial distribution is rather flat within one standard deviation of its mean, and decays exponentially after that, it is qualitatively somewhat like the uniform distribution over an interval; for this intuitive sketch it is helpful to think of $\mathbf{U}_i$ and $\mathbf{V}_i$ as actually being uniform distributions over intervals. We take the support of $\mathbf{V}_1$ to be an interval of length q/p, so that adjacent members of the support of $(p\mathbf{V}_1 \mod q)$ will be at distance p apart from each other. More generally, taking $\mathbf{V}_i$ to be uniform over an interval of length $2^{i-1}q/p$, the average gap between adjacent members of $\mathrm{supp}(p\mathbf{V}_i \mod q)$ is of length essentially $p/2^{i-1}$, and by a careful choice of p relative to q one might furthermore hope that the gaps would be "balanced", so that they are all of length roughly $p/2^{i-1}$. (This "careful choice" is the technical heart of our actual construction presented later.)

How does $\mathbf{U}_i$ enter the picture? The idea is to take each $\mathbf{U}_i$ to be uniform over a *short* interval, of length $3p/2^i$. This "fills in each gap" and additionally "fills in the first half of the following gap;" as a result, the first half of each gap ends up with twice the probability mass of the second half. (As a result, every two points have probability mass within a constant factor of each other under every distribution — in fact, any point under any one of our distributions has probability mass within a constant factor of that of any other point under any other one of our distributions. This gives the $D_{KL}(\mathbf{S}_i||\mathbf{S}_j) \leq O(1)$ upper bound mentioned above.) For example, recalling that the "gaps" in $\mathrm{supp}(p\mathbf{V}_1 \mod q)$ are of length p, choosing $\mathbf{U}_1$ to be uniform over $\{1, \ldots, 3p/2\}$ will fill in each gap along with the first half of the following gap. Intuitively, each $\mathbf{S}_i = \mathbf{U}_i + p\mathbf{V}_i$ is a "striped" distribution, with equal-width "light stripes" (of uniformly distributed smaller mass) and "dark stripes" (of uniformly distributed larger mass), and each $\mathbf{S}_{i+1}$ has stripes of width half of the $\mathbf{S}_i$-sum's stripes. Roughly speaking, two such distributions $\mathbf{S}_i$ and $\mathbf{S}_j$ "overlap enough" (by a constant fraction) so that they are difficult to distinguish; however they are also "distinct enough" that a successful learning algorithm must be able to distinguish which $\mathbf{S}_i$ its samples are drawn from in order to generate a high-accuracy hypothesis.

We now elaborate on the careful choice of p and q that was mentioned above. The critical part of this choice of p and q is that for $i \geq 1$, in order to get "evenly spaced gaps," the remainders of $p \cdot s$ modulo q where $s \in \{1, \ldots, 2^{i-1}q/p\}$ should be roughly evenly spaced, or *equidistributed*, in the group $\mathbb{Z}_q$. Here the notion of "evenly spaced" is with respect to the "wrap-around" distance (also known as the *Lee metric*) on the group $\mathbb{Z}_q$ (so, for example, the wrap-around distance between 1 and 2 is 1, whereas the wrap-around distance between $q-1$ and 1 is 2). Roughly speaking, we would like $p \cdot s$ modulo q to be equidistributed in $\mathbb{Z}_q$ when $s \in \{1, \ldots, 2^{i-1}q/p\}$, for a range of successive values of i (the more the better, since this means more distributions in our hard family and a stronger lower bound). Thus, qualitatively, we would like the

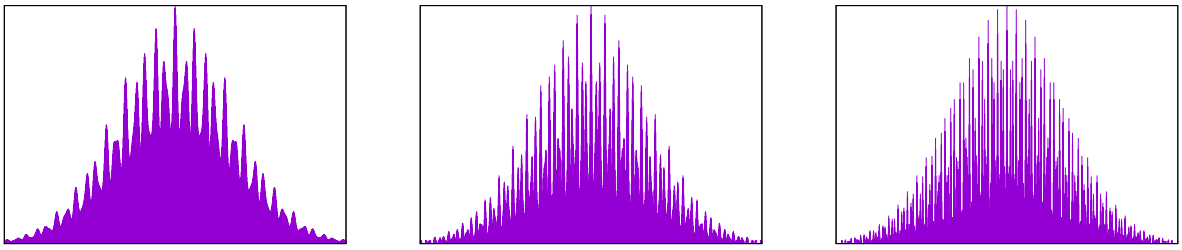
remainders of p modulo q to be *equidistributed at several scales.* We note that equidistribution phenomena are well studied in number theory and ergodic theory, see e.g. [Tao14].

While this connection to equidistribution phenomena is useful for providing visual intuition (at least to the authors), in our attempts to implement the construction using powers of two that was just sketched, it seemed that in order to control the errors that arise in fact a *doubly* exponential growth was required, leading to the construction of only $\Theta(\log\log q)$ such distributions and hence a $\Omega(\log\log\log q)$ sample complexity lower bound. Thus to achieve an $\Omega(\log\log q)$ sample complexity lower bound, our actual choice of p and q comes from the theory of continued fractions. In particular, we choose p and q so that p/q has a continued fraction representation with "many" ($\Theta(\log q)$, though for technical reasons we use only $\log^{\Theta(1)} q$ many) convergents that grow relatively slowly. These $T = \log^{\Theta(1)} q$ convergents translate into T distributions $\mathbf{S}_1, \ldots, \mathbf{S}_T$ in our "hard family" of distributions, and thus into an $\Omega(\log\log q)$ sample lower bound via Fano's inequality.

The key property that we use is a well-known fact in the theory of continued fractions: if g_i/h_i is the i^{th} convergent of a continued fraction for p/q, then $|g_i/h_i - p/q| \leq 1/(h_i \cdot h_{i+1})$. In other words, the i^{th} convergent g_i/h_i provides a non-trivially good approximation of p/q (note that getting an error of $1/h_i$ would have been trivial). From this property, it is not difficult to see that the remainders of $p \cdot \{1, \ldots, h_i\}$ are roughly equidistributed modulo q.

Thus, a more accurate description of our (still idealized) construction is that we choose $\mathbf{V}_i$ to be uniform on $\{1, \ldots, h_i\}$ and $\mathbf{U}_i$ to be uniform on roughly $\{1, \ldots, (3/2) \cdot (q/h_i)\}$. So as to have as many distributions as possible in our family, we would like $h_i \approx (q/p) \cdot c^i$ for some fixed $c > 1$. This can be ensured by choosing p, q such that all the numbers appearing in the continued fraction representation of p/q are bounded by an absolute constant; in fact, in the actual construction, we simply take p/q to be a convergent of $1/\phi$ where ϕ is the golden ratio. With this choice we have that the i^{th} convergent of the continued fraction representation of $1/\phi$ is g_i/h_i, where $h_i \approx ((\sqrt{5}+1)/2)^i$. This concludes our informal description of the choice of p and q.

Again, we note that in our actual construction (see Figure 1), we cannot use uniform distributions over intervals (since we need to use PBDs), but rather we have shifted binomial distributions. This adds some technical complication to the formal proofs, but the core ideas behind the construction are indeed as described above.



(a) (b) (c)

Figure 1: Examples of targets used in our lower bound construction for Theorem 5. Very roughly, the distribution in (b) has peaks where the distribution (a) does, plus a constant factor more peaks. To compensate, its peaks are thinner. The distribution (c) has still more, still thinner, peaks.

Proof overview of Theorem 6. As mentioned earlier, Theorem 6 also uses our reduction from the modular

learning problem. Taking $a_1 = 0$ and $a_3 \approx a_{\max}$ to be "known" to the learner, we show that any algorithm for learning a distribution of the form $(a_2\mathbf{S}_2 \mod a_3)$, where $0 < a_2 < a_3$ and a_2 is unknown to the learner and $\mathbf{S}_2$ is a PBD_N, must use $\Omega(\log a_3)$ samples. Like Theorem 5, we prove this using Fano's inequality, by constructing a "hard family" of $(a_3)^{\Omega(1)}$ many distributions of this type such that any two distinct distributions in the family have variation distance $\Omega(1)$ but KL-divergence $O(1)$.

We sketch the main ideas of our construction, starting with the upper bound on KL-divergence. The value a_3 is taken to be a prime. The same PBD_N distribution $\mathbf{S}_2$, which is simply a shifted binomial distribution and may be assumed to be "known" to the learner, is used for all of the distributions in the "hard family", so different distributions in this family differ only in the value of a_2. The shifted binomial distribution $\mathbf{S}_2$ is taken to have variance $\Theta((a_3)^2)$, so, very roughly, $\mathbf{S}_2$ assigns significant probability on $\Theta(a_3)$ distinct values. From this property, it is not difficult to show (similar to our earlier discussion) that any point in the domain $\{0, 1, \ldots, a_3 - 1\}$ under any one of our distributions has probability mass within a constant factor of that of any other point under any other one of our distributions (where the constant factor depends on the hidden constant in the $\Theta((a_3)^2)$). This gives the required $O(1)$ upper bound on KL-divergence.

It remains to sketch the $\Omega(1)$ lower bound on variation distance. As in our discussion of the Theorem 5 lower bound, for intuition it is convenient to think of the shifted binomial distribution $\mathbf{S}_2$ as being uniform over an interval of the domain $\{0, 1, \ldots, a_3 - 1\}$; by carefully choosing the variance and offset of this shifted binomial, we may think of this interval as being $\{0, 1, \ldots, r - 1\}$ for $r = \kappa a_3$ for some small constant $\kappa > 0$ (the constant κ again depends on the hidden constant in the $\Theta((a_3)^2)$) value of the variance). So for the rest of our intuitive discussion we view the distributions in the hard family as being of the form $(a_2 \cdot \mathbf{U}_r \mod a_3)$ where $\mathbf{U}_r$ is uniform over $\{0, 1, \ldots, r - 1\}$, $r = \kappa a_3$.

Recalling that a_3 is prime, it is clear that for any $0 < a_2 < a_3$, the distribution $(a_2 \cdot \mathbf{U}_r \mod a_3)$ is uniform over an $(r = \kappa a_3)$-element subset of $\{0, \ldots, a_3 - 1\}$. If a_2 and a_2' are two independent uniform random elements from $\{1, \ldots, a_3 - 1\}$, then since κ is a small constant, intuitively the overlap between the supports of $(a_2 \cdot \mathbf{U}_r \mod a_3)$ and $(a_2' \cdot \mathbf{U}_r \mod a_3)$ should be small, and consequently the variation distance between these two distributions should be large. This in turn suggests that by drawing a large random set of values for a_2, it should be possible to obtain a large family of distributions of the form $(a_2 \cdot \mathbf{U}_r \mod a_3)$ such that any two of them have large variation distance. We make this intuition precise using a number-theoretic equidistribution result of Shparlinski [Shp08] and a probabilistic argument showing that indeed a random set of $(a_3)^{1/3}$ choices of a_2 is likely to have the desired property. This gives a "hard family" of size $(a_3)^{1/3}$, leading to an $\Omega(\log a_3) = \Omega(\log a_{\max})$ lower bound via Fano's inequality. As before some technical work is required to translate these arguments for the uniform distribution over to the shifted binomial distributions that we actually have to work with, but we defer these technical details to Section 13.

4 Preliminaries

4.1 Basic notions and useful tools from probability.

Distributions. We will typically ignore the distinction between a random variable and its distribution. We use bold font $\mathbf{X}_i, \mathbf{S}$, etc. to denote random variables (and also distributions).

For a distribution $\mathbf{X}$ supported on the integers we write $\mathbf{X}(i)$ to denote the value $\mathbf{Pr}[\mathbf{X} = i]$ of the probability density function of $\mathbf{X}$ at point i, and $\mathbf{X}(\leq i)$ to denote the value $\mathbf{Pr}[\mathbf{X} \leq i]$ of the cumulative density function of $\mathbf{X}$ at point i. For $S \subseteq \mathbb{Z}$, we write $\mathbf{X}(S)$ to denote $\sum_{i \in S} \mathbf{X}(i)$ and $\mathbf{X}_S$ to denote the

conditional distribution of $\mathbf{X}$ restricted to S.

Total Variation Distance. Recall that the *total variation distance* between two distributions $\mathbf{X}$ and $\mathbf{Y}$ over a countable set D is

$$d_{\mathrm{TV}}(\mathbf{X}, \mathbf{Y}) := \frac{1}{2} \cdot \sum_{\alpha \in D} |\mathbf{X}(\alpha) - \mathbf{Y}(\alpha)| = \max_{S \subseteq D} [\mathbf{X}(S) - \mathbf{Y}(S)],$$

with analogous definitions for pairs of distributions over $\mathbb{R}$, over $\mathbb{R}^k$, etc. Similarly, if $\mathbf{X}$ and $\mathbf{Y}$ are two random variables ranging over a countable set, their total variation distance $d_{\mathrm{TV}}(\mathbf{X}, \mathbf{Y})$ is defined as the total variation distance between their distributions. We sometimes write "$\mathbf{X} \overset{\varepsilon}{\approx} \mathbf{Y}$" as shorthand for "$d_{\mathrm{TV}}(\mathbf{X}, \mathbf{Y}) \leq \varepsilon$".

For $\mathbf{X}$ and $\mathbf{Y}$ with $d_{\mathrm{TV}}(\mathbf{X}, \mathbf{Y}) \leq \varepsilon$, the following coupling lemma justifies thinking of a draw from $\mathbf{Y}$ as being obtained by making a draw from $\mathbf{X}$, and modifying it with probability at most ε.

Lemma 7 ([Lin02])**.** *For random variables $\mathbf{X}$ and $\mathbf{Y}$ with $d_{\mathrm{TV}}(\mathbf{X}, \mathbf{Y}) \leq \varepsilon$, there is a joint distribution whose marginals are $\mathbf{X}$ and $\mathbf{Y}$ such that, with probability at least $1 - \varepsilon$, $\mathbf{X} = \mathbf{Y}$.*

Shift-invariance. Let $\mathbf{X}$ be a finitely supported real-valued random variable. For an integer k we write $d_{\mathsf{shift},k}(\mathbf{X})$ to denote $d_{\mathrm{TV}}(\mathbf{X}, \mathbf{X} + k)$. We say that $\mathbf{X}$ is *α-shift-invariant at scale k* if $d_{\mathsf{shift},k}(\mathbf{X}) \leq \alpha$; if $\mathbf{X}$ is α-shift-invariant at scale 1 then we sometimes simply say that $\mathbf{X}$ is *α-shift-invariant*. We will use the following basic fact:

Fact 8. *1. If $\mathbf{X}, \mathbf{Y}$ are independent random variables then $d_{\mathsf{shift},k}(\mathbf{X} + \mathbf{Y}) \leq d_{\mathsf{shift},k}(\mathbf{X})$.*

2. Let $\mathbf{X}$ be α-shift-invariant at scale p and $\mathbf{Y}$ (independent from $\mathbf{X}$) be β-shift-invariant at scale q. Then $\mathbf{X} + \mathbf{Y}$ is both α-shift-invariant at scale p and β-shift-invariant at scale q.

Kolmogorov Distance and the DKW Inequality. Recall that the *Kolmogorov distance* $d_{\mathrm{K}}(\mathbf{X}, \mathbf{Y})$ between probability distributions over the integers is

$$d_{\mathrm{K}}(\mathbf{X}, \mathbf{Y}) := \max_{j \in \mathbb{Z}} |\mathbf{Pr}[\mathbf{X} \leq j] - \mathbf{Pr}[\mathbf{Y} \leq j]|,$$

and hence for any interval $I = \{a, a+1, \ldots, a+b\} \subset \mathbb{Z}$ we have that

$$|\mathbf{Pr}[\mathbf{X} \in I] - \mathbf{Pr}[\mathbf{Y} \in I]| \leq 2 d_{\mathrm{K}}(\mathbf{X}, \mathbf{Y}).$$

Learning any distribution with respect to the Kolmogorov distance is relatively easy, which follows from the *Dvoretzky-Kiefer-Wolfowitz* (DKW) inequality. Let $\widehat{\mathbf{X}}_m$ denote the empirical distribution of m i.i.d. samples drawn from $\mathbf{X}$. The DKW inequality states that for $m = \Omega((1/\varepsilon^2) \cdot \ln(1/\delta))$, with probability $1 - \delta$ (over the draw of m samples from $\mathbf{X}$) the empirical distribution $\widehat{\mathbf{X}}_m$ will be ε-close to $\mathbf{X}$ in Kolmogorov distance:

Theorem 9 ([DKW56, Mas90])**.** *Let $\widehat{\mathbf{X}}_m$ be an empirical distribution of m samples from distribution $\mathbf{X}$ over the integers. Then for all $\varepsilon > 0$, we have*

$$\mathbf{Pr}[d_{\mathrm{K}}(\mathbf{X}, \widehat{\mathbf{X}}_m) > \varepsilon] \leq 2e^{-2m\varepsilon^2}.$$

Convolving with an α-shift invariant distribution can "spread the butter" to transform distributions that are close w.r.t. Kolmogorov distance into distributions that are close with respect to the more demanding total variation distance. The following lemma makes this intuition precise:

Lemma 10 ([GMRZ11]). *Let $\mathbf{Y}, \mathbf{Z}$ be distributions supported on the integers and $\mathbf{X}$ be an α-shift invariant distribution that is independent of $\mathbf{Y}, \mathbf{Z}$. Then for any a, b such that $d_{\mathrm{K}}(\mathbf{Y}, \mathbf{Z}) \leq \alpha b$, we have*

$$d_{\mathrm{TV}}(\mathbf{Y}+\mathbf{X}, \mathbf{Z}+\mathbf{X}) = O(\sqrt{d_{\mathrm{K}}(\mathbf{Y}, \mathbf{Z}) \cdot \alpha \cdot b}) + \mathbf{Pr}[\mathbf{Y} \notin [a, a+b)] + \mathbf{Pr}[\mathbf{Z} \notin [a, a+b)].$$

We will also require a multidimensional generalization of Kolmogorov distance and of the DKW inequality. Given probability distributions $\mathbf{X}, \mathbf{Y}$ over $\mathbb{Z}^d$, the Kolmogorov distance between $\mathbf{X}$ and $\mathbf{Y}$ is

$$d_{\mathrm{K}}(\mathbf{X}, \mathbf{Y}) := \max_{(j_1,\ldots,j_d)\in\mathbb{Z}^d} |\,\mathbf{Pr}[\mathbf{X}_i \leq j_i \text{ for all } i \in [d]] - \mathbf{Pr}[\mathbf{Y} \leq j_i \text{ for all } i \in [d]]|,$$

and so for any axis-aligned rectangle $R = \prod_{i=1}^{d} \{a_i, \ldots, a_i + b_i\} \subset \mathbb{Z}^d$ we have

$$|\,\mathbf{Pr}[\mathbf{X} \in R] - \mathbf{Pr}[\mathbf{Y} \in R]| \leq 2^d d_{\mathrm{K}}(\mathbf{X}, \mathbf{Y}).$$

We will use the following generalization of the DKW inequality to the multidimensional setting.

Lemma 11 ([Tal94]). *Let $\widehat{\mathbf{X}}_m$ be an empirical distribution of m samples from distribution $\mathbf{X}$ over $\mathbb{Z}^d$. There are absolute constants c_1, c_2 and c_3 such that, for all $\varepsilon > 0$, for all $m \geq c_1 d/\varepsilon^2$,*

$$\mathbf{Pr}[d_{\mathrm{K}}(\mathbf{X}, \widehat{\mathbf{X}}_m) > \varepsilon] \leq c_2^d e^{-c_3 \varepsilon^2 m}.$$

Covers. Let $\mathcal{P}$ denote a set of distributions over the integers. Given $\delta > 0$, a set of distributions $\mathcal{Q}$ is said to be a *δ-cover of $\mathcal{P}$* (w.r.t. the total variation distance) if for every distribution $\mathbf{P}$ in $\mathcal{P}$ there exists some distribution $\mathbf{Q}$ in $\mathcal{Q}$ such that $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) \leq \delta$. We sometimes say that distributions $\mathbf{P}, \mathbf{Q}$ are *δ-neighbors* if $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) \leq \delta$, or that $\mathbf{P}$ and $\mathbf{Q}$ are *δ-close*.

Support and essential support. We write $\mathrm{supp}(\mathbf{P})$ to denote the support of distribution $\mathbf{P}$. Given a distribution $\mathbf{P}$ over the integers, we say that $\mathbf{P}$ is *τ-essentially supported on $S \subset \mathbb{Z}$* if $\mathbf{P}(S) \geq 1 - \tau$.

4.2 The distributions we work with.

We recall the definition of an $\mathcal{A}$-sum and give some related definitions. For $0 \leq a_1 < \ldots < a_k$ and $\mathcal{A} = \{a_1, \ldots, a_k\}$, a $\mathcal{A}$-sum is a distribution $\mathbf{S} = \sum_{i=1}^{N} \mathbf{X}_i$ where the $\mathbf{X}_i$'s are independent integer random variables (*not* assumed to be identically distributed) all of which are supported on the same set of integer values $a_1 < a_2 < \cdots < a_k \in \mathbb{Z}_{\geq 0}$. A *Poisson Binomial Distribution*, or PBD_N, is a $\{0,1\}$-sum.

A *weighted sum of PBDs* is a distribution $\mathbf{S} = a_2\mathbf{S}_2 + \cdots + a_k\mathbf{S}_k$ where each $\mathbf{S}_i$ is an independent PBD_{N_i} and $N_2 + \cdots + N_k = N$. Equivalently we have that $\mathbf{S} = \sum_{i=1}^{N} \mathbf{X}_i$ where N_2 of the $\mathbf{X}_i$'s are supported on $\{0, a_2\}$, N_3 are supported on $\{0, a_3\}$, and so on.

Let us say that a *signed PBD_N* is a random variable $\mathbf{S} = \sum_{i=1}^{N} \mathbf{X}_i$ where the $\mathbf{X}_i$'s are independent and each is either supported on $\{0,1\}$ or is supported on $\{0,-1\}$. We defined a weighted sum of signed PBDs analogously to the unsigned case.

Finally, we say that an integer valued random variable $\mathbf{X}$ has *mode 0* if $\mathbf{Pr}[\mathbf{X} = 0] \geq \mathbf{Pr}[\mathbf{X} = b]$ for all $b \in \mathbb{Z}$.

Translated Poisson Distributions and Discretized Gaussians. We will make use of the translated Poisson distribution for approximating signed PBDs with large variance.

Definition 12 ([RÖ07])**.** We say that an integer random variable $\mathbf{Y}$ is distributed according to the *translated Poisson distribution with parameters μ and σ^2*, denoted $TP(\mu, \sigma^2)$, iff $\mathbf{Y}$ can be written as

$$\mathbf{Y} = \lfloor \mu - \sigma^2 \rfloor + \mathbf{Z},$$

where $\mathbf{Z}$ is a random variable distributed according to $\mathrm{Poisson}(\sigma^2 + \{\mu - \sigma^2\})$, where $\{\mu - \sigma^2\}$ represents the fractional part of $\mu - \sigma^2$.

The following lemma gives a useful bound on the variation distance between a signed PBD and a suitable translated Poisson distribution.

Lemma 13. *Let $\mathbf{S}$ be a signed PBD$_N$ with mean μ and variance $\sigma^2 \geq 1$. Then*

$$d_{\mathrm{TV}}\left(\mathbf{S}, TP(\mu, \sigma^2)\right) \leq O(1/\sigma).$$

Proof. Without loss of generality we may suppose that $\mathbf{S} = \mathbf{X}_1 + \cdots + \mathbf{X}_N$ where $\mathbf{X}_1, \ldots, \mathbf{X}_M$ are supported on $\{0, -1\}$ with $\mathbf{E}[\mathbf{X}_i] = -p_i$ for $i \leq M$, and $\mathbf{X}_{M+1}, \ldots, \mathbf{X}_N$ are supported on $\{0, 1\}$ with $\mathbf{E}[\mathbf{X}_i] = p_i$ for $i > M$. Let $\mathbf{X}'_i = \mathbf{X}_i + 1$ for $1 \leq i \leq M$, so $\mathbf{S}' := \mathbf{X}'_1 + \cdots + \mathbf{X}'_M + \mathbf{X}_{M+1} + \cdots + \mathbf{X}_N$ are independent Bernoulli random variables where $\mathbf{E}[\mathbf{X}'_i] = 1 - p_i$ for $i \leq M$.

[RÖ07] (specifically equation (3.4)) shows that if $\mathbf{J}_1, \ldots, \mathbf{J}_N$ are independent Bernoulli random variables with $\mathbf{E}[\mathbf{J}_i] = p_i$, then

$$d_{\mathrm{TV}}\left(\sum_{i=1}^{N} \mathbf{J}_i, TP(\mu, \sigma^2)\right) \leq \frac{\sqrt{\sum_{i=1}^{N} p_i^3(1-p_i)} + 2}{\sum_{i=1}^{N} p_i(1-p_i)}$$

where $\mu = \sum_{i=1}^{N} p_i$. Applying this to $\mathbf{S}'$, we see that for $\mu' = \mathbf{E}[\mathbf{S}']$, we have

$$\begin{aligned} d_{\mathrm{TV}}\left(\mathbf{S}', TP(\mu', \sigma^2)\right) &\leq \frac{\sqrt{\sum_{i=1}^{M} p_i(1-p_i)^3 + \sum_{i=M+1}^{N} p_i^3(1-p_i)} + 2}{\sum_{i=1}^{N} p_i(1-p_i)} \\ &\leq \frac{\sqrt{\sum_{i=1}^{N} p_i(1-p_i)} + 2}{\sum_{i=1}^{N} p_i(1-p_i)} \leq O(1/\sigma). \end{aligned}$$

The claimed bound follows from this on observing that $\mathbf{S}'$ is a translation of $\mathbf{S}$ by M and $TP(\mu', \sigma^2)$ is likewise a translation of $TP(\mu, \sigma^2)$ by M. □

The following bound on the total variation distance between translated Poisson distributions will be useful.

Lemma 14 (Lemma 2.1 of [BL06])**.** *For $\mu_1, \mu_2 \in \mathbb{R}$ and $\sigma_1^2, \sigma_2^2 \in \mathbb{R}_+$ with $\lfloor \mu_1 - \sigma_1^2 \rfloor \leq \lfloor \mu_2 - \sigma_2^2 \rfloor$, we have*

$$d_{\mathrm{TV}}(TP(\mu_1, \sigma_1^2), TP(\mu_2, \sigma_2^2)) \leq \frac{|\mu_1 - \mu_2|}{\sigma_1} + \frac{|\sigma_1^2 - \sigma_2^2| + 1}{\sigma_1^2}.$$

We will also use discretized Gaussians, both real-valued and vector-valued (i.e. multidimensional). A draw from the *discretized Gaussian* $\mathcal{N}_D(\mu, \sigma^2)$ is obtained by making a draw from the normal distribution $\mathcal{N}(\mu, \sigma)$ and rounding to the nearest integer. We refer to μ and σ^2 respectively as the "underlying mean" and "underlying variance" of $\mathcal{N}_D(\mu, \sigma)$. Similarly, a draw from the *multidimensional discretized Gaussian*

$\mathcal{N}_D(\mu, \Sigma)$ is obtained by making a draw from the multidimensional Gaussian $\mathcal{N}(\mu, \Sigma)$ with mean vector μ and covariance matrix Σ and rounding each coordinate to the nearest integer. To avoid confusion we will always explicitly write "multidimensional" when dealing with a multidimensional Gaussian.

We recall some simple facts about the variation distance between different discretized Gaussian distributions (see Appendix B of the full version of [DDO+13]):

Lemma 15 (Proposition B.5 of [DDO+13])**.** *Let $\mathbf{G}$ be distributed as $\mathcal{N}(\mu, \sigma^2)$ and let $\lambda \in \mathbb{R}$. Then $d_{\mathrm{TV}}(\lfloor \mathbf{G} + \lambda \rfloor, \lfloor \mathbf{G} \rfloor + \lfloor \lambda \rfloor) \leq \frac{1}{2\sigma}$.*

The same argument that gives Lemma 15 also gives the following small extension:

Lemma 16. *Let $\mathbf{G}$ be distributed as $\mathcal{N}(\mu, \sigma^2)$ and let $\lambda \in \mathbb{R}, \rho \in \mathbb{Z}$. Then $d_{\mathrm{TV}}(\lfloor \mathbf{G} + \lambda \rfloor, \lfloor \mathbf{G} \rfloor + \rho) \leq \frac{|\rho - \lambda|}{2\sigma}$.*

We will use the following theorem about approximation of signed PBDs.

Theorem 17 ([CGS11] Theorem 7.1[3])**.** *For $\mathbf{S}$ a signed PBD, $d_{\mathrm{TV}}(\mathbf{S}, \mathcal{N}_D(\mu, \sigma^2)) \leq O(1/\sigma)$ where $\mu = \mathbf{E}[\mathbf{S}]$ and $\sigma^2 = \mathbf{Var}[\mathbf{S}]$.*

The following is a consequence of Theorem 17 and Lemma 16 which we explicitly record for later reference:

Fact 18. *Let $\mathbf{S}$ be a signed PBD with $\mathbf{Var}[\mathbf{S}] = \sigma_{\mathbf{S}}^2$. Then $\mathbf{S}$ is τ-shift-invariant at scale 1 for $\tau = O(1/\sigma_{\mathbf{S}})$, and hence for any integer c, the distribution $c\mathbf{S}$ is τ-shift-invariant at scale c.*

We also need a central limit theorem for multinomial distributions. We recall the following result, which is a direct consequence of the "size-free CLT" for Poisson Multinomial Distributions in [DDKT16]. (Below we write $\mathbf{e}_i$ to denote the real vector in $\{0,1\}^d$ that has a 1 only in the i-th coordinate.)

Theorem 19. *Let $\mathbf{X}_1, \ldots, \mathbf{X}_N$ be independent $\mathbb{Z}^d$-valued random variables where the support of each $\mathbf{X}_i$ is contained in the set $\{0, \pm\mathbf{e}_1, \ldots, \pm\mathbf{e}_d\}$. Let $\mathbf{M} = \mathbf{X}_1 + \ldots + \mathbf{X}_N$. Then we have*

$$d_{\mathrm{TV}}(\mathbf{M}, \mathcal{N}_D(\mu, \Sigma)) \leq O\left(\frac{d^{7/2}}{\sigma^{1/10}}\right),$$

where $\mu = \mathbf{E}[\mathbf{M}]$ is the mean and Σ is the $d \times d$ covariance matrix of $\mathbf{S}$, and σ^2 is the minimum eigenvalue of Σ.

Covers and structural results for PBDs. Our proof of Theorem 4, which is about learning PBDs that have been subject to an unknown shifting and scaling, uses the fact that for any ε there is a "small cover" for the set of all PBD_N distributions. We recall the following from [DP14]:

Theorem 20 (Cover for PBDs)**.** *Let $\mathbf{S}$ be any PBD_N distribution. Then for any $\varepsilon > 0$, we have that either*

- *$\mathbf{S}$ is ε-essentially supported on an interval of $O(1/\varepsilon^3)$ consecutive integers (in this case we say that $\mathbf{S}$ is in* sparse form*); or if not,*
- *$\mathbf{S}$ is ε-close to some distribution $u + \mathrm{Bin}(\ell, q)$ where $u, \ell \in \{0, 1, \ldots, N\}$, and $\mathbf{Var}[\mathrm{Bin}(\ell, q)] = \Omega(1/\varepsilon^2)$ (in this case we say that $\mathbf{S}$ is in $1/\varepsilon$-*heavy Binomial form*).*

[3] The theorem in [CGS11] is stated only for PBDs, but the result for signed PBDs is easily derived from the result for PBDs via a simple translation argument similar to the proof of Lemma 13.

We recall some well-known structural results on PBDs that are in $1/\varepsilon$-heavy Binomial form (see e.g. [CGS11], Theorem 7.1 and p. 231):

Fact 21. *Let $\mathbf{Y}$ be a PBD$_N$ distribution that is in $1/\varepsilon$-heavy Binomial form as described in Theorem 20. Then*

1. $d_{\mathrm{TV}}(\mathbf{Y}, \mathbf{Z}) = O(\varepsilon)$*, where $\mathbf{Z}$ is a discretized $\mathcal{N}(\mathbf{E}[\mathbf{Y}], \mathbf{Var}[\mathbf{Y}])$ Gaussian.*
2. $d_{\mathsf{shift},1}(\mathbf{Y}) = O(\varepsilon)$.

4.3 Extension of the Barbour-Xia coupling lemma

In [BX99], Barbour and Xia proved the following lemma concerning the shift-invariance of sums of independent integer random variables.

Lemma 22 ([BX99], Proposition 4.6)**.** *Let $\mathbf{X}_1, \ldots, \mathbf{X}_N$ be N independent integer valued random variables and let $\mathbf{S} = \mathbf{X}_1 + \ldots + \mathbf{X}_N$. Let $d_{\mathsf{shift},1}(\mathbf{X}_i) \leq 1 - \delta_i$. Then,*

$$d_{\mathsf{shift},1}(\mathbf{S}) \leq O\left(\frac{1}{\sqrt{\sum_{i=1}^{N} \delta_i}}\right).$$

We require a $d_{\mathsf{shift},p}$ analogue of this result. To obtain such an analogue we first slightly generalize the above lemma so that it does not require $\mathbf{X}_i$ to be supported on $\mathbb{Z}$. The proof uses a simple reduction to the integer case.

Claim 23. *Let $\mathbf{X}_1, \ldots, \mathbf{X}_N$ be N independent finitely supported random variables and let $\mathbf{S} = \mathbf{X}_1 + \ldots + \mathbf{X}_N$. Let $d_{\mathsf{shift},1}(\mathbf{X}_i) \leq 1 - \delta_i$. Then,*

$$d_{\mathsf{shift},1}(\mathbf{S}) \leq O\left(\frac{1}{\sqrt{\sum_{i=1}^{N} \delta_i}}\right).$$

Proof. Assume that for any i, the support of $\mathbf{X}_i$ is of size at most k and is supported in the interval $[-k, k]$. (By the assumption of finite support this must hold for some integer k.) Given any $\mathbf{X}_i$, create a new random variable $\mathbf{Y}_i$ which is defined as follows: First, let us partition the support of $\mathbf{X}_i$ by putting two outcomes into the same cell whenever the difference between them is an integer. Let $S_1^{(i)}, \ldots, S_{k'}^{(i)}$ be the non-empty cells, so $k' \leq k$, and, for each $S_j^{(i)}$, there is a real β_j such that $S_j^{(i)} \subseteq \{\beta_j + \ell : \ell \in \mathbb{Z}\}$. Let $\gamma_{j,i}$ denote the smallest element of $S_j^{(i)}$. Let us define integers $\{m_{j,i}\}_{1 \leq j \leq k, 1 \leq i \leq N}$ as follows: $m_{j,i} = (N \cdot k)^{k \cdot i + j}$. The random variable $\mathbf{Y}_i$ is defined as follows: For all $\ell \in \mathbb{Z}^+$, let the map M_i send $\gamma_{j,i} + \ell$ to $m_{j,i} + \ell$. The probability distribution of $\mathbf{Y}_i$ is the distributed induced by the map M_i when acting on $\mathbf{X}_i$, i.e. a draw from $\mathbf{Y}_i$ is obtained by drawing x_i from $\mathbf{X}_i$ and outputting $M_i(x_i)$. It is clear that $\mathbf{Y}_i$ is integer-valued and satisfies

$$d_{\mathsf{shift},1}(\mathbf{X}_i) = d_{\mathsf{shift},1}(\mathbf{Y}_i).$$

Now consider a sequence of outcomes $\mathbf{Y}_1 = y_1, \ldots, \mathbf{Y}_N = y_N$ and $\mathbf{Y}'_1 = y'_1, \ldots, \mathbf{Y}'_N = y'_N$ such that

$$\left|\sum_{i=1}^{N} (y_i - y'_i)\right| = 1.$$

We can write each y_i as $m_{\alpha_i,i}+\delta_i$ where each $1 \le \alpha_i \le k$ and each $-k \le \delta_i \le k$. Likewise, $y'_i = m_{\alpha'_i,i}+\delta'_i$ where each $1 \le \alpha'_i \le k$ and each $-k \le \delta'_i \le k$. Since $m_{j,i} = (N \cdot k)^{k \cdot i + j}$, it is easy to see that the following must hold:

$$\text{For all } i = 1, \ldots, N, \quad m_{\alpha_i,i} = m_{\alpha'_i,i} \quad \text{and} \quad \left| \sum_{i=1}^{N} (\delta_i - \delta'_i) \right| = 1.$$

This immediately implies that $d_{\mathsf{shift},1}\left(\sum_{i=1}^{N} \mathbf{X}_i\right) = d_{\mathsf{shift},1}\left(\sum_{i=1}^{N} \mathbf{Y}_i\right)$. Applying Lemma 22, we have that

$$d_{\mathsf{shift},1}(\sum_{i=1}^{N} \mathbf{Y}_i) \le O\left(\frac{1}{\sqrt{\sum_{i=1}^{N} \delta_i}}\right),$$

which finishes the upper bound. □

This immediately yields the following corollary.

Corollary 24. *Let $\mathbf{X}_1, \ldots, \mathbf{X}_N$ be finitely supported independent integer valued random variables. Let $d_{\mathsf{shift},p}(\mathbf{X}_i) \le 1 - \delta_i$. Then, for $\mathbf{S} = \sum_{i=1}^{N} \mathbf{X}_i$, we have*

$$d_{\mathsf{shift},p}(\mathbf{S}) = O\left(\frac{1}{\sqrt{\sum_{i=1}^{N} \delta_i}}\right).$$

Proof. Let $\mathbf{Y}_i = \mathbf{X}_i/p$ for all $1 \le i \le N$. Then for $\mathbf{S}' = \sum_{i=1}^{N} \mathbf{Y}_i$, it is clear that $d_{\mathsf{shift},1}(\mathbf{S}') = d_{\mathsf{shift},p}(\mathbf{S})$. Applying Claim 23, we get the corollary. □

4.4 Other background results on distribution learning.

Learning distributions with small essential support. We recall the following folklore result, which says that distributions over a small essential support can be learned efficiently:

Fact 25. *There is an algorithm A with the following performance guarantee: A is given a positive integer s, an accuracy parameter ε, a confidence parameter δ, and access to i.i.d. draws from an unknown distribution $\mathbf{P}$ over $\mathbb{Z}$ that is promised to be ε-essentially supported on some set S with $|S| = s$. Algorithm A makes $m = \mathrm{poly}(s, 1/\varepsilon, \log(1/\delta))$ draws from $\mathbf{P}$, runs for time $\mathrm{poly}(s, 1/\varepsilon, \log(1/\delta))$ and with probability at least $1 - \delta$ outputs a hypothesis distribution $\tilde{\mathbf{P}}$ such that $d_{\mathrm{TV}}(\mathbf{P}, \tilde{\mathbf{P}}) \le 2\varepsilon$.*

(The algorithm of Fact 25 simply returns the empirical distribution of its m draws from $\mathbf{P}$.) Note that by Fact 25, if $\mathbf{S}$ is a sum of $N < \mathrm{poly}(1/\varepsilon)$ integer random variables then there is a $\mathrm{poly}(1/\varepsilon)$-time, $\mathrm{poly}(1/\varepsilon)$-sample algorithm for learning $\mathbf{S}$, simply because the support of $\mathbf{S}$ is contained in a set of size $\mathrm{poly}(1/\varepsilon)$. Thus in the analysis of our algorithm for $k = 3$ we can (and do) assume that N is larger than any fixed $\mathrm{poly}(1/\varepsilon)$ that arises in our analysis.

4.4.1 Hypothesis selection and "guessing".

To streamline our presentation as much as possible, many of the learning algorithms that we present are described as "making guesses" for different values at various points in their execution. For each such algorithm our analysis will establish that with very high probability there is a "correct outcome" for the

guesses which, if it is achieved (guessed), results in an ε-accurate hypothesis. This leads to a situation in which there are multiple hypothesis distributions (one for each possible outcome for the guesses that the algorithm makes), one of which has high accuracy, and the overall learning algorithm must output (with high probability) a high-accuracy hypothesis. Such situations have been studied by a range of authors (see e.g. [Yat85, DK14, AJOS14, DDS12c, DDS15]) and a number of different procedures are known which can do this. For concreteness we recall one such result, Proposition 6 from [DDS15]:

Proposition 26. *Let $\mathbf{D}$ be a distribution over a finite set W and let $\mathcal{D}_\varepsilon = \{\mathbf{D}_j\}_{j=1}^M$ be a collection of M hypothesis distributions over W with the property that there exists $i \in [M]$ such that $d_{\mathrm{TV}}(\mathbf{D}, \mathbf{D}_i) \leq \varepsilon$. There is an algorithm $\mathrm{Select}^{\mathbf{D}}$ which is given ε and a confidence parameter δ, and is provided with access to (i) a source of i.i.d. draws from $\mathbf{D}$ and from $\mathbf{D}_i$, for all $i \in [M]$; and (ii) an "evaluation oracle" $\mathrm{eval}_{\mathbf{D}_i}$, for each $i \in [M]$, which, on input $w \in W$, deterministically outputs the value $\mathbf{D}_i(w)$. The $\mathrm{Select}^{\mathbf{D}}$ algorithm has the following behavior: It makes $m = O\left((1/\varepsilon^2) \cdot (\log M + \log(1/\delta))\right)$ draws from $\mathbf{D}$ and from each $\mathbf{D}_i$, $i \in [M]$, and $O(m)$ calls to each oracle $\mathrm{eval}_{\mathbf{D}_i}$, $i \in [M]$. It runs in time $\mathrm{poly}(m, M)$ (counting each call to an $\mathrm{eval}_{\mathbf{D}_i}$ oracle and draw from a $\mathbf{D}_i$ distribution as unit time), and with probability $1 - \delta$ it outputs an index $i^\star \in [M]$ that satisfies $d_{\mathrm{TV}}(\mathbf{D}, \mathbf{D}_{i^\star}) \leq 6\varepsilon$.*

We shall apply Proposition 26 via the following simple corollary (the algorithm A' described below works simply by enumerating over all possible outcomes of all the guesses and then running the $\mathrm{Select}^{\mathbf{D}}$ procedure of Proposition 26):

Corollary 27. *Suppose that an algorithm A for learning an unknown distribution $\mathbf{D}$ works in the following way: (i) it "makes guesses" in such a way that there are a total of M possible different vectors of outcomes for all the guesses; (ii) for each vector of outcomes for the guesses, it makes m draws from $\mathbf{D}$ and runs in time T; (iii) with probability at least $1 - \delta$, at least one vector of outcomes for the guesses results in a hypothesis $\tilde{\mathbf{D}}$ such that $d_{\mathrm{TV}}(\mathbf{D}, \tilde{\mathbf{D}}) \leq \varepsilon$, and (iv) for each hypothesis distribution $\mathbf{D}'$ corresponding to a particular vector of outcomes for the guesses, A can simulate a random draw from $\mathbf{D}'$ in time T' and can simulate a call to the evaluation oracle $\mathrm{eval}_{\mathbf{D}'}$ in time T'. Then there is an algorithm A' that makes $m + O\left((1/\varepsilon^2) \cdot (\log M + \log(1/\delta))\right)$ draws from $\mathbf{D}$; runs in time $O(TM) + \mathrm{poly}(m, M, T')$; and with probability at least $1 - 2\delta$ outputs a hypothesis distribution $\tilde{\mathbf{D}}$ such that $d_{\mathrm{TV}}(\mathbf{D}, \tilde{\mathbf{D}}) \leq 6\varepsilon$.*

We will often implicitly apply Corollary 27 by indicating a series of guesses and specifying the possible outcomes for them. It will always be easy to see that the space of all possible vectors of outcomes for all the guesses can be enumerated in the required time. In Appendix A we discuss the specific form of the hypothesis distributions that our algorithm produces and show that the time required to sample from or evaluate any such hypothesis is not too high (at most $1/\varepsilon^{2^{\mathrm{poly}(k)}}$ when $|\mathcal{A}| = 3$, hence negligible given our claimed running times).

4.5 Small error

We freely assume throughout that the desired error parameter ε is at most some sufficiently small absolute constant value.

4.6 Fano's inequality and lower bounds on distribution learning.

A useful tool for our lower bounds is Fano's inequality, or more precisely, the following extension of it given by Ibragimov and Khasminskii [IH81] and Assouad and Birge [AB83]:

Theorem 28 (Generalization of Fano's Inequality.)**.** *Let $\mathbf{P}_1, \ldots, \mathbf{P}_{t+1}$ be a collection of $t+1$ distributions such that for any $i \neq j \in [t+1]$, we have (i) $d_{\mathrm{TV}}(\mathbf{P}_i, \mathbf{P}_j) \geq \alpha/2$, and (ii) $D_{KL}(\mathbf{P}_i || \mathbf{P}_j) \leq \beta$, where D_{KL} denotes Kullback-Leibler divergence. Let A be a learning algorithm which is given samples from an unknown distribution $\mathbf{P}$ which is promised to be one of $\mathbf{P}_1, \ldots, \mathbf{P}_{t+1}$ and which outputs an index $i \in [t+1]$ specifying a distribution $\mathbf{P}_i$. Then, to achieve expected error $\mathbf{E}[d_{\mathrm{TV}}(\mathbf{P}, \mathbf{P}_i)] \leq \alpha/4$ (where the expectation is over the random samples from $\mathbf{P}$), algorithm A must have sample complexity $\Omega\left(\frac{\ln t}{\beta}\right)$.*

5 Tools for kernel-based learning

At the core of our actual learning algorithm is the well-known technique of learning via the "kernel method" (see [DL01]). In this section we set up some necessary machinery for applying this technique in our context.

The goal of this section is ultimately to establish Lemma 35, which we will use later in our main learning algorithm. Definition 29 and Lemma 35 together form a "self-contained take-away" from this section.

We begin with the following important definition.

Definition 29. Let $\mathbf{Y}, \mathbf{Z}$ be two distributions supported on $\mathbb{Z}$. We say that $\mathbf{Y}$ is (ε, δ)*-kernel learnable from* $T = T(\varepsilon, \delta)$ *samples using* $\mathbf{Z}$ if the following holds: Let $\widehat{Y} = \{y_1, \ldots, y_{T_1}\}$ be a multiset of $T_1 \geq T$ i.i.d. samples drawn from $\mathbf{Y}$ and let $\mathbf{U}_{\widehat{Y}}$ be the uniform distribution over $\widehat{Y}$. Then with probability $1 - \delta$ (over the outcome of $\widehat{Y}$) it is the case that $d_{\mathrm{TV}}(\mathbf{U}_{\widehat{Y}} + \mathbf{Z}, \mathbf{Y}) \leq \varepsilon$.

Intuitively, the definition says that convolving the empirical distribution $\mathbf{U}_{\widehat{Y}}$ with $\mathbf{Z}$ gives a distribution which is close to $\mathbf{Y}$ in total variation distance. Note that once T_1 is sufficiently large, $\mathbf{U}_{\widehat{Y}}$ is close to $\mathbf{Y}$ in Kolmogorov distance by the DKW inequality. Thus, convolving with $\mathbf{Z}$ smoothens $\mathbf{U}_{\widehat{Y}}$.

The next lemma shows that if $\mathbf{Y}$ is (ε, δ)-kernel learnable, then a mixtures of shifts of $\mathbf{Y}$ is also (ε, δ)-kernel learnable with comparable parameters (provided the number of components in the mixture is not too large).

Lemma 30. *Let $\mathbf{Y}$ be (ε, δ)-kernel learnable using $\mathbf{Z}$ from $T(\varepsilon, \delta)$ samples. If $\mathbf{X}$ is a mixture (with arbitrary mixing weights) of distributions $c_1 + \mathbf{Y}, \ldots, c_k + \mathbf{Y}$ for some integers $c_1, \ldots, c_k$, then $\mathbf{X}$ is $(7\varepsilon, 2\delta)$-kernel learnable from T' samples using $\mathbf{Z}$, provided that $T' \geq \max\left\{\frac{kT(\varepsilon, \delta/k)}{\varepsilon}, C \cdot \frac{k^2 \log(k/\delta)}{\varepsilon^2}\right\}$.*

Proof. Let π_j denote the weight of distribution $c_j + \mathbf{Y}$ in the mixture $\mathbf{X}$. We view the draw of a sample point from $\mathbf{X}$ as a two stage process, where in the first stage an index $1 \leq j \leq k$ is chosen with probability π_j and in the second stage a random draw is made from the distribution $c_j + \mathbf{Y}$.

Consider a draw of T' independent samples $x_1, \ldots, x_{T'}$ from $\mathbf{X}$. In the draw of x_i, let the index chosen in the first stage be denoted j_i (note that $1 \leq j_i \leq k$). For $j \in [k]$ define

$$S_j = \{1 \leq i \leq T' : j_i = j\}.$$

The idea behind Lemma 30 is simple. Those j such that π_j is small will have $|S_j|$ small and will not contribute much to the error. Those j such that π_j is large will have $|S_j|/T'$ very close to π_j so their cumulative contribution to the total error will also be small since each such $\mathbf{U}_{\{x_i : i \in S_j\}} + \mathbf{Z}$ is very close to the corresponding $c_j + \mathbf{Y}$. We now provide details.

Since $T' \geq O\left(\frac{k^2 \log(k/\delta)}{\varepsilon^2}\right)$, a simple Chernoff bound and union bound over all $j \in [k]$ gives that

$$\left|\frac{|S_j|}{T'} - \pi_j\right| \leq \varepsilon/k \quad \text{for all } j \in [k] \tag{1}$$

with probability at least $1-\delta$. For the rest of the analysis we assume that indeed (1) holds. We observe that even after conditioning on (1) and on the outcome of $j_1, \ldots, j_{T'}$, it is the case that for each $i \in [T']$ the value x_i is drawn independently from $c_{j_i} + \mathbf{Y}$.

Let Low denote the set $\{1 \le j \le k : T' \cdot (\pi_j - \varepsilon/k) \le T(\varepsilon, \delta/k)\}$, so each $j \notin \text{Low}$ satisfies $T' \cdot (\pi_j - \varepsilon/k) \ge T(\varepsilon, \delta/k)$. Fix any $j \notin \text{Low}$. From (1) and the definition of Low we have that $|S_j| \ge T' \cdot (\pi_j - \varepsilon/k) \ge T(\varepsilon, \delta/k)$, and since $c_j + \mathbf{Y}$ is $(\varepsilon, \delta/k)$-kernel learnable from $T(\varepsilon, \delta/k)$ samples using $\mathbf{Z}$, it follows that with probability at least $1 - \delta/k$ we have

$$d_{\mathrm{TV}}\left(\mathbf{U}_{\{x_i : i \in S_j\}} + \mathbf{Z}, c_j + \mathbf{Y}\right) \le \varepsilon,$$

and thus

$$\sum_{z \in \mathbb{Z}} \left| \frac{|S_j|}{T'} \mathbf{Pr}[\mathbf{U}_{\{x_i : i \in S_j\}} + \mathbf{Z} = z] - \pi_j \mathbf{Pr}[\mathbf{Y} = z] \right| \le \left| \frac{|S_j|}{T'} - \pi_j \right| + \max\left\{ \frac{|S_j|}{T'}, \pi_j \right\} \cdot \frac{\varepsilon}{2}. \tag{2}$$

By a union bound, with probability at least $1 - \delta$ the bound (2) holds for all $j \notin \text{Low}$. For $j \in \text{Low}$, we trivially have

$$\sum_{z \in \mathbb{Z}} \left| \frac{|S_j|}{T'} \mathbf{Pr}[\mathbf{U}_{\{x_i : i \in S_j\}} + \mathbf{Z} = z] - \pi_j \mathbf{Pr}[Y = z] \right| \le \frac{|S_j|}{T'} + \pi_j \le \left| \frac{|S_j|}{T'} - \pi_j \right| + 2 \cdot \pi_j.$$

Next, note that

$$\sum_{j \in \text{Low}} \pi_j \le \sum_{j \in \text{Low}} \left(\frac{T(\varepsilon, \delta/k)}{T'} + \varepsilon/k \right) \le \sum_{j \in \text{Low}} (\varepsilon/k + \varepsilon/k) \le 2\varepsilon.$$

Thus, we obtain that

$$\begin{aligned}
&\sum_{z \in \mathbb{Z}} \left| \sum_{j=1}^{k} \frac{|S_j|}{T'} \mathbf{Pr}[\mathbf{U}_{\{x_i : i \in S_j\}} + \mathbf{Z} = z] - \sum_{j=1}^{k} \pi_j \mathbf{Pr}[Y = z] \right| \\
&\le \sum_{j=1}^{k} \sum_{z \in \mathbb{Z}} \left| \frac{|S_j|}{T'} \mathbf{Pr}[\mathbf{U}_{\{x_i : i \in S_j\}} + \mathbf{Z} = z] - \sum_{j=1}^{k} \pi_j \mathbf{Pr}[Y = z] \right| \\
&\le \sum_{j=1}^{k} \left| \frac{|S_j|}{T'} - \pi_j \right| + \sum_{j \notin \text{Low}} \max\left\{ \frac{|S_j|}{T'}, \pi_j \right\} \cdot \frac{\varepsilon}{2} + 2 \sum_{j \in \text{Low}} \pi_j \le 7\varepsilon.
\end{aligned}$$

As $\mathbf{X}$ is obtained by mixing $c_1 + \mathbf{Y}, ..., c_k + \mathbf{Y}$ with weights $\pi_1, ..., \pi_k$ and $\mathbf{U}_{x_1, ..., x_{T'}}$ is obtained by mixing $\mathbf{U}_{\{x_i : i \in S_1\}}, ..., \mathbf{U}_{\{x_i : i \in S_k\}}$ with weights $\frac{|S_1|}{T'}, ..., \frac{|S_k|}{T'}$, the lemma is proved. □

The next lemma is a formal statement of the well-known robustness of kernel learning; roughly speaking, it says that if $\mathbf{X}$ is kernel learnable using $\mathbf{Z}$ then any $\mathbf{X}'$ which is close to $\mathbf{X}$ is likewise kernel learnable using $\mathbf{Z}$.

Lemma 31. *Let $\mathbf{X}$ be (ε, δ)-kernel learnable using $\mathbf{Z}$ from $T(\varepsilon, \delta)$ samples, and suppose that $0 < d_{\mathrm{TV}}(\mathbf{X}, \mathbf{X}') = \kappa < 1$. If $T_0 > \max\{T(\varepsilon, \delta), C \cdot \frac{\log(1/\delta)}{\varepsilon^2}\}$, then $\mathbf{X}'$ is $(2\varepsilon + 2\kappa, 2\delta)$-kernel learnable from T_0 samples using $\mathbf{Z}$.*

Proof. We establish some useful notation: let $\mathbf{X}_{\text{common}}$ denote the distribution defined by

$$\mathbf{Pr}[\mathbf{X}_{\text{common}} = i] = \frac{\min\{\mathbf{Pr}[\mathbf{X} = i], \mathbf{Pr}[\mathbf{X}' = i]\}}{\sum_i \min\{\mathbf{Pr}[\mathbf{X} = i], \mathbf{Pr}[\mathbf{X}' = i]\}},$$

let $\mathbf{X}_{\text{residual}}$ denote the distribution defined by

$$\mathbf{Pr}[\mathbf{X}_{\text{residual}} = i] = \frac{\mathbf{Pr}[\mathbf{X} = i] - \min\{\mathbf{Pr}[\mathbf{X} = i], \mathbf{Pr}[\mathbf{X}' = i]\}}{\sum_i (\mathbf{Pr}[\mathbf{X} = i] - \min\{\mathbf{Pr}[\mathbf{X} = i]\,\mathbf{Pr}[\mathbf{X}' = i]\})},$$

and likewise let $\mathbf{X}'_{\text{residual}}$ denote the distribution defined by

$$\mathbf{Pr}[\mathbf{X}'_{\text{residual}} = i] = \frac{\mathbf{Pr}[\mathbf{X}' = i] - \min\{\mathbf{Pr}[\mathbf{X} = i], \mathbf{Pr}[\mathbf{X}' = i]\}}{\sum_i (\mathbf{Pr}[\mathbf{X}' = i] - \min\{\mathbf{Pr}[\mathbf{X} = i]\,\mathbf{Pr}[\mathbf{X}' = i]\})}.$$

A draw from $\mathbf{X}$ (from $\mathbf{X}'$ respectively) may be obtained as follows: draw from $\mathbf{X}_{\text{common}}$ with probability $1 - \kappa$ and from $\mathbf{X}_{\text{residual}}$ (from $\mathbf{X}'_{\text{residual}}$ respectively) with the remaining κ probability. To see this, note that if $\tilde{\mathbf{X}}$ is a random variable generated according to this two-stage process and $\mathbf{C} \in \{0, 1\}$ is an indicator variable for whether the draw was from $\mathbf{X}_{\text{common}}$, then, since $\kappa = 1 - \sum_i \min\{\mathbf{Pr}[\mathbf{X} = i], \mathbf{Pr}[\mathbf{X}' = i]\}$, we have

$$\begin{aligned}
\mathbf{Pr}[\tilde{\mathbf{X}} = i] &= \mathbf{Pr}[\tilde{\mathbf{X}} = i \wedge \mathbf{C} = 1] + \mathbf{Pr}[\tilde{\mathbf{X}} = i \wedge \mathbf{C} = 0] \\
&= \frac{\min\{\mathbf{Pr}[\mathbf{X} = i], \mathbf{Pr}[\mathbf{X}' = i]\}}{1 - \kappa} \times (1 - \kappa) + \frac{\mathbf{Pr}[\mathbf{X}' = i] - \min\{\mathbf{Pr}[\mathbf{X} = i], \mathbf{Pr}[\mathbf{X}' = i]\}}{1 - (1 - \kappa)} \times \kappa \\
&= \mathbf{Pr}[\mathbf{X} = i].
\end{aligned}$$

We consider the following coupling of $(\mathbf{X}, \mathbf{X}')$: to make a draw of (x, x') from the coupled joint distribution $(\mathbf{X}, \mathbf{X}')$, draw x_{common} from $\mathbf{X}_{\text{common}}$, draw x_{residual} from $\mathbf{X}_{\text{residual}}$, and draw x'_{residual} from $\mathbf{X}'_{\text{residual}}$. With probability $1 - \kappa$ output $(x_{\text{common}}, x_{\text{common}})$ and with the remaining κ probability output $(x_{\text{residual}}, x'_{\text{residual}})$.

Let $((x_1, x'_1), \ldots, (x_{T_0}, x'_{T_0}))$ be a sample of T_0 pairs each of which is independently drawn from the coupling of $(\mathbf{X}, \mathbf{X}')$ described above. Let $\widehat{X} = (x_1, \ldots, x_{T_0})$ and $\widehat{X}' = (x'_1, \ldots, x'_{T_0})$ and observe that $\widehat{X}$ is a sample of T_0 i.i.d. draws from $\mathbf{X}$ and similarly for $\widehat{X}'$. We have

$$\begin{aligned}
d_{\text{TV}}(\mathbf{U}_{\widehat{X}'} + \mathbf{Z}, \mathbf{X}') &\le d_{\text{TV}}(\mathbf{U}_{\widehat{X}'} + \mathbf{Z}, \mathbf{U}_{\widehat{X}} + \mathbf{Z}) + d_{\text{TV}}(\mathbf{U}_{\widehat{X}} + \mathbf{Z}, \mathbf{X}) + d_{\text{TV}}(\mathbf{X}, \mathbf{X}') \\
&\le d_{\text{TV}}(\mathbf{U}_{\widehat{X}'}, \mathbf{U}_{\widehat{X}}) + \varepsilon + \kappa \qquad \text{(by the data processing inequality for } \ell_1\text{)},
\end{aligned}$$

where the second inequality holds with probability $1 - \delta$ over the draw of $\widehat{X}$ since $T_0 \ge T(\varepsilon, \delta)$. A simple Chernoff bound tells us that with probability at least $1 - \delta$, the fraction of the $T_0 \ge C \cdot \frac{\log(1/\delta)}{\varepsilon^2}$ pairs that are of the form $(x_{\text{residual}}, x'_{\text{residual}})$ is at most $\kappa + \varepsilon$. Given that this happens we have $d_{\text{TV}}(\mathbf{U}_{\widehat{X}'}, \mathbf{U}_{\widehat{X}}) \le \kappa + \varepsilon$, and the lemma is proved. □

To prove the next lemma (Lemma 33 below) we will need a multidimensional generalization of the usual coupling argument used to prove the correctness of the kernel method. This is given by the following proposition:

Proposition 32. *For all $1 \le j \le k$, let $a_j, b_j \in \mathbb{Z}$ with $b_j \ge 1$ and let $\mathcal{B}$ be the subset of $\mathbb{Z}^k$ given by $\mathcal{B} = [a_1, a_1 + b_1] \times \ldots \times [a_k, a_k + b_k]$. Let $\mathbf{X}, \mathbf{Y}$ be random variables supported on $\mathbb{Z}^k$ such that $\mathbf{Pr}[\mathbf{X} \notin \mathcal{B}], \mathbf{Pr}[\mathbf{Y} \notin \mathcal{B}] \le \delta$. Let $\mathbf{Z}$ be a random variable supported on $\mathbb{Z}^k$ such that for all $1 \le j \le k$, $d_{\mathrm{TV}}(\mathbf{Z}, \mathbf{Z} + \mathbf{e}_j) \le \beta_j$. If $d_{\mathrm{K}}(\mathbf{X}, \mathbf{Y}) \le \lambda$, $\beta_j \le \rho/b_j$ for all j (where $\rho \ge 1$), and $\mathbf{Z}$ is independent of $\mathbf{X}$ and $\mathbf{Y}$, then*

$$d_{\mathrm{TV}}(\mathbf{X} + \mathbf{Z}, \mathbf{Y} + \mathbf{Z}) \le 2\delta + O\left(4^k \lambda^{\frac{1}{k+1}} \rho^{1-\frac{1}{k+1}}\right).$$

Proof. Let $d_1 \le b_1, \ldots, d_k \le b_k$ be positive integers that we will fix later. Divide the box $\mathcal{B}$ into boxes of size at most $d_1 \times \ldots \times d_k$ by dividing each $[a_i, a_i + b_i]$ into intervals of size d_i (except possibly the last interval which may be smaller). Let $\mathcal{S}$ denote the resulting set of k-dimensional boxes induced by these intervals, and note that the number of boxes in $\mathcal{S}$ is $\ell_1 \times \ldots \times \ell_k$ where $\ell_j = \lceil b_j/d_j \rceil$.

Let $\mu_{\mathbf{X}}$ and $\mu_{\mathbf{Y}}$ be the probability measures associated with $\mathbf{X}$ and $\mathbf{Y}$, and let $\mu_{\mathbf{X},\mathcal{B}}$ and $\mu_{\mathbf{Y},\mathcal{B}}$ be the restrictions of $\mu_{\mathbf{X}}$ and $\mu_{\mathbf{Y}}$ to the box $\mathcal{B}$ (so $\mu_{\mathbf{X}}$ and $\mu_{\mathbf{Y}}$ assign value zero to any point not in $\mathcal{B}$). For a box $S \in \mathcal{S}$, let $\mu_{\mathbf{X},S}$ denote the restriction of $\mu_{\mathbf{X}}$ to S. Let $x_S = \mathbf{Pr}[\mathbf{X} \in S]$ and $y_S = \mathbf{Pr}[\mathbf{Y} \in S]$. Let $w_S = \min\{x_S, y_S\}$. Let $\mathbf{X}_S$ and $\mathbf{Y}_S$ be the random variables obtained by conditioning $\mathbf{X}$ and $\mathbf{Y}$ on S, and $\mu_{\mathbf{X}_S}$ and $\mu_{\mathbf{Y}_S}$ be their measures. Note that $\mu_{\mathbf{X},S} = x_S \cdot \mu_{\mathbf{X}_S}$ and $\mu_{\mathbf{Y},S} = y_S \cdot \mu_{\mathbf{Y}_S}$. With this notation in place, using $f * g$ to denote the convolution of the measures f and g, we now have

$$\begin{aligned}
d_{\mathrm{TV}}(\mathbf{X} + \mathbf{Z}, \mathbf{Y} + \mathbf{Z}) &= \frac{1}{2}\ell_1(\mu_{\mathbf{X}+\mathbf{Z}}, \mu_{\mathbf{Y}+\mathbf{Z}}) \\
&= \frac{1}{2}\ell_1(\mu_{\mathbf{X}} * \mu_{\mathbf{Z}}, \mu_{\mathbf{Y}} * \mu_{\mathbf{Z}}) \\
&\le \mathbf{Pr}[\mathbf{X} \notin \mathcal{B}] + \mathbf{Pr}[\mathbf{Y} \notin \mathcal{B}] + \frac{1}{2}\ell_1(\mu_{\mathbf{X},\mathcal{B}} * \mu_{\mathbf{Z}}, \mu_{\mathbf{Y},\mathcal{B}} * \mu_{\mathbf{Z}}) \\
&\le 2\delta + \frac{1}{2}\sum_{S \in \mathcal{S}} \ell_1(\mu_{\mathbf{X},S} * \mu_{\mathbf{Z}}, \mu_{\mathbf{Y},S} * \mu_{\mathbf{Z}}) \\
&\le 2\delta + \frac{1}{2}\sum_{S \in \mathcal{S}} \ell_1(x_S \mu_{\mathbf{X}_S} * \mu_{\mathbf{Z}}, y_S \mu_{\mathbf{Y}_S} * \mu_{\mathbf{Z}}) \\
&\le 2\delta + \frac{1}{2}\sum_{S \in \mathcal{S}} \ell_1(w_S \mu_{\mathbf{X}_S} * \mu_{\mathbf{Z}}, w_S \mu_{\mathbf{Y}_S} * \mu_{\mathbf{Z}}) + \sum_{S \in \mathcal{S}} |x_S - y_S| \\
&\le 2\delta + \frac{1}{2}\sum_{S \in \mathcal{S}} w_S \ell_1(\mu_{\mathbf{X}_S} * \mu_{\mathbf{Z}}, \mu_{\mathbf{Y}_S} * \mu_{\mathbf{Z}}) + |\mathcal{S}| \cdot 2^k \lambda \\
&\le 2\delta + \sum_{S \in \mathcal{S}} w_S d_{\mathrm{TV}}(\mathbf{X}_S + \mathbf{Z}, \mathbf{Y}_S + \mathbf{Z}) + |\mathcal{S}| \cdot 2^k \lambda.
\end{aligned}$$

Here the second to last inequality uses the fact that the definition of $d_{\mathrm{K}}(\mathbf{X}, \mathbf{Y})$ gives $\sup |x_S - y_S| \le 2^k \lambda$. Next, notice that since $d_{\mathrm{TV}}(\mathbf{Z}, \mathbf{Z} + \mathbf{e}_j) \le \rho/b_j$ and each box in $\mathcal{S}$ has size at most $d_1 \times \ldots \times d_k$, we get that $d_{\mathrm{TV}}(\mathbf{X}_S + \mathbf{Z}, \mathbf{Y}_S + \mathbf{Z}) \le \sum_{i=1}^k \beta_i(d_i - 1)$. Thus, using that $|\mathcal{S}| = \prod_{j=1}^k \lceil b_j/d_j \rceil$ and $\sum_{S \in \mathcal{S}} w_S \le 1$, we have

$$d_{\mathrm{TV}}(\mathbf{X} + \mathbf{Z}, \mathbf{Y} + \mathbf{Z}) \le 2\delta + \sum_{S \in \mathcal{S}} w_S \cdot \big(\sum_{i=1}^k \beta_i(d_i - 1)\big) + 4^k \lambda \cdot \prod_{j=1}^k (b_j/d_j).$$

Optimizing the parameters $d_1, \ldots, d_k$, we set each $d_i = \left\lceil \left(\frac{\lambda}{\rho}\right)^{\frac{1}{k+1}} b_i \right\rceil$ which yields

$$d_{\mathrm{TV}}(\mathbf{X}+\mathbf{Z}, \mathbf{Y}+\mathbf{Z}) \leq 2\delta + (k+4^k)\lambda^{\frac{1}{k+1}}\rho^{1-\frac{1}{k+1}}. \qquad \square$$

Now we can prove Lemma 33, which we will use to prove that a weighted sum of high-variance PBDs is kernel-learnable for appropriately chosen smoothening distributions.

Lemma 33. *Let independent random variables $\mathbf{X}_1, \ldots, \mathbf{X}_k$ over $\mathbb{Z}$, and $\rho \geq 1$, be such that*

1. *For $1 \leq j \leq k$, there exist $a_j, b_j \in \mathbb{Z}$, $\delta_j \geq 0$ such that $\mathbf{Pr}[\mathbf{X}_j \notin [a_j, a_j + b_j]] \leq \delta_j$,*
2. *For all $1 \leq j \leq k$, $d_{\mathsf{shift},1}(\mathbf{X}_j) \leq \beta_j \leq \rho/b_j$.*

Let $\mathbf{Y} = \sum_{j=1}^k p_j \cdot \mathbf{X}_j$ for some integers $p_1, \ldots, p_k$. Let $\mathbf{Z}_j$ be the uniform distribution on the set $\mathbb{Z} \cap [-c_j, c_j]$ where[4] $c_j \in \mathbb{Z}$ satisfies $c_j = \frac{\Theta(\varepsilon) b_j}{k \cdot \rho}$ and $1 \leq c_j \leq b_j$ and $\mathbf{Z}_1, \ldots, \mathbf{Z}_k$ are mutually independent and independent of $\mathbf{X}_1, \ldots, \mathbf{X}_k$. Define $\mathbf{Z} = \sum_{j=1}^k p_j \cdot \mathbf{Z}_j$. Then, $\mathbf{Y}$ is $(\varepsilon + 4(\delta_1 + \ldots + \delta_k), \delta)$-kernel learnable using $\mathbf{Z}$ from $T = \frac{\exp(O(k^2))}{\varepsilon^{O(k)}} \cdot \rho^{O(k)} \cdot \log(1/\delta) + \log(\frac{4k}{\delta}) \cdot \max_j 1/\delta_j^2$ samples.

Proof. We first observe that

$$\begin{aligned} d_{\mathrm{TV}}(\mathbf{Y}+\mathbf{Z}, \mathbf{Y}) &\leq \sum_{j=1}^k d_{\mathrm{TV}}(p_j \cdot \mathbf{X}_j + p_j \cdot \mathbf{Z}_j, p_j \cdot \mathbf{X}_j) \\ &= \sum_{j=1}^k d_{\mathrm{TV}}(\mathbf{X}_j + \mathbf{Z}_j, \mathbf{X}_j) \leq \sum_{j=1}^k \frac{\rho c_j}{b_j} = \Theta(\varepsilon) \end{aligned} \tag{3}$$

where the last inequality uses the fact that $\mathbf{Z}_j$ is supported on the interval $[-c_j, c_j]$ and $d_{\mathsf{shift},1}(\mathbf{X}_j) \leq \frac{\rho}{b_j}$. Now, consider a two-stage sampling process for an element $y \leftarrow \mathbf{Y}$: For $1 \leq j \leq k$, we sample $x_j^{(y)} \sim \mathbf{X}_j$ and then output $y = \sum_{j=1}^k p_j \cdot x_j^{(y)}$. Thus, for every sample y, we can associate a sample $x^{(y)} = (x_1^{(y)}, \ldots, x_k^{(y)})$. For $y_1, \ldots, y_T \leftarrow \mathbf{Y}$, let $x^{(y_1)}, \ldots, x^{(y_T)}$ denote the corresponding samples from $\mathbb{Z}^k$. Let $\mathbf{U}_{\widehat{X}}$ denote the uniform distribution over the multiset of T samples $x^{(y_1)}, \ldots, x^{(y_T)}$, and let $\mathbf{U}_{\widehat{Y}}$ denote the uniform distribution on $y_1, \ldots, y_T$. Let $\mathbf{X}_{\mathsf{multi}} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k)$. By Lemma 11, we get that if $T \geq c(k + \log(1/\delta))/\eta^2$ (for a parameter η we will fix later), then with probability $1 - \delta/2$ we have $d_{\mathrm{K}}(\mathbf{X}_{\mathsf{multi}}, \mathbf{U}_{\widehat{X}}) \leq \eta$; moreover, if $T \geq \log(\frac{4k}{\delta}) \cdot \max_j 1/\delta_j^2$, then by a Chernoff bound and a union bound we have that $\mathbf{Pr}[(\mathbf{U}_{\widehat{X}})_j \notin [a_j, a_j + b_j]] \leq 2\delta_j$ for $1 \leq j \leq k$ (which we will use later) with probability $1 - \delta/2$. In the rest of the argument we fix such an $\widehat{X}$ satisfying these conditions, and show that for the corresponding $\widehat{Y}$ we have $d_{\mathrm{TV}}(\mathbf{Y}, \mathbf{U}_{\widehat{Y}} + \mathbf{Z}) \leq 4(\delta_1 + \ldots + \delta_k) + \varepsilon$, thus establishing kernel learnability of $\mathbf{Y}$ using $\mathbf{Z}$.

Next, we define $\mathbf{Z}_{\mathsf{multi}} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_k)$, with the aim of applying Proposition 32. We observe that $d_{\mathrm{TV}}(\mathbf{Z}_{\mathsf{multi}}, \mathbf{Z}_{\mathsf{multi}} + \mathbf{e}_j) \leq \frac{1}{c_j}$ and as noted above, for $1 \leq j \leq k$ we have $\mathbf{Pr}[(\mathbf{U}_{\widehat{X}})_j \notin [a_j, a_j + b_j]] \leq 2\delta_j$.

[4] **Phil**: We might have considered setting $c_j = \frac{\Theta(\varepsilon)}{k \cdot \beta_j}$, analogously to before, but this could be inconsistent with $c_j \leq b_j$. I don't see where we use $c_j \leq b_j$ though. Even if we do set $c_j = \frac{\Theta(\varepsilon)}{k \cdot \beta_j}$, it seems that we will still need $\beta_i \leq \varepsilon^2/k$, will seems like it may require significant downstream changes.

Define the box $\mathcal{B} = [a_1, a_1 + b_1] \times \ldots \times [a_k, a_k + b_k]$. Applying Proposition 32, we get

$$d_{\mathrm{TV}}(\mathbf{X}_{\mathsf{multi}} + \mathbf{Z}_{\mathsf{multi}}, \mathbf{U}_{\widehat{X}} + \mathbf{Z}_{\mathsf{multi}}) \leq 4(\delta_1 + \ldots + \delta_k) + O\left(4^k \eta^{\frac{1}{k+1}} \left(\frac{k\rho}{\Theta(\varepsilon)}\right)^{1-\frac{1}{k+1}}\right)$$

since $\sum_j \beta_j \leq \varepsilon$. Taking an inner product with $\overline{p} = (p_1, \ldots, p_k)$, we get

$$\begin{aligned} d_{\mathrm{TV}}(\mathbf{Y} + \mathbf{Z}, \mathbf{U}_{\widehat{Y}} + \mathbf{Z}) &= d_{\mathrm{TV}}(\langle \overline{p}, \mathbf{X}_{\mathsf{multi}} + \mathbf{Z}_{\mathsf{multi}} \rangle, \langle \overline{p}, \mathbf{U}_{\widehat{X}} + \mathbf{Z}_{\mathsf{multi}} \rangle) \\ &\leq d_{\mathrm{TV}}(\mathbf{X}_{\mathsf{multi}} + \mathbf{Z}_{\mathsf{multi}}, \mathbf{U}_{\widehat{X}} + \mathbf{Z}_{\mathsf{multi}}) \\ &\leq 4(\delta_1 + \ldots + \delta_k) + O\left(4^k \eta^{\frac{1}{k+1}} \left(\frac{k\rho}{\Theta(\varepsilon)}\right)^{1-\frac{1}{k+1}}\right). \end{aligned}$$

Combining this with (3), we get that

$$d_{\mathrm{TV}}(\mathbf{Y}, \mathbf{U}_{\widehat{Y}} + \mathbf{Z}) \leq 4(\delta_1 + \ldots + \delta_k) + O\left(4^k \eta^{\frac{1}{k+1}} \left(\frac{k\rho}{\Theta(\varepsilon)}\right)^{1-\frac{1}{k+1}}\right) + \Theta(\varepsilon), \tag{4}$$

Setting $\eta = \frac{\varepsilon^{2k+1}}{4^{k(k+1)} k^k \rho^k}$, the condition $T \geq c(k + \log(1/\delta))/\eta^2$ from earlier becomes

$$T \geq c(k + \log(1/\delta))/\eta^2 = \frac{e^{O(k^2)}}{\varepsilon^{O(k)}} \cdot \rho^{2k} \cdot \log(1/\delta)$$

and we have $d_{\mathrm{TV}}(\mathbf{Y}, \mathbf{U}_{\widehat{Y}} + \mathbf{Z}) \leq 4(\delta_1 + \ldots + \delta_k) + \Theta(\varepsilon)$, proving the lemma. □

We specialize Lemma 33 to establish kernel learnability of weighted sums of signed PBDs as follows:

Corollary 34. *Let $\mathbf{S}_1, \ldots, \mathbf{S}_k$ be independent signed PBDs and let $\mathbf{Y} = \sum_{j=1}^k p_j \cdot \mathbf{S}_j$. Let $\sigma_j^2 = \mathbf{Var}[\mathbf{S}_j] = \omega(k^2/\varepsilon^2)$ and let $\mathbf{Z}_j$ be the uniform distribution on $[-c_j, c_j] \cap \mathbb{Z}$ where $c_j = \Theta(\varepsilon \cdot \sigma_j/k)$. Let $\mathbf{Z} = \sum_{j=1}^k p_j \cdot \mathbf{Z}_j$. Then $\mathbf{Y}$ is (ε, δ)-kernel learnable using $\mathbf{Z}$ from $T = \frac{e^{O(k^2)}}{\varepsilon^{O(k)}} \cdot \log(1/\delta)$ samples.*

Proof. Note that for $1 \leq j \leq k$, there are integers a_j such that for $b_j = O(\sigma_j \cdot \ln(k/\varepsilon))$, by Bernstein's inequality we have $\mathbf{Pr}[\mathbf{S}_j \notin [a_j, a_j + b_j]] \leq \varepsilon/k$. Also, recall from Fact 18 that $d_{\mathsf{shift},1}(\mathbf{S}_j) = \frac{O(1)}{\sigma_j}$. Since each c_j satisfies $1 \leq c_j \leq b_j$, we may apply Lemma 33 and we get that $\mathbf{Y}$ is $(O(\varepsilon), \delta)$-kernel learnable using $T = \frac{e^{O(k^2)}}{\varepsilon^{O(k)}} \cdot \log(1/\delta)$ samples. □

(It should be noted that while the previous lemma shows that a weighted sum of signed PBDs that have "large variance" are kernel learnable, the hypothesis $\mathbf{U}_{\widehat{Y}} + \mathbf{Z}$ is based on $\mathbf{Z}$ and thus constructing it requires knowledge of the variances $\sigma_1, \ldots, \sigma_j$; thus Lemma 33 does not immediately yield an efficient learning algorithm when the variances of the underlying PBDs are unknown. We will return to this issue of knowing (or guessing) the variances of the constituent PBDs later.)

Finally, we generalize Corollary 34 to obtain a robust version. Lemma 35 will play an important role in our ultimate learning algorithm.

Lemma 35. *Let $\mathbf{S}$ be κ-close to a distribution of the form $\mathbf{S}' = \mathbf{S}_{\mathit{offset}} + \sum_{j=1}^{K} p_j \cdot \mathbf{S}_j$, where $\mathbf{S}_{\mathit{offset}}, \mathbf{S}_1, \ldots, \mathbf{S}_K$ are all independent and $\mathbf{S}_1, \ldots, \mathbf{S}_K$ are signed PBDs. For $a \in [K]$ let $\sigma_a^2 = \mathbf{Var}[\mathbf{S}_a] = \omega(K^2/\varepsilon^2)$. Let $m = |\mathrm{supp}(\mathbf{S}_{\mathit{offset}})|$ and let $\gamma_1, \ldots, \gamma_K$ be such that for all $1 \le a \le K$ we have $\sigma_a \le \gamma_a \le 2\sigma_a$. Let $\mathbf{Z}_j$ be the uniform distribution on the interval $[-c_j, c_j] \cap \mathbb{Z}$ where $c_j = \Theta(\varepsilon \cdot \gamma_j / K)$. Then for $\mathbf{Z} = \sum_{a=1}^{K} p_a \cdot \mathbf{Z}_a$, the distribution $\mathbf{S}$ is $(O(\varepsilon + \kappa), O(\delta))$-kernel learnable using $\mathbf{Z}$ from $\frac{\exp(O(K^2))}{\varepsilon^{O(K)}} \cdot m^2 \cdot \log(m/\delta)$ samples.*

Proof. Applying Corollary 34 and Lemma 30, we first obtain that the distribution $\mathbf{S}'$ is $(O(\varepsilon), O(\delta))$-kernel-learnable using $\mathbf{Z}$ from $\frac{\exp(O(K^2))}{\varepsilon^{O(K)}} \cdot m^2 \cdot \log(m/\delta)$ samples. Now, applying Lemma 31, we obtain that $\mathbf{S}$ is $(O(\varepsilon + \kappa), O(\delta))$-learnable using $\mathbf{Z}$ from $\frac{\exp(O(K^2))}{\varepsilon^{O(K)}} \cdot m^2 \cdot \log(m/\delta)$ samples. □

6 Setup for the upper bound argument

Recall that an $\mathcal{A}$-sum is $\mathbf{S} = \mathbf{X}_1 + \cdots + \mathbf{X}_N$ where the $\mathbf{X}_1, \ldots, \mathbf{X}_N$ distributions are independent (but not identically distributed) and each $\mathbf{X}_i$ is supported on the set $\mathcal{A} = \{a_1, \ldots, a_k\}$ where $\{a_1, \ldots, a_k\} \subset \mathbb{Z}_{\ge 0}$ and $a_1 < \cdots < a_k$ (and $a_1, \ldots, a_k, N, \varepsilon$ are all given to the learning algorithm in the known-support setting).

For each $\mathbf{X}_i$ we define $\mathbf{X}_i'$ to be the "zero-moded" variant of $\mathbf{X}_i$, namely $\mathbf{X}_i' = \mathbf{X}_i - \mathrm{mode}(\mathbf{X}_i)$ where $\mathrm{mode}(\mathbf{X}_i) \in \{a_1, \ldots, a_k\}$ is a mode of $\mathbf{X}_i$ (i.e. $\mathrm{mode}(\mathbf{X}_i)$ satisfies $\mathbf{Pr}[\mathbf{X}_i = \mathrm{mode}(\mathbf{X}_i)] \ge \mathbf{Pr}[\mathbf{X}_i = a_{i'}]$ for all $i' \in [k]$). We define $\mathbf{S}'$ to be $\sum_{i=1}^{N} \mathbf{X}_i'$. It is clear that $\mathbf{S}' + V = \mathbf{S}$ where V is an (unknown) "offset" in $\mathbb{Z}$. Below we will give an algorithm that learns $\mathbf{S}' + V$ given independent draws from it.

For each $i \in [N]$ the support of random variable $\mathbf{X}_i'$ is contained in $\{0, \pm q_1, \ldots, \pm q_K\}$, where $K = O(k^2)$ and $\{q_1, \ldots, q_K\}$ is the set of all distinct values achieved by $|a_\ell - a_{\ell'}|, 1 \le \ell < \ell' \le k$. As noted above each $\mathbf{X}_i'$ has $\mathbf{Pr}[\mathbf{X}_i' = 0] \ge 1/k \ge 1/K$.

To help minimize confusion we will consistently use letters i, j, etc. for dummy variables that range over $1, \ldots, N$ and a, b, c, d etc. for dummy variables that range over $1, \ldots, K$.

We define the following probabilities and associated values:

$$\text{For } i \in [N] \text{ and } a \in [K]: \quad c_{q_a,i} = \mathbf{Pr}[\mathbf{X}_i' = \pm q_a] \tag{5}$$

$$\text{For } a \in [K]: \quad c_{q_a} = \sum_{i=1}^{N} c_{q_a,i}. \tag{6}$$

We may think of the value c_{q_a} as the "weight" of q_a in $\mathbf{S}'$.

It is useful for us to view $\mathbf{S}' = \sum_{i=1}^{N} \mathbf{X}_i'$ in the following way. Recall that the support of $\mathbf{X}_i'$ is contained in $\{0, \pm q_1, \ldots, \pm q_K\}$. For $i \in [N]$ we define a vector-valued random variable $\mathbf{Y}_i$ that is supported on $\{0, \pm \boldsymbol{e}_1, \ldots, \pm \boldsymbol{e}_K\}$ by

$$\mathbf{Pr}[\mathbf{Y}_i = 0] = \mathbf{Pr}[\mathbf{X}_i' = 0] \ge \frac{1}{K}, \qquad \mathbf{Pr}[\mathbf{Y}_i = \tau \boldsymbol{e}_a] = \mathbf{Pr}[\mathbf{X}_i' = \tau q_a] \text{ for } \tau \in \{-1, 1\}, a \in [K]. \tag{7}$$

We define the vector-valued random variable $\mathbf{M} = \sum_{i=1}^{N} \mathbf{Y}_i$, so we have $\mathbf{X}_i' = (q_1, \ldots, q_K) \cdot \mathbf{Y}_i$ for each i and $\mathbf{S}' = (q_1, \ldots, q_K) \cdot \mathbf{M}$. Summarizing for convenient later reference:

$$\mathbf{X}'_1, \ldots, \mathbf{X}'_N : \text{ independent, each supported in } \{0, \pm q_1, \ldots, \pm q_K\} \tag{8}$$
$$\mathbf{S}' = \mathbf{X}'_1 + \cdots + \mathbf{X}'_N : \text{ supported in } \mathbb{Z} \tag{9}$$
$$\mathbf{Y}_1, \ldots, \mathbf{Y}_N : \text{ independent, each supported in } \{0, \pm \boldsymbol{e}_1, \ldots, \pm \boldsymbol{e}_K\} \tag{10}$$
$$\mathbf{M} = \mathbf{Y}_1 + \cdots + \mathbf{Y}_N : \text{ supported in } \mathbb{Z}^k \tag{11}$$
$$\mathbf{S}' = (q_1, \ldots, q_K) \cdot \mathbf{M}. \tag{12}$$

From this perspective, in order to analyze $\mathbf{S}'$ it is natural to analyze the multinomial random variable $\mathbf{M}$, and indeed this is what we do in the next section.

Finally, we note that while it suffices to learn $\mathbf{S}'$ of the form captured in (8) and (9) for the K and $\mathbf{S}'$ that arise from our reduction to this case, our analysis will hold for all $K \in \mathbb{Z}^+$ and all $\mathbf{S}'$ of this form.

7 Useful structural results when all c_{q_a}'s are large

In this section we establish some useful structural results for dealing with a distribution $\mathbf{S}' = \sum_{i=1}^{N} \mathbf{X}'_i$ for which, roughly speaking, all the values $c_{q_1}, \ldots, c_{q_K}$ (as defined in Section 6) are "large." More formally, we shall assume throughout this section that each $c_{q_a} \geq R$, where the exact value of the parameter R will be set later in the context of our learning algorithm in (29) (we note here only that R will be set to a fixed "large" polynomial in K and $1/\varepsilon$). Looking ahead, we will later use the results of this section to handle whatever c_{q_a}'s are "large".

The high-level plan of our analysis is as follows: In Section 7.1 we show that the multinomial distribution $\mathbf{M}$ (recall (11) and (12)) is close in total variation distance to a suitable discretized multidimensional Gaussian. In Section 7.2 we show in turn that such a discretized multidimensional Gaussian is close to a vector-valued random variable that can be expressed in terms of independent signed PBDs. Combining these results, in Section 7.3 we show that $\mathbf{S}'$ is close in variation distance to a weighted sum of signed PBDs. The lemma stating this, Lemma 40 in Section 7.3, is one of the two main structural results in this section. The second main structural result in this section, Lemma 41, is stated and proved in Section 7.4. Roughly speaking, it shows that, for a weighted sum of signed PBDs, it is possible to replace the scaled sum of the "high-variance" PBDs by a single scaled PBD. This is useful later for learning since it leaves us in a situation where we only need to deal with scaled PBDs whose variance is "not too high."

We record some useful notation for this section: for $i \in [N]$, $a \in [K]$ and $\tau \in \{-1, 1\}$ let $p_{i,a,\tau}$ denote

$$p_{i,a,\tau} := \mathbf{Pr}[\mathbf{Y}_i = \tau \boldsymbol{e}_a] = \mathbf{Pr}[\mathbf{X}'_i = \tau q_a]. \tag{13}$$

7.1 From multinomials to discretized multidimensional Gaussians

The result of this subsection, Lemma 36, establishes that the multinomial distribution $\mathbf{M}$ is close in total variation distance to a discretized multidimensional Gaussian.

Lemma 36. *Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$ be as in (10), so each $\mathbf{Y}_i$ has $\mathbf{Pr}[\mathbf{Y}_i = 0] \geq 1/K$. Assume that $c_{q_a} \geq R$ for all $a \in [K]$. As in (11) let $\mathbf{M} = \mathbf{Y}_1 + \cdots + \mathbf{Y}_N$, and let $\tilde{\mu} = \mathbf{E}[\mathbf{M}]$ be the K-dimensional mean of $\mathbf{M}$ and $\tilde{\Sigma}$ be the $K \times K$ covariance matrix $\mathbf{Cov}(\mathbf{M})$. Then*

(1) Defining $\tilde{\sigma}^2$ to be the smallest eigenvalue of $\tilde{\Sigma}$, we have that $\tilde{\sigma}^2 \geq R/K$.

(2) $d_{\mathrm{TV}}(\mathbf{M}, \mathcal{N}_D(\tilde{\mu}, \tilde{\Sigma})) \leq O(K^{71/20}/R^{1/20})$.

Proof. Given part (1), Theorem 19 directly gives the claimed variation distance bound in part (2), so in the following we establish (1).

Since $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$ are independent we have that

$$\tilde{\Sigma} = \sum_{i=1}^{N} \tilde{\Sigma}_i, \quad \text{where } \tilde{\Sigma}_i = \mathbf{Cov}(\mathbf{Y}_i).$$

Fix $i \in [N]$. Recalling (13), we have that $\tilde{\Sigma}_i$ is the $K \times K$ matrix defined by

$$(\tilde{\Sigma}_i)_{a,b} = \begin{cases} (p_{i,a,1} + p_{i,a,-1})(1 - p_{i,a,1} - p_{i,a,-1}) + 4p_{i,a,1}p_{i,a,-1} & \text{if } a = b \\ -(p_{i,a,1} - p_{i,a,-1})(p_{i,b,1} - p_{i,b,-1}) & \text{if } a \neq b. \end{cases}$$

Hence we have

$$(\tilde{\Sigma})_{a,b} = \begin{cases} \sum_{i=1}^{N} (p_{i,a,1} + p_{i,a,-1})(1 - p_{i,a,1} - p_{i,a,-1}) + 4p_{i,a,1}p_{i,a,-1} & \text{if } a = b \\ \sum_{i=1}^{N} -(p_{i,a,1} - p_{i,a,-1})(p_{i,b,1} - p_{i,b,-1}) & \text{if } a \neq b. \end{cases} \tag{14}$$

For later reference (though we do not need it in this proof) we also note that the mean vector $\tilde{\mu}$ is defined by

$$\tilde{\mu}_a = \sum_{i=1}^{N} (p_{i,a,1} - p_{i,a,-1}). \tag{15}$$

Let $\delta_i = \mathbf{Pr}[\mathbf{Y}_i = 0] = 1 - p_{i,1,1} - p_{i,1,-1} - \cdots - p_{i,K,1} - p_{i,K,-1}$ and observe that by assumption we have $\delta_i \geq 1/K$ for all $i \in [N]$. We lower bound the smallest eigenvalue using the variational characterization. For any unit vector $\boldsymbol{x}$ in $\mathbb{R}^K$, we have

$$\begin{aligned} \boldsymbol{x}^T \cdot \tilde{\Sigma} \cdot \boldsymbol{x} \quad = \quad & \sum_{a=1}^{K} x_a^2 \left(\sum_{i=1}^{N} (p_{i,a,1} + p_{i,a,-1})(1 - p_{i,a,1} - p_{i,a,-1}) + 4p_{i,a,1}p_{i,a,-1} \right) \\ & - \sum_{a=1}^{K} \sum_{b \in [K], b \neq a} x_a x_b \left(\sum_{i=1}^{N} (p_{i,a,1} - p_{i,a,-1})(p_{i,b,1} - p_{i,b,-1}) \right). \end{aligned} \tag{16}$$

Let $p'_{i,a,1} = p_{i,a,1} + p_{i,a,-1}$. Recalling that each $p_{i,a,1}, p_{i,a,-1} \geq 0$, it is not difficult to see that then we have

$$(16) \quad \geq \quad \sum_{a=1}^{K} x_a^2 \left(\sum_{i=1}^{N} p'_{i,a,1}(1 - p'_{i,a,1}) \right) - \sum_{a=1}^{K} \sum_{b \in [K], b \neq a} |x_a| \cdot |x_b| \left(\sum_{i=1}^{N} p'_{i,a,1} p'_{i,b,1} \right), \tag{17}$$

so for the purpose of lower bounding (16) it suffices to lower bound (17). Rewriting $p'_{i,a,1}$ as $p_{i,a}$ for

notational simplicity, so now $\delta_i = 1 - p_{i,1} - \cdots - p_{i,K}$,we have

$$\begin{aligned}
(17) \quad &\geq \sum_{a=1}^{K} x_a^2 \sum_{i=1}^{N} p_{i,a}(1-p_{i,a}) - \sum_{a=1}^{K} \sum_{b\in[K],b\neq a} |x_a|\cdot|x_b| \sum_{i=1}^{N} p_{i,a}p_{i,b} \\
&= \sum_{i=1}^{N} \left(\sum_{a=1}^{K} p_{i,a}(1-p_{i,a})x_a^2 - \sum_{a=1}^{K} \sum_{b\in[K],b\neq a} p_{i,a}p_{i,b}|x_a|\cdot|x_b| \right) \\
&= \sum_{i=1}^{N} \left(\sum_{a=1}^{K} \delta_i p_{i,a} x_a^2 + \sum_{a=1}^{K} p_{i,a}x_a^2 \left(\sum_{b\in[K],b\neq a} p_{i,b} \right) - \sum_{a=1}^{K} \sum_{b\in[K],b\neq a} p_{i,a}p_{i,b}|x_a|\cdot|x_b| \right) \\
&= \sum_{i=1}^{N} \left(\delta_i \sum_{a=1}^{K} p_{i,a}x_a^2 + \sum_{a=1}^{K} \sum_{b\in[K],b\neq a} (p_{i,a}p_{i,b}x_a^2 - p_{i,a}p_{i,b}|x_a|\cdot|x_b|) \right) \\
&= \sum_{i=1}^{N} \left(\delta_i \sum_{a=1}^{K} p_{i,a}x_a^2 + \sum_{a=1}^{K} \sum_{b<a} p_{i,a}p_{i,b}(|x_a| - |x_b|)^2 \right) \\
&\geq \sum_{i=1}^{N} \delta_i \sum_{a=1}^{K} p_{i,a}x_a^2. \qquad (18)
\end{aligned}$$

Recalling that $\delta_i \geq 1/K$ for all $i \in [N]$ and $\sum_{i=1}^{n} p_{i,a} = c_{q_a} \geq R$ for all $a \in [K]$, we get

$$\sum_{i=1}^{N} \delta_i \sum_{a=1}^{K} p_{i,a}x_a^2 \geq \frac{1}{K} \sum_{a=1}^{K} x_a^2 \sum_{i=1}^{N} p_{i,a} \geq \frac{1}{K} \sum_{a=1}^{K} c_{q_a} x_a^2 \geq \frac{R}{K} \sum_{a=1}^{K} x_a^2 = \frac{R}{K},$$

so $\tilde{\sigma}^2 \geq R/K$ and the lemma is proved. □

7.2 From discretized multidimensional Gaussians to combinations of independent signed PBDs

The first result of this subsection, Lemma 37, is a technical lemma establishing that the discretized multidimensional Gaussian given by Lemma 36 is close to a vector-valued random variable in which each marginal (coordinate) is a (± 1)-weighted linear combination of independent discretized Gaussians, certain of which are promised to have large variance.

Lemma 37. *Under the assumptions of Lemma 36 the following items (1) and (2) both hold:*

(1) The pair $\tilde{\mu} \in \mathbb{R}^K, \tilde{\Sigma} \in \mathbb{R}^{K\times K}$, defined in (15) and (14), are such that there exist $\mu_{a,b} \in \mathbb{R}$, $1 \leq a \leq b \leq K$, satisfying

$$\tilde{\mu}_a = \mu_{a,a} + \sum_{c<a} \mu_{c,a} + \sum_{a<d} \mathrm{sign}(\tilde{\Sigma}_{a,d}) \cdot \mu_{a,d}, \qquad (19)$$

and there exist $\sigma_{a,b} \in \mathbb{R}$, $1 \leq a \leq b \leq K$, such that

$$\sigma_{a,b}^2 = |\tilde{\Sigma}_{a,b}| = |\tilde{\Sigma}_{b,a}| \ \text{ for all } a < b \qquad \text{and} \qquad \tilde{\Sigma}_{a,a} = \sigma_{a,a}^2 + \sum_{c<a} \sigma_{c,a}^2 + \sum_{a<d} \sigma_{a,d}^2. \qquad (20)$$

Furthermore, for all $a \in [K]$ we have $\sigma_{a,a}^2 \geq \sigma^2$, where we define $\sigma^2 := R/K$.

(2) Let $\mathbf{U}_{a,b}$, $1 \le a < b \le K$ be discretized Gaussians $\mathbf{U}_{a,b} = \mathcal{N}_D(\mu_{a,b}, \sigma_{a,b}^2)$ that are all mutually independent. For $a \in [K]$ let $\mathbf{X}_a$ be defined as

$$\mathbf{X}_a = \mathbf{U}_{a,a} + \sum_{c<a} \mathbf{U}_{c,a} + \sum_{a<d} \operatorname{sign}(\tilde{\Sigma}_{a,d}) \cdot \mathbf{U}_{a,d}.$$

Then

$$d_{\mathrm{TV}}\left((\mathbf{X}_1, \ldots, \mathbf{X}_K), \mathcal{N}_D(\tilde{\mu}, \tilde{\Sigma})\right) \le \frac{K^2}{\sigma} = \frac{K^{5/2}}{R^{1/2}}.$$

Proof. We first prove part (1). Existence of the desired $\mu_{a,b}$ values is immediate since for each $a \in [K]$ the variable $\mu_{a,a}$ appears in only one equation given by (19) (so we can select arbitrary values for each $\mu_{a,b}$ with $a < b$, and there will still exist a value of $\mu_{a,a}$ satisfying (19)). The first part of (20) is trivial since for $a < b$ we take $\sigma_{a,b}^2 = |\tilde{\Sigma}_{a,b}|$ (which of course equals $|\tilde{\Sigma}_{b,a}|$ since the covariance matrix $\tilde{\Sigma}$ is symmetric). For the second part we take $\sigma_{a,a}^2 = \tilde{\Sigma}_{a,a} - \sum_{c<a} \sigma_{c,a}^2 - \sum_{a<d} \sigma_{a,d}^2$ which we now proceed to lower bound.

$$\begin{aligned}
\sigma_{a,a}^2 &= \tilde{\Sigma}_{a,a} - \sum_{c<a} \sigma_{c,a}^2 - \sum_{a<d} \sigma_{a,d}^2 = \tilde{\Sigma}_{a,a} - \sum_{b \ne a} |\tilde{\Sigma}_{b,a}| \\
&\ge \tilde{\Sigma}_{a,a} - \sum_{b \ne a} \sum_{i=1}^{N} (p_{i,a,1} + p_{i,a,-1})(p_{i,b,1} + p_{i,b,-1}) && \text{(by (14))} \\
&= \sum_{i=1}^{N} (p_{i,a,1} + p_{i,a,-1}) \left((1 - p_{i,a,1} - p_{i,a,-1}) - \sum_{b \ne a} (p_{i,b,1} + p_{i,b,-1}) \right) + 4 p_{i,a,1} p_{i,a,-1} \\
& && \text{(again by (14))} \\
&\ge \sum_{i=1}^{N} (p_{i,a,1} + p_{i,a,-1}) \left((1 - p_{i,a,1} - p_{i,a,-1}) - \sum_{b \ne a} (p_{i,b,1} + p_{i,b,-1}) \right) \\
&= \sum_{i=1}^{N} (p_{i,a,1} + p_{i,a,-1}) \left(\delta_i + \sum_{b \ne a} (p_{i,b,1} + p_{i,b,-1}) - \sum_{b \ne a} (p_{i,b,1} + p_{i,b,-1}) \right) && \text{(by definition of } \delta_i) \\
&= \sum_{i=1}^{N} \delta_i (p_{i,a,1} + p_{i,a,-1}) \ge \frac{1}{K} \sum_{i=1}^{N} (p_{i,a,1} + p_{i,a,-1}) && \text{(since } \delta_i \ge \tfrac{1}{K}) \\
&\ge \frac{1}{K} c_{q_a} \ge \frac{R}{K}.
\end{aligned}$$

With $\mu_{a,b}$ and $\sigma_{a,b}$ in hand, now we turn to proving part (2) of the lemma. For $1 \le a \le b \le K$ let $\mathbf{U}'_{a,b}$ be the (non-discretized) univariate Gaussian $\mathcal{N}(\mu_{a,b}, \sigma_{a,b}^2)$ that $\mathbf{U}_{a,b}$ is based on, so $\mathbf{U}_{a,b} = \left\lfloor \mathbf{U}'_{a,b} \right\rfloor$ and the distributions $\mathbf{U}'_{a,b}$ are all mutually independent. For $a \in [K]$ we define random variables $\mathbf{V}'_{a,a}$, $\mathbf{V}_{a,a}$ as

$$\begin{aligned}
\mathbf{V}'_{a,a} &= \sum_{c<a} \mathbf{U}'_{c,a} + \sum_{a<d} \operatorname{sign}(\tilde{\Sigma}_{a,d}) \cdot \mathbf{U}'_{a,d}, \\
\mathbf{V}_{a,a} &= \sum_{c<a} \left\lfloor \mathbf{U}'_{c,a} \right\rfloor + \sum_{a<d} \left\lfloor \operatorname{sign}(\tilde{\Sigma}_{a,d}) \cdot \mathbf{U}'_{a,d} \right\rfloor = \sum_{c<a} \left\lfloor \mathbf{U}'_{c,a} \right\rfloor + \sum_{a<d} \operatorname{sign}(\tilde{\Sigma}_{a,d}) \cdot \left\lfloor \mathbf{U}'_{a,d} \right\rfloor \\
&= \sum_{c<a} \mathbf{U}_{c,a} + \sum_{a<d} \operatorname{sign}(\tilde{\Sigma}_{a,d}) \cdot \mathbf{U}_{a,d}.
\end{aligned}$$

Fix a possible outcome $(u'_{a,b})_{a<b}$ of $(\mathbf{U}'_{a,b})_{a<b}$ and for each $a < b$ let $u_{a,b} = \left\lfloor u'_{a,b} \right\rceil$ be the corresponding outcome of $\mathbf{U}_{a,b}$. For $a \in [K]$ let

$$v'_{a,a} = \sum_{c<a} u'_{c,a} + \sum_{a<d} \operatorname{sign}(\tilde{\Sigma}_{a,d}) \cdot u'_{a,d},$$
$$v_{a,a} = \sum_{c<a} \lfloor u'_{c,a} \rceil + \sum_{a<d} \left\lfloor \operatorname{sign}(\tilde{\Sigma}_{a,d}) \cdot u'_{a,d} \right\rceil = \sum_{c<a} u_{c,a} + \sum_{a<d} \operatorname{sign}(\tilde{\Sigma}_{a,d}) \cdot u_{a,d}.$$

Recalling Lemma 16, we have that

$$d_{\mathrm{TV}}(\lfloor \mathbf{U}'_{a,a} + v'_{a,a} \rceil, \lfloor \mathbf{U}'_{a,a} \rceil + v_{a,a}) \leq \frac{K}{\sigma}$$

for each $a \in [K]$, and hence by independence we get that

$$d_{\mathrm{TV}}\left((\lfloor \mathbf{U}'_{1,1} + v'_{1,1} \rceil, ..., \lfloor \mathbf{U}'_{K,K} + v'_{K,K} \rceil),\ (\lfloor \mathbf{U}'_{1,1} \rceil + v_{1,1}, ..., \lfloor \mathbf{U}'_{K,K} \rceil + v_{K,K})\right) \leq \frac{K^2}{\sigma}.$$

Averaging over all outcomes of $(u'_{a,b})_{a<b} \leftarrow (\mathbf{U}'_{a,b})_{a<b}$, we get that

$$d_{\mathrm{TV}}\left((\lfloor \mathbf{U}'_{1,1} + \mathbf{V}'_{1,1} \rceil, ..., \lfloor \mathbf{U}'_{K,K} + \mathbf{V}'_{K,K} \rceil),\ (\lfloor \mathbf{U}'_{1,1} \rceil + \mathbf{V}_{1,1}, ..., \lfloor \mathbf{U}'_{K,K} \rceil + \mathbf{V}_{K,K})\right) \leq \frac{K^2}{\sigma}.$$

To complete the proof it remains to show that the vector-valued random variable

$$(\lfloor \mathbf{U}'_{1,1} + \mathbf{V}'_{1,1} \rceil, \ldots, \lfloor \mathbf{U}'_{K,K} + \mathbf{V}'_{K,K} \rceil)$$

is distributed according to $\mathcal{N}_D(\tilde{\mu}, \tilde{\Sigma})$. It is straightforward to verify, using (19) and linearity of expectation, that $\mathbf{E}[\mathbf{U}'_{a,a} + \mathbf{V}'_{a,a}] = \tilde{\mu}_a$. For the covariance matrix, we first consider the diagonal terms: we have $\mathbf{Var}[\mathbf{U}'_{a,a} + \mathbf{V}'_{a,a}] = \tilde{\Sigma}_{a,a}$ by the second part of (20) and independence of the $\mathbf{U}'_{a,b}$ distributions. Finally, for the off-diagonal terms, for $a < b$ we have

$$\begin{aligned}
&\mathbf{Cov}(\mathbf{U}'_{a,a} + \mathbf{V}'_{a,a}, \mathbf{U}'_{b,b} + \mathbf{V}'_{b,b}) \\
&= \mathbf{Cov}\left(\mathbf{U}'_{a,a} + \sum_{c<a} \mathbf{U}'_{c,a} + \sum_{a<d} \operatorname{sign}(\tilde{\Sigma}_{a,d}) \cdot \mathbf{U}'_{a,d}, \mathbf{U}'_{b,b} + \sum_{c<b} \mathbf{U}'_{c,b} + \sum_{b<d} \operatorname{sign}(\tilde{\Sigma}_{b,d}) \cdot \mathbf{U}'_{b,d}\right) \\
&= \mathbf{Cov}(\operatorname{sign}(\tilde{\Sigma}_{a,b}) \cdot \mathbf{U}'_{a,b}, \mathbf{U}'_{a,b}) = \operatorname{sign}(\tilde{\Sigma}_{a,b}) \cdot \mathbf{Var}[\mathbf{U}'_{a,b}] = \operatorname{sign}(\tilde{\Sigma}_{a,b}) \cdot \sigma^2_{a,b} = \tilde{\Sigma}_{a,b} = \tilde{\Sigma}_{b,a}
\end{aligned}$$

as desired. □

We would like a variant of Lemma 37 where signed PBDs play the role of discretized Gaussians. This is given by the following lemma. (Note that the lemma also ensures that every nontrivial signed PBD has high variance; this will be useful later.)

Lemma 38. *Given the $\tilde{\mu} \in \mathbb{R}^K, \tilde{\Sigma} \in \mathbb{R}^{K \times K}$ from Lemma 36 and the $\mu_{a,b}$, $\sigma_{a,b}$ and σ^2 defined in Lemma 37, there exist signed PBDs $\mathbf{W}_{a,b}$, $1 \leq a \leq b \leq K$, each of which is either trivial (a constant random variable) or has $\mathbf{Var}[\mathbf{W}_{a,b}] \geq \sigma^{1/2} = R^{1/4}/K^{1/4}$, such that the random variables $\mathbf{S}_a$, $a \in [K]$, defined as*

$$\mathbf{S}_a = \mathbf{W}_{a,a} + \sum_{c<a} \mathbf{W}_{c,a} + \sum_{a<d} \operatorname{sign}(\tilde{\Sigma}_{a,d}) \mathbf{W}_{a,d}, \tag{21}$$

satisfy

$$d_{\mathrm{TV}}\left((\mathbf{S}_1, \ldots, \mathbf{S}_K), \mathcal{N}_D(\tilde{\mu}, \tilde{\Sigma})\right) \leq O\left(\frac{K^2}{\sigma^{1/4}}\right) = O\left(\frac{K^{17/8}}{R^{1/8}}\right).$$

Proof. Let $\mathbf{U}_{a,b}$, $\mathbf{X}_a$ be as defined in Lemma 37. We "swap out" each discretized Gaussian $\mathbf{U}_{a,b} = \mathcal{N}_D(\mu_{a,b}, \sigma^2_{a,b})$ in $\mathbf{X}_a$ for a signed PBD $\mathbf{W}_{a,b}$ as follows: Given $1 \le a \le b \le K$,

(I) If $\sigma^2_{a,b} \ge \sigma^{1/2}$, we define $\mathbf{W}_{a,b}$ to be a signed PBD that has $|\mathbf{E}[\mathbf{W}_{a,b}] - \mu_{a,b}| \le 1/2$ and $\mathbf{Var}[\mathbf{W}_{a,b}] = \sigma^2_{a,b}$. (To see that there exists such a signed PBD, observe that we can take N_1 many Bernoulli random variables each with expectation p, satisfying $N_1 p(1-p) = \sigma^2_{a,b}$, to exactly match the variance, and then take an additional N_2 many constant-valued random variables (each of which is 1 or -1 depending on whether $N_1 p$ is greater or less than $\mu_{a,b}$) to get the mean of the signed PBD to lie within an additive $1/2$ of $\mu_{a,b}$.)

(II) If $\sigma^2_{a,b} < \sigma^{1/2}$ we define $\mathbf{W}_{a,b}$ to be a trivial signed PBD that has $\mathbf{E}[\mathbf{W}_{a,b}] = \lfloor \mu_{a,b} \rceil$ and $\mathbf{Var}[\mathbf{W}_{a,b}] = 0$.

In the above definition all $\mathbf{W}_{a,b}$'s are independent of each other. We note that Lemma 37 implies that when $b = a$ the PBD $\mathbf{W}_{a,a}$ has $\mathbf{Var}[\mathbf{W}_{a,a}] = \sigma^2_{a,a} \ge \sigma^2 = R/K \gg \sigma^{1/2}$, and hence the PBD $\mathbf{W}_{a,a}$ falls into the "large-variance" Case (I) above.

The random variable $\mathbf{S}_a$ defined in Equation (21) is the analogue of $\mathbf{X}_a$ from Lemma 37 but with $\mathbf{W}_{a,b}$ replacing each $\mathbf{U}_{a,b}$. To establish the variation distance bound, fix $a \in [K]$; we first argue that the variation distance between $\mathbf{S}_a$ and $\mathbf{X}_a$ is small. We start by observing that since $\mathbf{Var}[\mathbf{W}_{a,a}] \ge \sigma^2$, Theorem 17 and Lemma 16 give

$$d_{\mathrm{TV}}(\mathbf{W}_{a,a}, \mathbf{U}_{a,a}) \le O(1/\sigma), \tag{22}$$

and moreover $\mathbf{W}_{a,a}$ is $O(1/\sigma)$-shift-invariant by Fact 18.

Now consider a $c < a$ such that $\mathbf{W}_{c,a}$ falls into Case (II). By the standard concentration bound for the Gaussian $\mathbf{U}'_{c,a} \sim \mathcal{N}(\mu_{c,a}, \sigma^2_{c,a})$ on which $\mathbf{U}_{c,a}$ is based, we have that $\mathbf{Pr}[\mathbf{U}'_{c,a} \notin [\mu_{c,a} - t\sigma_{c,a}, \mu_{c,a} + t\sigma_{c,a}]] \le 2e^{-t^2/2}$ for all $t > 0$. It follows from Claim 39 (stated and justified below) and the $O(1/\sigma)$-shift-invariance of $\mathbf{W}_{a,a}$ that

$$d_{\mathrm{TV}}(\mathbf{W}_{a,a} + \mathbf{W}_{c,a}, \mathbf{W}_{a,a} + \mathbf{U}_{c,a}) \le O\left(\frac{t\sigma_{c,a} + 1}{\sigma}\right) + 2e^{-t^2/2}.$$

Selecting $t = \sigma^{1/4}$ so that $t\sigma_{c,a} + 1 \le \sigma^{1/4} \cdot \sigma^{1/4} + 1 = O(\sigma^{1/2})$ and $e^{-t^2/2} = o(1/\sigma^{1/2})$, we get that

$$d_{\mathrm{TV}}(\mathbf{W}_{a,a} + \mathbf{W}_{c,a}, \mathbf{W}_{a,a} + \mathbf{U}_{c,a}) \le O\left(\frac{1}{\sigma^{1/2}}\right). \tag{23}$$

A similar argument holds for each $d > a$ such that $\mathbf{W}_{a,d}$ falls into Case (II), giving

$$d_{\mathrm{TV}}(\mathbf{W}_{a,a} + \mathrm{sign}(\tilde{\Sigma}_{a,d}) \cdot \mathbf{W}_{a,d}, \mathbf{W}_{a,a} + \mathrm{sign}(\tilde{\Sigma}_{a,d}) \cdot \mathbf{U}_{a,d}) \le O\left(\frac{1}{\sigma^{1/2}}\right). \tag{24}$$

Finally, for each $c < a$ such that $\mathbf{W}_{c,a}$ falls into Case (I), once again applying Theorem 17 and Lemma 16, we get

$$d_{\mathrm{TV}}(\mathbf{W}_{c,a}, \mathbf{U}_{c,a}) \le O(1/\sigma^{1/4}), \tag{25}$$

and similarly for $d > a$ such that $\mathbf{W}_{a,d}$ falls into Case (I) we have

$$d_{\mathrm{TV}}(\mathrm{sign}(\tilde{\Sigma}_{a,d}) \cdot \mathbf{W}_{a,d}, \mathrm{sign}(\tilde{\Sigma}_{a,d}) \cdot \mathbf{U}_{a,d}) \le O(1/\sigma^{1/4}). \tag{26}$$

Combining (22—25) and recalling the definitions of $\mathbf{S}_a$ and $\mathbf{X}_a$, by the triangle inequality for each $a \in [K]$ we have

$$d_{\mathrm{TV}}(\mathbf{S}_a, \mathbf{X}_a) \leq O\left(\frac{K}{\sigma^{1/4}}\right).$$

Finally, another application of the triangle inequality gives

$$d_{\mathrm{TV}}((\mathbf{S}_1, \ldots, \mathbf{S}_K), (\mathbf{X}_1, \ldots, \mathbf{X}_K)) \leq O\left(\frac{K^2}{\sigma^{1/4}}\right),$$

which with Lemma 37 gives the claimed bound. □

The following claim is an easy consequence of the definition of shift-invariance:

Claim 39. *Let $\mathbf{A}$ be an integer random variable that is α-shift-invariant, and let $\mathbf{B}$ be an integer random variable such that $\mathbf{Pr}[\mathbf{B} \notin [u, u+r]] \leq \delta$ for some integers u, r. Then for any integer $r' \in [u, u+r]$ we have $d_{\mathrm{TV}}(\mathbf{A}+\mathbf{B}, \mathbf{A}+r') \leq \alpha r + \delta$.*

7.3 $\mathbf{S}'$ is close to a shifted weighted sum of signed PBDs

Recall that $\mathbf{S}' = (q_1, \ldots, q_K) \cdot \mathbf{M}$ is as defined in (12). Combining Lemmas 36 and 38, and taking the dot-product with $(q_1, \ldots, q_K)$ to pass from $\mathbf{M}$ to $\mathbf{S}'$, we get that the variation distance between $\mathbf{S}' = (q_1, \ldots, q_K) \cdot \mathbf{M}$ and $(q_1, \ldots, q_K) \cdot (\mathbf{S}_1, \ldots, \mathbf{S}_K)$ is at most $O(K^{71/20}/R^{1/20})$. We can express $(q_1, \ldots, q_K) \cdot (\mathbf{S}_1, \ldots, \mathbf{S}_K)$ as

$$\begin{aligned}
&\sum_{a=1}^{K} q_a \left(\mathbf{W}_{a,a} + \sum_{c<a} \mathbf{W}_{c,a} + \sum_{a<d} \mathrm{sign}(\tilde{\Sigma}_{a,d}) \mathbf{W}_{a,d} \right) \\
&= \sum_{a=1}^{K} q_a \mathbf{W}_{a,a} + \sum_{1 \leq a < b \leq K} (q_b + \mathrm{sign}(\tilde{\Sigma}_{a,b}) \cdot q_a) \mathbf{W}_{a,b}.
\end{aligned}$$

Recalling that each $\mathbf{W}_{a,b}$ is either a constant random variable or a signed PBD with variance at least $\sigma^{1/2} = R^{1/4}/K^{1/4}$, that each $\mathbf{Var}[\mathbf{W}_{a,a}] \geq \sigma^2 > \sigma^{1/2}$, and that all of the distributions $\mathbf{W}_{a,a}, \mathbf{W}_{a,b}$ are mutually independent, we get the following result showing that $c_{q_1}, \ldots, c_{q_K} \geq R$ implies that $\mathbf{S}'$ is close to a weighted sum of signed PBDs.

Lemma 40. *Assume that $c_{q_1}, \ldots, c_{q_K} \geq R$. Then there is an integer V', a subset of pairs $A \subseteq \{(a,b) : 1 \leq a < b \leq K\}$, and a set of sign values $\{\tau_{a,b}\}_{(a,b) \in A}$ where each $\tau_{a,b} \in \{-1, 1\}$, such that $d_{\mathrm{TV}}(\mathbf{S}', \mathbf{B}) = O(K^{71/20}/R^{1/20})$, where $\mathbf{B}$ is a shifted sum of signed PBDs*

$$\mathbf{B} = V' + \sum_{a=1}^{K} q_a \mathbf{W}_{a,a} + \sum_{(a,b) \in A} (q_b + \tau_{a,b} \cdot q_a) \mathbf{W}_{a,b} \tag{27}$$

in which all the $\mathbf{W}_{a,a}$ and $\mathbf{W}_{a,b}$ distributions are independent signed PBDs with variance at least $R^{1/4}/K^{1/4}$.

7.4 A useful limit theorem: Simplifying by coalescing multiple large-variance scaled PBDs into one

Lemma 40 leads to consideration of distributions of the form $\mathbf{T} = r_1\mathbf{T}_1 + \cdots + r_D\mathbf{T}_D$, where $\mathbf{T}_1, \ldots, \mathbf{T}_D$ are independent signed large-variance PBDs. Let us consider for a moment the case that $D = 2$, so that $\mathbf{T} = r_1\mathbf{T}_1 + r_2\mathbf{T}_2$. (As we will see, to handle the case of general D it suffices to consider this case.) Since $\gcd(r_1, r_2)$ divides every outcome of $\mathbf{T}$, we may assume that $\gcd(r_1, r_2) = 1$ essentially without loss of generality. When $\gcd(r_1, r_2) = 1$, if the variance of $\mathbf{T}_2$ is large enough relative to r_1, then the gaps between multiples of r_1 are filled in, and $\mathbf{T}$ is closely approximated by a single PBD. This is helpful for learning, because it means that cases in which $\mathbf{Var}[\mathbf{T}_2]$ is this large are subsumed by cases in which there are fewer PBDs. This phenomenon is the subject of Lemma 41.

Lemma 41. *Let $\mathbf{T} = r_1\mathbf{T}_1 + \cdots + r_D\mathbf{T}_D$ where $\mathbf{T}_1, \ldots, \mathbf{T}_D$ are independent signed PBDs and $r_1, \ldots, r_D$ are nonzero integers with the following properties:*

- $\mathbf{Var}[r_1\mathbf{T}_1] \geq \frac{1}{D}\mathbf{Var}[\mathbf{T}]$*;*
- *For each $a \in \{2, \ldots, D\}$ we have $\mathbf{Var}[\mathbf{T}_a] \geq \max\{\sigma_{\min}^2, \left(\frac{r_1}{\varepsilon'}\right)^2\}$, where $\sigma_{\min}^2 \geq (1/\varepsilon')^8$.*

Let $r' = \gcd(r_1, \ldots, r_D)$. Then there is a signed PBD $\mathbf{T}'$ with $\mathbf{Var}[r'\mathbf{T}'] = \mathbf{Var}[\mathbf{T}]$ such that

$$d_{\mathrm{TV}}\left(\mathbf{T},\ r'\mathbf{T}'\right) \leq O(D\varepsilon').$$

Proof. **Reduction to the case that $D = 2$.** We begin by showing that the case $D = 2$ implies the general case by induction, and thus it suffices to prove the $D = 2$ case. Let us suppose that we have proved the lemma in the $D = 2$ case and in the $D = t - 1$ case; we now use these to prove the $D = t$ case. By the $D = 2$ case, there is an absolute constant $C > 0$ and a signed PBD $\mathbf{T}_{12}$ such that we have

$$d_{\mathrm{TV}}(r_1\mathbf{T}_1 + r_2\mathbf{T}_2, \gcd(r_1, r_2)\mathbf{T}_{12}) \leq C\varepsilon' \text{ and } \mathbf{Var}[r_1\mathbf{T}_1 + r_2\mathbf{T}_2] = \mathbf{Var}[\gcd(r_1, r_2)\mathbf{T}_{12}]. \tag{28}$$

Since for all $a = 3, ..., t$ we have

$$\mathbf{Var}[\mathbf{T}_a] \geq \left(\frac{r_1}{\varepsilon'}\right)^2 \geq \left(\frac{\gcd(r_1, r_2)}{\varepsilon'}\right)^2,$$

the $D = t - 1$ case implies that, if $\mathbf{T}_{12}, \mathbf{T}_3, ..., \mathbf{T}_T$ are mutually independent, then there is a PBD $\mathbf{T}'$ such that

$$d_{\mathrm{TV}}\left(\gcd(r_1, r_2)\mathbf{T}_{12} + \sum_{a=3}^{t} r_a\mathbf{T}_a,\ r'\mathbf{T}'\right) \leq C(t-1)\varepsilon'$$

and

$$\mathbf{Var}[r'\mathbf{T}'] = \mathbf{Var}\left[\gcd(r_1, r_2)\mathbf{T}_{12} + \sum_{a=3}^{t} r_a\mathbf{T}_a\right] = \mathbf{Var}[\gcd(r_1, r_2)\mathbf{T}_{12}] + \mathbf{Var}\left[\sum_{a=3}^{t} r_a\mathbf{T}_a\right],$$

which, combined with (28), completes the proof of the $D = t$ case. We thus subsequently focus on the $D = 2$ case.

Reduction to the case that $r' = 1$. Next, we note that we may assume without loss of generality that $r' = \gcd(r_1, r_2) = 1$, since dividing each r_a by r' scales down $\mathbf{Var}[r_1\mathbf{T}_1 + r_2\mathbf{T}_2]$ by $(r')^2$ and $d_{\mathrm{TV}}\left(r_1\mathbf{T}_1 + r_2\mathbf{T}_2,\ r'\mathbf{T}'\right)$ is easily seen to equal $d_{\mathrm{TV}}\left((r_1/r')\mathbf{T}_1 + (r_2/r')\mathbf{T}_2,\ \mathbf{T}'\right)$.

Main proof of the $D = 2, r' = 1$ case. Recall that $\mathbf{T} = r_1\mathbf{T}_1 + r_2\mathbf{T}_2$. Let μ denote $\mathbf{E}[\mathbf{T}]$, and let σ^2 denote $\mathbf{Var}[\mathbf{T}]$.

As in [GMRZ11], we will use shift-invariance to go from bounds on d_{K} to bounds on d_{TV}. Our first step is to give a bound on d_{K}. For this we will use the following well-known Berry-Esseen-like inequality, which can be shown using Lemma 13, Theorem 17 and Gaussian anti-concentration:

Lemma 42. *There is a universal constant c such*

$$d_{\mathrm{K}}(TP(\mu, \sigma^2), N(\mu, \sigma^2)) \leq \frac{c}{\sigma}$$

for all μ and all $\sigma^2 > 0$.

Now we are ready for our bound on the Kolmogorov distance:

Lemma 43. $d_{\mathrm{K}}(\mathbf{T}, TP(\mu, \sigma^2)) \leq O(1/\sigma_{\min})$.

Proof: Lemma 13 implies that for $a = 1, 2$ we have

$$d_{\mathrm{K}}(\mathbf{T}_a, TP(\mu(\mathbf{T}_a), \sigma(\mathbf{T}_a)^2)) \leq O(1/\sigma_{\min}),$$

which directly implies

$$d_{\mathrm{K}}(r_a\mathbf{T}_a, r_aTP(\mu(\mathbf{T}_a), \sigma(\mathbf{T}_a)^2)) \leq O(1/\sigma_{\min}).$$

Lemma 42 and the triangle inequality then give

$$d_{\mathrm{K}}(r_a\mathbf{T}_a, N(r_a\mu(\mathbf{T}_a), r_a^2\sigma(\mathbf{T}_a)^2)) \leq O(1/\sigma_{\min}),$$

and applying Lemma 42 and the triangle inequality again, we get

$$d_{\mathrm{K}}(r_a\mathbf{T}_a, TP(r_a\mu(\mathbf{T}_a), r_a^2\sigma(\mathbf{T}_a)^2)) \leq O(1/\sigma_{\min}).$$

The lemma follows from the fact that $d_{\mathrm{K}}(\mathbf{X} + \mathbf{Y}, \mathbf{X}' + \mathbf{Y}') \leq d_{\mathrm{K}}(\mathbf{X}, \mathbf{X}') + d_{\mathrm{K}}(\mathbf{Y}, \mathbf{Y}')$ when $\mathbf{X}, \mathbf{Y}$ are independent and $\mathbf{X}', \mathbf{Y}'$ are independent. □

Facts 8 and 18 together imply that $\mathbf{T}$ is $O(1/\sigma_{\min})$-shift invariant at scales r_1 and r_2, but, to apply Lemma 10, we need it to be shift-invariant at a smaller scale. Very roughly, we will do this by effecting a small shift using a few shifts with steps with sizes in $\{r_1, r_2\}$. The following generalization of Bézout's Identity starts to analyze our ability to do this.

Lemma 44. *Given any integer $0 \leq u < r_1 \cdot r_2$, there are integers v_1, v_2 such that $u = v_1 \cdot r_1 + v_2 \cdot r_2$ with $|v_1| < r_2, |v_2| < r_1$.*

Proof. By Bézout's Identity, there exist x_1 and x_2 with $|x_1| < r_2$ and $|x_2| < r_1$ such that

$$x_1r_1 + x_2r_2 = 1.$$

Let y_1 be obtained by adding r_2 to x_1 if x_1 is negative, and otherwise just taking x_1, and define y_2 similarly; i.e., $y_1 = x_1 + r_2\mathbf{1}[x_1 < 0]$ and $y_2 = x_2 + r_1\mathbf{1}[x_2 < 0]$. Then

$$y_1r_1 + y_2r_2 = 1 \mod (r_1r_2)$$

and $0 \leq y_1 < r_2$ and $0 \leq y_2 < r_1$. Thus

$$uy_1r_1 + uy_2r_2 = u \mod (r_1r_2).$$

This in turn implies that

$$u = uy_2r_2 \mod r_1 \text{ and } u = uy_1r_1 \mod r_2,$$

so if $z_1 \in \{0, 1, \ldots, r_2 - 1\}$ and $z_2 \in \{0, 1, \ldots, r_1 - 1\}$ satisfy $z_1 = uy_1 \bmod r_2$ and $z_2 = uy_2 \bmod r_1$, we get

$$(z_1r_1 + z_2r_2) = u \mod r_1$$

and

$$(z_1r_1 + z_2r_2) = u \mod r_1.$$

By the Chinese Remainder Theorem, $z_1r_1 + z_2r_2 = u \bmod (r_1r_2)$. Furthermore, as $0 \le z_1 < r_2$ and $0 \le z_2 < r_1$, we have $0 \le z_1r_1 + z_2r_2 < 2r_1r_2$. If $z_1r_1 + z_2r_2 < r_1r_2$, then we are done; we can set $v_1 = z_1$ and $v_2 = z_2$. If not, either $z_1 > 0$ or $z_2 > 0$. If $z_1 > 0$, setting $v_1 = z_1 - r_2$ and $v_2 = z_2$ makes $z_1r_1 + z_2r_2 = z_1r_1 + z_2r_2 - r_1r_2 = u$, and the corresponding modification of z_2 works if $z_2 > 0$. □

Armed with Lemma 44, we are now ready to work on the "local" shift-invariance of $\mathbf{T}$. The following more general lemma will do the job.

Lemma 45. *Let $\mathbf{X}, \mathbf{Y}$ be independent integer random variables where $\mathbf{X}$ is α-shift-invariant at scale 1 and $\mathbf{Y}$ is β-shift-invariant at scale 1. Let $\mathbf{Z} = r_1 \cdot \mathbf{X} + r_2 \cdot \mathbf{Y}$. Then for any positive integer d we have $d_{\mathrm{TV}}(\mathbf{Z}, \mathbf{Z} + d) \le r_2\alpha + r_1\beta + \min\left\{\frac{d}{r_1}\alpha, \frac{d}{r_2}\beta\right\}$.*

Proof. Note that $d = s \cdot r_1 \cdot r_2 + z$ where $0 \le z < r_1 \cdot r_2$, $0 \le s \le d/(r_1 \cdot r_2)$, and s is an integer. By using Lemma 44, we have that $d = s \cdot r_1 \cdot r_2 + v_1 \cdot r_1 + v_2 \cdot r_2$ where $|v_1| < r_2$ and $|v_2| < r_1$, so $d = (s \cdot r_2 + v_1) \cdot r_1 + v_2 \cdot r_2$. Note that

$$|s \cdot r_2 + v_1| \le |v_1| + s \cdot r_2 \le (r_2 - 1) + d/r_1.$$

Thus, $d = t_1 \cdot r_1 + t_2 \cdot r_2$ where t_1, t_2 are integers and $|t_1| \le r_2 - 1 + d/r_1$ and $|t_2| \le (r_1 - 1)$. Hence we have

$$\begin{aligned} d_{\mathrm{TV}}(\mathbf{Z}, \mathbf{Z} + d) &= d_{\mathrm{TV}}(r_1 \cdot \mathbf{X} + r_2 \cdot \mathbf{Y}, r_1 \cdot \mathbf{X} + r_2 \cdot \mathbf{Y} + t_1 \cdot r_1 + t_2 \cdot r_2) \\ &\le |t_1| \cdot \alpha + |t_2| \cdot \beta \\ &\le \left(\frac{d}{r_1} + r_2\right) \cdot \alpha + r_1 \cdot \beta. \end{aligned}$$

By swapping the roles of r_1 and r_2 in the above analysis, we get the stated claim. □

Now we have everything we need to prove Lemma 41 in the case that $D = 2$.

Let $\mathbf{V} = TP(\mu, \sigma^2)$. Let $\mathbf{U}_d$ denote the uniform distribution over $\{0, 1, \ldots, d - 1\}$, where d will be chosen later. We will bound $d_{\mathrm{TV}}(\mathbf{T} + \mathbf{U}_d, \mathbf{V} + \mathbf{U}_d)$, $d_{\mathrm{TV}}(\mathbf{T}, \mathbf{T} + \mathbf{U}_d)$, and $d_{\mathrm{TV}}(\mathbf{V}, \mathbf{V} + \mathbf{U}_d)$, and apply the triangle inequality via

$$d_{\mathrm{TV}}(\mathbf{T}, \mathbf{V}) \le d_{\mathrm{TV}}(\mathbf{T}, \mathbf{T}+\mathbf{U}_d)+d_{\mathrm{TV}}(\mathbf{T}+\mathbf{U}_d, \mathbf{V}) \le d_{\mathrm{TV}}(\mathbf{T}, \mathbf{T}+\mathbf{U}_d)+d_{\mathrm{TV}}(\mathbf{T}+\mathbf{U}_d, \mathbf{V}+\mathbf{U}_d)+d_{\mathrm{TV}}(\mathbf{V}, \mathbf{V}+\mathbf{U}_d).$$

First, recalling that $\mathbf{T} = r_1 \cdot \mathbf{T}_1 + r_2 \cdot \mathbf{T}_2$ and that (by Fact 18) $\mathbf{T}_1$ is $O(1/\sigma(\mathbf{T}_1))$-shift-invariant at scale 1 and $\mathbf{T}_2$ is $O(1/\sigma(\mathbf{T}_2))$-shift-invariant at scale 1, we have that

$$
\begin{aligned}
d_{\mathrm{TV}}(\mathbf{T}, \mathbf{T}+\mathbf{U}_d) &\le \mathbf{E}_{x\sim\mathbf{U}_d}\left[d_{\mathrm{TV}}(\mathbf{T}, \mathbf{T}+x)\right] \\
&\le \max_{x\in\mathrm{supp}(\mathbf{U}_d)} d_{\mathrm{TV}}(\mathbf{T}, \mathbf{T}+x) \\
&\le O\left(\frac{r_1}{\sigma(\mathbf{T}_2)} + \frac{r_2}{\sigma(\mathbf{T}_1)} + \min\left\{\frac{d}{r_1\sigma(\mathbf{T}_1)}, \frac{d}{r_2\sigma(\mathbf{T}_2)}\right\}\right) \qquad \text{(by Lemma 45)} \\
&\le O\left(\frac{r_1}{\sigma(\mathbf{T}_2)} + \frac{d}{r_1\sigma(\mathbf{T}_1)}\right) \qquad \text{(since } r_1\sigma(\mathbf{T}_1) > r_2\sigma(\mathbf{T}_2)\text{)} \\
&\le O(\varepsilon') + O\left(\frac{d}{r_1\sigma(\mathbf{T}_1)}\right),
\end{aligned}
$$

since $\sigma(\mathbf{T}_2) > r_1/\varepsilon'$.

Next, Fact 18 implies that $\mathbf{V}$ is $O\left(\frac{1}{r_1\sigma(\mathbf{T}_1)}\right)$-shift-invariant, so repeated application of the triangle inequality gives

$$
d_{\mathrm{TV}}(\mathbf{V}, \mathbf{V}+\mathbf{U}_d) \le O\left(\frac{d}{r_1\sigma(\mathbf{T}_1)}\right).
$$

Finally, we want to bound $d_{\mathrm{TV}}(\mathbf{T}+\mathbf{U}_d, \mathbf{V}+\mathbf{U}_d)$. Observe that $\mathbf{Pr}[|\mathbf{T}-\mu| < \sigma/\varepsilon']$ and $\mathbf{Pr}[|\mathbf{V}-\mathbf{E}[\mathbf{V}]| \le \sigma/\varepsilon']$ are both $2^{-\mathrm{poly}(1/\varepsilon')}$. Hence applying Lemma 10 and recalling that $r_1\sigma(\mathbf{T}_1) > r_2\sigma(\mathbf{T}_2)$, we get

$$
d_{\mathrm{TV}}(\mathbf{T}+\mathbf{U}_d, \mathbf{V}+\mathbf{U}_d) \le o(\varepsilon') + O\left(\sqrt{(1/\sigma_{\min})\cdot((r_1\sigma(\mathbf{T}_1))/\varepsilon')\cdot(1/d)}\right).
$$

Combining our bounds, we get that

$$
d_{\mathrm{TV}}(\mathbf{T}, \mathbf{V}) \le O(\varepsilon') + O\left(\sqrt{(1/\sigma_{\min})\cdot((r_1\sigma(\mathbf{T}_1))/\varepsilon')\cdot(1/d)} + \frac{d}{r_1\sigma(\mathbf{T}_1)}\right).
$$

Taking $d = r_1\sigma(\mathbf{T}_1)/(\sigma_{\min}\varepsilon')^{1/3}$, we get

$$
d_{\mathrm{TV}}(\mathbf{T}, \mathbf{V}) \le O(\varepsilon') + 1/(\sigma_{\min}\varepsilon')^{1/3} = O(\varepsilon')
$$

since $(1/\sigma_{\min})^2 > (1/\varepsilon')^8$.

Finally, let $\mathbf{T}'$ be a signed PBD that has $|\mathbf{E}[\mathbf{T}'] - \mu| \le 1/2$ and $\mathbf{Var}[\mathbf{T}'] = \sigma^2$. (The existence of such a signed PBD can be shown as in (I) in the proof of Lemma 38.) Lemmas 13 and 14 imply that $d_{\mathrm{TV}}(\mathbf{T}_{\mathrm{MIX}'}, \mathbf{V}) \le 1/\sigma(\mathbf{V}) \le \varepsilon'$, completing the proof. □

8 The learning result: Learning when $|\mathcal{A}| \ge 4$

With the kernel-based learning results from Section 5 and the structural results from Section 7 in hand, we are now ready to learn a distribution $\mathbf{S}^*$ that is $c\varepsilon$-close to a distribution $\mathbf{S} = \mathbf{S}' + V$, where $\mathbf{S}'$ is described in Section 6. We give two distinct learning algorithms, one for each of two mutually exclusive cases. The overall learning algorithm works by running both algorithms and using the hypothesis selection procedure, Proposition 26, to construct one final hypothesis.

The high-level idea is as follows. In Section 8.1 we first easily handle a special case in which all the c_{q_a} values are "small," essentially using a brute-force algorithm which is not too inefficient since all c_{q_a}'s

are small. We then turn to the remaining general case, which is that some c_{q_a} are large while others may be small.

The idea of how we handle this general case is as follows. First, via an analysis in the spirit of the "Light-Heavy Experiment" from [DDO+13], we approximate the distribution $\mathbf{S}' + V$ as a sum of two independent distributions $\mathbf{S}_{\text{light}} + \mathbf{S}_{\text{heavy}}$ where intuitively $\mathbf{S}_{\text{light}}$ has "small support" and $\mathbf{S}_{\text{heavy}}$ is a 0-moded $\mathcal{A}$-sum supported on elements all of which have large weight (this is made precise in Lemma 46). Since $\mathbf{S}_{\text{light}}$ has small support, it is helpful to think of $\mathbf{S}_{\text{light}} + \mathbf{S}_{\text{heavy}}$ as a mixture of shifts of $\mathbf{S}_{\text{heavy}}$. We then use structural results from Section 7 to approximate this distribution in turn by a mixture of not-too-many shifts of a weighted sum of signed PBDs, whose component independent PBDs satisfy a certain technical condition on their variances (see Corollary 48). Finally, we exploit the kernel-based learning tools developed in Section 5 to give an efficient learning algorithm for this mixture distribution. Very roughly speaking, the final $\log \log a_k$ sample complexity dependence (ignoring other parameters such as ε and k) comes from making $O(\log a_k)$ many "guesses" for parameters (variances) of the weighted sum of signed PBDs; this many guesses suffice because of the technical condition alluded to above.

We now proceed to the actual analysis. Let us reorder the sequence $q_1, \ldots, q_K$ so that $c_{q_1} \leq \ldots \leq c_{q_K}$. Let us now define the sequence $t_1, \ldots, t_K$ as $t_a = (1/\varepsilon)^{2^a}$. (For intuition on the conceptual role of the t_i's, the reader may find it helpful to review the discussion given in the "Our analysis" subsection of Section 2.1.) Define the "largeness index" of the sequence $c_{q_1} \leq \ldots \leq c_{q_K}$ as the minimum $\ell \in [K]$ such that $c_{q_\ell} > t_\ell$, and let ℓ_0 denote this value. If there is no $\ell \in [K]$ such that $c_{q_\ell} > t_\ell$, then we set $\ell_0 = K + 1$.

We first deal with the easy special case that $\ell_0 = K + 1$ and then turn to the main case.

8.1 Learning when $\ell_0 = K + 1$

Intuitively, in this case all of $c_{q_1}, \ldots, c_{q_K}$ are "not too large" and we can learn via brute force. More precisely, since each $c_{q_a} \leq 1/\varepsilon^{2^K}$, in a draw from $\mathbf{S}'$ the expected number of random variables $\mathbf{X}'_1, \ldots, \mathbf{X}'_N$ that take a nonzero value is at most K/ε^{2^K}, and a Chernoff bound implies that in a draw from $\mathbf{S}'$ we have $\mathbf{Pr}[\text{more than } \text{poly}(K/\varepsilon^{2^K}) \text{ of the } \mathbf{X}'_i\text{'s take a nonzero value}] \leq \varepsilon$. Note that for any M, there are at most $M^{O(K)}$ possible outcomes for $\mathbf{S} = \mathbf{S}' + V$ that correspond to having at most M of the $\mathbf{X}'_i$'s take a nonzero value. Thus it follows that in this case the random variable $\mathbf{S}'$ (and hence $\mathbf{S}$) is ε-essentially supported on a set of size at most $M^{O(K)} = (K/\varepsilon^{2^K})^{O(K)}$. Thus $\mathbf{S}^*$ is $O(\varepsilon)$-essentially supported on a set of the same size. Hence the algorithm of Fact 25 can be used to learn $\mathbf{S}^*$ to accuracy $O(\varepsilon)$ in time $\text{poly}(1/\varepsilon^{O(2^K)}) = \text{poly}(1/\varepsilon^{2^{O(k^2)}})$.

8.2 Learning when $\ell_0 \leq K$.

Now we turn to the main case, which is when $\ell_0 \leq K$. The following lemma is an important component of our analysis of this case. Roughly speaking, it says that $\mathbf{S}'$ is close to a sum of two independent random variables, one of which ($\mathbf{S}_{\text{light}}$) has small support and the other of which ($\mathbf{S}_{\text{heavy}}$) is the sum of 0-moded random variables that all have large weight.

Lemma 46. *Suppose that $\ell_0 \leq K$. Then there exists $\tilde{\mathbf{S}} = \mathbf{S}_{\text{heavy}} + \mathbf{S}_{\text{light}}$ such that $d_{\text{TV}}(\tilde{\mathbf{S}}, \mathbf{S}') \leq O(\varepsilon)$ and the following hold:*

1. *$\mathbf{S}_{\text{heavy}}$ and $\mathbf{S}_{\text{light}}$ are independent of each other;*
2. *The random variable $\mathbf{S}_{\text{light}}$ is $\mathbf{S}_{\text{light}} = \sum_{1 \leq b < \ell_0} q_b \cdot \mathbf{S}_b$ where for each $1 \leq b < \ell_0$, $\mathbf{S}_b$ is supported on the set $[-(1/\varepsilon) \cdot t_{\text{cutoff}}, (1/\varepsilon) \cdot t_{\text{cutoff}}] \cap \mathbb{Z}$ where $t_{\text{cutoff}} = (t_1 + \ldots + t_{\ell_0 - 1})$ and the $\{\mathbf{S}_b\}$ are not necessarily independent of each other;*

3. *The random variable* $\mathbf{S}_{\text{heavy}}$ *is the sum of* 0*-moded random variables supported in* $\{0, \pm q_{\ell_0}, \ldots, \pm q_K\}$*. Further, for all* $b \geq \ell_0$*, we have* $c_{q_b,\text{heavy}} > \frac{t_{\ell_0}}{2}$ *where* $c_{q_b,\text{heavy}}$ *is defined as in Section 6 but now with respect to* $\mathbf{S}_{\text{heavy}}$ *rather than with respect to* $\mathbf{S}'$*.*

Proof. The proof follows the general lines of the proof of Theorem 4.3 of [DDO+13]. Let $\mathcal{L} = \{\pm q_1, \ldots, \pm q_{\ell_0 - 1}\}$ and $\mathcal{H} = \{0, \pm q_{\ell_0}, \ldots, \pm q_K\}$. (It may be helpful to think of $\mathcal{L}$ as the "light" integers, and $\mathcal{H}$ as "heavy" ones.) We recall the following experiment that can be used to make a draw from $\mathbf{S}'$, referred to in [DDO+13] as the "Light-Heavy Experiment":

1. [Stage 1]: Informally, sample from the conditional distributions given membership in $\mathcal{L}$. Specifically, independently we sample for each $i \in [N]$ a random variable $\underline{\mathbf{X}}'_i \in \mathcal{L}$ as follows:

$$\text{for each } b \in \mathcal{L}, \quad \underline{\mathbf{X}}'_i = b, \text{with probability } \frac{\mathbf{Pr}[\mathbf{X}'_i = b]}{\mathbf{Pr}[\mathbf{X}'_i \in \mathcal{L}]};$$

   i.e. $\underline{\mathbf{X}}'_i$ is distributed according to the conditional distribution of $\mathbf{X}'_i$, conditioning on $\mathbf{X}'_i \in \mathcal{L}$. In the case that $\mathbf{Pr}[\mathbf{X}'_i \in \mathcal{L}] = 0$ we define $\underline{\mathbf{X}}'_i = 0$ with probability 1.

2. [Stage 2]: Sample analogously for $\mathcal{H}$. Independently we sample for each $i \in [N]$ a random variable $\overline{\mathbf{X}}'_i \in \mathcal{H}$ as follows:

$$\text{for each } b \in \mathcal{H}, \quad \overline{\mathbf{X}}'_i = b, \text{with probability } \frac{\mathbf{Pr}[\mathbf{X}'_i = b]}{\mathbf{Pr}[\mathbf{X}'_i \in \mathcal{H}]};$$

   i.e. $\overline{\mathbf{X}}'_i$ is distributed according to the conditional distribution of $\mathbf{X}'_i$, conditioning on $\mathbf{X}'_i \in \mathcal{H}$.

3. [Stage 3]: Choose which $\mathbf{X}'_i$ take values in $\mathcal{L}$: sample a random subset $\mathbf{L} \subseteq [N]$, by independently including each i into $\mathbf{L}$ with probability $\mathbf{Pr}[\mathbf{X}'_i \in \mathcal{L}]$.

After these three stages we output $\sum_{i \in \mathbf{L}} \underline{\mathbf{X}}'_i + \sum_{i \notin \mathbf{L}} \overline{\mathbf{X}}'_i$ as a sample from $\mathbf{S}'$, where $\sum_{i \in \mathbf{L}} \underline{\mathbf{X}}'_i$ represents "the contribution of $\mathcal{L}$" and $\sum_{i \notin \mathbf{L}} \overline{\mathbf{X}}'_i$ "the contribution of $\mathcal{H}$." Roughly, Stages 1 and 2 provide light and heavy options for each $\mathbf{X}'_i$, and Stage 3 chooses among the options. We note that the two contributions are not independent, but they are independent conditioned on the outcome of $\mathbf{L}$. Thus we may view a draw of $\mathbf{S}'$ as a mixture, over all possible outcomes L of $\mathbf{L}$, of the distributions $\sum_{i \in L} \underline{\mathbf{X}}'_i + \sum_{i \notin L} \overline{\mathbf{X}}'_i$; i.e. we have $\mathbf{S}' = \text{Mix}_{L \leftarrow \mathbf{L}}(\sum_{i \in L} \underline{\mathbf{X}}'_i + \sum_{i \notin L} \overline{\mathbf{X}}'_i)$. This concludes the definition of the Light-Heavy Experiment.

Let $t_{\text{cutoff}} = t_1 + \ldots + t_{\ell_0 - 1}$. Note that $\mathbf{E}[|\mathbf{L}|] \leq t_{\text{cutoff}}$. Let Bad denote the set of all outcomes L of $\mathbf{L}$ such that $|L| > (1/\varepsilon) \cdot t_{\text{cutoff}}$. A standard application of the Hoeffding bound implies that $\mathbf{Pr}[\mathbf{L} \in \mathsf{Bad}] = \mathbf{Pr}[|\mathbf{L}| > (1/\varepsilon) \cdot t_{\text{cutoff}}] \leq 2^{-\Omega(1/\varepsilon)}$. It follows that if we define the distribution $\mathbf{S}''$ to be an outcome of the Light-Heavy Experiment conditioned on $\mathbf{L} \notin \mathsf{Bad}$, i.e. $\mathbf{S}'' = \text{Mix}_{L \leftarrow \mathbf{L} \mid \mathbf{L} \notin \mathsf{Bad}}(\sum_{i \in L} \underline{\mathbf{X}}'_i + \sum_{i \notin L} \overline{\mathbf{X}}'_i)$, we have that $d_{\text{TV}}(\mathbf{S}'', \mathbf{S}') \leq 2^{-\Omega(1/\varepsilon)}$. Consequently it suffices to show the existence of $\tilde{\mathbf{S}}$ satisfying the properties of the lemma such that $d_{\text{TV}}(\tilde{\mathbf{S}}, \mathbf{S}'') \leq \varepsilon$.

We will now show that for any $L_1, L_2 \notin \mathsf{Bad}$, the random variables $\mathbf{S}_{L_j} = \sum_{i \notin L_j} \overline{\mathbf{X}}'_i$ (for $j \in \{1, 2\}$) are close to each other in total variation distance. (If we think of L_1 and L_2 as different possibilities for the final step in the process of sampling from the distribution of $\mathbf{S}'$, recall that the values of $\overline{\mathbf{X}}'_i$ are *always* in $\mathcal{H}$ – loosely speaking, during the first sample from $\mathbf{S}'$ the values of $\overline{\mathbf{X}}'_i$ for $i \in L_1$ are not used, and during the second sample, the values for $i \in L_2$ are not used.) Let $L_{\text{union}} = L_1 \cup L_2$. Note that by definition

$$|L_2 \setminus L_1|, \ |L_1 \setminus L_2| \leq (1/\varepsilon) \cdot t_{\text{cutoff}}.$$

Define

$$\mathbf{S}_{L_{\text{union}}} = \sum_{i \notin L_{\text{union}}} \overline{\mathbf{X}}'_i \;=\; \mathbf{S}_{L_1} - \sum_{i \in L_2 - L_1} \overline{\mathbf{X}}'_i \;=\; \mathbf{S}_{L_2} - \sum_{i \in L_1 - L_2} \overline{\mathbf{X}}'_i.$$

Choose $b \geq \ell_0$. We have that

$$\begin{aligned} d_{\mathsf{shift},q_b}(\overline{\mathbf{X}}'_i) &= 1 - \sum_j (\min\{\mathbf{Pr}[\overline{\mathbf{X}}'_i = j], \mathbf{Pr}[\overline{\mathbf{X}}'_i = j + q_b\}) \\ &\leq 1 - \min\{\mathbf{Pr}[\overline{\mathbf{X}}'_i = 0], \mathbf{Pr}[\overline{\mathbf{X}}'_i = q_b]\} - \min\{\mathbf{Pr}[\overline{\mathbf{X}}'_i = -q_b], \mathbf{Pr}[\overline{\mathbf{X}}'_i = 0]\} \\ &= 1 - \mathbf{Pr}[\overline{\mathbf{X}}'_i = -q_b] - \mathbf{Pr}[\overline{\mathbf{X}'}_i = q_b], \end{aligned}$$

since $\overline{\mathbf{X}'}_i$ is 0-moded. By Corollary 24, this implies that

$$d_{\mathsf{shift},q_b}(\mathbf{S}_{L_{\text{union}}}) \leq \frac{O(1)}{\sqrt{\sum_{i \notin L_{\text{union}}} \mathbf{Pr}[\overline{\mathbf{X}}'_i = \pm q_b]}} \leq \frac{O(1)}{\sqrt{c_{q_b} - |L_{\text{union}}|}} \leq \sqrt{\frac{2}{t_{\ell_0}}}.$$

Here the penultimate inequality uses the fact that

$$\sum_{i \notin L_{\text{union}}} \mathbf{Pr}[\overline{\mathbf{X}}'_i = \pm q_b] = \sum_{i \in [n]} \mathbf{Pr}[\overline{\mathbf{X}}'_i = \pm q_b] - \sum_{i \in L_{\text{union}}} \mathbf{Pr}[\overline{\mathbf{X}}'_i = \pm q_b] \geq c_{q_b} - |L_{\text{union}}|.$$

The last inequality uses that

$$c_{q_b} - |L_{\text{union}}| \geq c_{q_b} - |L_1| - |L_2| \geq c_{q_b} - 2 \cdot (1/\varepsilon) \cdot t_{\text{cutoff}} \geq t_b - 2 \cdot (1/\varepsilon) \cdot t_{\text{cutoff}} \geq \frac{t_{\ell_0}}{2}.$$

As each of the summands in the sum $\sum_{i \in L_2 \setminus L_1} \overline{\mathbf{X}}'_i$ is supported on the set $\{0, \pm q_{\ell_0}, \ldots, \pm q_K\}$, viewing $\mathbf{S}_{L_1}$ as a mixture of distributions each of which is obtained by shifting $\mathbf{S}_{L_{\text{union}}}$ at most $|L_2 \setminus L_1|$ many times, each time by an element of $\{0, \pm q_{\ell_0}, \ldots, q_K\}$, we immediately obtain that

$$d_{\text{TV}}(\mathbf{S}_{L_{\text{union}}}, \mathbf{S}_{L_1}) \leq |L_2 \setminus L_1| \cdot \sqrt{\frac{2}{t_{\ell_0}}} \leq 2(1/\varepsilon) \cdot t_{\text{cutoff}} \cdot \sqrt{\frac{2}{t_{\ell_0}}} \leq O(\varepsilon).$$

Choose any $L^* \notin \mathsf{Bad}$ arbitrarily, and define $\mathbf{S}_{\text{heavy}} := \sum_{i \notin L^*} \overline{\mathbf{X}}'_i$. By the above analysis, for any $L' \notin \mathsf{Bad}$ it holds that $d_{\text{TV}}(\sum_{i \notin L'} \overline{\mathbf{X}}'_i, \mathbf{S}_{\text{heavy}}) = O(\varepsilon)$. Thus, for any outcome $L \notin \mathsf{Bad}$, we have $d_{\text{TV}}(\sum_{i \in L} \underline{\mathbf{X}}'_i + \sum_{i \notin L} \overline{\mathbf{X}}'_i, \sum_{i \in L} \underline{\mathbf{X}}'_i + \mathbf{S}_{\text{heavy}}) = O(\varepsilon)$. Define $\mathbf{S}_{\text{light}} := \text{Mix}_{L \leftarrow \mathbf{L} \mid \mathbf{L} \notin \mathsf{Bad}} \sum_{i \in L} \underline{\mathbf{X}}'_i$.

We now verify that the above-defined $\mathbf{S}_{\text{heavy}}$ and $\mathbf{S}_{\text{light}}$ indeed satisfies the claimed properties. Note that for each $L \notin \mathsf{Bad}$, $\sum_{i \in L} \underline{\mathbf{X}}'_i$ is supported on the set

$$\left\{ \sum_{b \leq \ell_0} q_b \cdot S_b : S_b \in [-(1/\varepsilon) \cdot t_{\text{cutoff}}, (1/\varepsilon) \cdot t_{\text{cutoff}}] \right\},$$

so the second property holds as well. Likewise, $\mathbf{S}_{\text{heavy}}$ is a sum of 0-moded random variables with support in $\{0, \pm q_{\ell_0}, \ldots, \pm q_K\}$. Note that we have already shown that $\sum_{i \notin L^*} \mathbf{Pr}[\overline{\mathbf{X}}'_i = \pm q_b] \geq t_{\ell_0}/2$, giving the third property. Finally, combining the fact that $\mathbf{Pr}[\mathbf{L} \in \mathsf{Bad}] \leq \varepsilon$ with $d_{\text{TV}}(\sum_{i \in L} \underline{\mathbf{X}}'_i + \sum_{i \notin L} \overline{\mathbf{X}}_i, \sum_{i \in L} \underline{\mathbf{X}}_i + \mathbf{S}_{\text{heavy}}) = O(\varepsilon)$, we obtain the claimed variation distance bound $d_{\text{TV}}(\tilde{\mathbf{S}}, \mathbf{S}') \leq O(\varepsilon)$, finishing the proof. □

With Lemma 46 in hand, we now apply Lemma 40 to the distribution $\mathbf{S}_{\text{heavy}}$ with

$$R = K^{25}/\varepsilon^{32}. \tag{29}$$

This gives the following corollary:

Corollary 47. *The distribution $\mathbf{S}'$ is δ-close in total variation distance to a distribution $\mathbf{S}'' = \mathbf{S}'_{\text{light}} + \sum_{a=\ell_0}^{K} q_a \cdot \mathbf{W}_{a,a} + \sum_{(a,b)\in A}(q_b + \tau_{a,b}\cdot q_a)\mathbf{W}_{a,b}$ where $\delta = O(K^{71/20}\cdot\varepsilon^{1/20})$ and*

1. *$A \subseteq \{(a,b) : \ell_0 \le a < b \le K\}$, $\tau_{a,b} \in \{-1,1\}$, and $\mathbf{W}_{a,a}$, $\mathbf{W}_{a,b}$ are signed PBDs.*
2. *$\mathbf{Var}[\mathbf{W}_{a,a}]$, $\mathbf{Var}[\mathbf{W}_{a,b}] \ge (R/K)^{-1/4} > K^6/\varepsilon^8$,*
3. *The random variables $\mathbf{S}'_{\text{light}}$, $\mathbf{W}_{a,a}$ and $\mathbf{W}_{a,b}$ are all independent of each other, and*
4. *$\mathbf{S}'_{\text{light}}$ is supported on a set of cardinality $M \le (2t_K/\varepsilon)^K$.*

Proof. By Lemma 46 $\mathbf{S}'$ is $O(\varepsilon)$-close to $\tilde{\mathbf{S}} = \mathbf{S}_{\text{heavy}} + \mathbf{S}_{\text{light}}$ where the decomposition of $\tilde{\mathbf{S}}$ is as described in that lemma. Applying Lemma 40, we further obtain that $\mathbf{S}_{\text{heavy}}$ is $\delta = O(K^{71/20}\cdot\varepsilon^{1/20})$-close to a distribution of the form

$$\sum_{a=\ell_0}^{K} q_a \cdot \mathbf{W}_{a,a} + \sum_{(a,b)\in A}(q_b + \tau_{a,b}\cdot q_a)\mathbf{W}_{a,b} + V',$$

with $\mathbf{W}_{a,a}$ and $\mathbf{W}_{a,b}$ satisfies the conditions stated in the corollary. Defining $\mathbf{S}'_{\text{light}} = V + \mathbf{S}_{\text{light}}$ will satisfy all the required conditions. We note that the size of the support of $\mathbf{S}'_{\text{light}}$ is the same as the size of the support of $\mathbf{S}_{\text{light}}$, so applying item (2) of Lemma 46, we get that the size of the support of $\mathbf{S}'_{\text{light}}$ is bounded by $(2t_K/\varepsilon)^K$. □

Let us look at the structure of the distribution

$$\mathbf{S}'' = \mathbf{S}'_{\text{light}} + \sum_{a=\ell_0}^{K} q_a \cdot \mathbf{W}_{a,a} + \sum_{(a,b)\in A}(q_b + \tau_{a,b}\cdot q_a)\mathbf{W}_{a,b}.$$

For $(a,b) \in A$ let $q_{(a,b)}$ denote $q_b + \tau_{a,b}\cdot q_a$, and let B denote the set $B = \{\ell_0,\dots,K\}\cup A$. For any $B' \subseteq B$, let us write $\gcd(B')$ to denote $\gcd(\{q_\alpha\}_{\alpha\in B})$. Let α^* denote the element of B for which $\mathbf{Var}[q_{\alpha^*}\cdot\mathbf{W}_{\alpha^*}]$ is largest (breaking ties arbitrarily) and let MIX denote the following subset of B:

$$\text{MIX} = \{\alpha^*\} \cup \{\alpha \in B : \mathbf{Var}[\mathbf{W}_\alpha] \ge \max\{1/\varepsilon^8, q_{\alpha^*}^2/\varepsilon^2\}\}. \tag{30}$$

By applying Lemma 41 to the distribution $\sum_{\alpha\in\text{MIX}} q_\alpha\cdot\mathbf{W}_\alpha$, with its $\sigma^2_{\min}$ set to K^6/ε^8 and its ε' set to ε, noting that $|B|\varepsilon' = O(K^2\varepsilon) = o(K^{71/20}\cdot\varepsilon^{1/20})$ we obtain the following corollary:

Corollary 48. *The distribution $\mathbf{S}'$ is $\delta' = O(K^{71/20}\cdot\varepsilon^{1/20})$-close in total variation distance to a distribution $\mathbf{S}^{(2)}$ of the following form, where $\emptyset \subsetneq \text{MIX} \subset B$ is as defined in (30):*

$$\mathbf{S}^{(2)} = \mathbf{S}'_{\text{light}} + q_{\text{MIX}}\cdot\mathbf{S}_{\text{MIX}} + \sum_{q_\alpha\in B\setminus\text{MIX}} q_\alpha\cdot\mathbf{S}_\alpha,$$

where $q_{\text{MIX}} = \gcd(\text{MIX})$ and the following properties hold:

1. *The random variables* $\mathbf{S}'_{\text{light}}$, $\mathbf{S}_{\text{MIX}}$ *and* $\{\mathbf{S}_\alpha\}_{q_\alpha \in B \setminus \text{MIX}}$ *are independent of each other.*

2. $\mathbf{S}'_{\text{light}}$ *is supported on a set of at most* M *integers, where* $M \le (2t_K/\varepsilon)^K$.

3. $\mathbf{S}_{\text{MIX}}$ *and* $\{\mathbf{S}_\alpha\}_{q_\alpha \in B \setminus \text{MIX}}$ *are signed PBDs such that for all* $q_\alpha \in B \setminus \text{MIX}$, *we have* $K^6/\varepsilon^8 \le \mathbf{Var}[S_\alpha] \le r^2/\varepsilon^2$, *where* $r = \max_{q_\alpha \in B} |q_\alpha|$. *Moreover* $\mathbf{Var}[\mathbf{S}_{\text{MIX}}] \ge K^6/\varepsilon^8$.

4. $\mathbf{Var}[q_{\text{MIX}} \cdot \mathbf{S}_{\text{MIX}}] = c \cdot \Big(\mathbf{Var}[q_{\text{MIX}} \cdot \mathbf{S}_{\text{MIX}}] + \sum_{q_\alpha \in B \setminus \text{MIX}} \mathbf{Var}[q_\alpha \cdot \mathbf{S}_\alpha]\Big)$ *for some* $c \in [\frac{1}{K^2}, 1]$.

The above corollary tells us that our distribution $\mathbf{S}'$ is close to a "nicely structured" distribution $\mathbf{S}^{(2)}$; we are now ready for our main learning result, which uses kernel-based tools developed in Section 5 to learn such a distribution. The following theorem completes the $\ell_0 \le K$ case:

Theorem 49. *There is a learning algorithm and a positive constant* c *with the following properties: It is given as input* N, *values* $\varepsilon, \delta > 0$, *and integers* $0 \le a_1 < \cdots < a_k$, *and can access draws from an unknown distribution* $\mathbf{S}^*$ *that is* $c\varepsilon$*-close to a* $\{a_1, \dots, a_k\}$*-sum* $\mathbf{S}$. *The algorithm runs in time* $(1/\varepsilon)^{2^{O(k^2)}} \cdot (\log a_k)^{\text{poly}(k)}$ *and uses* $(1/\varepsilon)^{2^{O(k^2)}} \cdot \log\log a_k$ *samples, and has the following property: Suppose that for the zero-moded distribution* $\mathbf{S}'$ *such that* $\mathbf{S}' + V = \mathbf{S}$ *(as defined in Section 6), the largeness index* ℓ_0 *(as defined at the beginning of this section) is at most* K *(again recall Section 6). Then with probability* $1 - o(1)$ *the algorithm outputs a hypothesis distribution* $\mathbf{H}$ *with* $d_{\text{TV}}(\mathbf{H}, \mathbf{S}) \le O(K^{71/20} \cdot \varepsilon^{1/20})$.

(To obtain an $O(\varepsilon)$-accurate hypothesis, simply run the learning algorithm with its accuracy parameter set to $\varepsilon' = \varepsilon^{20}/K^{71}$.)

Proof. The high level idea of the algorithm is as follows: The algorithm repeatedly samples two points from the distribution $\mathbf{S}^*$ and, for each pair, uses those two points to guess (approximately) parameters of the distribution

$$\mathbf{S}_{\text{pure}} := q_{\text{MIX}} \cdot \mathbf{S}_{\text{MIX}} + \sum_{q_\alpha \in B \setminus \text{MIX}} q_\alpha \cdot \mathbf{S}_\alpha$$

from Corollary 48. The space of possible guesses will be of size $(1/\varepsilon)^{2^{O(k^2)}} \cdot (\log a_k)^{\text{poly}(k)}$, which leads to a $\text{poly}(2^{k^2}, \log(1/\varepsilon)) \cdot \log\log a_k$ factor in the sample complexity by Corollary 27. For each choice of parameters in this space, Lemma 35 allows us to produce a candidate hypothesis distribution (this lemma leads to a $\exp(\text{poly}(k))/\varepsilon^{2^{O(k)}}$ factor in the sample complexity); by the guarantee of Lemma 35, for the (approximately) correct choice of parameters the corresponding candidate hypothesis distribution will be close to the target distribution $\mathbf{S}'$. Given that there is a high-accuracy candidate hypothesis distribution in the pool of candidates, by Corollary 27 (which details how our algorithms can "make guesses"), the algorithm of that corollary will with high probability select a high-accuracy hypothesis distribution $\mathbf{H}$ from the space of candidates.

We now give the detailed proof. To begin, the algorithm computes K and the values $q_1, \dots, q_K$. It guesses an ordering of $q_1, \dots, q_K$ such that $c_{q_1} \le \cdots \le c_{q_K}$ ($K! = 2^{\text{poly}(k)}$ possibilities), guesses the value of the largeness index ℓ_0 ($O(K) = \text{poly}(k)$ possibilities), guesses the subset $A \subseteq \{(a,b) : \ell_0 \le a < b \le K\}$ and the associated bits $(\tau_{a,b})_{(a,b)\in A}$ from Corollary 47 ($2^{\text{poly}(k)}$ possibilities) , and guesses the subset $\text{MIX} \subseteq B$ from Corollary 48 ($2^{\text{poly}(k)}$ possibilities). The main portion of the algorithm consists of the following three steps:

First main step of the algorithm: Estimating the variance of $\mathbf{S}^{(2)}$. In the first main step, the algorithm constructs a space of $1/\varepsilon^{2^{O(k^2)}}$ many guesses, one of which with very high probability is a multiplicatively

accurate approximation of $\sqrt{\mathbf{Var}[\mathbf{S}_{\rm pure}]}$. This is done as follows: the algorithm makes two independent draws from $\mathbf{S}^*$. Since $\mathbf{S}^*$ is $c\varepsilon$-close to $\mathbf{S} = \mathbf{S}' + V$, by Corollary 48 and Lemma 7, the distribution over these two draws could be obtained by sampling twice independently from $\mathbf{S}^{(2)}$, and modifying the result with probability $O(K^{71/20} \cdot \varepsilon^{1/20})$. Let us write these two draws as $s^{(j)} = s^{(j)}_{\rm light} + s^{(j)}_{\rm pure}$ where $j \in \{1,2\}$ and $s^{(j)}_{\rm light} \sim \mathbf{S}_{\rm light}$ and $s^{(j)}_{\rm pure} \sim \mathbf{S}_{\rm pure}$ (where $s^{(1)}_{\rm light}$, $s^{(1)}_{\rm pure}$, $s^{(2)}_{\rm light}$, $s^{(2)}_{\rm pure}$ are all independent draws). By part (2) of Corollary 48, with probability at least $1/|\mathbf{S}_{\rm light}| \geq 1/M \geq (\varepsilon/2t_K)^K = \varepsilon^{2^{O(k^2)}}$, it is the case that $s^{(1)}_{\rm light} = s^{(2)}_{\rm light}$. In that event, with probability at least $1/2^{\rm poly(K)}$, we have

$$\frac{1}{2} \cdot \sqrt{\mathbf{Var}[\mathbf{S}_{\rm pure}]} \leq |s^{(2)} - s^{(1)}| \leq 2 \cdot \sqrt{\mathbf{Var}[\mathbf{S}_{\rm pure}]}. \tag{31}$$

To see this, observe that since each of the $O(K^2)$ independent constituent PBDs comprising $\mathbf{S}_{\rm pure}$ has variance at least K^6, for each one with probability at least $\frac{1}{\Theta(K^2)}$ the difference between two independent draws will lie between $(1 - \frac{1}{\Theta(K^2)})$ and $(1 + \frac{1}{\Theta(K^2)})$ times the square root of its variance. If this happens then we get (31). By repeating $2^{\rm poly(k)} \cdot \varepsilon^{2^{O(k^2)}}$ times, the algorithm can obtain $2^{\rm poly(k)}/\varepsilon^{2^{O(k^2)}}$ many guesses, one of which will, with overwhelmingly high probability, be a quantity $\gamma_{\rm pure}$ that is a multiplicative 2-approximation of $\sqrt{\mathbf{Var}[\mathbf{S}_{\rm pure}]}$.

Second main step of the algorithm: Gridding in order to approximate variances. Consider the set J defined as

$$J = \bigcup_{j=-1}^{1+\log(K)} \{2^j \cdot \gamma_{\rm pure}/q_{\rm MIX}\}.$$

Given that $\gamma_{\rm pure}$ is within a factor of two of $\sqrt{\mathbf{Var}[\mathbf{S}_{\rm pure}]}$ (by (31)) and given part (4) of Corollary 48, it is easy to see that there is an element $\gamma_{\rm MIX} \in J$ such that $\gamma_{\rm MIX}$ is within a multiplicative factor of 2 of $\sqrt{\mathbf{Var}[\mathbf{S}_{\rm MIX}]}$. Likewise, for each $\alpha \in B \setminus {\rm MIX}$, define the set J_{q_α} as

$$J_{q_\alpha} = \bigcup_{j=-1}^{1+\log(\sqrt{\max\{1/\varepsilon^8, r^2/\varepsilon^2\}})} \{2^j \cdot (\varepsilon \cdot K)^{-1/4}/q_\alpha\},$$

where, as in Corollary 48, $r = \max_{q_\alpha \in B} |q_\alpha|$. By part (3) of Corollary 48, for each $q_\alpha \in B \setminus {\rm MIX}$ there is an element $\gamma_{q_\alpha} \in J_{q_\alpha}$ such that γ_{q_α} is within a multiplicative factor of two of $\mathbf{Var}[\mathbf{S}_{q_\alpha}]$. These elements of J and of J_{q_α} are the guesses for the values of $\sqrt{\mathbf{Var}[\mathbf{S}_{\rm MIX}]}$ and of $\sqrt{\mathbf{Var}[\mathbf{S}_{q_\alpha}]}$ that are used in the final main step described below. We note that the space of possible guesses here is of size at most $O(\log k) \cdot (\log(a_k/\varepsilon))^{\rm poly(k)}$.

Third main step of the algorithm: Using guesses for the variances to run the kernel-based learning approach. For each outcome of the guesses described above (denote a particular such outcome by $\overline{\gamma}$; note that a particular outcome for $\overline{\gamma}$ comprises an element of J and an element of J_{q_α} for each $\alpha \in B \setminus {\rm MIX}$), let us define the distribution $\mathbf{Z}_{{\rm MIX},\overline{\gamma}}$ to be uniform on the set $[-(c\varepsilon \cdot \gamma_{\rm MIX})/K, (c\varepsilon \cdot \gamma_{\rm MIX})/K] \cap \mathbb{Z}$ and $\mathbf{Z}_{q_\alpha,\overline{\gamma}}$ to be uniform on the set $[-(c\varepsilon \cdot \gamma_{q_\alpha})/K, (c\varepsilon \cdot \gamma_{q_\alpha})/K] \cap \mathbb{Z}$, where c is the hidden constant in the definition of c_j in Lemma 35. Applying Lemma 35, we can draw $\frac{\exp({\rm poly}(K))}{\varepsilon^{\rm poly(K)}} \cdot m^2 \cdot \log(m/\delta)$ samples from $\mathbf{S}$, where $m = (1/\varepsilon)^{2^{O(k^2)}} \geq (2t_K/\varepsilon)^K \geq |\mathbf{S}_{\rm light}|$, and we get a hypothesis $\mathbf{H}_{\overline{\gamma}}$ resulting from this outcome of the guesses and this draw of samples from $\mathbf{S}$. The guarantee of Lemma 35 ensures that for the outcome $\overline{\gamma}$ all of whose components are factor-of-two accurate as ensured in the previous step, the resulting hypothesis $\mathbf{H}_{\overline{\gamma}}$

satisfies $d_{\mathrm{TV}}(\mathbf{H}_{\overline{\gamma}}, \mathbf{S}') \le O(K^{71/20} \cdot \varepsilon^{1/20} + \varepsilon) = O(K^{71/20} \cdot \varepsilon^{1/20})$ with probability at least $1 - \delta$. Finally, an application of Corollary 27 concludes the proof. □

9 Learning $\{a_1, a_2, a_3\}$-sums

In this section we show that when $|\mathcal{A}| = 3$ the learning algorithm can be sharpened to have no dependence on a_1, a_2, a_3 at all. Recall Theorem 1:

Theorem 1 (Learning when $|\mathcal{A}| = 3$ with known support)**.** *There is an algorithm and a positive constant c with the following properties: The algorithm is given N, an accuracy parameter $\varepsilon > 0$, distinct values $a_1 < a_2 < a_3 \in \mathbb{Z}_{\ge 0}$, and access to i.i.d. draws from an unknown random variable $\mathbf{S}^*$ that is $c\varepsilon$-close to an $\{a_1, a_2, a_3\}$-sum $\mathbf{S}$. The algorithm uses $\mathrm{poly}(1/\varepsilon)$ draws from $\mathbf{S}^*$, runs in $\mathrm{poly}(1/\varepsilon)$ time, and with probability at least $9/10$ outputs a concise representation of a hypothesis distribution $\mathbf{H}$ such that $d_{\mathrm{TV}}(\mathbf{H}, \mathbf{S}^*) \le \varepsilon$.*

The high-level approach we take follows the approach for general k; as in the general case, a sequence of transformations will be used to get from the initial target to a "nicer" distribution (whose exact form depends on the precise value of the "largeness index") which we learn using the kernel-based approach. (Lemmas 50 and 51, which establish learning results for distributions in two of these nicer forms, are deferred to later subsections.) Intuitively, the key to our improved independent-of-a_3 bound is a delicate analysis that carefully exploits extra additive structure that is present when $k = 3$, and which lets us avoid the "gridding" over $O(\log a_k)$ many multiplicatively spaced guesses for variances that led to our $\log \log a_k$ dependence in the general-k case.

To describe this additive structure, let us revisit the framework established in Section 6, now specializing to the case $k = 3$, so $\mathbf{S}$ is an $\{a_1, a_2, a_3\}$-sum with $a_1 < a_2 < a_3$. We now have that for each $i \in [N]$ the support of the zero-moded random variable $\mathbf{X}'_i$ is contained in $\{0\} \cup Q$ where $Q = \{\pm q_1, \pm q_2, \pm q_3\}$ where $q_1 = a_2 - a_1, q_2 = a_3 - a_2$, and $q_3 = a_3 - a_1$. Further, the support size of each $\mathbf{X}'_i$ is 3 and hence it includes at most two of the elements from the set $\{q_1, q_2, q_3\}$. The fact that $q_3 = q_1 + q_2$ is the additive structure that we shall crucially exploit. Note that in the case $k = 3$ we have $K = 3$ as well, and $\mathbf{Pr}[\mathbf{X}'_i = 0] \ge 1/k = 1/K = 1/3$ for each $i \in [N]$.

Recalling the framework from the beginning of Section 8, we reorder q_1, q_2, q_3 so that $c_{q_1} \le c_{q_2} \le c_{q_3}$. We define the "largeness index" $\ell_0 \in \{1, 2, 3, 4\}$ analogously to the definition at the beginning of Section 8, but with a slight difference in parameter settings: we now define the sequence $t_1, \ldots, t_K$ as $t_\ell = (1/\varepsilon)^{C^\ell}$ where C is a (large) absolute constant to be fixed later. Define the "largeness index" of the sequence $c_{q_1} \le \ldots \le c_{q_K}$ as the minimum $\ell \in [K]$ such that $c_{q_\ell} > t_\ell$, and let ℓ_0 denote this value. If there is no $\ell \in \{1, 2, 3\}$ such that $c_{q_\ell} > t_\ell$, then we set $\ell_0 = 4$.

Viewing $\mathbf{S}$ as $\mathbf{S}' + V$ as before, our analysis now involves four distinct cases, one for each possible value of ℓ_0.

9.1 The case that $\ell_0 = 4$.

This case is identical to Section 8.1 specialized to $K = 3$, so we can easily learn to accuracy $O(\varepsilon)$ in $\mathrm{poly}(1/\varepsilon^{C^3}) = \mathrm{poly}(1/\varepsilon)$ time.

9.2 The case that $\ell_0 = 3$.

In this case we have $c_{q_1} \leq (1/\varepsilon)^C$ and $c_{q_2} \leq (1/\varepsilon)^{C^2}$ but $c_{q_3} \geq (1/\varepsilon)^{C^3}$. By Lemma 46, we have that $d_{\mathrm{TV}}(\tilde{\mathbf{S}}, \mathbf{S}') \leq O(\varepsilon)$ where $\tilde{\mathbf{S}} = \mathbf{S}_{\text{heavy}} + \mathbf{S}_{\text{light}}$, $\mathbf{S}_{\text{heavy}}$ and $\mathbf{S}_{\text{light}}$ are independent of each other, $\mathbf{S}_{\text{light}}$ is supported on a set of $O(1/\varepsilon^{2C^2+2})$ integers, and $\mathbf{S}_{\text{heavy}}$ is simply $q_3\mathbf{S}_3$ where $\mathbf{S}_3 = \sum_{i=1}^N \mathbf{Y}_i$ is a signed PBD with $\sum_{i=1}^N \mathbf{Pr}[\mathbf{Y}_i = \pm 1] \geq 1/(2\varepsilon^{C^3})$. Given this constrained structure, the $\mathrm{poly}(1/\varepsilon)$-sample and running time learnability of $\mathbf{S}^*$ follows as a special case of the algorithm given in the proof of Theorem 49. In more detail, as described in that proof, two points drawn from $\mathbf{S}^*$ can be used to obtain, with at least $\mathrm{poly}(\varepsilon)$ probability, a multiplicative factor-2 estimate of $\sqrt{\mathbf{Var}[\mathbf{S}_{\text{heavy}}]}$. Given such an estimate no gridding is required, as it is possible to learn $\mathbf{S}^*$ to accuracy $O(\varepsilon)$ simply by using the $K = 1$ case of the kernel learning result Lemma 35 (observe that, crucially, having an estimate of $\mathbf{Var}[\mathbf{S}_{\text{heavy}}]$ provides the algorithm with the value γ_1 in Lemma 35 which is required to construct $\mathbf{Z}$ and thereby carry out the kernel learning of $\mathbf{S}' + V$ using $\mathbf{Z}$).

9.3 The case that $\ell_0 = 2$.

In this case we have $c_{q_1} \leq (1/\varepsilon)^C$ while $c_{q_3}, c_{q_2} \geq (1/\varepsilon)^{C^2}$. As earlier we suppose that $q_1 + q_2 = q_3$. (This is without loss of generality as the other two cases are entirely similar; for example, if instead we had $q_1 + q_3 = q_2$, then we would have $q_3 = -q_1 + q_2$, and it is easy to check that replacing q_1 by $-q_1$ everywhere does not affect our arguments.)

Lemma 46 now gives us a somewhat different structure, namely that $d_{\mathrm{TV}}(\tilde{\mathbf{S}}, \mathbf{S}') \leq O(\varepsilon)$ where $\tilde{\mathbf{S}} = \mathbf{S}_{\text{heavy}} + \mathbf{S}_{\text{light}}$, $\mathbf{S}_{\text{heavy}}$ and $\mathbf{S}_{\text{light}}$ are independent of each other, $\mathbf{S}_{\text{light}} = q_1\mathbf{S}_1$ where $\mathbf{S}_1$ is supported on $[-O(1/\varepsilon^{C+1}), O(1/\varepsilon^{C+1})] \cap \mathbb{Z}$, and $\mathbf{S}_{\text{heavy}}$ is a sum of 0-moded integer random variables over $\{\pm q_2, \pm q_3\}$, and which satisfies $c_{q_2,\text{heavy}}, c_{q_3,\text{heavy}} > 1/(2\varepsilon^{C^2})$. Applying Lemma 40 to $\mathbf{S}_{\text{heavy}}$, we get that $d_{\mathrm{TV}}(\mathbf{S}_{\text{heavy}}, \mathbf{B}) = O(\varepsilon^{C^2/20})$ where either

$$\mathbf{B} = V' + q_2\mathbf{W}_2 + q_3\mathbf{W}_3 \tag{32}$$

(if the set A from Lemma 40 is empty) or

$$\mathbf{B} = V' + q_2\mathbf{W}_2 + q_3\mathbf{W}_3 + (q_3 + \tau_{2,3}q_2)\mathbf{W}_{2,3} \tag{33}$$

(if $A = \{(2,3)\}$), where all the distributions $\mathbf{W}_2, \mathbf{W}_3$ (and possibly $\mathbf{W}_{2,3}$) are independent signed PBDs with variance at least $\Omega(1/\varepsilon^{C^2/4})$ and $\tau_{2,3} \in \{-1, 1\}$.

Let us first suppose that (32) holds, so $\mathbf{S}' + V$ is $O(\varepsilon^{C^2/20})$-close to

$$V'' + q_1\mathbf{S}_1 + q_2\mathbf{W}_2 + q_3\mathbf{W}_3, \tag{34}$$

where $V'' = V + V'$. Since, by Fact 18, $q_2\mathbf{W}_2$ is $O(\varepsilon^{C^2/8})$-shift-invariant at scale q_2, recalling the support of $\mathbf{S}_1$ we get that $\mathbf{S}' + V$ is $(O(\varepsilon^{C^2/20}) + O(\varepsilon^{C^2/8}/\varepsilon^{C+1}))$-close (note that this is $O(\varepsilon^{C^2/20})$ for sufficiently large constant C) to

$$V'' + (q_1 + q_2)\mathbf{S}_1 + q_2\mathbf{W}_2 + q_3\mathbf{W}_3 = V'' + q_2\mathbf{W}_2 + q_3(\mathbf{W}_3 + \mathbf{S}_1).$$

Again using the support bound on $\mathbf{S}_1$ and Fact 18 (but now on $q_3\mathbf{W}_3$), we get that $\mathbf{S}' + V$, and therefore $\mathbf{S}^*$, is $O(\varepsilon^{C^2/20})$-close to

$$V'' + q_2\mathbf{W}_2 + q_3\mathbf{W}_3. \tag{35}$$

We can now apply the algorithm in Lemma 50 to semi-agnostically learn the distribution $V'' + q_2\mathbf{W}_2 + q_3\mathbf{W}_3$ with $\mathrm{poly}(1/\varepsilon)$ samples and time complexity.

Next, let us consider the remaining possibility in this case which is that (33) holds. If $\tau_{2,3} = -1$, then $\mathbf{S}' + V$ is $O(\varepsilon^{C^2/20})$-close to

$$V' + q_1\mathbf{S}_1 + q_2\mathbf{W}_2 + q_3\mathbf{W}_3 + (q_3 - q_2)\mathbf{W}_{2,3} = V' + q_1\mathbf{S}_1 + q_2\mathbf{W}_2 + q_3\mathbf{W}_3 + q_1\mathbf{W}_{2,3},$$

and using Fact 18 as earlier, we get that $\mathbf{S}' + V$ is $O(\varepsilon^{C^2/20})$-close to

$$V'' + q_2\mathbf{W}_2 + q_3\mathbf{W}_3 + q_1\mathbf{W}_{2,3}. \tag{36}$$

On the other hand, if $\tau_{2,3} = 1$ then $\mathbf{S}' + V$ is $O(\varepsilon^{C^2/20})$-close to

$$V'' + q_1\mathbf{S}_1 + q_2\mathbf{W}_2 + q_3\mathbf{W}_3 + (q_3 + q_2)\mathbf{W}_{2,3},$$

and by the analysis given between (34) and (35) we get that $\mathbf{S}' + V$ is $O(\varepsilon^{C^2/20})$-close to

$$V'' + q_2\mathbf{W}_2 + q_3\mathbf{W}_3 + (q_3 + q_2)\mathbf{W}_{2,3}, \tag{37}$$

In either case (37) or (36), we can use Lemma 51 to learn the target distribution with $\mathrm{poly}(1/\varepsilon)$ samples and running time.

9.4 The case that $\ell_0 = 1$.

In this case we have $c_{q_1}, c_{q_2}, c_{q_3} \geq (1/\varepsilon)^C$. Assuming that $C \geq 96$, we appeal to Lemma 53 to obtain that there are independent signed PBDs $\mathbf{S}_1$, $\mathbf{S}_2$ and $\mathbf{S}_3$, each with variance at least $1/\varepsilon^2$, such that

$$d_{\mathrm{TV}}(\mathbf{S}', V + q_1\mathbf{S}_1 + q_2\mathbf{S}_2 + q_3\mathbf{S}_3) \leq O(\varepsilon^2).$$

As before, we can appeal to Lemma 51 to learn the target distribution with $\mathrm{poly}(1/\varepsilon)$ samples and running time.

9.5 Deferred proofs and learning algorithms from the earlier cases

9.5.1 Learning algorithm for weighted sums of two PBDs

Lemma 50. *There is a universal constant C_1 such that the following holds: Let $\mathbf{S}^{2,\mathrm{high}}$ be a distribution of the form $p \cdot \mathbf{S}^{(p)} + q \cdot \mathbf{S}^{(q)} + V$, where both $\mathbf{S}^{(p)}$ and $\mathbf{S}^{(q)}$ are independent PBD$_N$ distributions with variance at least $1/\varepsilon^{C_1}$ and $V \in \mathbb{Z}$. Let $\mathbf{S}$ be a distribution with $d_{\mathrm{TV}}(\mathbf{S}, \mathbf{S}^{2,\mathrm{high}}) \leq \varepsilon$. There is an algorithm with the following property: The algorithm is given ε, p, q and access to i.i.d. draws from $\mathbf{S}$. The algorithm makes $\mathrm{poly}(1/\varepsilon)$ draws, runs in $\mathrm{poly}(1/\varepsilon)$ time, and with probability 999/1000 outputs a hypothesis distribution $\mathbf{H}$ satisfying $d_{\mathrm{TV}}(\mathbf{H}, \mathbf{S}) \leq O(\varepsilon)$.*

Proof. The high level idea of the algorithm is similar to Theorem 49. First, assume that $\mathbf{Var}[p \cdot \mathbf{S}^{(p)}] \geq \mathbf{Var}[q \cdot \mathbf{S}^{(q)}]$ (the other case is identical, and the overall algorithm tries both possibilities and does hypothesis testing). Let $\sigma_p^2 = \mathbf{Var}[\mathbf{S}^{(p)}]$, $\sigma_q^2 = \mathbf{Var}[\mathbf{S}^{(q)}]$ and $\sigma_{2,\mathrm{high}}^2 = \mathbf{Var}[\mathbf{S}^{2,\mathrm{high}}]$. We consider three cases depending upon the value of σ_q and show that in each case the kernel based approach (i.e. Lemma 35) can be used to learn the target distribution $\mathbf{S}^{2,\mathrm{high}}$ with $\mathrm{poly}(1/\varepsilon)$ samples (this suffices, again by hypothesis testing). We now provide details.

Estimating the variance of $\mathbf{S}^{2,\text{high}}$: The algorithm first estimates the variance of $\mathbf{S}^{2,\text{high}}$. This is done by sampling two elements $s^{(1)}, s^{(2)}$ from $\mathbf{S}^{2,\text{high}}$ and letting $|s^{(1)} - s^{(2)}| = \widehat{\sigma}_{2,\text{high}}$. Similar to the analysis of Theorem 49, it is easy to show that with probability $\Omega(1)$, we have

$$\frac{1}{\sqrt{2}} \cdot \sigma_{2,\text{high}} \leq \widehat{\sigma}_{2,\text{high}} \leq \sqrt{2} \cdot \sigma_{2,\text{high}}. \tag{38}$$

Guessing the dominant variance term and the relative magnitudes: Observe that

$$\mathbf{Var}[\mathbf{S}^{2,\text{high}}] = \mathbf{Var}[p \cdot \mathbf{S}^{(p)}] + \mathbf{Var}[q \cdot \mathbf{S}^{(q)}].$$

The algorithm next guesses whether $p \cdot \sigma_p \geq q \cdot \sigma_q$ or vice-versa. Let us assume that it is the former possibility. The algorithm then guesses one of the three possibilities: (i) $\sigma_q \leq \varepsilon \cdot p$, (ii) $\varepsilon \cdot p \leq \sigma_q < p/\varepsilon$, (iii) $\sigma_q > p/\varepsilon$. The chief part of the analysis is in showing that in each of these cases, the algorithm can draw $O(1/\varepsilon^2)$ samples from $\mathbf{S}$ and (with the aid of Lemma 35) can produce a hypothesis $\mathbf{H}$ such that $d_{\text{TV}}(\mathbf{H}, \mathbf{S}^{2,\text{high}}) = O(\varepsilon)$.

(i) In this case, we assume $\sigma_q \leq \varepsilon \cdot p$. This case is the *crucial point of difference* where we save the factor of $\log \log p$ as opposed to the case $k > 3$; this is done by working modulo p to estimate σ_q. (This is doable in this case because σ_q is so small relative to p.) The algorithm samples two points $s^{(3)}, s^{(4)} \sim \mathbf{S}$; note that with probability $1 - O(\varepsilon)$ these points are distributed exactly as if they were drawn from $\mathbf{S}^{2,\text{high}}$, so we may analyze the points as if they were drawn from $\mathbf{S}^{2,\text{high}}$. Let us assume that $s^{(3)} = p \cdot s_p^{(3)} + q \cdot s_q^{(3)} + V$, $s^{(4)} = p \cdot s_p^{(4)} + q \cdot s_q^{(4)} + V$ where $s_p^{(3)}, s_p^{(4)}$ are i.i.d. draws from $\mathbf{S}^{(p)}$ and similarly for $s_q^{(3)}, s_q^{(4)}$. Then, note that with probability at least $1/10$, we have

$$\frac{1}{\sqrt{2}} \cdot \sigma_q \leq |s_q^{(4)} - s_q^{(3)}| \leq \sqrt{2} \cdot \sigma_q.$$

This immediately implies that if we define $\widehat{\sigma}_q = q^{-1} \cdot (s^{(3)} - s^{(4)}) \pmod p$, then $\widehat{\sigma}_q = |s_q^{(4)} - s_q^{(3)}|$, and thus

$$\frac{1}{\sqrt{2}} \cdot \sigma_q \leq \widehat{\sigma_q} \leq \sqrt{2} \cdot \sigma_q.$$

This gives one of the estimates required by Lemma 35; for the other one, we observe that defining $\widehat{\sigma}_p := \widehat{\sigma}_{2,\text{high}}/p$, having $p\sigma_p \in [\frac{\mathbf{Var}[\mathbf{S}^{2,\text{high}}]}{2}, \mathbf{Var}[\mathbf{S}^{2,\text{high}}]]$ and (38) together give that

$$\frac{1}{2} \cdot \sigma_p \leq \widehat{\sigma}_p \leq 2\sigma_p.$$

We can now apply Lemma 35 to get that using $\text{poly}(1/\varepsilon)$ samples, we can produce a hypothesis distribution $\mathbf{H}_{\text{low}}$ such that $d_{\text{TV}}(\mathbf{H}_{\text{low}}, \mathbf{S}) = O(\varepsilon)$.

(ii) In this case, we assume $\varepsilon \cdot p < \sigma_q \leq p \cdot (1/\varepsilon)$. In this case we simply guess one of the $O(\log(1/\varepsilon)/\varepsilon)$ many values

$$\widehat{\sigma}_q \in \left\{ \frac{p}{(1+\varepsilon/10)^i} \right\}_{i \in \{-O(\ln(1/\varepsilon)/\varepsilon), \ldots, O(\ln(1/\varepsilon)/\varepsilon)\}}$$

and one of these guesses $\widehat{\sigma}_q$ for σ_q will be $(1+\varepsilon/10)$-multiplicatively accurate. For each of these values of $\widehat{\sigma_q}$, as in case (ii) we can get a multiplicatively accurate estimate $\widehat{\sigma}_p$ of σ_p, so again by invoking Lemma 35 we can create a hypothesis distribution $\mathbf{H}_{\text{med},i}$, and for the right guess we will have that $d_{\text{TV}}(\mathbf{H}_{\text{med},i}, \mathbf{S}) = O(\varepsilon)$.

(iii) In this case, we invoke Lemma 41 to get that there is a signed PBD $\mathbf{S}'$ such that $d_{\mathrm{TV}}(\mathbf{S}', p \cdot \mathbf{S}^{(p)} + q \cdot \mathbf{S}^{(q)}) = O(\varepsilon)$. This also yields that there is a signed PBD $\mathbf{S}'' = \mathbf{S}' + V$ such that $d_{\mathrm{TV}}(\mathbf{S}'', \mathbf{S}^{2,\mathrm{high}}) = O(\varepsilon)$. By a trivial application of Lemma 35, using $\mathrm{poly}(1/\varepsilon)$ samples, we obtain a hypothesis $\mathbf{H}_{\mathrm{high}}$ such that $d_{\mathrm{TV}}(\mathbf{H}_{\mathrm{high}}, \mathbf{S}) = O(\varepsilon)$.

Finally, invoking the Select procedure from Proposition 26 on the hypothesis distributions

$$\mathbf{H}_{\mathrm{low}}, \{\mathbf{H}_{\mathrm{med},i}\}_{i \in \{-O(\ln(1/\varepsilon)/\varepsilon), \ldots, O(\ln(1/\varepsilon)/\varepsilon)\}} \text{ and } \mathbf{H}_{\mathrm{high}},$$

we can use an additional $\mathrm{poly}(1/\varepsilon)$ samples to output a distribution $\mathbf{H}$ such that $d_{\mathrm{TV}}(\mathbf{H}, \mathbf{S}) = O(\varepsilon)$. □

9.5.2 Learning algorithm for weighted sums of three PBDs

We now give an algorithm for learning a distribution of the form $p \cdot \mathbf{S}^{(p)} + q \cdot \mathbf{S}^{(q)} + r \cdot \mathbf{S}^{(r)} + V$ where $r = p + q$.

Lemma 51. *There is a universal constant C_1 such that the following holds: Let $\mathbf{S}^{3,\mathrm{high}}$ be a distribution of the form $p \cdot \mathbf{S}^{(p)} + q \cdot \mathbf{S}^{(q)} + r \cdot \mathbf{S}^{(r)} + V$, where $\mathbf{S}^{(p)}, \mathbf{S}^{(q)}$ and $\mathbf{S}^{(r)}$ are independent PBD$_N$ distributions with variance at least $1/\varepsilon^C$ and $V \in \mathbb{Z}$ and $r = q + p$. Let $\mathbf{S}$ be a distribution with $d_{\mathrm{TV}}(\mathbf{S}, \mathbf{S}^{3,\mathrm{high}}) \le \varepsilon$. There is an algorithm with the following property: The algorithm is given ε, p, q, r and access to i.i.d. draws from $\mathbf{S}$. The algorithm makes $\mathrm{poly}(1/\varepsilon)$ draws, runs in $\mathrm{poly}(1/\varepsilon)$ time, and with probability 999/1000 outputs a hypothesis distribution $\mathbf{H}$ satisfying $d_{\mathrm{TV}}(\mathbf{H}, \mathbf{S}) \le O(\varepsilon)$.*

Proof. The algorithm begins by sampling two points $s^{(1)}, s^{(2)}$ from $\mathbf{S}$. Similar to the preceding proof, with probability $\Omega(1)$ we have

$$\frac{1}{\sqrt{2}} \cdot \sigma_{3,\mathrm{high}} \le \widehat{\sigma}_{3,\mathrm{high}} \le \sqrt{2} \cdot \sigma_{3,\mathrm{high}},$$

where $\sigma^2_{3,\mathrm{high}} = \mathbf{Var}[\mathbf{S}^{3,\mathrm{high}}]$. Having obtained an estimate of $\sigma_{3,\mathrm{high}}$, let us now assume (without loss of generality, via hypothesis testing) that $\sigma_p \ge \sigma_q \ge \sigma_r$. Similar to Lemma 50, we consider various cases, and for each case (and relevant guesses) we run Lemma 35 and obtain a hypothesis distribution for each of these guesses. Finally, we will use procedure Select (Proposition 26) on the space of these hypotheses to select one. Let us now consider the cases:

1. $\sigma_r \ge \varepsilon^5 \cdot \sigma_p$: In this case, note that given $\widehat{\sigma}_{3,\mathrm{high}}$, we can construct a grid J of $\mathrm{poly}(1/\varepsilon)$ many triples such that there exists $\overline{\gamma} = (\gamma_p, \gamma_q, \gamma_r) \in J$ such that for $\alpha \in \{p, q, r\}$,

$$\frac{1}{\sqrt{2}} \cdot \sigma_\alpha \le \gamma_\alpha \le \sqrt{2} \cdot \sigma_\alpha.$$

   For each such possibility $\overline{\gamma}$, we can apply Lemma 35 which uses $\mathrm{poly}(1/\varepsilon)$ samples; as before, for the right guess, we will obtain a hypothesis $\mathbf{H}_{\overline{\gamma}}$ such that $d_{\mathrm{TV}}(\mathbf{H}_{\overline{\gamma}}, \mathbf{S}) = O(\varepsilon)$.

2. $\sigma_r \le \varepsilon^5 \cdot \sigma_p$: In this case, since $r = p + q$,

$$p \cdot \mathbf{S}^{(p)} + q \cdot \mathbf{S}^{(q)} + r \cdot \mathbf{S}^{(r)} = p \cdot \mathbf{S}^{(p)} + q \cdot \mathbf{S}^{(q)} + (p + q) \cdot \mathbf{S}^{(r)}.$$

   As $\sigma_r \le \varepsilon^5 \cdot \sigma_p$, using the $O(1/\sigma_p)$-shift-invariance of $p \cdot \mathbf{S}^{(p)}$ at scale p that follows from Fact 18 and a Chernoff bound on $\mathbf{S}^{(r)}$, we get that for some integer V',

$$d_{\mathrm{TV}}(p \cdot \mathbf{S}^{(p)} + q \cdot \mathbf{S}^{(q)} + r \cdot \mathbf{S}^{(r)}, p \cdot (\mathbf{S}^{(p)} + V') + q \cdot (\mathbf{S}^{(q)} + \mathbf{S}^{(r)})) = O(\varepsilon^4).$$

Thus we have

$$d_{\mathrm{TV}}(\mathbf{S}^{3,\mathrm{high}}, V'' + p \cdot \mathbf{S}^{(p)} + q \cdot (\mathbf{S}^{(q)} + \mathbf{S}^{(r)})) = O(\varepsilon^4)$$

for some integer V''. However, now we are precisely in the same case as Lemma 50. Thus, using $\mathrm{poly}(1/\varepsilon)$ samples, we can now obtain $\mathbf{H}_{(2)}$ such that $d_{\mathrm{TV}}(\mathbf{H}_{(2)}, \mathbf{S}) = O(\varepsilon)$.

Finally, we apply Select (Proposition 26) on $\mathbf{H}_{(2)}$ and $\{\mathbf{H}_{\overline{\gamma}}\}_{\overline{\gamma} \in J}$. This finishes the proof. □

9.5.3 Structural lemma for decomposing a heavy distribution into a sum of weighted sums of PBDs

Our goal in this subsection is to prove Lemma 53. To do this, we will need a slightly more detailed version of Lemma 40 in the case that $K = 2$, which is implicit in the proof of that lemma (using the case that $A = \emptyset$ for (39) and the case that $A = \{(1, 2)\}$ for (40)).

Lemma 52. *Under the assumptions of Lemma 40 in the case that $K = 2$, there is an integer V', and independent signed PBDs $\mathbf{W}_{1,1}$, $\mathbf{W}_{2,2}$ and $\mathbf{W}_{1,2}$, all with variance at least $\Omega(R^{1/4})$, such that either*

$$d_{TV}(\mathbf{S}', V' + q_1\mathbf{W}_{1,1} + q_2\mathbf{W}_{2,2}) = O(R^{-1/20}), \tag{39}$$

or

$$d_{TV}(\mathbf{S}', V' + q_1\mathbf{W}_{1,1} + q_2\mathbf{W}_{2,2} + (q_2 + \mathrm{sign}(\mathbf{Cov}(M_1, M_2))q_1)\mathbf{W}_{1,2}) = O(R^{-1/20}), \tag{40}$$

where $\mathbf{M} = (\mathbf{M}_1, \mathbf{M}_2)$ is as defined in (11).

Let $\mathbf{S}' = \sum_{i=1}^N \mathbf{X}'_i$ where each $\mathbf{X}'_i$ is 0-moded and supported on $\{0, \pm p, \pm q, \pm r\}$ where $r = q + p$. Each random variable $\mathbf{X}'_i$ has a support of size 3, and by inspection of how $\mathbf{X}'_i$ is obtained from $\mathbf{X}_i$, we see that each $\mathbf{X}'_i$ is supported either on $\{0, p, r\}$, or on $\{-p, 0, q\}$, or on $\{-r, -q, 0\}$. If for $\alpha \in \{p, q, r\}$, we define $c_\alpha = \sum_{i=1}^N \mathbf{Pr}[X'_i = \pm\alpha]$, then we have the following lemma.

Lemma 53. *Let $\mathbf{S}' = \sum_{i=1}^N \mathbf{X}'_i$ as described above where $c_p, c_q, c_r > 1/\varepsilon^C$. Then we have*

$$d_{\mathrm{TV}}(\mathbf{S}', V + p \cdot \mathbf{S}^{(p)} + q \cdot \mathbf{S}^{(q)} + r \cdot \mathbf{S}^{(r)}) \leq O(\varepsilon^{C_1}), \tag{41}$$

where V is a constant, $C_1 = C/48$ and $\mathbf{S}^{(p)}, \mathbf{S}^{(q)}$ and $\mathbf{S}^{(r)}$ are mutually independent PBD$_N$ distributions each of which has variance at least $1/(\varepsilon^{C_1})$.

Proof. We first prove that

$$d_{\mathsf{shift},p}(\mathbf{S}') \leq \varepsilon^{C/2};\ d_{\mathsf{shift},q}(\mathbf{S}') \leq \varepsilon^{C/2};\ d_{\mathsf{shift},r}(\mathbf{S}') \leq \varepsilon^{C/2}. \tag{42}$$

To see this, note that, as we showed in the proof of Lemma 46, $d_{\mathsf{shift},p}(\mathbf{X}'_i) \leq 1 - \mathbf{Pr}[\mathbf{X}'_i = \pm p]$. By applying Corollary 24, we get that

$$d_{\mathsf{shift},p}(\mathbf{S}') = O\left(\frac{1}{\sqrt{\sum_i \mathbf{Pr}[\mathbf{X}'_i = \pm p]}}\right) = O(\varepsilon^{C/2}).$$

Likewise, we also get the other components of (42).

Let us next consider three families of i.i.d. random variables $\{\mathbf{Y}'_i\}_{i=1}^m$, $\{\mathbf{Z}'_i\}_{i=1}^m$ and $\{\mathbf{W}'_i\}_{i=1}^m$ defined as follows: for $1 \leq i \leq m$,

$$\mathbf{Pr}[\mathbf{Y}'_i = 0] = \mathbf{Pr}[\mathbf{Y}'_i = p] = 1/2;\ \mathbf{Pr}[\mathbf{Z}'_i = 0] = \mathbf{Pr}[\mathbf{Z}'_i = q] = 1/2; \mathbf{Pr}[\mathbf{W}'_i = 0] = \mathbf{Pr}[\mathbf{W}'_i = r] = 1/2;$$

Let $m = \varepsilon^{-C/4}$. Let $\sum_{i=1}^{m} \mathbf{Y}'_i = \mathbf{S}^{(y)}$, $\sum_{i=1}^{m} \mathbf{Z}'_i = \mathbf{S}^{(z)}$ and $\sum_{i=1}^{m} \mathbf{W}'_i = \mathbf{S}^{(w)}$. Let $\mathbf{S}_e = \mathbf{S}^{(y)} + \mathbf{S}^{(z)} + \mathbf{S}^{(w)}$, and note that $\mathbf{S}_e$ is supported on $\{i \cdot p + j \cdot q + k \cdot r\}$ where $0 \le i, j, k \le m$. Using (42), we have

$$d_{\mathrm{TV}}(\mathbf{S}', \mathbf{S}' + \mathbf{S}_e) \le m \cdot \varepsilon^{C/2} = O(\varepsilon^{C/4}).$$

Thus, it suffices to prove

$$d_{\mathrm{TV}}(\mathbf{S}' + \mathbf{S}_e, V + p \cdot \mathbf{S}^{(p)} + q \cdot \mathbf{S}^{(q)} + r \cdot \mathbf{S}^{(r)}) \le O(\varepsilon^{C_1}). \tag{43}$$

We assign each random variable $\mathbf{X}'_i$ to one of three different types:

- **Type 1:** The support of $\mathbf{X}'_i$ is $\{0, p, r\}$.
- **Type 2:** The support of $\mathbf{X}'_i$ is $\{-p, 0, q\}$.
- **Type 3:** The support of $\mathbf{X}'_i$ is $\{-r, -q, 0\}$.

Let the set of Type 1 variables be given by $\mathcal{S}_1$. We will show that there exists independent signed PBDs $\mathbf{S}^{(p,1)}$, $\mathbf{S}^{(q,1)}$ and $\mathbf{S}^{(r,1)}$ and a constant V_1 such that the variances of $\mathbf{S}^{(p,1)}$ and $\mathbf{S}^{(r,1)}$ are each at least ε^{-C_1}, and $\mathbf{S}^{(q,1)}$ is either constant (when (39) holds) or has variance at least ε^{-C_1} (when (40) holds), and that satisfy

$$d_{\mathrm{TV}}\left(\sum_{i \in \mathcal{S}_1} \mathbf{X}'_i + \sum_{i=1}^{m/2} \mathbf{Y}'_i + \sum_{i=1}^{m/2} \mathbf{W}'_i, V_1 + p \cdot \mathbf{S}^{(p,1)} + q \cdot \mathbf{S}^{(q,1)} + r \cdot \mathbf{S}^{(r,1)} \right) = O(\varepsilon^{C/48}). \tag{44}$$

If we can prove (44), then we can analogously prove the symmetric statements that

$$d_{\mathrm{TV}}\left(\sum_{i \in \mathcal{S}_2} \mathbf{X}'_i + \sum_{i=m/2+1}^{m} \mathbf{Y}'_i + \sum_{i=1}^{m/2} \mathbf{Z}'_i, V_2 + p \cdot \mathbf{S}^{(p,2)} + q \cdot \mathbf{S}^{(q,2)} + r \cdot \mathbf{S}^{(r,2)} \right) = O(\varepsilon^{C/48})$$

and

$$d_{\mathrm{TV}}\left(\sum_{i \in \mathcal{S}_3} \mathbf{X}'_i + \sum_{i=m/2+1}^{m} \mathbf{W}'_i + \sum_{i=m/2+1}^{m} \mathbf{Z}'_i, V_3 + p \cdot \mathbf{S}^{(p,3)} + q \cdot \mathbf{S}^{(q,3)} + r \cdot \mathbf{S}^{(r,3)} \right) = O(\varepsilon^{C/48}),$$

with analogous conditions on the variances of $\mathbf{S}^{(p,2)}, \mathbf{S}^{(q,2)}, \mathbf{S}^{(r,2)}, \mathbf{S}^{(p,3)}, \mathbf{S}^{(q,3)}, \mathbf{S}^{(r,3)}$, and combining these bounds with (44) will imply the desired inequality (43).

Thus it remains to prove (44). Let $\gamma_1 = \sum_{i \in \mathcal{S}_1} \mathbf{Pr}[\mathbf{X}'_i = p]$ and $\delta_1 = \sum_{i \in \mathcal{S}_1} \mathbf{Pr}[\mathbf{X}'_i = r]$. We consider the following cases:

Case (I): Assume γ_1 and $\delta_1 \ge \varepsilon^{-C/8}$. Since the possibilities $\mathbf{X}'_i = p$ and $\mathbf{X}'_i = r$ are mutually exclusive, for each i we have that $\mathbf{Cov}(\mathbf{1}_{\mathbf{X}'_i = p}, \mathbf{1}_{\mathbf{X}'_i = r}) \le 0$, which implies that $\mathbf{Cov}(\sum_i \mathbf{1}_{\mathbf{X}'_i = p}, \sum_i \mathbf{1}_{\mathbf{X}'_i = r}) \le 0$. Applying Lemma 52 to the distribution $\sum_{i \in \mathcal{S}_1} \mathbf{X}'_i$, we obtain (44) in this case.

Case (II): Now let us assume that at least one of γ_1 or δ_1 is less than $\varepsilon^{-C/8}$, without loss of generality, that $\gamma_1 < \varepsilon^{-C/8}$. For each variable $\mathbf{X}'_i$, let us consider a corresponding random variable $\tilde{\mathbf{X}}'_i$ defined by replacing the p-outcomes of $\mathbf{X}'_i$ with 0's. If $\mathbf{Z}$ is any distribution such that $d_{\mathsf{shift},p}(\mathbf{Z}) \le \kappa$, then

$$d_{\mathrm{TV}}(\mathbf{Z} + \mathbf{X}'_i, \mathbf{Z} + \tilde{\mathbf{X}}'_i) \le \kappa \cdot \mathbf{Pr}[\tilde{\mathbf{X}}'_i = p].$$

Applying this observation iteratively, we have

$$d_{\mathrm{TV}}\left(\sum_{i\in\mathcal{S}_1}\mathbf{X}'_i+\sum_{i=1}^{m/2}\mathbf{Y}'_i+\sum_{i=1}^{m/2}\mathbf{W}'_i,\sum_{i\in\mathcal{S}_1}\tilde{\mathbf{X}}'_i+\sum_{i=1}^{m/2}\mathbf{Y}'_i+\sum_{i=1}^{m/2}\mathbf{W}'_i\right)=O(\gamma_1\cdot\varepsilon^{C/4})=O(\varepsilon^{C/8}).$$

However, now note that $\sum_{i\in\mathcal{S}_1}\tilde{\mathbf{X}}'_i+\sum_{i=1}^{m/2}\mathbf{Y}'_i+\sum_{i=1}^{m/2}\mathbf{W}'_i$ can be expressed as $p\cdot\mathbf{S}^{(p,1)}+r\cdot\mathbf{S}^{(r,1)}$ for independent signed PBDs $\mathbf{S}^{(p,1)}$, and $\mathbf{S}^{(r,1)}$, and that the variances of $\sum_{i=1}^{m/2}\mathbf{Y}'_i$ and $\sum_{i=1}^{m/2}\mathbf{W}'_i$ ensure that $\mathbf{Var}[\mathbf{S}^{(p,1)}],\mathbf{Var}[\mathbf{S}^{(r,1)}]\geq\varepsilon^{-C/2}$. This establishes (44) in this case, completing the proof of the lemma. □

10 Unknown-support algorithms: Proof of Theorems 3 and 4

We begin by observing that the hypothesis selection procedure described in Section 4.4.1 provides a straightforward reduction from the case of unknown-support to the case of known-support. More precisely, it implies the following:

Observation 54. *For any k, let A be an algorithm that semi-agnostically learns $\{a_1,...,a_k\}$-sums, with $0\leq a_1<\cdots<a_k$, using $m(a_1,\ldots,a_k,\varepsilon,\delta)$ samples and running in time $T(a_1,\ldots,a_k,\varepsilon,\delta)$ to learn to accuracy ε with probability at least $1-\delta$, outputting a hypothesis distribution from which it is possible to generate a draw or evaluate the hypothesis's p.m.f. on a given point in time $T'(a_1,\ldots,a_k)$. Then there is an algorithm A' which semi-agnostically learns $(a_{\max},k)$-sums using*

$$\left(\max_{0\leq a_1<\cdots<a_k\leq a_{\max}}m(a_1,\ldots,a_k,\varepsilon/6,\delta/2)\right)+O(k\log(a_{\max})/\varepsilon^2+\log(1/\delta)/\varepsilon^2)$$

samples, and running in time

$$\mathrm{poly}((a_{\max})^k,1/\varepsilon)\cdot\left(\max_{0\leq a_1<\cdots<a_k\leq a_{\max}}(T(a_1,\ldots,a_k,\varepsilon))+T'(a_1,\ldots,a_k)\right),$$

and, with probability at least $1-\delta$, outputting a hypothesis with error at most 6ε.

The algorithm A' works as follows: first, it tries all (at most $(a_{\max})^k$) possible vectors of values for $(a_1,\ldots,a_k)$ as the parameters for algorithm A, using the same set of $\max_{0\leq a_1<\cdots<a_k\leq a_{\max}}m(a_1,\ldots,a_k,\varepsilon,\delta/2)$ samples as the input for each of these runs of A. Having done this, A' has a list of candidate hypotheses such that with probability at least $1-\delta/2$, at least one of the candidates is ε-accurate. Then A' runs the Select procedure from Proposition 26 on the resulting hypothesis distributions.

Together with Theorems 1 and 2, Observation 54 immediately yields Theorem 3 (learning $(a_{\max},3)$-sums). It also yields a result for the unknown-support $k=2$ case, but a sub-optimal one because of the $\log(a_{\max})$ dependence. In the rest of this section we show how the sharper bound of Theorem 4, with no dependence on $a_{\max}$, can be obtained by a different (but still simple) approach.

Recall Theorem 4:

Theorem 4. Learning $(a_{\max},2)$-sums *There is an algorithm and a positive constant c with the following properties: The algorithm is given N, accuracy and confidence parameters $\varepsilon,\delta>0$, an upper bound $a_{\max}\in\mathbb{Z}_+$, and access to i.i.d. draws from an unknown random variable $\mathbf{S}^*$ that is $c\varepsilon$-close to an $\{a_1,a_2\}$-sum $\mathbf{S}$, where $0\leq a_1\leq a_2\leq a_{\max}$. The algorithm uses $\mathrm{poly}(1/\varepsilon)$ draws from $\mathbf{S}^*$, runs in $\mathrm{poly}(1/\varepsilon,\log(1/\delta))$ time, and with probability $1-\delta$ outputs a (concise representation of a) hypothesis distribution $\mathbf{H}$ such that $d_{\mathrm{TV}}(\mathbf{H},\mathbf{S}^*)\leq\varepsilon$.*

Proof. Let the $\{a_1, a_2\}$-sum $\mathbf{S}$ over (unknown values) $\{a_1, a_2\}$ be $\mathbf{S} = \sum_{i=1}^{N} \mathbf{X}_i$ where $\mathbf{X}_i(a_1) = 1 - p_i$, $\mathbf{X}_i(a_2) = p_i$. Let $\mathbf{X}'_i$, $i \in [N]$ be independent Bernoulli random variables with $\mathbf{X}'_i(1) = p_i$. The distribution $\mathbf{S}$ is identical to $a_1 N + (a_2 - a_1)\mathbf{S}'$ where $\mathbf{S}'$ is the PBD $\mathbf{S}' = \sum_{i=1}^{N} \mathbf{X}'_i$.

Intuitively, if the values a_1 and a_2 (hence a_1 and $a_2 - a_1$) were known then it would be simple to learn using the algorithm for learning a PBD_N. The idea of what follows is that either (i) learning is easy because the essential support is small, or (ii) it is possible to infer the value of $a_2 - a_1$ (basically by taking the gcd of a few sample points) and, given this value, to reduce to the problem of learning a PBD_N. Details follow.

If the PBD $\mathbf{S}'$ is in sparse form (see Theorem 20), then it (and hence $\mathbf{S}$) is ε-essentially-supported on a set of size $O(1/\varepsilon^3)$. In this case the algorithm A' of Fact 25 can learn $\mathbf{S}^*$ to accuracy $O(\varepsilon)$ in $\mathrm{poly}(1/\varepsilon)$ time using $\mathrm{poly}(1/\varepsilon)$ samples. Thus we may subsequently assume that $\mathbf{S}'$ is not in sparse form. (The final overall algorithm will run both A' and the learning algorithm described below, and use the hypothesis selection procedure from Section 4.4.1 to choose the overall final hypothesis from these two.)

Since $\mathbf{S}'$ is not in sparse form, by the last part of Theorem 20 it is in $1/\varepsilon$-heavy binomial form. We will require the following proposition:

Proposition 55. *Let $a \in \mathbb{Z}_+, b \in \mathbb{Z}$ be arbitrary constants. For all small enough $\varepsilon > 0$, if $\mathbf{S}' = \sum_{i=1}^{N} \mathbf{S}'_i$ is a PBD in $1/\varepsilon$-heavy binomial form, then with probability at least $1 - O(\sqrt{\varepsilon})$, the $\gcd$ of $m = \Omega(1/\sqrt{\varepsilon})$ i.i.d. draws $v_1, \ldots, v_m$ from $a(\mathbf{S}' - b)$ is equal to a.*

Proof. $\mathbf{S}'$ is ε-close to a translated Binomial distribution $\mathbf{Y}$ as described in the second bullet of Theorem 20, and (by Fact 21) $\mathbf{Y}$ is $O(\varepsilon)$-close to $\mathbf{Z}$, a discretized $N(\mu_{\mathbf{Y}}, \sigma^2_{\mathbf{Y}})$ Gaussian where $\sigma^2_{\mathbf{Y}} = \Omega(1/\varepsilon^2)$. By Lemma 7, a collection of m i.i.d. draws from $\mathbf{S}^*$ is distributed exactly as a collection of m i.i.d. draws from $\mathbf{Z}$ except with failure probability $O(m\varepsilon)$. Incurring this failure probability, we may suppose that $v_1, \ldots, v_m$ are i.i.d. draws from $\mathbf{Z}$. Except with an additional failure probability $2^{-\Omega(m)}$, at least $\Omega(m)$ of these draws lie within $\pm\sigma_{\mathbf{Y}}$ of $\mu_{\mathbf{Y}}$, so we additionally suppose that this is the case. Next, since any two points within one standard deviation of the mean of a discretized Gaussian have probability mass within a constant factor of each other, with an additional failure probability of at most $2^{-\Omega(m)}$ we may suppose that the multiset $\{v_1, \ldots, v_m\}$ contains at least $\ell = \Omega(m)$ points that are distributed uniformly and independently over the integer interval $I := [\mu_{\mathbf{Y}} - \sigma_{\mathbf{Y}}, \mu_{\mathbf{Y}} + \sigma_{\mathbf{Y}}] \cap \mathbb{Z}$. Thus, to establish the proposition, it suffices to prove that for any $b \in \mathbb{Z}$, with high probability the $\gcd$ of ℓ points drawn uniformly and independently from the shifted interval $I - b$ is 1.

The $\gcd$ is 1 unless there is some prime p such that all ℓ draws from $I - b$ are divisible by p. Since ε is at most some sufficiently small constant, we have that $|I|$ is at least (say) 100; since $|I| \geq 100$, for any prime p at most a $1.02/p$ fraction of the points in I are divisible by p, so $\mathbf{Pr}[\text{all } \ell \text{ draws are divisible by } p] \leq (1.02/p)^\ell$. Thus a union bound gives

$$\mathbf{Pr}[\gcd > 1] \leq \sum_{\text{prime p}} \mathbf{Pr}[\text{ all } \ell \text{ draws are divisible by } p] \leq \sum_{\text{prime p}} (1.02/p)^\ell < \sum_{n \geq 2} (1.02/n)^\ell < (2/3)^\ell,$$

and the proposition is proved. □

We now describe an algorithm to learn $\mathbf{S}^*$ when $\mathbf{S}'$ is in $1/\varepsilon$-heavy binomial form. The algorithm first makes a single draw from $\mathbf{S}^*$ call this the "reference draw". With probability at least $9/10$, it is from $\mathbf{S}$; let us assume from for the rest of the proof that this is the case, and let its value be $v = a_1(N - r) + a_2 r$. Next, the algorithm makes $m = \Omega(1/\sqrt{\varepsilon})$ i.i.d. draws $u_1, \ldots, u_m$ from $\mathbf{S}^*$. Since $d_T V(\mathbf{S}, \mathbf{S}^*) < c\varepsilon$ for a small positive constant c, and ε is at most a small constant, a union bound implies that, except for a failure probability $O((1/\sqrt{\varepsilon})\varepsilon) < 1/10$, all of these draws come from $\mathbf{S}$. Let us assume from here on that this is

the case. For each i, the algorithm sets $v_i = u_i - v$, and computes the gcd of $v_1, \ldots, v_m$. Each u_i equals $a_1(N - n_i) + a_2 n_i$ where n_i is drawn from the PBD $\mathbf{S}'$, so we have that

$$v_i = a_1(N - n_i) + a_2 n_i - a_1(N - r) - a_2 r = (a_2 - a_1)(n_i - r),$$

and Proposition 55 gives that with failure probability at most $O(\sqrt{\varepsilon})$, the gcd of $n_1 - r, ..., n_m - r$ is 1, so that the gcd of $v_1, ..., v_m$ is equal to $a_2 - a_1$.

With the value of $a_2 - a_1$ in hand, it is straightforward to learn $\mathbf{S}$. Dividing each draw from $\mathbf{S}$ by $a_2 - a_1$, we get draws from $\frac{a_1 N}{a_2 - a_1} + \mathbf{S}'$ where $\mathbf{S}'$ is the PBD$_N$ described above. Such a "shifted PBD" can be learned easily as follows: if $\frac{a_1 N}{a_2 - a_1}$ is an integer then this is a PBD$_{(a_1+1)N}$, hence is a PBD$_{(a_{\max}+1)N}$, and can be learned using the algorithm for learning a PBD$_{N'}$ given the value of N'. If $\frac{a_1 N}{a_2 - a_1}$ is not an integer, then its non-integer part can be obtained from a single draw, and subtracting the non-integer part we arrive at the case handled by the previous sentence.

The algorithm described above has failure probability $O(\sqrt{\varepsilon})$, but by standard techniques (see Lemma 3.4 of [HKLW91] and Proposition 26) this failure probability can be reduced to an arbitrary δ at the cost of a $\log(1/\delta)$ factor increase in sample complexity. This concludes the proof of Theorem 4. □

11 A reduction for weighted sums of PBDs

Below we establish a reduction showing that an efficient algorithm for learning sums of weighted PBDs with weights $\{0 = a_1, \ldots, a_k\}$ implies the existence of an efficient algorithm for learning sums of weighted PBDs with weights $\{0 = a_1, \ldots, a_{k-1}\} \mod a_k$. Here by "learning sums of weighted PBDs with weights $\{0 = a_1, \ldots, a_{k-1}\} \mod a_k$" we mean an algorithm which is given access to i.i.d. draws from the distribution $\mathbf{S}' := (\mathbf{S} \mod a_k)$ where $\mathbf{S}$ is a weighted sum of PBDs with weights $\{0 = a_1, \ldots, a_{k-1}\}$, and should produce a high-accuracy hypothesis distribution for $\mathbf{S}'$ (which is is supported over $\{0, 1, \ldots, a_k - 1\}$); so both the hypothesis distribution *and the samples provided to the learning algorithm* are reduced mod a_k. Such a reduction will be useful for our lower bounds because it enables us to prove a lower bound for learning a weighted sum of PBDs with k unknown weights by proving a lower bound for learning mod a_k with $k - 1$ weights.

The formal statement of the reduction is as follows:

Theorem 56. *Suppose that A is an algorithm with the following properties: A is given N, an accuracy parameter $\varepsilon > 0$, a confidence parameter $\delta > 0$, and distinct non-negative integers $0 = a_1, \ldots, a_k$. A is provided with access to i.i.d. draws from a distribution $\mathbf{S}$ where $\mathbf{S} = a_2\mathbf{S}_2 + \cdots + a_k\mathbf{S}_k$ and each $\mathbf{S}_i$ is an unknown PBD$_N$. For all N, A makes $m(a_1, \ldots, a_k, \varepsilon, \delta)$ draws from $\mathbf{S}$ and with probability at least $1 - \delta$ outputs a hypothesis $\tilde{\mathbf{S}}$ such that $d_{\mathrm{TV}}(\mathbf{S}', \tilde{\mathbf{S}}) \le \varepsilon$.*

Then there is an algorithm A' with the following properties: A' is given $N, 0 = a_1, \ldots, a_k, \varepsilon, \delta$ and is provided with access to i.i.d. draws from $\mathbf{T}' := (\mathbf{T} \mod a_k)$ where $\mathbf{T} = a_2\mathbf{T}_2 + \cdots + a_{k-1}\mathbf{T}_{k-1}$ where in turn each $\mathbf{T}_i$ is a PBD$_N$. A' makes $m' = m(a_1, \ldots, a_k, \varepsilon, \delta/2)$ draws from $\mathbf{T}'$ and with probability $1 - \delta$ outputs a hypothesis $\tilde{\mathbf{T}}'$ such that $d_{\mathrm{TV}}(\mathbf{T}', \tilde{\mathbf{T}}') \le \varepsilon$.

Proof. The high-level idea is simple; in a nutshell, we leverage the fact that the algorithm A works with sample complexity $m(a_1, \ldots, a_k, \varepsilon, \delta)$ independent of N for all N to construct a data set suitable for algorithm A from a "mod a_k" data set that is the input to algorithm A'.

In more detail, as above suppose the target distribution $\mathbf{T}'$ is $(\mathbf{T} \mod a_k)$ where $\mathbf{T} = a_2\mathbf{T}_2 + \cdots + a_{k-1}\mathbf{T}_{k-1}$ and each $\mathbf{T}_i$ is an independent PBD$_N$. Algorithm A' works as follows: First, it makes m' draws

$v'_1, \ldots, v'_{m'}$ from $\mathbf{T}'$, the j-th of which is equal to some value $(a_2 N_{2,j} + \cdots + a_{k-1} N_{k-1,j}) \mod a_k$. Next, using its own internal randomness it makes m' draws $N_{k,1}, \ldots, N_{k,m'}$ from the $\mathrm{PBD}_{N^\star}$ distribution $\mathbf{T}_k := \mathrm{Bin}(N^\star, 1/2)$ (we specify $N^\star$ below, noting here only that $N^\star \gg N$) and constructs the "synthetic" data set of m' values whose j-th element is

$$u_j := v'_j + a_k N_{k,j}.$$

Algorithm A' feeds this data set of values to the algorithm A, obtains a hypothesis $\mathbf{H}$, and outputs $(\mathbf{H} \mod a_k)$ as its final hypothesis.

To understand the rationale behind this algorithm, observe that if each value v'_j were an independent draw from $\mathbf{T}$ rather than $\mathbf{T}'$ (i.e., if it were not reduced mod a_k), then each u_j would be distributed precisely as a draw from $\mathbf{T}^\star := a_2 \mathbf{T}_2 + \cdots + a_{k-1} \mathbf{T}_{k-1} + a_k \mathbf{T}_k$ (observe that each PBD_{N_i} is also a $\mathrm{PBD}_{N^\star}$, simply by having the "missing" $N^\star - N_i$ Bernoulli random variables all trivially output zero). In this case we could invoke the performance guarantee of algorithm A when it is run on such a target distribution. The issue, of course, is that v'_j is a draw from $\mathbf{T}'$ rather than $\mathbf{T}$, i.e. v'_j equals $(v_j \mod a_k)$ where v_j is some draw from $\mathbf{T}$. We surmount this issue by observing that since $a_k \mathbf{T}_k$ is shift-invariant at scale a_k, by taking $\mathbf{T}_k$ to have sufficiently large variance, we can make the variation distance between the distribution of each v'_j and the original v_j sufficiently small that so it is as if the values v'_j actually were drawn from $\mathbf{T}$ rather than $\mathbf{T}'$.

In more detail, let us view v'_j as the reduction mod a_k of a draw v_j from $\mathbf{T}$ as just discussed; i.e., let $v'_j \in \{0, ..., a_k - 1\}$ satisfy $v'_j = v_j + a_k c_j$ for $c_j \in \mathbb{Z}$. We observe that each c_j satisfies $|c_j| < a_{\max} \cdot N$. Recalling that $\mathbf{T}_k = \mathrm{Bin}(N^\star, 1/2)$ has $\mathbf{Var}[\mathbf{T}_k] = N^\star/4$, by Fact 18 we have that $d_{\mathrm{TV}}(a_k \mathbf{T}_k + \mathbf{T}, a_k \mathbf{T}_k + \mathbf{T}') \leq O(1/\sqrt{N^\star}) \cdot a_{\max} \cdot N$. Hence the variation distance between $(a_k \mathbf{T}_k + \mathbf{T}')^{m'}$ (from which the sample $u_1, \ldots, u_{m'}$ is drawn) and $(a_k \mathbf{T}_k + \mathbf{T})^{m'} = (\mathbf{T}^\star)^{m'}$ (what we would have gotten if each v'_j were replaced by v_j) is at most $O(1/\sqrt{N^\star}) \cdot a_{\max} \cdot N \cdot m'$. By taking $N^\star = \Theta((a_{\max} \cdot N \cdot m'/\delta)^2)$, this is at most $\delta/2$, so at the cost of a $\delta/2$ failure probability we may assume that the $m' = m(a_1, \ldots, a_k, \varepsilon, \delta/2)$-point sample $u_1, \ldots, u_{m'}$ is drawn from $\mathbf{T}^\star$. Then with probability $1 - \delta/2$ algorithm A outputs an ε-accurate hypothesis, call it $\tilde{\mathbf{T}}^\star$ (this is the $\mathbf{H}$ mentioned earlier), for the target distribution $\mathbf{T}^\star$ from which its input sample was drawn, so $d_{\mathrm{TV}}(\tilde{\mathbf{T}}^\star, \mathbf{T}^\star) \leq \varepsilon$. Taking $\tilde{\mathbf{T}}'$ to be $(\tilde{\mathbf{T}}^\star \mod a_k)$ and observing that $(\mathbf{T}^\star \mod a_k) \equiv \mathbf{T}'$, we have that

$$d_{\mathrm{TV}}(\tilde{\mathbf{T}}', \mathbf{T}') = d_{\mathrm{TV}}((\tilde{\mathbf{T}}^\star \mod a_k), (\mathbf{T}^\star \mod a_k)) \leq d_{\mathrm{TV}}(\tilde{\mathbf{T}}^\star, \mathbf{T}^\star) \leq \varepsilon,$$

and the proof is complete. □

12 Known-support lower bound for $|\mathcal{A}| = 4$: Proof of Theorem 5

Recall Theorem 5:

Theorem 5. ($k = 4$, known-support lower bound) *Let A be any algorithm with the following properties: algorithm A is given N, an accuracy parameter $\varepsilon > 0$, distinct values $0 \leq a_1 < a_2 < a_3 < a_4 \in \mathbb{Z}$, and access to i.i.d. draws from an unknown $\{a_1, a_2, a_3, a_4\}$-sum and with probability at least $9/10$ algorithm A outputs a hypothesis distribution $\tilde{\mathbf{S}}$ such that $d_{\mathrm{TV}}(\tilde{\mathbf{S}}, \mathbf{S}) \leq \varepsilon$. Then there are infinitely many quadruples (a_1, a_2, a_3, a_4) such that for sufficiently large N, A must use $\Omega(\log \log a_4)$ samples even when run with ε set to a (suitably small) positive absolute constant.*

12.1 Proof of Theorem 5

Fix $a_1 = 0$ and $a_2 = 1$. (It suffices to prove a lower bound for this case.) To reduce clutter in the notation, let $p = a_3$ and $q = a_4$. Applying Theorem 56, it suffices to prove that $\Omega(\log \log q)$ examples are needed

to learn distributions of random variables of the form $\mathbf{S} = \mathbf{U} + p\mathbf{V} \mod q$, where $\mathbf{U}$ and $\mathbf{V}$ are unknown PBDs over $\Theta(N)$ variables. We do this in the rest of this section.

Since an algorithm that achieves a small error with high probability can be used to achieve small error in expectation, we may use Lemma 28, which provides lower bounds on the number of examples needed for small expected error, to prove Theorem 5. To apply Lemma 28, we must find a set of distributions $\mathbf{S}_1, \ldots, \mathbf{S}_i, \ldots$, where $\mathbf{S}_i = \mathbf{U}_i + p\mathbf{V}_i \mod q$, that are separated enough that they must be distinguished by a successful learning algorithm (this is captured by the variation distance lower bound of Lemma 28), but close enough (as captured by the Kullback-Liebler divergence upper bound) that this is difficult. We sketched the ideas behind our construction of these distributions $\mathbf{S}_1, \ldots, \mathbf{S}_T$, $T = \log^{\Theta(1)} q$, earlier in Section 3, so we now proceed to the actual construction and proof.

The first step is to choose p and q. The choice is inspired by the theory of rational approximation of irrational numbers. The core of the construction requires us to use an irrational number which is hard to approximate as a ratio of small integers but such that, expressed as a continued fraction, the convergents do not grow very rapidly. For concreteness, we will consider the inverse of the golden ratio ϕ:

$$\frac{1}{\phi} = \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cdots}}}$$

Let $f_0 = 1, f_1 = 1, f_2 = 2, \ldots$ denote the Fibonacci numbers; It is easy to see that the t^{th} convergent of $1/\phi$ is given by f_{t-1}/f_t. We take $p = f_L$, $q = f_{L+1}$ where we think of L as an asymptotically large parameter (and of course $q = f_{L+1}$ implies $L = \Theta(\log q)$). Looking ahead, the two properties of $1/\phi$ which will be useful will be: (a) For any t, f_t/f_{t+1} is a very good approximation of $1/\phi$, and moreover, (b) the approximations obtained by these convergents are essentially the best possible.

Definition 57. Let $\rho_q(a, b)$ be the Lee metric on $\mathbb{Z}_q$, i.e., the minimum of $|j|$ over all j such that $a = b + j \mod q$.

The following lemma records the properties of p and q that we will use. To interpret this lemma, it may be helpful to imagine starting at 0, taking steps of size p through $[q]$, wrapping around when you get to the end, and dropping a breadcrumb after each step. Because p and q are relatively prime, after q steps, each member of $[q]$ has a breadcrumb. Before this, however, the lemma captures two ways in which the breadcrumbs are "distributed evenly" (in fact, within constant factors of optimal) throughout the walk: (a) they are pairwise well-separated, and (b) all positions have a breadcrumb nearby.

Lemma 58. *There are absolute constants $c_1, c_2 > 0$ such that, for all integers $v \neq v', v, v' \in (-c_2 q, c_2 q)$, we have*

$$\rho_q(pv, pv') > \frac{c_1 q}{\max\{|v|, |v'|\}}. \tag{45}$$

Furthermore, for any $i \in [q]$, for any $t \leq L$, there is a $v \in \{-f_t, ..., f_t\}$ such that

$$\rho_q(i, pv) \leq \frac{3q}{f_t}. \tag{46}$$

To prove the first part of Lemma 58, we need the following lemma on the difficulty of approximating $1/\phi$ by rationals.

Lemma 59 ([HWHB+08]). *There is a constant $c_3 > 0$ such that for all positive integers m and n,*

$$\left|\frac{m}{n} - \frac{1}{\phi}\right| \geq \frac{c_3}{n^2}.$$

We will also use the fact that $\frac{f_{t-1}}{f_t}$ is a good approximation.

Lemma 60 ([HWHB+08]). *For all t,*

$$\left|\frac{f_{t-1}}{f_t} - \frac{1}{\phi}\right| < \frac{1}{f_t^2}.$$

Proof of Lemma 58. Assume without loss of generality that $v' < v$. By the definition of ρ_q, there is an integer u such that

$$|pv - pv' - uq| = \rho_q(pv, pv').$$

Dividing both sides by $q(v - v')$, we get

$$\left|\frac{p}{q} - \frac{u}{v - v'}\right| = \frac{\rho_q(pv, pv')}{q(v - v')}.$$

Hence we have

$$\frac{c_3}{(v - v')^2} \leq \left|\frac{1}{\phi} - \frac{u}{v - v'}\right| < \frac{\rho_q(pv, pv')}{q(v - v')} + \frac{1}{q^2},$$

where we have used Lemma 59 for the first inequality and Lemma 60 for the second.

If $|v| \leq \sqrt{c_3/4} \cdot q$ and $|v'| \leq \sqrt{c_3/4} \cdot q$, then we get

$$\frac{\rho_q(pv, pv')}{q(v - v')} > \frac{c_3}{2(v - v')^2},$$

so that

$$\rho_q(pv, pv') > \frac{c_3 q}{2(v - v')} \geq \frac{c_3 q}{4\max\{|v|, |v'|\}},$$

completing the proof of (45).

Now we turn to (46). Two applications of Lemma 60 and the triangle inequality together imply

$$\left|\frac{f_{t-1}}{f_t} - \frac{p}{q}\right| = \left|\frac{f_{t-1}}{f_t} - \frac{f_L}{f_{L+1}}\right| \leq \left|\frac{f_{t-1}}{f_t} - \frac{1}{\phi}\right| + \left|\frac{f_L}{f_{L+1}} - \frac{1}{\phi}\right| \leq \frac{1}{f_t^2} + \frac{1}{f_{L+1}^2} \leq \frac{2}{f_t^2}.$$

Thus, for all integers j with $|j| \leq f_t$, we have

$$\left|\frac{jf_{t-1}}{f_t} - \frac{jp}{q}\right| \leq \frac{2}{f_t}. \tag{47}$$

Since f_{t-1} and f_t are relatively prime, the elements $\left\{\frac{jf_{t-1}}{f_t} \mod 1 : j \in [f_t]\right\}$ are exactly equally spaced over $[0, 1]$, so there is some $\tilde{j} \in [q]$ such that the fractional part of $\frac{\tilde{j}f_{t-1}}{f_t}$ is within $\pm\frac{1}{f_t}$ of i/q. Applying (47), the fractional part of $\frac{\tilde{j}p}{q}$ is within $\pm\frac{3}{f_t}$ of i/q, and scaling up by q yields (46), completing the proof of Lemma 58. □

Let $\ell = \lfloor \sqrt{L} \rfloor$. Now that we have p and q, we turn to defining $(\mathbf{U}_1, \mathbf{V}_1), ..., (\mathbf{U}_\ell, \mathbf{V}_\ell)$. The distributions that are hard to distinguish will be chosen from among

$$\mathbf{S}_1 = \mathbf{U}_1 + p\mathbf{V}_1 \mod q, \ldots, \mathbf{S}_\ell = \mathbf{U}_\ell + p\mathbf{V}_\ell \mod q.$$

For a positive integer a let $\mathrm{Bin}(a^2, 1/2)$ be the Binomial distribution which is a sum of a^2 independent Bernoulli random variables with expectation $1/2$, and let $\mathbf{W}(a) = \mathrm{Bin}(2a^2, 1/2) - a^2$. Let $C = \lceil a^2/q \rceil$ so that $Cq - a^2 \geq 0$, and observe that

$$\mathbf{W}(a) + Cq \mod q \quad \text{is identical to} \quad \mathbf{W}(a) \mod q,$$

and that $\mathbf{W}(a) + Cq$ is a $\mathrm{PBD}_{\Theta(q+a^2)}$ distribution.

We will need a lemma about how the $\mathbf{W}$ random variables "behave like discretized Gaussians" that is a bit stronger in some cases than the usual Chernoff-Hoeffding bounds. We will use the following bound on binomial coefficients:

Lemma 61 ([Ros99])**.** *If $k = o(n^{3/4})$, then*

$$\frac{1}{2^n} \cdot \binom{n}{\lfloor n/2 \rfloor + k} = O\left(\frac{1}{\sqrt{n}} \exp\left(\frac{-2k^2}{n}\right)\right).$$

Now for our bound regarding $\mathbf{W}$.

Lemma 62. *There is a constant $c_4 > 0$ such that for all k,*

$$\mathbf{Pr}[\mathbf{W}(a) = k] \leq \frac{c_4 \exp\left(-\frac{k^2}{2a^2}\right)}{a}.$$

Proof. If $|k| \leq a^{4/3}$, the lemma follows directly from Lemma 61. If $|k| > a^{4/3}$, we may suppose w.l.o.g. that k is positive. Then Hoeffing's inequality implies that

$$\begin{aligned}
&\mathbf{Pr}[\mathbf{W}(a) = k] \leq \mathbf{Pr}[\mathbf{W}(a) \geq k] \leq \exp\left(-\frac{k^2}{a^2}\right) \\
&\leq \exp\left(-\frac{k^2}{2a^2}\right) \exp\left(-\frac{k^2}{2a^2}\right) \leq \exp\left(-\frac{k^2}{2a^2}\right) \exp\left(-\frac{a^{2/3}}{2}\right) = O\left(\exp\left(-\frac{k^2}{2a^2}\right)/a\right),
\end{aligned}$$

completing the proof. □

The following lemma may be considered a standard fact about binomial coefficients, but for completeness we give a proof below.

Lemma 63. *For every $c > 0$, there exists $c' > 0$ such that for all integers x, a with $|x| < c \cdot a$,*

$$\mathbf{Pr}[\mathbf{W}(a) = x] \geq \frac{c'}{a}.$$

Proof. Note that

$$\mathbf{Pr}[\mathbf{W}(a) = x] = \frac{1}{2^{2a^2}} \cdot \binom{2a^2}{a^2 + x}$$

Thus,

$$\begin{aligned}\mathbf{Pr}[\mathbf{W}(a) = x] &\geq \frac{1}{2^{2a^2}} \cdot \binom{2a^2}{a^2} \cdot \frac{(a^2)!(a^2)!}{(a^2+x)!(a^2-x)!} \\ &\geq \frac{1}{10 \cdot a} \cdot \prod_{j=1}^{x} \frac{a^2 - j + 1}{a^2 + j} = \frac{1}{10 \cdot a} \cdot \prod_{j=1}^{x} \frac{1 - \frac{j-1}{a^2}}{1 + \frac{j}{a^2}} \\ &\geq \frac{1}{10 \cdot a} \cdot \prod_{j=1}^{x} e^{\frac{-10 \cdot j}{a^2}} \geq \frac{1}{10 \cdot a} \cdot e^{-\frac{10 \cdot x^2}{a^2}}.\end{aligned}$$

Here the second inequality uses that $2^{-2n} \cdot \binom{2n}{n} \geq \frac{1}{10 \cdot \sqrt{n}}$ and the third inequality uses that for $j \leq a^2/2$, $e^{\frac{-10 \cdot j}{a^2}} \leq (1 - \frac{j-1}{a^2})/(1 + \frac{j}{a^2})$. The bound on $\mathbf{Pr}[\mathbf{W}(a) = x]$ follows immediately. □

Now, let $\mathbf{U}_t = \mathbf{W}(\lfloor \frac{p}{c_5 f_t} \rfloor)$, where c_5 is a constant that we will set in the analysis, and let $\mathbf{V}_t = \mathbf{W}(f_t)$. Recall that $\mathbf{S}_t = \mathbf{U}_t + p\mathbf{V}_t \mod q$. Let $\ell' = \lfloor L^{1/4} \rfloor$ and recall that $\ell = \lfloor L^{1/2} \rfloor$. Let $\mathcal{S} = \{\mathbf{S}_{\ell'}, ..., \mathbf{S}_\ell\}$ This is the set of $\Omega(\log^{1/2} q)$ distributions to which we will apply Lemma 28.

Now, to apply Lemma 28 we need upper bounds on the KL-divergence between pairs of elements of $\mathcal{S}$, and lower bounds on their total variation distance. Intuitively, the upper bound on the KL-divergence will follow from the fact that each $\mathbf{S}_t$ "spaces apart by $\Theta(p/f_t)$ PBDs with a lot of measure in a region of size p/f_t" (i.e. the translated $\mathbf{U}_t$ distributions), so the probability never gets very small between consecutive "peaks" in the distribution; consequently, all of the probabilities in all of the distributions are within a constant factor of one another. The following lemma makes this precise:

Lemma 64. *There is a constant $c_6 > 0$ such that, for large enough q, if $\lfloor L^{1/4} \rfloor < t < L^{1/2}$, for all $i \in [q]$,*

$$\frac{1}{c_6 q} \leq \mathbf{Pr}[\mathbf{S}_t = i] \leq \frac{c_6}{q}.$$

Proof. Fix an arbitrary t for which $\lfloor L^{1/4} \rfloor < t < L^{1/2}$. Since t is fixed, we drop it from all subscripts.

First, let us work on the lower bound. Roughly, we will show that a random $v \sim \mathbf{V}$ has a good chance of translating $\mathbf{U}$ within $\Theta(\sigma(\mathbf{U}))$ of i. Specifically, (46) implies that there is a $u \in [-3\lfloor q/f_t \rfloor, 3\lfloor q/f_t \rfloor]$ and a $v \in [-f_t, f_t]$ such that $i = u + pv \mod q$. Thus

$$\mathbf{Pr}[\mathbf{S} = i] \geq \mathbf{Pr}[\mathbf{U} = u] \cdot \mathbf{Pr}[\mathbf{V} = v] \geq \Omega\left(\frac{f_t}{p} \cdot \frac{1}{f_t}\right) \geq \Omega\left(\frac{1}{p}\right) = \Omega\left(\frac{1}{q}\right),$$

where the second inequality follows from an application of Lemma 63 (recalling that $q = \Theta(p)$).

Now for the upper bound. We have

$$\begin{aligned}\mathbf{Pr}[\mathbf{S} = i] &= \sum_v \mathbf{Pr}[\mathbf{S} = i | \mathbf{V} = v] \, \mathbf{Pr}[\mathbf{V} = v] \\ &= \sum_v \mathbf{Pr}[\mathbf{U} = i - pv \mod q] \, \mathbf{Pr}[\mathbf{V} = v] \\ &< o(1/q) + \sum_{v : |v| \leq \sigma(V) \ln q} \mathbf{Pr}[\mathbf{U} = i - pv \mod q] \, \mathbf{Pr}[\mathbf{V} = v],\end{aligned}$$

since $\mathbf{Pr}[|v| > \sigma(V)\ln q] = o(1/q)$. Let $\mathcal{V}_1 = [-\sigma(\mathbf{V}), \sigma(\mathbf{V})]$, and, for each $j > 1$, let $\mathcal{V}_j = [-j\sigma(\mathbf{V}), j\sigma(\mathbf{V})] - \mathcal{V}_{j-1}$. Then

$$\begin{aligned}\mathbf{Pr}[\mathbf{S} = i] &\leq o(1/q) + \sum_{j=1}^{\lfloor \ln q \rfloor} \sum_{v \in \mathcal{V}_j} \mathbf{Pr}[\mathbf{U} = i - pv \mod q]\,\mathbf{Pr}[\mathbf{V} = v] \\ &\leq o(1/q) + O(1) \cdot \sum_{j=1}^{\lfloor \ln q \rfloor} \frac{e^{-j^2/2}}{\sigma(\mathbf{V})} \sum_{v \in \mathcal{V}_j} \mathbf{Pr}[\mathbf{U} = i - pv \mod q], \end{aligned} \tag{48}$$

by Lemma 62.

Now fix a $j \leq \ln q$. Let $(v'_k)_{k=1,2,...}$ be the ordering of the elements of $\mathcal{V}_j$ by order of increasing ρ_q-distance from pv'_k to i. Since each $|v'_k| \leq j \cdot \sigma(\mathbf{V}) \ll c_2 q$, Lemma 58 implies that ρ_q-balls of radius $\Omega\left(\frac{q}{j\sigma(\mathbf{V})}\right)$ centered at members of $pv'_1, ..., pv'_k$ are disjoint, so

$$k \cdot \Omega\left(\frac{q}{j\sigma(\mathbf{V})}\right) < 2\rho_q(pv'_k, i) + 1.$$

Since $\sigma(\mathbf{U})\sigma(\mathbf{V}) = \Theta(q)$, we get

$$\rho_q(pv'_k, i) = \Omega\left(\frac{k\sigma(\mathbf{U})}{j}\right). \tag{49}$$

Applying Lemma 62, we get

$$\sum_{v \in \mathcal{V}_j} \mathbf{Pr}[\mathbf{U} = i - pv \mod q] \leq \frac{1}{\sigma(\mathbf{U})} \sum_{k>0} \exp\left(-\Omega\left(\frac{k^2 \cdot \sigma^2(\mathbf{U})}{j^2\sigma^2(\mathbf{U})}\right)\right) = O\left(\frac{j}{\sigma(\mathbf{U})}\right).$$

Combining with (48), we get

$$\mathbf{Pr}[\mathbf{S} = i] = \sum_{j=1}^{\infty} O\left(\frac{j}{\sigma(\mathbf{U})}\right) \cdot O\left(\frac{1}{\sigma(\mathbf{V})}\right) \cdot e^{-j^2/2} = O\left(\frac{1}{\sigma(\mathbf{U}) \cdot \sigma(\mathbf{V})}\right) = O\left(\frac{1}{q}\right).$$

This finishes the upper bound on $\mathbf{Pr}[\mathbf{S} = i]$, concluding our proof. □

We have the following immediate corollary.

Lemma 65. *There is a constant c_7 such that, for all $i, j \in \{\ell', ..., \ell\}$, we have $D_{KL}(\mathbf{S}_i || \mathbf{S}_j) \leq c_7$.*

It remains only to give a lower bound on the total variation distance.

Lemma 66. *There is a positive constant c_8 such that, for large enough q, for $\ell \geq i > j$, we have $d_{TV}(\mathbf{S}_i, \mathbf{S}_j) > c_8$.*

Proof. Let $\mathcal{W}$ be the union of two integer intervals

$$\mathcal{W} = [-f_i, ..., -f_{j+1}] \cup [f_{j+1}, ..., f_i].$$

It may be helpful to think of $\mathcal{W}$ as being comprised of v such that pv is the location of a "peak" in $\mathbf{S}_i$, but not in $\mathbf{S}_j$. We will show that $\mathbf{S}_i$ assigns significantly more probability to points close to pv than $\mathbf{S}_j$ does.

Choose $v \in \mathcal{W}$, and u with $|u| \leq \frac{p}{c_5 f_i}$. (We will later set $c_5 > 0$ to be a sufficiently large absolute constant.) For large enough q, standard facts about binomial coefficients give that

$$\mathbf{Pr}[\mathbf{S}_i = pv + u \mod q] \geq \mathbf{Pr}[\mathbf{V}_i = v] \cdot \mathbf{Pr}[\mathbf{U}_i = u] \geq \frac{1}{5f_i} \cdot \frac{c_5 f_i}{5p} = \frac{c_5}{25p}. \tag{50}$$

Now, let us upper bound $\mathbf{Pr}[\mathbf{S}_j = pv + u \mod q]$. Let $a \in [q]$ be such that $pv + u = a \mod q$. We have

$$\begin{aligned}\mathbf{Pr}[\mathbf{S}_j = a] &= \sum_v \mathbf{Pr}[\mathbf{S}_j = a \mid \mathbf{V}_j = v]\, \mathbf{Pr}[\mathbf{V}_j = v] \\ &< o(1/q^2) + \sum_{v : |v| \leq \sigma(\mathbf{V}_j) \ln q} \mathbf{Pr}[\mathbf{S}_j = a \mid \mathbf{V}_j = v]\, \mathbf{Pr}[\mathbf{V}_j = v].\end{aligned}$$

As before, let $\mathcal{V}_1 = [-\sigma(\mathbf{V}_j), \sigma(\mathbf{V}_j)]$, and, for each $h > 1$, let $\mathcal{V}_h = [-h\sigma(\mathbf{V}_j), h\sigma(\mathbf{V}_j)] - \mathcal{V}_{h-1}$, so that Lemma 62 implies

$$\mathbf{Pr}[\mathbf{S}_j = a] \leq o(1/q^2) + \sum_{h=1}^{\lfloor \ln q \rfloor} \sum_{v \in \mathcal{V}_h} \mathbf{Pr}[\mathbf{U}_j = a - pv \mod q]\, \mathbf{Pr}[\mathbf{V}_j = v] \tag{51}$$

$$\leq o(1/q^2) + \sum_{h=1}^{\lfloor \ln q \rfloor} \frac{c_4 e^{-(h-1)^2/2}}{\sigma(\mathbf{V}_j)} \sum_{v \in \mathcal{V}_h} \mathbf{Pr}[\mathbf{U}_j = a - pv \mod q]. \tag{52}$$

Let $(v'_k)_{k=1,2,\ldots}$ be the ordering of the elements of $\mathcal{V}_h$ by order of increasing ρ_q distance from a to pv'_k. Since each $|v'_k| \leq h \cdot \sigma(\mathbf{V}) \ll c_2 q$, Lemma 58 implies that ρ_q-balls of radius $\frac{c_1 q}{2hf_j}$ centered at members of $pv'_1, \ldots, pv'_k$ are disjoint, so

$$k \cdot \frac{c_1 q}{h f_j} < 2\rho_q(a, pv'_k) + \frac{c_1 q}{h f_j}.$$

so, for large enough q, we have

$$\rho_q(a, pv'_k) > \frac{c_1 (k-1) q}{5 h f_j}.$$

Using this with Lemma 62, we get that, for large enough q,

$$\begin{aligned}\sum_{v \in \mathcal{V}_h} \mathbf{Pr}[\mathbf{U}_j = a - pv \mod q] &\leq \frac{1}{\sigma(\mathbf{U}_j)} \sum_k c_4 \exp\left(-\frac{(k-1)^2 c_1^2 q^2 c_5^2}{100 h^2 p^2}\right) \\ &\leq 2 \cdot \frac{c_4 c_5 f_j}{p} \sum_k \exp\left(-\frac{(k-1)^2 c_1^2 c_5^2}{100 h^2}\right) \\ &\leq \frac{c_4 c_5 f_j}{p} \cdot \frac{40(h+1)}{c_1 c_5} = \frac{40(h+1) c_4 f_j}{c_1 p}.\end{aligned}$$

Plugging back into (52), we get

$$\mathbf{Pr}[\mathbf{S}_j = a] \leq o(1/q^2) + 40 \sum_{h=1}^{\lfloor \ln q \rfloor} \frac{(h+1) e^{-(h-1)^2/2} c_4^2}{c_1 p}.$$

Thus, if c_5 is a large enough absolute constant, there is a constant c_7 such that

$$\mathbf{Pr}[\mathbf{S}_i = pv + u \mod q] - \mathbf{Pr}[\mathbf{S}_j = pv + u \mod q] > \frac{c_7}{p},$$

for all $v \in \mathcal{W}$, and u with $|u| \leq \frac{p}{c_5 f_i}$. Lemma 58 implies that, if c_5 is large enough, the resulting values of $pv + u$ are distinct elements of $[q]$, and the number of such pairs is at least $(f_{i+1} - f_i) \cdot \lfloor \Omega(\frac{p}{f_i}) \rfloor$ which is $\Omega(p)$, which completes the proof. □

13 Lower bound for $(a_{\max}, 3)$-sums: Proof of Theorem 6

Theorem 6 follows from the following stronger result, which gives a lower bound for learning a weighted sum of $PBDs$ with weights $\{0 = a_1, a_2, a_3\}$ even if the largest support value a_3 is known.

Theorem 67 ($k = 3$, strengthened unknown-support lower bound)**.** *Let A be any algorithm with the following properties: algorithm A is given N, an accuracy parameter $\varepsilon > 0$, a value $0 < a_{\max} \in \mathbb{Z}$, and access to i.i.d. draws from an unknown $\mathbf{S} = a_2\mathbf{S}_2 + a_3\mathbf{S}_3$, where a_3 is the largest prime that is at most $a_{\max}\}$ and $0 < a_2 < a_3$. (So the values $a_1 = 0$ and a_3 are "known" to the learning algorithm A, but the value of a_2 is not.) Suppose that A is guaranteed to output, with probability at least $9/10$, a hypothesis distribution $\tilde{\mathbf{S}}$ satisfying $d_{\mathrm{TV}}(\mathbf{S}, \tilde{\mathbf{S}}) \leq \varepsilon$. Then for sufficiently large N, A must use $\Omega(\log a_{\max})$ samples even when run with ε set to a (suitably small) positive absolute constant.*

Via our reduction, Theorem 56, we obtain Theorem 67 from the following lower bound for learning a single scaled PBD mod a_3 when the scaling factor is unknown:

Theorem 68 (lower bound for learning mod a_3)**.** *Let A be any algorithm with the following properties: algorithm A is given N, an accuracy parameter $\varepsilon > 0$, a value $0 < a_{\max} \in \mathbb{Z}$, and access to i.i.d. draws from from a distribution $\mathbf{S}' = (a_2\mathbf{S}_2 \mod a_3)$ where $\mathbf{S}_2$ is a PBD$_N$, a_3 is the largest prime that is at most $a_{\max}$, and $a_2 \in \{1, \ldots, a_3 - 1\}$ is "unknown" to A. Suppose that A is guaranteed to output a hypothesis distribution $\tilde{\mathbf{S}}'$ satisfying $\mathbf{E}[d_{\mathrm{TV}}(\mathbf{S}', \tilde{\mathbf{S}}')] \leq \varepsilon$ (where the expectation is over the random samples drawn from $\mathbf{S}'$ and any internal randomness of A). Then for sufficiently large N, A must use $\Omega(\log a_{\max})$ samples when run with ε set to some sufficiently small absolute constant.*

While Theorem 68 lower bounds the expected error of the hypothesis produced by a learning algorithm that uses too few samples, such a lower bound is easily seen to imply an (ε, δ)-type bound as well. Thus to prove Theorem 67 (and thus Theorem 6) it suffices to prove Theorem 68.

13.1 Proof of Theorem 68

Recall that by the Bertrand-Chebychev theorem we have $a_3 = \Theta(a_{\max})$; throughout what follows we view a_3 as a "sufficiently large" prime number. Let $\mathbf{S}_2$ be the distribution $\mathrm{Bin}(N', \frac{1}{2}) + a_3 - \left(\frac{N'}{2} - \frac{c\sqrt{N'}}{2}\right)$, where $N' = \lceil \left(\frac{a_3}{cK}\right)^2 \rceil$ and $c, K > 0$ are absolute constants that will be specified later. (It is helpful to think of c as being a modest number like, say, 10, and to think of K as being extremely large relative to c.) Note that $\mathbf{S}_2$ is a PBD$_N$ for $N = \mathrm{poly}(a_3)$, and that $\mathbf{S}_2$ has $\mathbf{Var}[\mathbf{S}_2] = N'/4 = \sigma_{\mathbf{S}_2}^2$ where $\sigma_{\mathbf{S}_2} = (a_3)/(cK) + O(1)$. Note further that nothing is "unknown" about $\mathbf{S}_2$ — the only thing about $\mathbf{S}' = a_2\mathbf{S}_2$ that is unknown to the learner is the value of a_2.

Remark 69. For intuition, it is useful to consider the distribution $\mathbf{S}_2$ mod a_3, and to view it as a coarse approximation of the distribution $\mathbf{U}$ which is uniform over the interval $\{1, \ldots, c\sqrt{N'}\}$ where $c\sqrt{N'} \approx a_3/K$; we will make this precise later.

The lower bound of Theorem 68 is proved by considering distributions $\mathbf{S}'_r$, $1 \le r \le a_3 - 1$, which are defined as $\mathbf{S}'_r := (r \cdot \mathbf{S}_2 \mod a_3)$. The theorem is proved by applying the generalized Fano's Inequality (Theorem 28) to a subset of the distributions $\{\mathbf{S}'_r\}_{r\in[a_3-1]}$; recall that this requires both an upper bound on the KL divergence and a lower bound on the total variation distance. The following technical lemma will be useful for the KL divergence upper bound:

Lemma 70. *For any $1 \le r_1 \ne r_2 \le a_3 - 1$ and any $j \in \{0, 1, \dots, a_3 - 1\}$, the ratio $\mathbf{S}'_{r_1}(j)/\mathbf{S}'_{r_2}(j)$ lies in $[1/C, C]$ where C is a constant (that is independent of a_3 but depends on c, K).*

Proof. Recalling that a_3 is prime, for any $r \in [a_3 - 1]$ and any $j \in \{0, 1, \dots, a_3 - 1\}$, if $r^{-1} \in [a_3]$ is such that $r^{-1}r = 1 \mod a_3$, we have

$$\mathbf{S}'_r(j) = \mathbf{Pr}[r \cdot \mathbf{S}_2 \equiv j \mod a_3] = \mathbf{Pr}[\mathbf{S}_2 \equiv jr^{-1} \mod a_3] = \Theta(1) \cdot \mathbf{S}_2(M),$$

where $M \in \{0, 1, \dots, N\}$ is the integer in $a_3\mathbb{Z} + jr^{-1}$ that is closest to $N/2$. Since $|M - N/2| \le a_3/2 = \Theta(\sqrt{N})$ and $\mathbf{S}_2 = \mathrm{Bin}(N, 1/2)$, standard facts about binomial coefficients imply that $\mathbf{S}_2(M) = \binom{N}{M}/2^N = \Theta(1)/\sqrt{N}$, from which the lemma follows. □

From this, recalling the definition of KL-divergence $D_{KL}(\mathbf{S}'_{r_1}||\mathbf{S}'_{r_2}) = \sum_i \mathbf{S}'_{r_1}(i) \ln \frac{\mathbf{S}'_{r_1}(i)}{\mathbf{S}'_{r_2}(i)}$, we easily obtain

Corollary 71. *For any $1 \le r_1 \ne r_2 \le a_3 - 1$ we have $D_{KL}(\mathbf{S}'_{r_1}||\mathbf{S}'_{r_2}) = O(1)$.*

Next we turn to a lower bound on the variation distance; for this we will have to consider only a restricted subset of the distributions $\{\mathbf{S}'_r\}_{r\in[a_3-1]}$, and use a number theoretic equidistribution result of Shparlinski. To apply this result it will be most convenient for us to work with the distribution $\mathbf{U}$ instead of $\mathbf{S}_2$ (recall Remark 69) and to bring $\mathbf{S}_2$ and the $\mathbf{S}'_r$ distributions into the picture later (in Section 13.1.2) once we have established an analogue of our desired result for some distributions related to $\mathbf{U}$.

13.1.1 Equidistribution of scaled modular uniform distributions $\mathbf{U}'_r$.

For $1 \le r \le a_3 - 1$ we consider the distributions $\mathbf{U}'_r := (r \cdot \mathbf{U} \mod a_3)$ (note the similarity with the distributions $\mathbf{S}'_r$). Since for each $r \in [a_3 - 1]$ the distribution $\mathbf{U}'_r$ is uniform on a $\Theta(1/K)$-fraction of the domain $\{0, 1, \dots, a_3 - 1\}$, it is natural to expect that $d_{TV}(\mathbf{U}'_{r_1}, \mathbf{U}'_{r_2})$ is large for $r_1 \ne r_2$. To make this intuition precise, we make the following definition.

Definition 72. Given integers r, p, Y, Z and a set $\mathcal{X}$ of integers, we define

$$N_{r,p}(\mathcal{X}, Y, Z) := \left|\left\{(x, y) \in \mathcal{X} \times [Z + 1, Z + Y] : r \cdot x \equiv y \pmod{p}\right\}\right|.$$

We will use the following, which is due to Shparlinski [Shp08].

Lemma 73 ([Shp08])**.** *For all integers p, Z, X, Y such that $p \ge 2$ and $0 < X, Y < p$, for any $\mathcal{X} \subseteq \{1, ..., X\}$, we have*

$$\sum_{r=1}^{p} \left| N_{r,p}(\mathcal{X}, Y, Z) - \frac{|\mathcal{X}| \cdot Y}{p} \right|^2 \le |\mathcal{X}| \cdot (X + Y) \cdot p^{o(1)}.$$

We shall use the following corollary. Set $\mathcal{X} = \{1, ..., X\}$. Let us define the quantity

$$\mathcal{N}_{r,X} = |\{(x, y) : x, y \in \mathcal{X}, r \cdot x \equiv y \pmod{p}\}|.$$

Taking $Z = 0$ and $Y = X$, we get:

Corollary 74. *For all integers p and X such that $p > 0$ and $0 < X < p$, we have*

$$\sum_{r=1}^{p} \left| \mathcal{N}_{r,X} - \frac{X^2}{p} \right|^2 \leq X^2 \cdot p^{o(1)}.$$

This easily yields

$$\mathbf{E}_{r \in [p]} \left[\left| \mathcal{N}_{r,X} - \frac{X^2}{p} \right|^2 \right] \leq \frac{X^2}{p^{1-o(1)}}$$

which in turn implies

$$\mathbf{Pr}_{r \in [p]} \left[\mathcal{N}_{r,X} \geq \frac{2X^2}{p} \right] \leq \frac{p^{1+o(1)}}{X^2}.$$

We specialize this bound by setting X to be $\lceil c\sqrt{N'} \rceil$ and $p = a_3$ which gives $\frac{X^2}{p} = \frac{a_3}{K^2} + O(1)$, from which we get that

$$\mathbf{Pr}_{r \in [a_3]} \left[\mathcal{N}_{r,X} \geq \frac{2a_3}{K^2} \right] \leq \frac{a_3^{o(1)}}{a_3}. \tag{53}$$

Using (53) it is straightforward to show that there is a large subset of the distributions $\{\mathbf{U}'_r\}_{r \in [a_3]}$ any two of which are very far from each other in total variation distance:

Theorem 75. *Given any sufficiently large prime a_3, there is a subset of $t \geq a_3^{1/3}$ many values $\{q_1, \ldots, q_t\} \subset [a_3]$ such that for any $q_i \neq q_j$ we have $d_{\mathrm{TV}}(\mathbf{U}'_{q_i}, \mathbf{U}'_{q_j}) \geq 1 - \frac{3}{K}$.*

Proof. To prove the theorem it suffices to show that if q_1, q_2 are chosen randomly from $[a_3]$ then $d_{\mathrm{TV}}(\mathbf{U}'_{q_i}, \mathbf{U}'_{q_j}) \geq 1 - \frac{3}{K}$ with probability $1 - \frac{a_3^{o(1)}}{a_3}$. (From there, the theorem follows from a standard deletion argument [ASE92].) Since a_3 is prime, to show this it suffices to prove that for a randomly chosen $r \in [a_3]$ we have that $d_{\mathrm{TV}}(\mathbf{U}, \mathbf{U}'_r) \geq 1 - \frac{3}{K}$ with probability $1 - \frac{a_3^{o(1)}}{a_3}$. Observe that for a given outcome of r, since both $\mathbf{U}$ and $\mathbf{U}'_r$ are uniform distributions over their domains ($\mathcal{X}$ and $(r \cdot \mathcal{X} \mod a_3)$ respectively) which are both of size $X = \lceil c\sqrt{N'} \rceil$, we have that $d_{\mathrm{TV}}(\mathbf{U}, \mathbf{U}'_r) = 1 - \frac{|\mathcal{X} \cap (r \cdot \mathcal{X} \mod a_3)|}{X}$. Moreover, we have that

$$\mathcal{N}_{r,X} = |\{(x, (rx \mod a_3)) : x, (rx \mod a_3) \in \mathcal{X}\}| = |\mathcal{X} \cap (r \cdot \mathcal{X} \mod a_3)|,$$

so

$$d_{\mathrm{TV}}(\mathbf{U}, \mathbf{U}'_r) = 1 - \frac{\mathcal{N}_{r,X}}{X} = 1 - \frac{\mathcal{N}_{r,X}}{\lceil c\sqrt{N'} \rceil} = 1 - \frac{\mathcal{N}_{r,X}}{a_3/K + O(1)},$$

which is at least $1 - \frac{3}{K}$ provided that $\mathcal{N}_{r,X} < \frac{2a_3}{K^2}$. So by (53) we get that $d_{\mathrm{TV}}(\mathbf{U}, \mathbf{U}'_r) \geq 1 - \frac{3}{K}$ with probability $1 - \frac{a_3^{o(1)}}{a_3}$ over a random r, as desired, and we are done. □

13.1.2 Concluding the proof of Theorem 68.

Given Theorem 75 it is not difficult to argue that the related family of distributions $\{\mathbf{S}'_{q_1}, \ldots, \mathbf{S}'_{q_t}\}$ are all pairwise far from each other with respect to total variation distance. First, recall that $\mathbf{S}_2$ is a $\mathrm{Bin}(N', \frac{1}{2})$ distribution (mod a_3) which has been shifted so that its mode is in the center of $\mathrm{supp}(\mathbf{U})$ and so that the left and right endpoints of $\mathrm{supp}(\mathbf{U})$ each lie $c/2$ standard deviations away from its mode. From this, the definition of $\mathbf{S}'_{q_i}$, and well-known tail bounds on the Binomial distribution it is straightforward to verify that a $1 - e^{-\Theta(c^2)}$ fraction of the probability mass of $\mathbf{S}'_{q_i}$ lies on $\mathrm{supp}(\mathbf{U}'_{q_i})$, the support of $\mathbf{U}'_{q_i}$. Moreover, standard bounds on the Binomial distribution imply that for any two points $\alpha, \beta \in \mathrm{supp}(\mathbf{U}'_{q_i})$, we have that

$$\frac{1}{\Gamma(c)} \leq \frac{\mathbf{S}'_{q_i}(\alpha)}{\mathbf{S}'_{q_j}(\beta)} \leq \Gamma(c) \tag{54}$$

where $\Gamma(c)$ is a function depending only on c. Let $\mathbf{S}''_r$ denote $(\mathbf{S}'_r)_{\mathrm{supp}(\mathbf{U}'_r)}$, i.e. the conditional distribution of $\mathbf{S}'_r$ restricted to the domain $\mathrm{supp}(\mathbf{U}'_r)$. Recalling Theorem 75 and the fact that $d_{\mathrm{TV}}(\mathbf{U}'_{q_i}, \mathbf{U}'_{q_j}) = 1 - \frac{|\mathrm{supp}(\mathbf{U}'_{q_i}) \cap \mathrm{supp}(\mathbf{U}'_{q_j})|}{|\mathrm{supp}(\mathbf{U}'_{q_i})|}$, by (54) we see that by choosing K to be suitably large relative to $\Gamma(c)$, we can ensure that $d_{\mathrm{TV}}(\mathbf{S}''_{q_i}, \mathbf{S}''_{q_j})$ is at least $9/10$. Since $d_{\mathrm{TV}}(\mathbf{S}''_r, \mathbf{S}'_r) \leq e^{-\Theta(c^2)}$, taking c to be a modest positive constant like 10, we get that $d_{\mathrm{TV}}(\mathbf{S}'_{q_i}, \mathbf{S}'_{q_j})$ is at least $8/10$ (with room to spare). Thus we have established:

Theorem 76. *Given any sufficiently large prime a_3, there is a subset of $t \geq a_3^{1/3}$ many values $\{q_1, \ldots, q_t\} \subset [a_3]$ such that for any $q_i \neq q_j$ we have $d_{\mathrm{TV}}(\mathbf{S}'_{q_i}, \mathbf{S}'_{q_j}) \geq 4/5$.*

All the pieces are now in place to apply Fano's Inequality. In the statement of Theorem 28 we may take $\alpha = 1$ (by Theorem 76), $\beta = O(1)$ (by Corollary 71), and ε to be an absolute constant, and Theorem 28 implies that any algorithm achieving expected error at most ε must use $\Omega(\ln t) = \Omega(\ln a_3)$ samples. This concludes the proof of Theorem 68. □

A Time complexity of evaluating and sampling from our hypotheses

Inspection of our learning algorithms reveals that any possible hypothesis distribution $\mathbf{H}$ that the algorithms may construct, corresponding to any possible vector of outcomes for the guesses that the algorithms may make, must have one of the following two forms:

(a) (see Sections 8.1 and 9.1) $\mathbf{H}$ is uniform over a multiset S of at most $1/\varepsilon^{2^{\mathrm{poly}(k)}}$ many integers (see the comment immediately after Fact 25; note that the algorithm has S).

(b) (see Lemma 35 and Definition 29) $\mathbf{H}$ is of the form $\mathbf{U}_{\widehat{Y}} + \mathbf{Z}$ where $\widehat{Y}$ is a multiset of at most $1/\varepsilon^{2^{\mathrm{poly}(k)}}$ integers (note that the algorithm has $\widehat{Y}$) and $\mathbf{Z} = \sum_{a=1}^{K=\mathrm{poly}(k)} p_a \mathbf{Z}_a$ where $\mathbf{Z}_a$ is the uniform distribution on the interval $[-c_a, c_a] \cap \mathbb{Z}$ (and the algorithm has K, the p_a's, and the c_a's).

It is easy to see that a draw from a type-(a) distribution can be simulated in time $1/\varepsilon^{2^{\mathrm{poly}(k)}}$, and likewise it is easy to simulate an $\mathrm{eval}_{\mathbf{H}}$ oracle for such a distribution in the same time. It is also easy to see that a draw from a type-(b) distribution can be simulated in time $1/\varepsilon^{2^{\mathrm{poly}(k)}}$. The main result of this appendix is Theorem 77, stated below. Given this theorem it is easy to see that a type-(b) $\mathrm{eval}_{\mathbf{H}}$ oracle can be simulated in time $1/\varepsilon^{2^{\mathrm{poly}(k)}}$, which is the final piece required to establish that our hypotheses can be efficiently sampled and evaluated as required by Corollary 27.

Theorem 77. *Let $\mathbf{H}$ be a distribution which is of the form $\mathbf{H} = \mathbf{Y} + \sum_{a=1}^{K} p_a \cdot \mathbf{Z}_a$ where the distributions $\mathbf{Y}, \mathbf{Z}_1, \ldots, \mathbf{Z}_K$ are independent integer valued random variables. Further,*

1. *Each $\mathbf{Z}_a$ is uniform on the integer intervals $[-\gamma_a, \ldots, \gamma_a]$.*
2. *$\mathbf{Y}$ is supported on a set $A_{\mathbf{Y}} \subseteq \mathbb{Z}$ of size m with the probabilities given by $\{\alpha_V\}_{V \in A_{\mathbf{Y}}}$.*

Given as input the sets $A_{\mathbf{Y}}$, $\{p_a\}_{a=1}^{K}$, $\{\alpha_V\}_{V \in A_{\mathbf{Y}}}$ and $\{\gamma_a\}_{a=1}^{K}$ and a point $x \in \mathbb{Z}$, we can evaluate $\mathbf{Pr}[\mathbf{H} = x]$ in time $\mathcal{L}^{O(K)}$ where $\mathcal{L}$ is the length of the input.

Our chief technical tool to prove this will be the following remarkable theorem of Barvinok [Bar94], which shows that the number of integer points in a rational polytope can be computed in polynomial time when the dimension is fixed:

Theorem 78. *[Barvinok] There is an algorithm with the following property: given as input a list of m pairs $(a_1, b_1), \ldots, (a_m, b_m)$ where each $a_i \in \mathbb{Q}^d, b_i \in \mathbb{Q}$, specifying a polytope $X \subseteq \mathbb{R}^d$, $X = \{x \in \mathbb{R}^d : \langle a_i, x \rangle \leq b_i\}_{i=1}^{m}$, the algorithm outputs the number of integer points in X in time $\mathcal{L}^{O(d)}$ where $\mathcal{L}$ is the length of the input, i.e. the description length of $\{a_i\}_{i=1}^{m}$ and $\{b_i\}_{i=1}^{m}$.*

We will use this algorithm via the following lemma.

Lemma 79. *Let $\mathbf{H}'$ be a distribution which is of the form $\mathbf{H}' = V + \sum_{a=1}^{K} p_a \cdot \mathbf{Z}_a$ where $V, p_1, \ldots, p_K \in \mathbb{Z}$ and $\mathbf{Z}_1, \ldots, \mathbf{Z}_K$ are independent integer valued random variables and for $1 \leq a \leq K$, $\mathbf{Z}_a$ is uniform on $[-\gamma_a, \ldots, \gamma_a]$. Then, given any point $x \in \mathbb{Z}$, $\{p_a\}_{a=1}^{K}$, $\{\gamma_a\}_{a=1}^{K}$ and V, the value $\mathbf{Pr}[\mathbf{H}' = x]$ can be computed in time $\mathcal{L}^{O(K)}$ where $\mathcal{L}$ is the description size of the input.*

Proof. Consider the polytope defined by the following inequalities:

$$\text{for } 1 \leq a \leq k, \quad -\gamma_a \leq y_a \leq \gamma_a, \qquad \text{and} \qquad x - V - 0.1 \leq \sum_{a=1}^{K} p_a \cdot y_a \leq x - V + 0.1.$$

Then it is easy to see that if $\mathcal{N}_x$ is the number of integer points in the above polytope, then

$$\mathbf{Pr}[\mathbf{H}' = x] = \mathcal{N}_x \cdot \prod_{a=1}^{K} \frac{1}{2\gamma_a + 1}.$$

Combining this observation with Theorem 78 proves the lemma. □

Proof of Theorem 77. Let $V \in A_{\mathbf{Y}}$. Then, using Lemma 79, we obtain that for $\mathbf{H}_V = V + \sum_{a=1}^{K} p_a \cdot \mathbf{Z}_a$, $\mathbf{Pr}[\mathbf{H}_V = x]$ can be computed in time $\mathcal{L}^{O(K)}$. Now, observe that

$$\mathbf{Pr}[\mathbf{H} = x] = \sum_{V \in A_{\mathbf{Y}}} \mathbf{Pr}[\mathbf{H}_V = x] \cdot \alpha_V.$$

As each term of the above sum can be computed in time $\mathcal{L}^{O(K)}$, hence the total time required to compute the above sum is bounded by $\mathcal{L}^{O(K)}$ (note that $\mathcal{L} \geq |A_{\mathbf{Y}}|$). □

Acknowledgments

We would like to thank Igor Shparlinski and Aravindan Vijayaraghavan for useful discussions, and the JMLR reviewers for many helpful comments.

References

[AB83] P. Assouad and C. Birge. Deux remarques sur l'estimation. *C. R. Acad. Sci. Paris Sér. I*, 296:1021–1024, 1983. 22

[AD15] Jayadev Acharya and Constantinos Daskalakis. Testing poisson binomial distributions. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms,SODA 2015*, pages 1829–1840, 2015. 1

[ADK15] Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pages 3591–3599, 2015. 1

[AJOS14] J. Acharya, A. Jafarpour, A. Orlitsky, and A.T. Suresh. Near-optimal-sample estimators for spherical gaussian mixtures, 19 Feb 2014. 22

[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. 1

[ASE92] N. Alon, J. Spencer, and P. Erdos. *The Probabilistic Method*. Wiley-Interscience, New York, 1992. 66

[Bar94] A. Barvinok. A Polynomial Time Algorithm for Counting Integral Points in Polyhedra When the Dimension is Fixed. *Mathematics of Operations Research*, 19(4):769–779, 1994. 68

[Bar15] A. D. Barbour. Personal communication, 2015. 11, 12

[BB85] A.A. Borovkov and A.V. Balakrishnan. *Advances in probability theory: limit theorems for sums of random variables*. Trudy Instituta matematiki. 1985. 1

[BL06] A. D. Barbour and T. Lindvall. Translated poisson approximation for markov chains. *Journal of Theoretical Probability*, 19, 2006. 18

[BS10] M. Belkin and K. Sinha. Polynomial learning of distribution families. In *FOCS*, pages 103–112, 2010. 1

[BX99] A. Barbour and A. Xia. Poisson perturbations. *European Series in Applied and Industrial Mathematics. Probability and Statistics*, 3:131–150, 1999. 9, 20

[Can15] Clément L. Canonne. Big data on the rise? - testing monotonicity of distributions. In *Automata, Languages, and Programming - 42nd International Colloquium, ICALP 2015*, pages 294–305, 2015. 1

[Can16] Clément L. Canonne. Are few bins enough: Testing histogram distributions. In *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2016*, pages 455–463, 2016. 1

[CDGR16] Clément L. Canonne, Ilias Diakonikolas, Themis Gouleakis, and Ronitt Rubinfeld. Testing shape restrictions of discrete distributions. In *33rd Symposium on Theoretical Aspects of Computer Science, STACS 2016*, pages 25:1–25:14, 2016. 1

[CDS17] Y. Cheng, I. Diakonikolas, and A. Stewart. Playing Anonymous Games Using Simple Strategies. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '17, pages 616–631, 2017. 2

[CGS11] L. Chen, L. Goldstein, and Q.-M. Shao. *Normal Approximation by Stein's Method*. Springer, 2011. 2, 19, 20

[Das99] S. Dasgupta. Learning mixtures of Gaussians. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science*, pages 634–644, 1999. 1

[DDKT16] Constantinos Daskalakis, Anindya De, Gautam Kamath, and Christos Tzamos. A size-free CLT for poisson multinomials and its applications. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, pages 1074–1086, 2016. 1, 2, 3, 8, 19

[DDO+13] C. Daskalakis, I. Diakonikolas, R. O'Donnell, R. Servedio, and L.-Y. Tan. Learning Sums of Independent Integer Random Variables. In *FOCS*, pages 217–226, 2013. 1, 2, 3, 4, 6, 7, 8, 9, 19, 41, 42

[DDS12a] C. Daskalakis, I. Diakonikolas, and R.A. Servedio. Learning k-modal distributions via testing. In *SODA*, pages 1371–1385, 2012. 1

[DDS12b] C. Daskalakis, I. Diakonikolas, and R.A. Servedio. Learning Poisson Binomial Distributions. In *Proceedings of the 44th Symposium on Theory of Computing*, pages 709–728, 2012. 1, 2, 3

[DDS12c] C. Daskalakis, I. Diakonikolas, and R.A. Servedio. Learning Poisson Binomial Distributions. In *STOC*, pages 709–728, 2012. 3, 6, 8, 22

[DDS14] Constantinos Daskalakis, Ilias Diakonikolas, and Rocco A. Servedio. Learning k-Modal Distributions via Testing. *Theory of Computing*, 10:535–570, 2014. 6

[DDS15] A. De, I. Diakonikolas, and R. Servedio. Learning from Satisfying Assignments. In *Proc. ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 478–497, 2015. 8, 22

[De18] A. De. Boolean Function Analysis Meets Stochastic Optimization: An Approximation Scheme for Stochastic Knapsack. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1286–1305, 2018. 2

[DK14] C. Daskalakis and G. Kamath. Faster and sample near-optimal algorithms for proper learning mixtures of gaussians. In *COLT*, pages 1183–1213, 2014. 22

[DKS16a] Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Optimal Learning via the Fourier Transform for Sums of Independent Integer Random Variables. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, pages 831–849, 2016. 1, 3, 4

[DKS16b] Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Properly Learning Poisson Binomial Distributions in Almost Polynomial Time. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, pages 850–878, 2016. 2, 3

[DKS16c] Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. The fourier transform of poisson multinomial distributions and its algorithmic applications. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, pages 1060–1073, 2016. 1, 2, 3

[DKT15] C. Daskalakis, G. Kamath, and C. Tzamos. On the Structure, Covering, and Learning of Poisson Multinomial Distributions. To appear in FOCS 2015. Available at http://arxiv.org/pdf/1504.08363v2.pdf, 2015. 2, 3, 6

[DKW56] A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Mathematical Statistics*, 27(3):642–669, 1956. 16

[DL01] L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. Springer Series in Statistics, Springer, 2001. 8, 9, 23

[DP14] C. Daskalakis and C. Papadimitriou. Sparse Covers for Sums of Indicators. PTRF, to appear. Available at http://arxiv.org/abs/1306.1265, 2014. 19

[GK54] BV Gnedenko and AN Kolmogorov. *Independent Random Variables*. Cambridge, Massachusetts: Addison-Wesley, 1954. 1

[GMRZ11] P. Gopalan, R. Meka, O. Reingold, and D. Zuckerman. Pseudorandom generators for combinatorial shapes. In *STOC*, pages 253–262, 2011. 11, 17, 38

[HKLW91] David Haussler, Michael J. Kearns, Nick Littlestone, and Manfred K. Warmuth. Equivalence of models for polynomial learnability. *Information and Computation*, 95(2):129–161, 1991. 56

[HWHB+08] G. H. Hardy, E. M. Wright, R. Heath-Brown, J. Silverman, and A. Wiles. *An introduction to the theory of numbers*. Oxford University Press, 2008. 59

[IH81] I.A. Ibragimov and R.Z. Has'minskii. *Statistical estimation, asymptotic theory (Applications of Mathematics, vol. 16)*. Springer-Verlag, New York, 1981. 22

[Kle14] O. Klesov. *Limit Theorems for Multi-Indexed Sums of Random Variables*, volume 71. 01 2014. 1

[KMR+94] M. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R. Schapire, and L. Sellie. On the learnability of discrete distributions. In *Proceedings of the 26th Symposium on Theory of Computing*, pages 273–282, 1994. 1

[KMV10] A. T. Kalai, A. Moitra, and G. Valiant. Efficiently learning mixtures of two Gaussians. In *STOC*, pages 553–562, 2010. 1

[Lin02] Torgny Lindvall. *Lectures on the coupling method*. Courier Corporation, 2002. 16

[LRSS15] Jian Li, Yuval Rabani, Leonard J. Schulman, and Chaitanya Swamy. Learning arbitrary statistical mixtures of discrete distributions. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015*, 2015. 1

[Mas90] P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Annals of Probability*, 18(3):1269–1283, 1990. 16

[MV10] A. Moitra and G. Valiant. Settling the polynomial learnability of mixtures of Gaussians. In *FOCS*, pages 93–102, 2010. 1

[Pet75] V. V. Petrov. *Sums of Independent Random Variables*. Springer-Verlag Berlin Heidelberg, 1975. Translated by A.A. Brown. 1

[Pet95] V. Petrov. *Limit Theorems of Probability Theory: Sequences of Independent Random Variables*. Oxford University Press, 1995. 1

[PS00] Yu. V. Prokhorov and V. Statulevicius. *Limit Theorems of Probability Theory*, volume 71. 2000. 1

[RÖ7] A. Röllin. Translated Poisson Approximation Using Exchangeable Pair Couplings. *Annals of Applied Probability*, 17(5/6):1596–1614, 2007. 11, 18

[Roo00] B. Roos. Binomial approximation to the Poisson binomial distribution: The Krawtchouk expansion. *Theory Probab. Appl.*, 45:328–344, 2000. 2

[Ros99] K.H. Rosen. *Handbook of Discrete and Combinatorial Mathematics*. Discrete Mathematics and Its Applications. Taylor & Francis, 1999. 60

[RSS14] Yuval Rabani, Leonard J. Schulman, and Chaitanya Swamy. Learning mixtures of arbitrary distributions over large discrete domains. In *Innovations in Theoretical Computer Science, ITCS 2014*, pages 207–224, 2014. 1

[Shp08] I. E. Shparlinski. Distribution of modular inverses and multiples of small integers and the sato-tate conjecture on average. *Michigan Mathematical Journal*, 56(1):99–111, 2008. 15, 65

[Tal94] Michel Talagrand. Sharper bounds for gaussian and empirical processes. *The Annals of Probability*, pages 28–76, 1994. 17

[Tao14] T. Tao. *Higher Order Fourier Analysis*. Number 1 in Graduate Texts in Mathematics. American Mathematical Society, 2014. 14

[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. 1

[WY12] Avi Wigderson and Amir Yehudayoff. Population Recovery and Partial Identification. In *53rd Annual IEEE Symposium on Foundations of Computer Science, FOCS 2012, New Brunswick, NJ*, pages 390–399, 2012. 6

[Yat85] Y. G. Yatracos. Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *Annals of Statistics*, 13:768–774, 1985. 22