Learning Sums of Independent Integer Random Variables

Constantinos Daskalakis[*]
MIT

Ilias Diakonikolas[†]
University of Edinburgh

Ryan O'Donnell[‡]
Carnegie Mellon University

Rocco A. Servedio[§]
Columbia University

Li-Yang Tan[¶]
Columbia University

April 3, 2013

Abstract

Let $\boldsymbol{S} = \boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n$ be a sum of n independent integer random variables $\boldsymbol{X}_i$, where each $\boldsymbol{X}_i$ is supported on $\{0, 1, \ldots, k-1\}$ but otherwise may have an arbitrary distribution (in particular the $\boldsymbol{X}_i$'s need not be identically distributed). How many samples are required to learn the distribution $\boldsymbol{S}$ to high accuracy? In this paper we show that the answer is *completely independent of* n, and moreover we give a computationally efficient algorithm which achieves this low sample complexity. More precisely, our algorithm learns any such $\boldsymbol{S}$ to ϵ-accuracy (with respect to the total variation distance between distributions) using $\text{poly}(k, 1/\epsilon)$ samples, independent of n. Its running time is $\text{poly}(k, 1/\epsilon)$ in the standard word RAM model. Thus we give a broad generalization of the main result of [DDS12b] which gave a similar learning result for the special case $k = 2$ (when the distribution $\boldsymbol{S}$ is a Poisson Binomial Distribution).

Prior to this work, no nontrivial results were known for learning these distributions even in the case $k = 3$. A key difficulty is that, in contrast to the case of $k = 2$, sums of independent $\{0, 1, 2\}$-valued random variables may behave very differently from (discretized) normal distributions, and in fact may be rather complicated — they are not log-concave, they can be $\Theta(n)$-modal, there is no relationship between Kolmogorov distance and total variation distance for the class, etc. Nevertheless, the heart of our learning result is a new limit theorem which characterizes what the sum of an arbitrary number of arbitrary independent $\{0, 1, \ldots, k-1\}$-valued random variables may look like. Previous limit theorems in this setting made strong assumptions on the "shift invariance" of the random variables $\boldsymbol{X}_i$ in order to force a discretized normal limit. We believe that our new limit theorem, as the first result for truly arbitrary sums of independent $\{0, 1, \ldots, k-1\}$-valued random variables, is of independent interest.

[*] `costis@csail.mit.edu`.

[†] `ilias.d@ed.ac.uk`. Part of this work was done while the author was at UC Berkeley supported by a Simons Postdoctoral Fellowship.

[‡] `odonnell@cs.cmu.edu`. Supported by NSF grants CCF-0747250 and CCF-1116594, a Sloan fellowship, and a grant from the MSR–CMU Center for Computational Thinking.

[§] `rocco@cs.columbia.edu`. Supported by NSF grant CCF-1115703.

[¶] `liyang@cs.columbia.edu`. Supported by NSF grant CCF-1115703. Part of this research was completed while visiting CMU.

1 Introduction

We study the problem of learning an unknown random variable given access to independent samples drawn from it. This is essentially the problem of *density estimation*, which has received significant attention in the probability and statistics literature over the course of several decades (see e.g. [DG85, Sil86, Sco92, DL01] for introductory books). More recently many works in theoretical computer science have also considered problems of this sort, with an emphasis on developing computationally efficient algorithms (see e.g. [KMR$^+$94, Das99, FM99, DS00, AK01, VW02, CGG02, BGK04, DHKS05, MR05, FOS05, FOS06, BS10, KMV10, MV10, DDS12a, DDS12b, RSS12, AHK12]).

In this paper we work in the following standard learning framework: the learning algorithm is given access to independent samples drawn from the unknown random variable $\boldsymbol{S}$, and it must output a hypothesis random variable $\widetilde{\boldsymbol{S}}$ such that with high probability the total variation distance $d_{\mathrm{TV}}(\boldsymbol{S}, \widetilde{\boldsymbol{S}})$ between $\boldsymbol{S}$ and $\widetilde{\boldsymbol{S}}$ is at most ϵ. This is a natural extension of the well-known PAC learning model for learning Boolean functions [Val84] to the unsupervised setting of learning an unknown random variable (i.e. probability distribution).

While density estimation has been well studied by several different communities of researchers as described above, both the number of samples and running time required to learn are not yet well understood, even for some surprisingly simple types of discrete random variables. Below we describe a simple and natural class of random variables — *sums of independent integer-valued random variables* — for which we give the first known results, both from an information-theoretic and computational perspective, characterizing the complexity of learning such random variables.

1.1 Sums of independent integer random variables.

Perhaps the most basic discrete distribution learning problem imaginable is learning an unknown random variable $\boldsymbol{X}$ that is supported on the k-element finite set $\{0, 1, \ldots, k-1\}$. Throughout the paper we refer to such a random variable as a *k-IRV* (for "Integer Random Variable"). Learning an unknown k-IRV is of course a well understood problem: it has long been known that a simple histogram-based algorithm can learn such a random variable to accuracy ϵ using $\Theta(k/\epsilon^2)$ samples, and that $\Omega(k/\epsilon^2)$ samples are necessary for any learning algorithm.

A natural extension of this problem is to learn a *sum* of n *independent* such random variables, i.e. to learn $\boldsymbol{S} = \boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n$ where the $\boldsymbol{X}_i$'s are independent k-IRVs (which, we stress, need not be identically distributed and may have arbitrary distributions supported on $\{0, 1, \ldots, k-1\}$). We call such a random variable a *k-SIIRV* (for "Sum of Independent Integer Random Variables"); learning an unknown k-SIIRV is the problem we solve in this paper.

Since every k-SIIRV is supported on $\{0, 1, \ldots, n(k-1)\}$ any such distribution can be learned using $O(nk/\epsilon^2)$ samples, but of course this simple observation does not use any of the k-SIIRV structure. On the other hand, it is clear (even when $n = 1$) that $\Omega(k/\epsilon^2)$ samples are necessary for learning k-SIIRVs. [1] A priori it is not clear how many samples (as a function of n and k) are information-theoretically sufficient to learn k-SIIRVs, even ignoring issues of computational efficiency. The $k = 2$ case of this problem (i.e., Poisson Binomial Distributions, or "PBDs") was only solved last year in [DDS12b], which gave an efficient algorithm using $\widetilde{O}(1/\epsilon^3)$ samples (independent of n) to learn any Poisson Binomial Distribution.

We stress that k-SIIRVs for general k may have a much richer structure than Poisson Binomial Distributions; even 3-SIIRVs are qualitatively very different from 2-SIIRVs. As a simple example of this more intricate structure, consider the 3-SIIRV $\boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n$ with $n = 50$ depicted in Figure 1, in which $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n-1}$ are identically distributed and uniform over $\{0, 2\}$ while $\boldsymbol{X}_n$ puts probability 2/3 on 0 and 1/3 on 1. It is easy to see from this simple example that even 3-SIIRVs can have significantly more daunting structure than any PBD; in particular, they can be $\Theta(1)$-far from every log-concave distribution; can be $\Theta(1)$-far from every Binomial distribution; and can have $\Theta(n)$ modes (and be $\Theta(1)$-far from every unimodal distribution). They thus dramatically fail to have all three kinds of structure (unimodality, log-concavity, and closeness to Binomial) that were exploited in the recent works [DDS12b, CDSS13] on learning PBDs.

[1] It should be noted that while all our results in the paper hold for all settings of n and k, intuitively one should think of n as a "large" asymptotic parameter and $k \ll n$ as a "small" fixed parameter. If $k \geq n$ then the trivial approach described above learns using $O(nk/\epsilon^2) = O(k^2/\epsilon^2)$ samples.

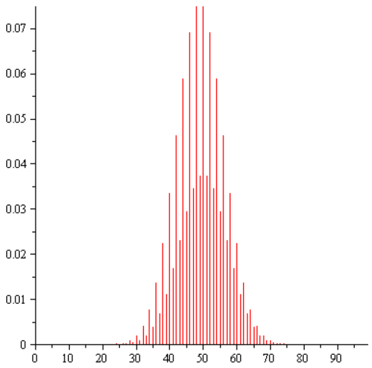

Figure 1: The probability mass function of a certain 3-SIIRV with $n = 50$.

The main learning result. Our main learning result is that both the sample complexity (number of samples required for learning) and the computational complexity (running time) of learning k-SIIRVs is polynomial in k and $1/\epsilon$, and *completely independent of* n. [2]

Theorem 1.1. *[Main Learning Result] There is a learning algorithm for k-SIIRVs with the following properties: Let* $\boldsymbol{S} = \boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n$ *be any sum of* n *independent (not necessarily identically distributed) random variables* $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ *each supported on* $\{0, \ldots, k-1\}$*. The algorithm uses* $\mathrm{poly}(k/\epsilon)$ *samples from* $\boldsymbol{S}$*, runs in time* $\mathrm{poly}(k/\epsilon)$*, and with probability at least* $9/10$ *outputs a (succinct description of a) random variable* $\widetilde{\boldsymbol{S}}$ *such that* $\mathrm{d_{TV}}(\boldsymbol{S}, \widetilde{\boldsymbol{S}}) \leq \epsilon$*.*

(Note that since even learning a single k-IRV requires $\Omega(k/\epsilon^2)$ samples as noted above, this $\mathrm{poly}(k, 1/\epsilon)$ complexity is best possible up to the specific degree of the polynomial.) We give a detailed description of the "succinct description" of our hypothesis random variable $\widetilde{\boldsymbol{S}}$ in Section 1.2, after we describe the new structural theorem that underlies our learning results.

1.2 Prior work and our techniques.

As noted above, Theorem 1.1 is a broad generalization of the main learning result of [DDS12b], which established it in the special case of $k = 2$. A key ingredient in the [DDS12b] learning result is a structural theorem of Daskalakis and Papadimitriou [DP11] which states that any Poisson Binomial Distribution must be either ϵ-close to a *sparse* distribution (supported on $\mathrm{poly}(1/\epsilon)$ consecutive integers), or ϵ-close to a *translated Binomial* distribution. In our current setting of working with k-SIIRVs for general k, structural results of this sort (giving arbitrary-accuracy approximation for an arbitrary k-SIIRV) were not previously known. Our main technical contribution is proving such a structural result (see Theorem 1.2 below); given this structural result, it is relatively straightforward for us to obtain our main learning result, Theorem 1.1, using algorithmic ingredients for learning probability distributions from the recent works [DDS12b, CDSS13].

There is a fairly long line of research on approximate limit theorems for sums of independent integer random variables, dating back several decades (see e.g. [Pre83, Kru86, BHJ92]). Our main structural result employs some of the latest results in this area [CL10, CGS11]; however, we need to extend these results beyond what is currently known. Known approximation theorems for sums of integer random variables (which are generally proved using Stein's method) typically give bounds on the variation distance between a SIIRV $\boldsymbol{S}$ and various specific types of "nice" (Gaussian-like) random variables such as translated/compound Poisson random variables or *discretized normals* (as described in Definition 2.3). However, it is easy to see that in general a k-SIIRV may be very far in variation distance from any discretized normal distribution; see for example the 3-SIIRVs discussed in Figure 1, or the discussion following Corollary 4.5 in [BX99]. To evade this difficulty, limit theorems in the literature typically put strong restrictions on the SIIRVs they apply to

[2] We work in the standard "word RAM" model in which basic arithmetic operations on $O(\log n)$-bit integers are assumed to take constant time. We give more model details in Section 5.

so that a normal-like distribution is forced. Specifically, they bound the total variation distance between $\boldsymbol{S}$ and a "nice" distribution using an error term involving the "shift-distance" (see Definition 2.4) of certain random variables closely related to $\boldsymbol{S}$; see for example Theorem 4.3 of [BX99], Theorem 7.4 of [CGS11], or Theorem 1.3 of [Fan12] for results of this sort. However it is easy to see that for general k-SIIRVs these shift-distances can be very large — large enough that no nontrivial bound is obtained. Thus previous bounds from the literature do not provide structural results that characterize general k-SIIRVs up to arbitrary accuracy.

Another approach to analyzing k-SIIRVs arises from the recent work of Valiant and Valiant [VV11]. They gave a limit theorem for sums $\vec{\boldsymbol{S}}$ of independent $\mathbb{Z}^k$-valued random variables supported on $\{0, e_1, e_2, \ldots, e_k\}$, where e_i denotes the vector $(0, \ldots, 0, 1, 0, \ldots, 0)$ with the 1 in the ith coordinate. Specifically, they bounded the total variation distance of such $\vec{\boldsymbol{S}}$ from the appropriate discretized k-dimensional normal. Note that these $\mathbb{Z}^k$-valued random sums effectively generalize k-SIIRVs, because any k-SIIRV can be obtained as the dot-product $\langle \vec{\boldsymbol{S}}, (0, 1, \ldots, k-1)\rangle$. Unfortunately we cannot use their work for two reasons. The first reason is technical: their error bound has a dependence on n, namely $\Theta(\log^{2/3} n)$, which we do not want to pay. The second reason is more conceptual; as in previous theorems their limiting distribution is a (discretized) normal, which means it cannot capture general SIIRVs. This issue manifests itself in their error term, which is large if the covariance matrix of the k-dimensional normal has a small eigenvalue. Indeed, the covariance matrix will have an $o_n(1)$ eigenvalue for k-SIIRVs of the sort illustrated in Figure 1.

Despite these difficulties, we are able to leverage prior partial results on SIIRVs to give a new structural result showing that *any* k-SIIRV can be approximated to *arbitrarily* high accuracy by a relatively "simple" random variable. More precisely, our result shows that every k-SIIRV is either close to a "sparse" random variable, or else is close to a random variable $c\boldsymbol{Z} + \boldsymbol{Y}$ which decomposes nicely into an arbitrary "local" component $\boldsymbol{Y}$ and a highly structured "global" component $c\boldsymbol{Z}$ (where as above $\boldsymbol{Z}$ is a discretized normal):

Theorem 1.2. *[Main Structural Result] Let $\boldsymbol{S} = \boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n$ be a sum of n independent (not necessarily identically distributed) random variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ each supported on $\{0, \ldots, k-1\}$. Then for any $\epsilon > 0$, $\boldsymbol{S}$ is either*

1. *$O(\epsilon)$-close to a random variable which is supported on at most $\frac{k^9}{\epsilon^4}$ consecutive integers; or*
2. *$O(\epsilon)$-close to a random variable of the form $c\boldsymbol{Z}+\boldsymbol{Y}$ for some $1 \le c \le k-1$, where $\boldsymbol{Y}$, $\boldsymbol{Z}$ are independent random variables such that:*
   - *(a) $\boldsymbol{Y}$ is a c-IRV, and*
   - *(b) $\boldsymbol{Z}$ is a discretized normal random variable with parameters $\frac{\mu}{c}, \frac{\sigma^2}{c^2}$ where $\mu = \mathbf{E}[\boldsymbol{S}]$ and $\sigma^2 = \mathbf{Var}[\boldsymbol{S}]$.*

An alternative statement of our main structural result is the following: for $\boldsymbol{S}$ a k-SIIRV with variance $\mathbf{Var}[\boldsymbol{S}] = \sigma^2$, there is a value $1 \le c \le k-1$ and independent random variables $\boldsymbol{Y}, \boldsymbol{Z}$ as specified in part (2) above, such that $\mathrm{d}_{\mathrm{TV}}(\boldsymbol{S}, c\boldsymbol{Z} + \boldsymbol{Y}) \le \mathrm{poly}(k, 1/\sigma)$. (See Corollary 4.5 for a more detailed statement.) Given this detailed structural characterization of an arbitrary k-SIIRV, it is not difficult to establish our main learning result, Theorem 1.1; see Section 5.

We believe this approximation theorem for arbitrary k-SIIRVs should be of independent interest. One potential direction for future application comes from the field of pseudorandomness. A classic problem in this area is to find pseudorandom generators with short seed length which fool "combinatorial rectangles". A notable recent work by Gopalan et al. [GMRZ11] made new progress on this problem, as well as a generalization they described as fooling "combinatorial shapes". A combinatorial shape is nothing more than a 2-SIIRV (in which the sample space of each $\boldsymbol{X}_i$ is $[m]$ for some integer m). Indeed, much of the technical work in [GMRZ11] goes into giving a new proof methodology for 2-SIIRV limit theorems, one which is more amenable to derandomization. It seems possible that our new limit theorem for k-SIIRVs may be useful in generalizing the [GMRZ11] derandomization results from 2-SIIRVs to k-SIIRVs.

We conclude this subsection by providing the high-level idea in the proof of our structural result, as well as the structure of the hypothesis output by our learning algorithm.

The idea behind Theorem 1.2. The two cases (1) and (2) of Theorem 1.2 correspond to $\boldsymbol{S}$ having "small" versus "large" variance respectively. The easier case is when $\mathbf{Var}(\boldsymbol{S})$ is "small": in this case it is straightforward to show that $\boldsymbol{S}$ must have almost all its probability mass on values in a small interval, and (1) follows easily from this.

The more challenging case is when $\mathbf{Var}(\boldsymbol{S})$ is "large." Intuitively, in order for $\mathbf{Var}(\boldsymbol{S})$ to be large it must be the case that at least one of the $k-1$ values $1, 2, \ldots, k-1$ makes a "large contribution" to $\mathbf{Var}(\boldsymbol{S})$. (This is made precise by working with "0-moded" SIIRVs and analyzing the "b-weight" of the SIIRV for $b \in \{1, \ldots, k-1\}$; see Definition 2.2 for details.)

It is useful to first consider the special case that all $k-1$ values $\{1, \ldots, k-1\}$ make a "large contribution" to $\mathbf{Var}(\boldsymbol{S})$; we do this in Section 3. To analyze this case it is useful to view a draw of the random variable $\boldsymbol{S}$ as taking place in two stages as follows: First (stage 1) we independently choose for each $\boldsymbol{X}_i$ a value $r_i \in \{1, \ldots, k-1\}$. Then (stage 2) for each i we independently choose whether $\boldsymbol{X}_i$ will be set to 0 or to r_i. Using this perspective on $\boldsymbol{S}$, it can be shown that with high probability over the stage-1 outcomes, the resulting random variable that is sampled in Stage 2 is of the form $\sum_{j=1}^{k-1} j \cdot \boldsymbol{Y}_j$ where each $\boldsymbol{Y}_j$ is a large-variance PBD. Given this, using Theorem 7.4 of [CGS11] (see Theorem 3.2) it is not difficult to show that the overall distribution of $\boldsymbol{S}$ is close to a discretetized normal distribution. (This is the $c = 1$ case of case (2) of Theorem 1.2.)

In the general case it may be the case that some of the $k-1$ values contribute very little to $\mathbf{Var}(\boldsymbol{S})$. (This is what happened in the example illustrated in Figure 1.) Let $\mathcal{L} \subset \{1, \ldots, k-1\}$ denote the set of values that make a "small" contribution to $\mathbf{Var}(\boldsymbol{S})$ (observe that $\mathcal{L}$ is nonempty by assumption in this case, or else we are in the special case of the previous paragraph) and let $\mathcal{H} \cup \{0\}$ denote the remaining values in $\{0, 1, \ldots, k-1\}$ (observe that $\mathcal{H}$ is nonempty since otherwise $\mathbf{Var}(\boldsymbol{S})$ would be small as noted earlier). In this general case it is useful to consider a *different* decomposition of the random variable $\boldsymbol{S}$. As before we view a draw of $\boldsymbol{S}$ as taking place in stages, but now the stages are as follows: First (stage 1) for each $i \in [N]$ we independently select whether $\boldsymbol{X}_i$ will be "light" (i.e. will take a value in $\mathcal{L}$) or will be "heavy" (will take a value in $\mathcal{H} \cup \{0\}$). Then (stage 2) for each $\boldsymbol{X}_i$ that has been designated to be "light" we independently choose which particular value in $\mathcal{L}$ it will take, and similarly (stage 3) for each $\boldsymbol{X}_i$ that has been designated "heavy" we independently choose an element of $\mathcal{H} \cup \{0\}$ for it.

The key advantage of the above decomposition is that conditioned on the stage 1 outcome, stages 2 and 3 are *independent* of each other. Using this decomposition, our analysis shows that the contribution from the "light" IRVs (stage 2) is close to a sparse random variable, and the contribution from the "heavy" random variables (stage 3) is close to a gcd($\mathcal{H}$)-scaled discretized normal distribution (this uses the special case, sketched earlier, in which all values make a "large contribution" to the variance). This essentially gives case (2) of Theorem 1.2, where as sketched above, the value "c" is gcd($\mathcal{H}$).

The structure of our hypotheses. The "succinct description" of the hypothesis random variable that our learning algorithm outputs naturally reflects the structure of the approximating random variable given by Theorem 1.2 above. Some terminology will be useful here: we say that an IRV $\boldsymbol{A}$ is *t-flat* if $\boldsymbol{A}$ is supported on a union of $t' \le t$ disjoint intervals $I_1 \cup \cdots \cup I_{t'}$, and for each fixed $1 \le j \le t'$ all points $x_1, x_2 \in I_j$ have $\mathbf{Pr}[\boldsymbol{X} = x_1] = \mathbf{Pr}[\boldsymbol{X} = x_2] = p_j$ for some $p_j > 0$ (so $\boldsymbol{A}$ is piecewise-constant across each interval I_j). An *explicit description* of a t-flat IRV $\boldsymbol{A}$ is a list of pairs $(I_1, p_1), \ldots, (I_{t'}, p_{t'})$, for some $t' \le t$.

There are two possible forms for the output hypothesis random variable $\widetilde{\boldsymbol{S}}$ of Theorem 1.1, corresponding to the two cases of Theorem 1.2 above. The first possible form is simply a list of pairs $(r, p_0), \ldots, (r + \ell, p_\ell)$ where the pair (s, p) indicates that $\mathbf{Pr}[\widetilde{\boldsymbol{S}} = s] = p$ and $\sum_{j=0}^{\ell} p_j = 1$ and $\ell = k^9/\epsilon^4$. The second possible form of the hypothesis is as two lists $(I_1, p_1), \ldots, (I_\ell, p_t)$ and $(0, q_0), \ldots, (c-1, q_{c-1})$, where $I_1, \ldots, I_t$ are disjoint intervals and $\sum_{j=1}^{t} |I_j| p_j = \sum_{i=0}^{c-1} q_i = 1$. The list $(I_1, p_1), \ldots, (I_t, p_t)$ specifies a t-flat random variable $\boldsymbol{Z}'$ and the list $(0, q_0), \ldots, (c-1, q_{c-1})$ specifies a c-IRV $\boldsymbol{Y}'$. The hypothesis distribution in this case is $\widetilde{\boldsymbol{S}} = c\boldsymbol{Z}' + \boldsymbol{Y}'$.

1.3 Discussion: Learning independent sums of more general random variables?

It is natural to ask whether our highly efficient poly(k/ϵ)-sample (independent of n) learning algorithm for k-SIIRVs can be extended to n-way independent sums $\boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n$ of more general types of integer-valued random variables $\boldsymbol{X}_i$ than k-IRVs. Here we note that no such efficient learning results are possible for several natural generalizations of k-IRVs.

One natural generalization is to consider integer random variables which are supported on k values which need not be consecutive. Let us say that $\boldsymbol{X}$ is a *k-support IRV* if $\boldsymbol{X}$ is an IRV supported on at most k values. It turns out that sums of independent k-support IRVs can be quite difficult to learn; in particular, Theorem 3 of [DDS12b] give an information-theoretic argument showing that even for $k = 2$, any algorithm that learns a sum of n 2-support IRVs must use $\Omega(n)$ samples (even if the i-th IRV is constrained to have support $\{0, i\}$).

A different generalization of k-IRVs to consider in this context is a class that we denote as *(c, k)-moment IRVs*. A (c, k)-moment IRV is an integer-valued random variable $\boldsymbol{X}$ such that the c-th absolute moment $\mathbf{E}[|\boldsymbol{X}|^c]$ lies in $[0, (k-1)^c]$.

It is clear that any k-IRV is a (c, k)-moment IRV for all c. Moreover, it is easy to show (using Markov's inequality) that for any fixed $c > 0$, any single (c, k)-moment IRV $\boldsymbol{X}$ can be learned to accuracy ϵ using $\mathrm{poly}(k/\epsilon)$ samples. However, our sample complexity bounds for learning *sums* of n independent k-IRVs provably cannot be extended to sums of n independent (c, k)-moment IRVs, in a strong sense: any learning algorithm for such sums must (information-theoretically) use at least $\mathrm{poly}(n)$ samples.

Observation 1.3. *Fix any integer $c \geq 1$. Let $\boldsymbol{S} = \boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n$ be a sum of n $(c, 2)$-moment IRVs. Let L be any algorithm which, given n and access to independent samples from $\boldsymbol{S}$, with probability at least $e^{-o(n^{1/c})}$ outputs a hypothesis distribution $\widetilde{\boldsymbol{S}}$ such that $\mathrm{d_{TV}}(\boldsymbol{S}, \widetilde{\boldsymbol{S}}) < 1/41$. Then L must use at least $n^{1/c}/10$ samples.*

The argument is a simple modification of the lower bound given by Theorem 3 of [DDS12b] and is given in Appendix A.

2 Definitions and Basic Tools

In this section we give some necessary definitions and recall some useful tools from probability.

2.1 Definitions

We begin with a formal definition of *total variation distance*, which we specialize to the case of integer-valued random variables. For two distributions $\mathbb{P}$ and $\mathbb{Q}$ supported on $\mathbb{Z}$, their total variation distance is defined to be

$$\mathrm{d_{TV}}(\mathbb{P}, \mathbb{Q}) = \sup_{A \subseteq \mathbb{Z}} |\mathbb{P}(A) - \mathbb{Q}(A)| = \frac{1}{2} \sum_{j \in \mathbb{Z}} |\mathbb{P}(\{j\}) - \mathbb{Q}(\{j\})|.$$

If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are integer random variables, their total variation distance, $\mathrm{d_{TV}}(\boldsymbol{X}, \boldsymbol{Y})$, is defined to be the total variation distance of their distributions. Throughout the paper we are casual about the distinction between a random variable and a distribution. For example, when we say "draw a sample from random variable $\boldsymbol{X}$" we formally mean "draw a sample from the distribution of $\boldsymbol{X}$", etc.

We proceed to discuss the most basic random variables that we will be interested in, namely IRVs, k-IRVs and $\pm k$-IRVs, and sums of these random variables:

Definition 2.1. An *IRV* is an integer-valued random variable. For an integer $k \geq 2$, a *k-IRV* is an IRV supported on $\{0, 1, \ldots, k-1\}$. (Note that a 2-IRV is the same as a Bernoulli random variable.) A *$\pm k$-IRV* is an IRV supported on $\{-k+1, -k+2, \ldots, k-2, k-1\}$. We say that an IRV $\boldsymbol{X}_i$ has *mode* 0 if $\mathbf{Pr}[\boldsymbol{X}_i = 0] \geq \mathbf{Pr}[\boldsymbol{X}_i = b]$ for all $b \in \mathbb{Z}$.

Definition 2.2. A *SIIRV* (Sum of Independent IRVs) is any random variable $\boldsymbol{S} = \boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n$ where the $\boldsymbol{X}_i$'s are independent IRVs. We define k-SIIRVs and $\pm k$-SIIRVs similarly; a 2-SIIRV is also called a *PBD* (Poisson Binomial Distribution). For $b \in \mathbb{Z}$ we say that the *b-weight* of the SIIRV is $\sum_{i=1}^{n} \mathbf{Pr}[\boldsymbol{X}_i = b]$. Finally, we say that a SIIRV is *0-moded* if each $\boldsymbol{X}_i$ has mode 0.

As a notational convention, we will typically use $\boldsymbol{X}$ to denote an IRV, and $\boldsymbol{S}$ to denote a SIIRV.

Discretized normal distributions will play an important role in our technical results, largely because of known theorems in probability which assert that under suitable conditions sums of independent integer random variables converge in total variation distance to discretized normal distributions (see e.g. Theorem 3.2). We now define these distributions:

Definition 2.3. Let $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}^{\geq 0}$. We let $Z(\mu, \sigma^2)$ denote the *discretized normal* distribution. The definition of $\boldsymbol{Z} \sim Z(\mu, \sigma^2)$ is that we first draw a normal $\boldsymbol{G} \sim \mathrm{N}(\mu, \sigma^2)$ and then we set $\boldsymbol{Z} = \lfloor \boldsymbol{G} \rceil$; i.e., $\boldsymbol{G}$ rounded to the nearest integer.

We note that in the "large-variance" regime that we shall be concerned with, discretized normals are known to be close in variation distance to other types of distributions such as Binomial distributions and Translated Poisson distributions (see e.g. [Rö7, RR12]); however we shall not need to work with these other distributions.

Some of our arguments will use the following notion of *shift-distance* of a random variable:

Definition 2.4. For $\boldsymbol{X}$ a random variable we define its *shift-distance* to be $\mathrm{d}_{\mathrm{shift}}(\boldsymbol{X}) = \mathrm{d}_{\mathrm{TV}}(\boldsymbol{X}, \boldsymbol{X}+1)$.

Finally, for completeness we record the following:

Definition 2.5. Given a sequence or set C of nonzero integers we define $\gcd(C)$ to be the greatest common divisor of the absolute values of the integers in C. We adopt the convention that $\gcd(\emptyset) = 0$.

2.2 Basic results from probability

Our proofs use various basic results from probability; these include bounds on total variation distance, results on normal and discretized normal distributions, bounds on shift-distance, and uniform convergence bounds. We give these results in Appendix B.

3 A useful special case: each $b \in \{1, \ldots, k-1\}$ has large weight

In this section we prove a useful special case of our our desired structural theorem for k-SIIRVs. In later sections we will use this special case to prove the general result.

Recall that for $b \in \{0, 1, \ldots, k-1\}$ the *b-weight* of a k-SIIRV $\boldsymbol{S} = \boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n$ is $\sum_{i=1}^n \mathbf{Pr}[\boldsymbol{X}_i = b]$. The special case we consider in this section is that every $b \in \{1, \ldots, k-1\}$ has large b-weight. The result we prove in this special case is that $\boldsymbol{S}$ is close to a discretized normal distribution:

Theorem 3.1. *Let $\boldsymbol{S} = \boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n$ be a 0-moded $\pm k$-SIIRV, and assume no $\boldsymbol{X}_i$ is constantly 0. For each nonzero integer c with $|c| < k$, let M_c denote the c-weight of $\boldsymbol{S}$ and let $C = \{c \in \mathbb{Z} : c \neq 0, |c| < k,$ and $M_c > 0\} \neq \emptyset$. Further assume $\gcd(C) = 1$ and $M_c \geq M$ for all $c \in C$ where $M = \omega(k \log k)$. Let $\boldsymbol{Z} \sim Z(\mu, \sigma^2)$, where $\mu = \mathbf{E}[\boldsymbol{S}]$ and $\sigma^2 = \mathbf{Var}[\boldsymbol{S}]$. Then $\mathrm{d}_{\mathrm{TV}}(\boldsymbol{S}, \boldsymbol{Z}) \leq O(k^{3.5}/\sqrt{M})$ and $\sigma^2 \geq M/8k$.*

Observe that Theorem 3.1 corresponds to Case (2) of Theorem 1.2 with $c = 1$ (so $\boldsymbol{Y}$ is a 1-IRV, i.e. the constant-0 random variable).

In Section 3.1 we record some ingredients from the probability literature and the two main tools we need for Theorem 3.1. We prove Theorem 3.1 in Section 3.2.

3.1 Ingredients from probability

We will need a bound on the distance of a SIIRV to a discrete Gaussian in terms of the shift-invariances of the "leave-one-out" partial sums of $n-1$ of the n random variables. We will use the following formulation which appears (in a slightly more quantitative form) in [CGS11, Theorem 7.4], where it is credited to Chen and Leong. It is proved by Stein's Method.

Theorem 3.2. *Let $\boldsymbol{S} = \boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n$ be a SIIRV. Write $\mu_i = \mathbf{E}[\boldsymbol{X}_i]$, $\sigma_i^2 = \mathbf{Var}[\boldsymbol{X}_i]$, $\beta_i = \mathbf{E}[|\boldsymbol{X}_i - \mu_i|^3]$, $\mu = \sum_i \mu_i$, $\sigma^2 = \sum_i \sigma_i^2$, and $\beta = \sum_i \beta_i$. Further, assume*

$$\mathrm{d}_{\mathrm{shift}}(\boldsymbol{X}_1 + \cdots + \boldsymbol{X}_{i-1} + \boldsymbol{X}_{i+1} + \cdots + \boldsymbol{X}_n) \leq \delta \quad \forall i \in [n].$$

Then for $\boldsymbol{Z} \sim Z(\mu, \sigma^2)$ we have

$$\mathrm{d}_{\mathrm{TV}}(\boldsymbol{S}, \boldsymbol{Z}) \leq O(1/\sigma) + O(\delta) + O(\beta/\sigma^3) + O(\delta\beta/\sigma^2).$$

In particular, if $\boldsymbol{S}$ is a $\pm k$-SIIRV then $\beta_i \leq k\sigma_i^2$ for each i; hence $\beta \leq k\sigma^2$ and so

$$\mathrm{d}_{\mathrm{TV}}(\boldsymbol{S}, \boldsymbol{Z}) \leq O(k)(1/\sigma + \delta).$$

The other main tool we need for Theorem 3.1 is a bound on the shift-invariance of the weighted sum of large-variance PBDs, which we will use in conjunction with Theorem 3.2 to establish closeness to a discrete Gaussian. Before proving this fact we first recall a few facts from the probability literature. The following theorem is proved via a coupling argument:

Theorem 3.3. *([MR07, Cor. 1.6]; see also [BX99, Prop. 4.6].) Let $\boldsymbol{S} = \boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n$ be any SIIRV. Then*

$$\mathrm{d}_{\mathrm{shift}}(\boldsymbol{S}) \leq \frac{\sqrt{2/\pi}}{\sqrt{\frac{1}{4} + \sum_{i=1}^{n}(1 - \mathrm{d}_{\mathrm{shift}}(\boldsymbol{X}_i))}}.$$

Corollary 3.4. *Let $\boldsymbol{S} = \boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n$ be any PBD with variance σ^2. Then $\mathrm{d}_{\mathrm{shift}}(\boldsymbol{S}) \leq O(1/\sigma)$.*

Proof. Let $\delta = \min(\mathbf{Pr}[\boldsymbol{X}_1 = 0], \mathbf{Pr}[\boldsymbol{X}_1 = 1])$. Then $\mathbf{Var}[\boldsymbol{X}_1] = \delta(1-\delta) \leq \delta = 1 - \mathrm{d}_{\mathrm{shift}}(\boldsymbol{X}_1)$. Using the analogous inequality for each $\boldsymbol{X}_i$ in Theorem 3.3 we get

$$\mathrm{d}_{\mathrm{shift}}(\boldsymbol{S}) \leq \frac{\sqrt{2/\pi}}{\sqrt{\frac{1}{4} + \sum_{i=1}^{n}\mathbf{Var}[\boldsymbol{X}_i]}} = \frac{\sqrt{2/\pi}}{\sqrt{\frac{1}{4} + \sigma^2}} = O(1/\sigma). \qquad \square$$

Theorem 3.5. *Let $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_m$ be independent IRVs each satisfying $\mathrm{d}_{\mathrm{shift}}(\boldsymbol{S}_i) \leq \epsilon$. Let $c_1, \ldots, c_m$ be a sequence of nonzero integers with* gcd 1 *and assume $\sum_i |c_i| \leq B$. Then*

$$\mathrm{d}_{\mathrm{shift}}\left(\sum_{i=1}^{m} c_i \boldsymbol{S}_i\right) \leq B\epsilon.$$

Proof. Since $\gcd(|c_1|, \ldots, |c_m|) = 1$, Bézout's identity says that there are integers $b_1, \ldots, b_m$ satisfying $\sum_i b_i |c_i| = 1$; it is also possible [Bru12] to ensure that $\sum_i |b_i| \leq \sum_i |c_i| \leq B$. Negating b_i's if necessary we can obtain $\sum_i b_i c_i = 1$. Then

$$\begin{aligned}
\mathrm{d}_{\mathrm{shift}}\left(\sum_{i=1}^{m} c_i \boldsymbol{S}_i\right) &= \mathrm{d}_{\mathrm{TV}}\left(\sum_{i=1}^{m} c_i \boldsymbol{S}_i, 1 + \sum_{i=1}^{m} c_i \boldsymbol{S}_i\right) \\
&= \mathrm{d}_{\mathrm{TV}}\left(\sum_{i=1}^{m} c_i \boldsymbol{S}_i, \sum_{i=1}^{m} c_i(\boldsymbol{S}_i + b_i)\right) \\
&\leq \sum_{i=1}^{m} \mathrm{d}_{\mathrm{TV}}(c_i \boldsymbol{S}_i, c_i(\boldsymbol{S}_i + b_i)) \\
&= \sum_{i=1}^{m} \mathrm{d}_{\mathrm{TV}}(\boldsymbol{S}_i, \boldsymbol{S}_i + b_i) \\
&\leq \sum_{i=1}^{m} |b_i| \mathrm{d}_{\mathrm{shift}}(\boldsymbol{S}_i) \leq \epsilon \sum_{i=1}^{m} |b_i|,
\end{aligned}$$

where the first inequality uses Proposition B.2 and the second uses Fact B.7. The result now follows from $\sum_i |b_i| \leq B$. □

Corollary 3.6. *Let $\emptyset \neq C \subseteq \{-k+1, -k+2, \ldots, -1, 1, \ldots, k-1\}$ satisfy $\gcd(C) = 1$. Let $\boldsymbol{J}_1, \ldots \boldsymbol{J}_n$ be independent Bernoulli random variables and fix a sequence $c_1, \ldots, c_n$ of values from C. For each $c \in C$ define $\sigma_c^2 = \sum_{i:c_i = c} \mathbf{Var}[\boldsymbol{J}_i]$. Then $\mathrm{d}_{\mathrm{shift}}(\sum_{i=1}^{n} c_i \boldsymbol{J}_i) \leq O(k^2)/\min_{c \in C}\{\sigma_c\}$.*

Proof. We may write the random variable $\sum_{i=1}^{n} c_i \boldsymbol{J}_i$ as $\sum_{c \in C} c \boldsymbol{Y}_c$, where the $\boldsymbol{Y}_c$'s are independent PBDs satisfying

$$\mathbf{Var}[\boldsymbol{Y}_i] = \sum_{i:c_i = c} \mathbf{Var}[\boldsymbol{J}_i].$$

The claim now follows from Theorem 3.5 and Corollary 3.4. □

3.2 Proof of Theorem 3.1

The high-level idea behind the proof of Theorem 3.1 is as follows. We view a draw of $\boldsymbol{S}_i$ as taking place in two stages: in the first stage, we choose for each $\boldsymbol{X}_i$ its value in $\{1, \ldots, k-1\}$ conditioned on a nonzero outcome, and in the second stage we decide whether each $\boldsymbol{X}_i$ will in fact attain a nonzero value. That is, we view each $\boldsymbol{X}_i$ as $\boldsymbol{C}_i \boldsymbol{J}_i$, where $\boldsymbol{C}_i$ is $\boldsymbol{X}_i$ conditioned on a nonzero outcome and $\boldsymbol{J}_i$ is a indicator random variable for the event that $\boldsymbol{X}_i$ attains a nonzero variable; note that $\boldsymbol{S}$ conditioned on an outcome of the first stage is simply a weighted sum of independent Bernoulli random variables. This is a useful view because we can then show that with high probability over the stage-one outcomes $\boldsymbol{S}$ is the weighted sum of large-variance PBDs, which is in turn close to a discrete Gaussian via Theorem 3.2 and Corollary 3.6.

Theorem 3.7. *Let $\boldsymbol{S} = \boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n$ be a 0-moded $\pm k$-SIIRV, and assume no $\boldsymbol{X}_i$ is constantly 0. For each nonzero integer c with $|c| < k$, let M_c denote the c-weight of $\boldsymbol{S}$ and let $C = \{c \in \mathbb{Z} : c \neq 0, |c| < k,$ and $M_c > 0\} \neq \emptyset$. Further assume* $\gcd(C) = 1$ *and $M_c \geq M$ for all $c \in C$ where $M = \omega(k \log k)$. Then* $\mathrm{d}_{\mathrm{shift}}(\boldsymbol{S}) \leq O(k^{2.5}/\sqrt{M})$.

Proof. We introduce a sequence $\boldsymbol{\mathcal{C}}$ of independent IRVs $\boldsymbol{C}_1, \ldots, \boldsymbol{C}_n$, supported on C, defined by

$$\mathbf{Pr}[\boldsymbol{C}_i = c] = \frac{\mathbf{Pr}[\boldsymbol{X}_i = c]}{\mathbf{Pr}[\boldsymbol{X}_i \neq 0]} \quad \text{for each } c \neq 0.$$

Thus $\boldsymbol{C}_i$ is $\boldsymbol{X}_i$ conditioned on a nonzero outcome, and this is well-defined since we assume that no $\boldsymbol{X}_i$ is constantly 0. Further introduce independent Bernoulli random variables $\boldsymbol{J}_1, \ldots, \boldsymbol{J}_n$, with $\mathbf{Pr}[\boldsymbol{J}_i = 1] = \mathbf{Pr}[\boldsymbol{X}_i \neq 0]$. We can now view the $\boldsymbol{X}_i$'s as being constructed via $\boldsymbol{X}_i = \boldsymbol{C}_i \boldsymbol{J}_i$. Consider now a particular outcome for $\boldsymbol{\mathcal{C}}$; say, $\boldsymbol{C}_1 = c_1, \ldots, \boldsymbol{C}_n = c_n$, which we denote as $\boldsymbol{\mathcal{C}} = \overline{c} = (c_1, \ldots, c_n)$. The conditional distribution $\boldsymbol{S} \mid \boldsymbol{\mathcal{C}} = \overline{c}$ is as $\sum_{i=1}^n c_i \boldsymbol{J}_i$. Now for each $c \in C$ define

$$\boldsymbol{\sigma}_c^2 = \sum_{i : c_i = c} \mathbf{Var}[\boldsymbol{J}_i] = \sum_{i : \boldsymbol{C}_i = c} \mathbf{Pr}[\boldsymbol{X}_i = 0]\, \mathbf{Pr}[\boldsymbol{X}_i \neq 0].$$

These quantities are random variables depending on the outcome of $\boldsymbol{\mathcal{C}}$. From Corollary 3.6 it follows that

$$\mathrm{d}_{\mathrm{shift}}(\boldsymbol{S} \mid \boldsymbol{\mathcal{C}} = \overline{c}) \leq \min\{1, O(k^2)/\min_{c \in C}\{\boldsymbol{\sigma}_c\}\}.$$

Recalling Proposition B.8, we can complete the proof by establishing that

$$\mathop{\mathbf{E}}_{\boldsymbol{\mathcal{C}}}[\min\{1, 1/\min_{c \in C}\{\boldsymbol{\sigma}_c\}\}] \leq O(\sqrt{k/M}). \tag{1}$$

Note that for each $c \in C$ the random variable $\boldsymbol{\sigma}_c^2$ is the sum of independent random variables $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n$, where $\boldsymbol{V}_i$ is $\mathbf{Pr}[\boldsymbol{X}_i = 0]\, \mathbf{Pr}[\boldsymbol{X}_i \neq 0]$ with probability $\mathbf{Pr}[\boldsymbol{X}_i = c] / \mathbf{Pr}[\boldsymbol{X}_i \neq 0]$ and is 0 otherwise. The expected value of $\boldsymbol{\sigma}_c^2$ is therefore $\sum_{i=1}^n \mathbf{Pr}[\boldsymbol{X}_i = 0]\, \mathbf{Pr}[\boldsymbol{X}_i = c] \geq M_c/2k \geq M/2k$, where we used $\mathbf{Pr}[\boldsymbol{X}_i = 0] \geq 1/2k$ since $\boldsymbol{S}$ is 0-moded. A multiplicative Chernoff bound tells us that $\mathbf{Pr}[\boldsymbol{\sigma}_c^2 < M/4k] \leq \exp(-M/16k)$. Thus except with probability at most $2k \exp(-M/16k)$ over the outcome of $\boldsymbol{\mathcal{C}}$ we have $\boldsymbol{\sigma}_c^2 \geq M/4k$ for all $c \in C$. It follows that

$$\mathbf{E}[\min\{1, 1/\min_{c \in C}\{\boldsymbol{\sigma}_c\}\}] \leq \sqrt{4k/M} + 2k \exp(-M/16k).$$

Recalling that $M = \omega(k \log k)$, this gives (1). □

Proof of Theorem 3.1. For each $i \in [n]$ we have $\mathrm{d}_{\mathrm{shift}}(\boldsymbol{X}_1 + \cdots + \boldsymbol{X}_{i-1} + \boldsymbol{X}_{i+1} + \cdots + \boldsymbol{X}_n) \leq O(k^{2.5}/\sqrt{M})$, by Theorem 3.7 — to compensate for $\boldsymbol{X}_i$ dropping out we only need to change "M" to "$M-1$", which doesn't affect the asymptotics since $M = \omega(k \log k)$. Applying Theorem 3.2 we deduce that $\mathrm{d}_{\mathrm{shift}}(\boldsymbol{S}) \leq O(k/\sigma) + O(k^{3.5}/\sqrt{M}))$. We will show the latter dominates the former by proving that $\sigma^2 = \Omega(M/2k)$. To see this, select any $c \in C$. Note that $\mathbf{Var}[\boldsymbol{X}_i]$ is at least $\min\{\mathbf{Pr}[\boldsymbol{X}_i = c](c/2)^2, \mathbf{Pr}[\boldsymbol{X}_i = 0](c/2)^2\} \geq (c^2/4)\, \mathbf{Pr}[\boldsymbol{X}_i = c]\, \mathbf{Pr}[\boldsymbol{X}_i = 0] \geq \mathbf{Pr}[\boldsymbol{X}_i = c]/8k$, where the last inequality is because $\boldsymbol{S}$ is 0-moded. Thus $\sigma^2 = \sum_i \mathbf{Var}[\boldsymbol{X}_i] \geq M_c/8k \geq M/8k$ as needed. □

4 Proof of Main Structural Result

4.1 Intuition and Preparatory Work

Throughout this section we let δ be an error parameter, and $M = M(k, 1/\delta)$ a large enough polynomial in k and $1/\delta$ to be determined later. With respect to M we make the following definition:

Definition 4.1 (Light integers and heavy integers)**.** Let $\boldsymbol{S} = \boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n$ be a 0-moded $\pm k$-SIIRV. We say that a non-zero integer $|b| < k$ is *M-light* if the b-weight of $\boldsymbol{S}$ is at most M; otherwise we call it *M-heavy*. We denote by $\mathcal{L}$ the set of M-light integers, and by $\mathcal{H}$ the set of M-heavy integers.

The bulk of the work towards proving Theorem 1.2 is showing that any 0-moded $\pm k$-SIIRV $\boldsymbol{S}$ is close to the sum of a sparse distribution (supported on some poly(k/δ) consecutive integers) and a discretized normal random variable scaled by $\gcd(\mathcal{H})$. To see why this may be true, we can distinguish the following cases:

- If $\mathcal{H} = \emptyset$ then $\boldsymbol{S}$ should be close to a sparse random variable, by Markov's inequality and $\mathbf{E}[|\boldsymbol{S}|] \leq \mathbf{E}[\sum_i |\boldsymbol{X}_i|] \leq \sum_i k \, \mathbf{Pr}[\boldsymbol{X}_i \in \mathcal{L}] = \sum_i k \sum_{j \in \mathcal{L}} \mathbf{Pr}[\boldsymbol{X}_i = j] \leq 2k^2 M$, where the last inequality holds because there are at most $2k$ integers in $\mathcal{L}$ and each of them is M-light.
- On the other hand, if $\mathcal{L} = \emptyset$, then Theorem 3.1 is readily applicable, showing that $\boldsymbol{S}$ is close to a discretized normal random variable with the same mean and variance as $\boldsymbol{S}$.
- The remaining possibility is that $\mathcal{L}$, $\mathcal{H} \neq \emptyset$. If we condition on the event $\boldsymbol{X}_i \notin \mathcal{L}$ for all i, then the conditional distribution of $\boldsymbol{S}$ is still, by Theorem 3.1, close to a discretized normal random variable, except that this discretized normal random variable is now scaled by $\gcd(\mathcal{H})$. (Indeed the conditioning only boosts the b-weight of integers $b \in \mathcal{H}$.) But "$\boldsymbol{X}_i \notin \mathcal{L}$ for all i" may be a rare event. Regardless, a typical sample from $\boldsymbol{S}$ shouldn't have a large set of indices $\boldsymbol{L} := \{i \mid \boldsymbol{X}_i \in \mathcal{L}\}$, because $\mathbf{E}[|\boldsymbol{L}|] \leq 2kM$. Indeed, one would expect that, conditioning on a typical $\boldsymbol{L}$, the b-weight of $\sum_{i \notin \boldsymbol{L}} \boldsymbol{X}_i$ for $b \in \mathcal{H}$ is still very large. Hence, conditioning on typical $\boldsymbol{L}$'s, $\sum_{i \notin \boldsymbol{L}} \boldsymbol{X}_i$ should still be close to a discretized normal (scaled by $\gcd(\mathcal{H})$). Moreover, we may hope that the normals arising by conditioning on different typical $\boldsymbol{L}$'s are close to a fixed "typical" discretized normal. Indeed, the fluctuations in the mean and variance of $\sum_{i \notin \boldsymbol{L}} \boldsymbol{X}_i$, conditioned on typical $\boldsymbol{L}$'s, should not be severe, since $\boldsymbol{L}$ is small. These considerations suggest that $\boldsymbol{S}$ is close to the sum of a sparse random variable, "the contribution of $\mathcal{L}$", and a $\gcd(\mathcal{H})$-scaled discretized normal, "the contribution of $\mathcal{H} \cup \{0\}$".

The last case is clearly the hardest and is handled by Theorem 4.3 in the next section. Before proceeding, let us make our intuition a bit more precise. First, let us formally disentangle $\boldsymbol{S}$ into the contributions of $\mathcal{L}$ and $\mathcal{H} \cup \{0\}$, by means of the following alternate sampling procedure for $\boldsymbol{S}$.

Definition 4.2. ["The Light-Heavy Experiment".] Let $\boldsymbol{S} = \boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n$ be a 0-moded $\pm k$-SIIRV. We define here an alternate experiment for sampling the random variable $\boldsymbol{S}$, called the "Light-Heavy Experiment". There are three stages:

1. [Stage 1]: We first sample a random subset $\boldsymbol{L} \subseteq [n]$, by independently including each i into $\boldsymbol{L}$ with probability $\mathbf{Pr}[\boldsymbol{X}_i \in \mathcal{L}]$.
2. [Stage 2]: Independently we sample for each $i \in [n]$ a random variable $\underline{\boldsymbol{X}}_i \in \mathcal{L}$ as follows:
$$\underline{\boldsymbol{X}}_i = b, \text{with probability } \frac{\mathbf{Pr}[\boldsymbol{X}_i = b]}{\mathbf{Pr}[\boldsymbol{X}_i \in \mathcal{L}]};$$
i.e. $\underline{\boldsymbol{X}}_i$ is distributed according to the conditional distribution of $\boldsymbol{X}_i$, conditioning on $\boldsymbol{X}_i \in \mathcal{L}$. In the exceptional case that $\mathbf{Pr}[\boldsymbol{X}_i \in \mathcal{L}] = 0$, we define $\underline{\boldsymbol{X}}_i = 0$ with probability 1.
3. [Stage 3]: Independently we sample for each $i \in [n]$ a random variable $\overline{\boldsymbol{X}}_i \in \mathcal{H} \cup \{0\}$ as follows:
$$\overline{\boldsymbol{X}}_i = b, \text{with probability } \frac{\mathbf{Pr}[\boldsymbol{X}_i = b]}{\mathbf{Pr}[\boldsymbol{X}_i \notin \mathcal{L}]};$$
i.e. $\overline{\boldsymbol{X}}_i$ is distributed according to the conditional distribution of $\boldsymbol{X}_i$, conditioning on $\boldsymbol{X}_i \notin \mathcal{L}$.

After these three stages we output $\sum_{i\in\boldsymbol{L}} \underline{\boldsymbol{X}}_i + \sum_{i\notin\boldsymbol{L}} \overline{\boldsymbol{X}}_i$ as a sample from $\boldsymbol{S}$, where $\sum_{i\in\boldsymbol{L}} \underline{\boldsymbol{X}}_i$ represents "the contribution of $\mathcal{L}$" and $\boldsymbol{S}_{\boldsymbol{L}} := \sum_{i\notin\boldsymbol{L}} \overline{\boldsymbol{X}}_i$ "the contribution of $\mathcal{H}\cup\{0\}$." We note that the two contributions are not independent, but they are independent conditioned on the outcome of $\boldsymbol{L}$. This concludes the definition of the Light-Heavy Experiment.

Coming back to our proof strategy, we aim to argue that:

(i) The contribution of $\mathcal{L}$ is close to a sparse random variable. This is clear from the definition of $\mathcal{L}$, since $\mathbf{E}[|\sum_{i\in\boldsymbol{L}} \underline{\boldsymbol{X}}_i|] \le \mathbf{E}[\sum_{i\in\boldsymbol{L}} |\underline{\boldsymbol{X}}_i|] \le k\sum_{i=1}^{n} \mathbf{Pr}[\boldsymbol{X}_i \in \mathcal{L}] = k\sum_{j\in\mathcal{L}}\sum_{i=1}^{n} \mathbf{Pr}[\boldsymbol{X}_i = j] \le 2k^2 M$.

(ii) With probability close to 1 (with respect to $\boldsymbol{L}$), the contribution of $\mathcal{H}\cup\{0\}$ is close to a fixed, $\gcd(\mathcal{H})$-scaled discretized normal random variable $\boldsymbol{Z}$, which is independent of $\boldsymbol{S}$. Showing this is the heart of our argument in the proof of Theorem 4.3, in the next section.

Given (i) and (ii), we can readily finish the proof of Theorem 4.3 using Proposition B.3: Indeed, if we set $\boldsymbol{X}$ to be the contribution of $\mathcal{L}$ and $\boldsymbol{Y}$ to be the contribution of $\mathcal{H}\cup\{0\}$, we get that $\boldsymbol{S}$ is close to the sum of $\boldsymbol{X}$ times the indicator that $\boldsymbol{L}$ is typical (which is close to a sparse random variable) and a discretized normal independent of $\boldsymbol{X}$, scaled by $\gcd(\mathcal{H})$.

4.2 The Structural Result

We make the intuition of the previous section precise, by providing the proof of the following.

Theorem 4.3. *Let $\boldsymbol{S} = \boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n$ be a 0-moded $\pm k$-SIIRV with mean μ and variance $\sigma^2 \ge 15k^4\log(1/\delta)\cdot M$, where $1 \le M = \omega(k\log k)$ and $\delta \in (0, \frac{1}{10})$ are parameters. Let also $c = \gcd(\mathcal{H})$, where $\mathcal{L}$, $\mathcal{H}$ are defined in terms of M and k as in Definition 4.1. Then there are independent random variables $\boldsymbol{Y}$ and $\boldsymbol{Z}$ such that:*

- *$\boldsymbol{Y}$ is a $\pm(kM')$-IRV, where $M' = 4k\log(1/\delta)\cdot M$;*
- *$\boldsymbol{Z} \sim Z(\frac{\mu}{c}, \frac{\sigma^2}{c^2})$;*
- *$\mathrm{d}_{\mathrm{TV}}(\boldsymbol{S}, \boldsymbol{Y} + c\boldsymbol{Z}) \le \delta + 2k\exp(-M/8) + O(k^{3.5}/\sqrt{M}) + O(k^2\log(1/\delta)M/\sigma)$.*

In particular, taking $M = k^7/\delta^2$ the total variation bound becomes $O(\delta + (k^9/\delta^2)\log(1/\delta)/\sigma)$.

Proof. Throughout we will assume that $\boldsymbol{S}$ is drawn according to the Light-Heavy Experiment from Definition 4.2. We use that definition's notation: $\boldsymbol{L}$, $\underline{\boldsymbol{X}}_i$, and $\overline{\boldsymbol{X}}_i$. For each outcome L of $\boldsymbol{L}$, we introduce the notation

$$\boldsymbol{S}_L = \sum_{i\notin L} \overline{\boldsymbol{X}}_i.$$

Note that each random variable $\boldsymbol{S}_L$ is a $\pm k$-SIIRV. Finally, for each $i \in [n]$ we introduce the shorthand $\ell_i = \mathbf{Pr}[\boldsymbol{X}_i \in \mathcal{L}]$. Note that $\sum_{i=1}^{n} \ell_i \le 2kM$ and $\ell_i < 1 - \frac{1}{2k}$ (since $\boldsymbol{X}_i$ has mode 0).

Understanding Typical $\boldsymbol{L}$'s. We study typical outcomes of $\boldsymbol{L}$. First, we argue that typical $\boldsymbol{L}$'s have small cardinality. Indeed, let us define the following event:

$$BAD_0 = \{\text{outcomes } L \text{ for } \boldsymbol{L} \text{ having } |L| \ge M'\}.$$

Since $\mathbf{E}[|\boldsymbol{L}|] = \sum_i \ell_i \le 2kM$, our choice of M' and a multiplicative Chernoff bound imply that $\mathbf{Pr}[BAD_0] \le \delta$.

Next, we argue that, conditioning on typical outcomes L for $\boldsymbol{L}$, the random variable $\boldsymbol{S}_L$ has b-weight at least $M/2$ for each $b \in \mathcal{H}$. In particular, for each $b \in \mathcal{H}$ define the event

$$BAD_b = \{\text{outcomes } L \text{ for } \boldsymbol{L} \text{ in which the } b\text{-weight of the } \pm k\text{-SIIRV } \boldsymbol{S}_L \text{ is less than } M/2\}.$$

Notice that the b-weight of $\boldsymbol{S}_{\boldsymbol{L}}$ is the sum of independent random variables $\boldsymbol{W}_i$, where $\boldsymbol{W}_i = 0$ with probability ℓ_i and $\boldsymbol{W}_i = \mathbf{Pr}[\boldsymbol{X}_i = b]/(1-\ell_i) \le 1$ with probability $1 - \ell_i$. Thus the expectation (over $\boldsymbol{L}$) of the b-weight of $\boldsymbol{S}_{\boldsymbol{L}}$ is simply the b-weight of $\boldsymbol{S}$; since this is at least M, a multiplicative Chernoff bound implies that $\mathbf{Pr}[BAD_b] \le \exp(-M/8)$. Defining BAD to be the union of all the bad events, we conclude that

$$\mathbf{Pr}[BAD] \le \delta + 2k\exp(-M/8). \tag{2}$$

Concluding the Proof. Let $\boldsymbol{Z} \sim Z(\frac{\mu}{c}, \frac{\sigma^2}{c^2})$ be independent of all other random variables, as in the statement of the theorem. The remainder of the proof will be devoted to showing that

$$\text{for every } L \notin BAD, \quad \mathrm{d}_{\mathrm{TV}}(\boldsymbol{S}_L, c\boldsymbol{Z}) \le O(k^{3.5}/\sqrt{M}) + O(k^2 \log(1/\delta) M/\sigma). \tag{3}$$

Given (3) we can conclude the proof by applying Proposition B.3, with $\sum_{i\in \boldsymbol{L}} \underline{\boldsymbol{X}}_i$ playing the role of “$\boldsymbol{X}$”, $\boldsymbol{S}_L$ playing the role of “$\boldsymbol{Y}$”, $c\boldsymbol{Z}$ playing the role of “$\boldsymbol{Z}$”, and “G” being the complement of BAD. Note that $(\sum_{i\in \boldsymbol{L}} \underline{\boldsymbol{X}}_i) \cdot \mathbf{1}_{\boldsymbol{L}\notin BAD}$ is indeed a $\pm kM'$-IRV.

Establishing (3). Notice first that if $\mathcal{H} = \emptyset$, then (3) is trivially true as then $\boldsymbol{Z} = 0$. Otherwise, $\boldsymbol{Z} \sim Z(\frac{\mu}{c}, \frac{\sigma^2}{c^2})$ and let us fix an arbitrary outcome $L \notin BAD$. Write $\mu_L = \mathbf{E}[\boldsymbol{S}_L]$, $\sigma_L^2 = \mathbf{Var}[\boldsymbol{S}_L]$, and define $\boldsymbol{S}' = \frac{1}{c}\boldsymbol{S}_L$. (Also, delete any identically zero summands from $\boldsymbol{S}'$.) By virtue of $L \notin BAD_b$ for all $b \in \mathcal{H}$ we are in a position to apply Theorem 3.1 to $\boldsymbol{S}'$ (except with $M/2$ in place of M). We deduce that for $\boldsymbol{Z}' \sim Z(\frac{\mu_L}{c}, \frac{\sigma_L^2}{c^2})$ we have

$$\mathrm{d}_{\mathrm{TV}}(\boldsymbol{S}', \boldsymbol{Z}') \le O(k^{3.5}/\sqrt{M}).$$

If we can furthermore show

$$\mathrm{d}_{\mathrm{TV}}(\boldsymbol{Z}', \boldsymbol{Z}) \le O(k^2 \log(1/\delta) M/\sigma) \tag{4}$$

then we will have established (3).

It therefore remains to show (4); i.e., to show that

$$\mathrm{d}_{\mathrm{TV}}\left(Z\left(\frac{\mu_L}{c}, \frac{\sigma_L^2}{c^2}\right), Z\left(\frac{\mu}{c}, \frac{\sigma^2}{c^2}\right)\right) \le O(k^2 \log(1/\delta) M/\sigma).$$

This is in turn implied by the claim

$$\mathrm{d}_{\mathrm{TV}}\left(\mathrm{N}(\mu_L, \sigma_L^2), \mathrm{N}(\mu, \sigma^2)\right) \le O(k^2 \log(1/\delta) M/\sigma). \tag{5}$$

To establish (5) we claim the following:

Claim 4.4. *For $L \notin BAD$,*

$$|\mu - \mu_L| \le 4k^2(\log(1/\delta) + 1) \cdot M; \tag{6}$$

$$|\sigma^2 - \sigma_L^2| \le 14k^4 \log(1/\delta) \cdot M. \tag{7}$$

Proof of Claim 4.4. (Bounding the Mean Difference.) We have:

$$\begin{aligned} |\mu - \mu_L| &= \left| \sum_{i\in L} \mathbf{E}[\boldsymbol{X}_i] + \sum_{i\notin L} (\mathbf{E}[\boldsymbol{X}_i] - \mathbf{E}[\overline{\boldsymbol{X}}_i]) \right| \\ &\le \sum_{i\in L} \mathbf{E}[|\boldsymbol{X}_i|] + \sum_{i\notin L} |\,\mathbf{E}[\boldsymbol{X}_i] - \mathbf{E}[\overline{\boldsymbol{X}}_i]|. \end{aligned}$$

Notice that $\sum_{i\in L} \mathbf{E}[|\boldsymbol{X}_i|] \le k|L| \le kM' \le 4k^2 \log(1/\delta) \cdot M$. Next we bound each difference $|\,\mathbf{E}[\boldsymbol{X}_i] - \mathbf{E}[\overline{\boldsymbol{X}}_i]|$ separately, using the law of total expectation. Namely, if $\boldsymbol{I}$ is the indicator for $\boldsymbol{X}_i \in \mathcal{L}$, we have

$$\mathbf{E}[\boldsymbol{X}_i] = \mathbf{E}[\mathbf{E}[\boldsymbol{X}_i \mid \boldsymbol{I}]] = \mathbf{Pr}[\boldsymbol{I}] E[\underline{\boldsymbol{X}}_i] + (1 - \mathbf{Pr}[\boldsymbol{I}]) E[\overline{\boldsymbol{X}}_i].$$

Hence,

$$|\,\mathbf{E}[\boldsymbol{X}_i] - \mathbf{E}[\overline{\boldsymbol{X}}_i]| = \mathbf{Pr}[\boldsymbol{I}] |E[\underline{\boldsymbol{X}}_i] - E[\overline{\boldsymbol{X}}_i]| \le \ell_i 2k$$

and consequently $\sum_{i\notin L} |\,\mathbf{E}[\boldsymbol{X}_i] - \mathbf{E}[\overline{\boldsymbol{X}}_i]| \le 2k \sum_i \ell_i \le 4k^2 M$. (6) follows.

(Bounding the Variance Difference.) We have

$$\begin{aligned} |\sigma^2-\sigma_L^2| &= \left|\sum_{i\in L}\mathbf{Var}[\boldsymbol{X}_i] + \sum_{i\notin L}(\mathbf{Var}[\boldsymbol{X}_i]-\mathbf{Var}[\overline{\boldsymbol{X}}_i])\right| \\ &\le \sum_{i\in L}\mathbf{Var}[\boldsymbol{X}_i]+\sum_{i\notin L}|\,\mathbf{Var}[\boldsymbol{X}_i]-\mathbf{Var}[\overline{\boldsymbol{X}}_i]|. \end{aligned} \tag{8}$$

The first term is at most $k^2\cdot|L|\le k^2\cdot M'=4k^3\log(1/\delta)\cdot M$. To bound the second term we bound each $|\,\mathbf{Var}[\boldsymbol{X}_i]-\mathbf{Var}[\overline{\boldsymbol{X}}_i]|$ using the law of total variance. Letting $\boldsymbol{I}$ be the indicator for $\boldsymbol{X}_i\in\mathcal{L}$ we have

$$\begin{aligned} \mathbf{Var}[\boldsymbol{X}_i] &= \mathbf{E}[\mathbf{Var}[\boldsymbol{X}_i\mid\boldsymbol{I}]]+\mathbf{Var}[\mathbf{E}[\boldsymbol{X}_i\mid\boldsymbol{I}]] \\ &= \mathbf{Pr}[\boldsymbol{I}]\,\mathbf{Var}[\underline{\boldsymbol{X}}_i]+(1-\mathbf{Pr}[\boldsymbol{I}])\,\mathbf{Var}[\overline{\boldsymbol{X}}_i]+\mathbf{Pr}[\boldsymbol{I}](1-\mathbf{Pr}[\boldsymbol{I}])(\mathbf{E}[\underline{\boldsymbol{X}}_i]-\mathbf{E}[\overline{\boldsymbol{X}}_i])^2 \end{aligned} \tag{9}$$

From this we get:

$$\mathbf{Var}[\boldsymbol{X}_i]\le\ell_i\cdot k^2+\mathbf{Var}[\overline{\boldsymbol{X}}_i]+\ell_i\cdot 4k^2 \implies \mathbf{Var}[\boldsymbol{X}_i]-\mathbf{Var}[\overline{\boldsymbol{X}}_i]\le\ell_i\cdot 5k^2.$$

For a lower bound, we get from (9):

$$\begin{aligned} (1-\mathbf{Pr}[\boldsymbol{I}])(\mathbf{Var}[\boldsymbol{X}_i]-\mathbf{Var}[\overline{\boldsymbol{X}}_i]) &= \mathbf{Pr}[\boldsymbol{I}](\mathbf{Var}[\underline{\boldsymbol{X}}_i]-\mathbf{Var}[\boldsymbol{X}_i])+\mathbf{Pr}[\boldsymbol{I}](1-\mathbf{Pr}[\boldsymbol{I}])(\mathbf{E}[\underline{\boldsymbol{X}}_i]-\mathbf{E}[\overline{\boldsymbol{X}}_i])^2 \\ &\ge -\mathbf{Pr}[\boldsymbol{I}]\,\mathbf{Var}[\boldsymbol{X}_i]. \end{aligned}$$

Hence, $\mathbf{Var}[\boldsymbol{X}_i]-\mathbf{Var}[\overline{\boldsymbol{X}}_i]\ge-\frac{\ell_i}{1-\ell_i}\mathbf{Var}[\boldsymbol{X}_i]\ge-\frac{\ell_i}{1-\ell_i}k^2\ge-2k^3\ell_i$.

Thus $|\,\mathbf{Var}[\boldsymbol{X}_i]-\mathbf{Var}[\overline{\boldsymbol{X}}_i]|\le 5k^3\ell_i$ and so the second sum in (8) is at most $5k^3\sum_i\ell_i\le 10k^4\cdot M$.

We conclude

$$|\sigma^2-\sigma_L^2|\le 14k^4\log(1/\delta)\cdot M. \qquad \square$$

Given Claim 4.4, (5) follows from (6), (7), Proposition B.4, and our assumption that $\sigma^2\ge 15k^4\log(1/\delta)\cdot M$. This completes the proof of Theorem 4.3. □

Corollary 4.5. *Let $\boldsymbol{S}$ be a k-SIIRV with mean μ and variance σ^2. Moreover, let $0<\delta<\frac{1}{10}$ and assume $\sigma^2\ge 15(k^{18}/\delta^6)\log^2(1/\delta)$. Then, for some integer c with $1\le c\le k-1$, we have that $\mathrm{d_{TV}}(\boldsymbol{S},\boldsymbol{Y}+c\boldsymbol{Z})\le O(\delta)$, where $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are independent, $\boldsymbol{Y}$ is a c-IRV and $\boldsymbol{Z}\sim Z(\frac{\mu}{c},\frac{\sigma^2}{c^2})$.*

Proof. The claim is trivial for $k=1$ so we assume that $k\ge 2$. By subtracting an appropriate integer constant from each component $\boldsymbol{X}_i$ of the k-SIIRV $\boldsymbol{S}$, we can obtain a 0-moded $\pm k$-SIIRV $\boldsymbol{S}'$ such that $\boldsymbol{S}=\boldsymbol{S}'+m$ for some $m\in\mathbb{Z}$. Note that $\boldsymbol{S}'$ has mean $\mu-m$ and variance σ^2. Now apply Theorem 4.3 to $\boldsymbol{S}'$ with $M=k^7/\delta^2$, calling the obtained random variables $\boldsymbol{Y}'$ and $\boldsymbol{Z}'$. (We leave it to the reader to verify, with the aid of Proposition 4.7 below, that the lower bound on σ means there is at least one M-heavy integer and hence the obtained c is nonzero.) $\boldsymbol{Y}'$ and $\boldsymbol{Z}'$ are independent, $\boldsymbol{Z}'\sim Z(\frac{\mu-m}{c},\frac{\sigma^2}{c^2})$, and $\boldsymbol{Y}'$ is a $\pm M''$-IRV, where $M''=4(k^9/\delta^2)\log(1/\delta)$. Moreover,

$$\mathrm{d_{TV}}(\boldsymbol{S}',\boldsymbol{Y}'+c\boldsymbol{Z}')\le O(\delta+(k^9/\delta^2)\log(1/\delta)/\sigma)\le O(\delta), \tag{10}$$

where the second inequality is by our assumed lower bound on σ.

Next, write $m=qc+r$ for some integers q,r with $|r|\le c/2\le k$. Defining $\boldsymbol{Y}''=\boldsymbol{Y}'+r$, clearly $\boldsymbol{Y}''$ is a $\pm(M''+k)$-IRV. Moreover, it follows from Proposition B.5 that $\mathrm{d_{TV}}(\boldsymbol{Z}'+q,\boldsymbol{Z})\le 1/\sigma$. So assuming $\boldsymbol{Z}$ is independent of $\boldsymbol{Y}''$, using the triangle inequality, Proposition B.2, and (10), we obtain:

$$\mathrm{d_{TV}}(\boldsymbol{S},\boldsymbol{Y}''+c\boldsymbol{Z})\le O(\delta). \tag{11}$$

Finally, define two dependent random variables $\boldsymbol{Y}=\boldsymbol{Y}(\boldsymbol{Y}'')$ and $\boldsymbol{Q}=\boldsymbol{Q}(\boldsymbol{Y}'')$ such that $\boldsymbol{Y}=\boldsymbol{Y}''\bmod c$, which is a c-IRV, and $\boldsymbol{Y}''=c\boldsymbol{Q}+\boldsymbol{Y}$, so that $\boldsymbol{Q}$ is a $\pm\lfloor\frac{M''+k}{c}\rfloor$-IRV. With this definition, we have $\boldsymbol{Y}''+c\boldsymbol{Z}=\boldsymbol{Y}+c(\boldsymbol{Z}+\boldsymbol{Q})$. The proof is concluded by noting the following:

Claim 4.6. $\mathrm{d_{TV}}(\boldsymbol{Y} + c(\boldsymbol{Z} + \boldsymbol{Q}), \boldsymbol{Y} + c\boldsymbol{Z}) \leq \frac{\lfloor \frac{M''+k}{c} \rfloor}{2\sigma}$.

Proof of Claim 4.6. First, by iterating Proposition B.6 and using the triangle inequality, we have that for any integer λ, $\mathrm{d_{TV}}(\boldsymbol{Z}, \boldsymbol{Z} + \lambda) \leq \frac{\lambda}{2\sigma}$.

For the following derivation, whenever $\boldsymbol{X}$ is a random variable, we write $f_{\boldsymbol{X}}$ for the probability density function of $\boldsymbol{X}$. We have:

$$\begin{aligned}
&\mathrm{d_{TV}}(\boldsymbol{Y} + c(\boldsymbol{Z} + \boldsymbol{Q}), \boldsymbol{Y} + c\boldsymbol{Z}) \\
&= \frac{1}{2}\int_{-\infty}^{+\infty} |f_{\boldsymbol{Y}+c(\boldsymbol{Z}+\boldsymbol{Q})}(x) - f_{\boldsymbol{Y}+c\boldsymbol{Z}}(x)| dx \\
&= \frac{1}{2}\int_{-\infty}^{+\infty} |f_{\boldsymbol{Y}+c(\boldsymbol{Z}+\boldsymbol{Q})}(x) - f_{\boldsymbol{Y}+c\boldsymbol{Z}}(x)| dx \\
&= \frac{1}{2}\int_{-\infty}^{+\infty} \left| \int_{-\infty}^{+\infty} f_{\boldsymbol{Y}(y'')+c(\boldsymbol{Z}+\boldsymbol{Q}(y''))}(x) f_{\boldsymbol{Y}''}(y'') dy'' - f_{\boldsymbol{Y}(y'')+c\boldsymbol{Z}}(x) f_{\boldsymbol{Y}''}(y'') dy'' \right| dx \\
&\leq \int_{-\infty}^{+\infty} \frac{1}{2}\int_{-\infty}^{+\infty} \left| f_{\boldsymbol{Y}(y'')+c(\boldsymbol{Z}+\boldsymbol{Q}(y''))}(x) - f_{\boldsymbol{Y}(y'')+c\boldsymbol{Z}}(x) \right| dx f_{\boldsymbol{Y}''}(y'') dy'' \\
&\leq \int_{-\infty}^{+\infty} \mathrm{d_{TV}}(\boldsymbol{Y}(y'') + c(\boldsymbol{Z} + \boldsymbol{Q}(y'')), \boldsymbol{Y}(y'') + c\boldsymbol{Z}) f_{\boldsymbol{Y}''}(y'') dy'' \\
&\leq \int_{-\infty}^{+\infty} \mathrm{d_{TV}}(\boldsymbol{Z} + \boldsymbol{Q}(y''), \boldsymbol{Z}) f_{\boldsymbol{Y}''}(y'') dy'' \\
&\leq \frac{\lfloor \frac{M''+k}{c} \rfloor}{2\sigma}. \qquad \square
\end{aligned}$$

Using the triangle inequality, Claim 4.6 and (11), we obtain: $\mathrm{d_{TV}}(\boldsymbol{S}, \boldsymbol{Y} + c\boldsymbol{Z}) \leq O(\delta) + \frac{\lfloor \frac{M''+k}{c} \rfloor}{2\sigma} = O(\delta)$, by our lower bound on σ. □

Proposition 4.7. *Let $\boldsymbol{X}$ be a $\pm k$-IRV with mode 0. Let $w = \mathbf{Pr}[\boldsymbol{X} \neq 0]$. Then $\frac{1}{8}w \leq \mathbf{Var}[\boldsymbol{X}] \leq k^2 w$.*

Proof. For the upper bound we have $\mathbf{Var}[\boldsymbol{X}] \leq \mathbf{E}[\boldsymbol{X}^2] \leq \mathbf{Pr}[\boldsymbol{X} \neq 0]k^2 = k^2 w$. As for the lower bound, let $\mu = \mathbf{E}[\boldsymbol{X}]$ and write $m = \lfloor \mu \rceil$. If $m = 0$ then whenever $\boldsymbol{X} \neq 0$ we have $|\boldsymbol{X} - \mu| \geq \frac{1}{2}$; hence $\mathbf{Var}[\boldsymbol{X}] = \mathbf{E}[(\boldsymbol{X} - \mu)^2] \geq (\frac{1}{2})^2 w \geq \frac{1}{8}w$. If $m \neq 0$ then we have $\mathbf{Pr}[\boldsymbol{X} \neq m] \geq \frac{1}{2}$ (else m would be the mode of $\boldsymbol{X}$); hence $\mathbf{E}[(\boldsymbol{X} - \mu)^2] \geq \frac{1}{2}(\frac{1}{2})^2 = \frac{1}{8} \geq \frac{1}{8}w$. □

We conclude this section with the following corollary, which is another way of stating Theorem 1.2.

Corollary 4.8. *Let $\boldsymbol{S} = \boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n$ be a k-SIIRV for some positive integer k. Let μ and σ^2 be respectively the mean and variance of $\boldsymbol{S}$. Then, for all $\epsilon > 0$, the distribution of $\boldsymbol{S}$ is $O(\epsilon)$-close in total variation distance to one of the following:*

1. *a random variable supported on $\frac{k^9}{\epsilon^4}$ consecutive integers; or*
2. *the sum of two independent random variables $\boldsymbol{S}_1 + c\boldsymbol{S}_2$, where c is some positive integer $1 \leq c \leq k-1$, $\boldsymbol{S}_2$ is distributed according to $Z(\mu, \sigma^2)$, and $\boldsymbol{S}_1$ is a c-IRV; in this case, $\sigma^2 = \Omega\left(\frac{k^{18}}{\epsilon^6}\log^2(1/\epsilon)\right)$.*

Proof. Assume $\epsilon < 1/10$. We distinguish two cases depending on whether σ^2 is $<$ or $\geq 15(k^{18}/\epsilon^6)\log^2(1/\epsilon)$. In the former case, we have by Chebyshev's inequality that $\boldsymbol{S}$ is ϵ-close to a random variable supported on $O(\frac{k^9}{\epsilon^4})$ consecutive integers, as in the first case of the statement. In the latter case, we can apply Corollary 4.5 to get that $\boldsymbol{S}$ is close to $\boldsymbol{S}_1 + c\boldsymbol{S}_2$ as in the second case of the statement. □

5 Learning Sums of Independent Integer Random Variables

In this section we apply our main structural result, Corollary 4.8, to prove our main learning result, Theorem 1.1. We do this using ideas and tools from previous work on learning discrete distributions [DDS12b, CDSS13].

The algorithm of Theorem 1.1 works by first running two different learning procedures, corresponding to the "small variance" and "large variance" cases of Corollary 4.8 respectively. It then does hypothesis testing to select a final hypothesis from the hypotheses thus obtained. We first describe the two learning procedures and analyze their performance, then describe the hypothesis testing routine and its performance, and finally put the pieces together to prove Theorem 1.1.

Before entering into the descriptions of our algorithms we briefly specify the details of the word RAM model within which they operate. As is standard, we assume that registers are of size $O(\log n)$ bits and that the basic operations of comparison, addition, subtraction, multiplication and integer division take unit time for values that fit into a single register.[3]

We remind the reader that as discussed in the footnote in Section 1.1, we may assume that $k \leq n$ since otherwise the desired learning result of $\mathrm{poly}(k, 1/\epsilon)$ samples and $\mathrm{poly}(k, 1/\epsilon)$ time is trivial. Thus we may assume that the target SIIRV $\boldsymbol{S}$ is an n^2-IRV and hence that each sample point drawn from $\boldsymbol{S}$ fits in a single $O(\log n)$-bit register.

5.1 The low-variance case.

The first procedure, `Learn-Sparse`, is useful when the variance σ^2 of $\boldsymbol{S}$ is small. We use the following result which is implicit in [DDS12b] (see Lemma 3):

Lemma 5.1. *There is a procedure* `Learn-Sparse` *with the following properties: It takes as input a size parameter L, an accuracy parameter $\epsilon' > 0$, and a confidence parameter $\delta' > 0$, as well as access to samples from a* $\mathrm{poly}(n)$*-IRV $\boldsymbol{S}$. Let a be the largest integer such that $\mathbf{Pr}[\boldsymbol{S} \leq a] \leq \epsilon'$, and let b be the smallest integer such that $\mathbf{Pr}[\boldsymbol{S} \geq b] \leq \epsilon'$.* `Learn-Sparse` *uses $O((L/\epsilon'^2) \cdot \log(1/\delta'))$ samples from $\boldsymbol{S}$, runs in time $\tilde{O}((L/\epsilon'^2) \cdot \log(1/\delta'))$ and has the following performance guarantee: If $b - a \leq L$ then with probability $1 - \delta'$* `Learn-Sparse` *outputs an explicit description of a hypothesis random variable $\boldsymbol{H}$ supported on $[a, \dots, b]$ such that $\mathrm{d_{TV}}(\boldsymbol{H}, \boldsymbol{S}) \leq O(\epsilon')$ (note that such an $\boldsymbol{H}$ is $(b - a + 1)$-flat).*

We note that the algorithm `Learn-Sparse` is quite simple: it truncates $O(\epsilon')$ of the probability mass from each end of $\boldsymbol{S}$, and then (assuming the non-truncated middle region contains at most L+1 points) it learns $\boldsymbol{S}$ by outputting the empirical distribution over this middle region (see Appendix A of [DDS12b] for details). It is straightforward to verify that the necessary operations can be performed in time that is nearly linear in the number of samples.

5.2 The high-variance case.

The second procedure, `Learn-Heavy`, is useful when the variance σ^2 of $\boldsymbol{S}$ is large.

Lemma 5.2. *There is a procedure* `Learn-Heavy` *with the following properties: It takes as input a value $c \in \{1, \dots, k-1\}$, an accuracy parameter $\epsilon' > 0$, a variance parameter $\sigma^2 = \Omega(k^2/\epsilon')$, and a confidence parameter $\delta' > 0$, as well as access to samples from a* $\mathrm{poly}(n)$*-IRV $\boldsymbol{S}$.* `Learn-Heavy` *uses $m = \tilde{O}(1/\epsilon'^4) + O((1/\epsilon'^2)(c + \log(1/\delta')))$ samples from $\boldsymbol{S}$, runs in time $\tilde{O}(m)$, and has the following performance guarantee:*

Suppose that $\mathrm{d_{TV}}(\boldsymbol{S}, c\boldsymbol{Z} + \boldsymbol{Y}) \leq \epsilon'$ where $\boldsymbol{Z}$ is a discretized normal random variable distributed as $Z(\frac{\mu'}{c}, \frac{\sigma'^2}{c^2})$ for some $\sigma'^2 \geq \sigma^2$, $\boldsymbol{Y}$ is a c-IRV, and $\boldsymbol{Z}$ and $\boldsymbol{Y}$ are independent. Then `Learn-Heavy` *outputs a hypothesis variable $\boldsymbol{H}_c$ such that $\mathrm{d_{TV}}(\boldsymbol{S}, \boldsymbol{H}_c) \leq O(\epsilon')$. (More precisely, the form of the output is a pair $\widehat{\boldsymbol{Y}}, \widehat{\boldsymbol{Z}}$ where $\widehat{\boldsymbol{Y}}$ is an (explicitly described) c-IRV and $\widehat{\boldsymbol{Z}}$ is an explicitly described $O(1/\epsilon'^2)$-flat IRV (independent from $\widehat{\boldsymbol{Y}}$); the hypothesis $\boldsymbol{H}_c$ is $c\widehat{\boldsymbol{Z}} + \widehat{\boldsymbol{Y}}$.)*

[3] In fact we only need multiplication and integer division by values that are at most k, and k can be assumed to be $o(\log \log n)$ bits as otherwise we can multiply and divide in $\mathrm{poly}(k)$-time in any model.

Proof. The procedure works in the obvious way, by learning $\boldsymbol{Y}$ and $\boldsymbol{Z}$ in separate stages.

To learn $\boldsymbol{Y}$, it draws $m_1 = O((1/\epsilon'^2)(c + \log(1/\delta')))$ samples from $\boldsymbol{S}$ and reduces each one to its residue mod c. For $0 \le i < c$ let γ_i denote the fraction of the m_1 samples that have value i mod c. The c-IRV $\widehat{\boldsymbol{Y}}$ is defined by $\mathbf{Pr}[\widehat{\boldsymbol{Y}} = i] = \gamma_i$. In other words, given samples from $\boldsymbol{S}$, it simulates samples of the random variable $\boldsymbol{Y}' = (\boldsymbol{S} \mod c)$ and outputs its *empirical distribution* $\widehat{\boldsymbol{Y}}$.

We claim that with probability at least $1-\delta'/2$, we will have $\mathrm{d_{TV}}(\widehat{\boldsymbol{Y}}, \boldsymbol{Y}) \le 2\epsilon'$. The argument is standard, but we include it here for the sake of completeness. First, since $\mathrm{d_{TV}}(\boldsymbol{S}, c\boldsymbol{Z} + \boldsymbol{Y}) \le \epsilon'$, the data processing inequality for the total variation distance (Proposition B.1) implies that $\mathrm{d_{TV}}(\boldsymbol{Y}', \boldsymbol{Y}) \le \epsilon'$. Hence, by the triangle inequality, it suffices to show that with probability at least $1 - \delta'/2$, we will have $\mathrm{d_{TV}}(\widehat{\boldsymbol{Y}}, \boldsymbol{Y}') \le \epsilon'$. Now consider the random variable $\boldsymbol{X} = \mathrm{d_{TV}}(\widehat{\boldsymbol{Y}}, \boldsymbol{Y}')$. Since $\boldsymbol{Y}'$ is a c-IRV, Theorem B.9 implies that

$$\mathbf{E}[\boldsymbol{X}] = O(\sqrt{c/m_1}) \le \epsilon'/2.$$

Moreover, Theorem B.10 for $\eta = \epsilon'/2$ implies that

$$\mathbf{Pr}[\boldsymbol{X} > \epsilon'] \le \mathbf{Pr}[\boldsymbol{X} - \mathbf{E}[\boldsymbol{X}] > \eta] \le e^{-2m\eta^2} \le \delta'/2$$

where the first inequality uses the upper bound on $\mathbf{E}[\boldsymbol{X}]$ and the last inequality follows by our choice of m_1.

To learn $\boldsymbol{Z}$, the procedure draws $m_2 = \tilde{O}(1/\epsilon'^4) + O((1/\epsilon'^2)\log(1/\delta'))$ samples from $\boldsymbol{S}$ and replaces each value v thus obtained with the value $\lfloor v/c \rfloor$. In other words, given samples from $\boldsymbol{S}$ it simulates samples of the random variable $\boldsymbol{Z}' = \lfloor \boldsymbol{S}/c \rfloor$. Since $\mathrm{d_{TV}}(\boldsymbol{S}, c\boldsymbol{Z} + \boldsymbol{Y}) \le \epsilon'$, the data processing inequality for the total variation distance implies that $\mathrm{d_{TV}}(\boldsymbol{Z}', \boldsymbol{Z}) \le \epsilon'$. We now require the following:

Claim 5.3. *$\boldsymbol{Z}$ is $O(\epsilon')$-close to a t-flat distribution $\boldsymbol{Z}''$, for $t = \tilde{O}(1/\epsilon')$.*

Proof. To show this, it is clearly sufficient to show that $\boldsymbol{Z}_\ell$ is $O(\epsilon')$-close to a t-flat distribution $\boldsymbol{Z}''$, where $\boldsymbol{Z}_\ell$ is distributed as $Z(\mu'', \frac{\sigma'^2}{c^2})$ and $\mu'' = \ell + \frac{\mu'}{c}$ is an integer translate by $\ell \in \mathbb{Z}$ of $\frac{\mu'}{c}$. We show this as follows: Let $\bar{\mu} \in \mathbb{R}^+$ be chosen such that there is a positive integer $n > 0$ and a value $0 < p < 1$ satisfying

$$\bar{\mu} = np, \quad \frac{\sigma'^2}{c^2} = np(1-p).$$

(Using the fact that $\frac{\sigma'^2}{c^2} = \Omega(1/\epsilon')$ and the observation that $p(1-p)$ may take any value in $[\frac{1}{5}, \frac{1}{4}]$ it is easy to see that there exists a value of $\bar{\mu}$ as desired.) Now we recall Theorem 7.1 of [CGS11]:

Theorem 7.1 of [CGS11] *Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ be independent $\{0,1\}$ variables with distribution $\mathbf{Pr}[\boldsymbol{X}_i = 1] = p_i$, and let $\boldsymbol{S} = \sum_{i=1}^n \boldsymbol{X}_i$, $\mu = \sum_{i=1}^n p_i$ and $\sigma^2 = \sum_{i=1}^n p_i(1-p_i)$. Then $\mathrm{d_{TV}}(\boldsymbol{S}, Z(\mu, \sigma^2)) \le O(1)/\sigma$.*

Taking each $p_i = p$, we get that $\boldsymbol{S}$ is a PBD which has $\mathrm{d_{TV}}(\boldsymbol{S}, Z(\bar{\mu}, \frac{\sigma'^2}{c^2})) = O(\epsilon')$. For a suitable integer ℓ, we have $|\bar{\mu} - \mu''| \le 1$. Proposition B.4 gives that $\mathrm{d_{TV}}(\mathrm{N}(\bar{\mu}, \frac{\sigma'^2}{c^2}), \mathrm{N}(\mu'', \frac{\sigma'^2}{c^2})) \le O(\epsilon')$, and hence the data processing inequality for total variation distance (Proposition B.1) gives that $\mathrm{d_{TV}}(Z(\bar{\mu}, \frac{\sigma'^2}{c^2}), Z(\mu'', \frac{\sigma'^2}{c^2})) \le O(\epsilon')$ as well. So the triangle inequality gives that $\mathrm{d_{TV}}(\boldsymbol{S}, Z(\mu'', \frac{\sigma'^2}{c^2})) \le O(\epsilon')$, i.e. $\mathrm{d_{TV}}(\boldsymbol{S}, \boldsymbol{Z}_\ell) \le O(\epsilon')$. Since $\boldsymbol{S}$ is a PBD it is a discrete log-concave distribution [KG71], and (as shown in [CDSS13], Theorem 4.2) any discrete log-concave distribution is ϵ'-close to a t-flat distribution for $t = \tilde{O}(1/\epsilon')$. Thus we have that $\boldsymbol{Z}_\ell$ (and hence $\boldsymbol{Z}$) is $O(\epsilon')$-close to a t-flat distribution, as was claimed. □

Given Claim 5.3, the triangle inequality implies that $\boldsymbol{Z}'$ is $O(\epsilon')$-close to a t-flat distribution. We now recall the procedure `Learn-Unknown-Decomposition` from [CDSS13] which efficiently learns distributions that are close to being t-flat:

Theorem 3.3 from [CDSS13] *Suppose that* `Learn-Unknown-Decomposition`$(\boldsymbol{D}', t, \epsilon', \delta')$ *is run on $O(t/\epsilon'^3 + \log(1/\delta')/\epsilon'^2)$ samples from a* poly(n)*-IRV $\boldsymbol{D}'$ which is ϵ'-close (in total variation distance) to some random variable $\boldsymbol{D}$ that is t-flat. Then* `Learn-Unknown-Decomposition` *runs in time $\tilde{O}(t/\epsilon'^3 + \log(1/\delta')/\epsilon'^2)$ and with probability $1 - \delta'/2$ outputs a $O(t/\epsilon')$-flat hypothesis random variable that is $O(\epsilon')$-close to $\boldsymbol{D}'$.*

By running the procedure Learn-Unknown-Decomposition using the m_2 transformed samples, we output a distribution $\widehat{\boldsymbol{Z}}$ such that $\mathrm{d_{TV}}(\widehat{\boldsymbol{Z}}, \boldsymbol{Z}') = O(\epsilon')$ and therefore $\mathrm{d_{TV}}(\widehat{\boldsymbol{Z}}, \boldsymbol{Z}) = O(\epsilon')$.

Thus with overall probability at least $1-\delta'$ we have that both $\mathrm{d_{TV}}(\widehat{\boldsymbol{Y}}, \boldsymbol{Y}) \le O(\epsilon')$ and $\mathrm{d_{TV}}(\widehat{\boldsymbol{Z}}, \boldsymbol{Z}) \le O(\epsilon')$. Since all these random variables are independent, by Proposition B.2 we consequently have $\mathrm{d_{TV}}(c\boldsymbol{Z}+\boldsymbol{Y}, c\widehat{\boldsymbol{Z}}+\widehat{\boldsymbol{Y}}) \le O(\epsilon')$. By the triangle inequality, this gives $\mathrm{d_{TV}}(\boldsymbol{S}, \boldsymbol{H}_c) \le O(\epsilon')$ as was to be shown. □

5.3 Hypothesis testing

We recall the hypothesis testing procedure from [DDS12b]. The following lemma is implicit in [DDS12b] (see Lemmas 5 and 11):

Lemma 5.4. *There is an algorithm* Hypothesis-Testing *with the following properties:* Hypothesis-Testing *is given an accuracy parameter $\epsilon' > 0$, a confidence parameter $\delta' > 0$, access to samples from an $N = \mathrm{poly}(n)$-IRV $\boldsymbol{S}$ and explicit descriptions of ℓ t-flat hypothesis IRVs $\boldsymbol{H}_1, \ldots, \boldsymbol{H}_\ell$ over $\{0, 1, \ldots, N-1\}$.* Choose-Hypothesis *draws $O(\log(\ell)\log(1/\delta')/\epsilon'^2)$ samples from $\boldsymbol{S}$ and runs in time $O((t\ell^2/\epsilon'^2)\log(\ell/\delta'))$. It has the following performance guarantee: If some $\boldsymbol{H}_i$ has $\mathrm{d_{TV}}(\boldsymbol{S}, \boldsymbol{H}_i) \le \epsilon'$ then with probability at least $1-\delta'$* Hypothesis-Testing *outputs a hypothesis $\boldsymbol{H}_{i'}$ such that $\mathrm{d_{TV}}(\boldsymbol{S}, \boldsymbol{H}_{i'}) \le 6\epsilon'$.*

For the sake of completeness, we now sketch the algorithm Hypothesis-Testing in tandem with an analysis of its running time. The basic primitive of the algorithm Hypothesis-Testing is a routine Choose-Hypothesis$(\boldsymbol{H}_1, \boldsymbol{H}_2, \epsilon, \delta)$ which, on input $\epsilon, \delta > 0$, sample access to $\boldsymbol{S}$, and explicit descriptions of two t-flat hypothesis IRVs $\boldsymbol{H}_1\boldsymbol{H}_2$, draws $O(\log(1/\delta)/\epsilon^2)$ samples from $\boldsymbol{S}$ and runs in time $O((t/\epsilon^2)\log(1/\delta))$. The routine Choose-Hypothesis$(\boldsymbol{H}_1, \boldsymbol{H}_2, \epsilon, \delta)$ has the following performance guarantee: If one of $\boldsymbol{H}_1, \boldsymbol{H}_2$ has $\mathrm{d_{TV}}(\boldsymbol{S}, \boldsymbol{H}_i) \le \epsilon$ then with probability at least $1-\delta$ it outputs an $i \in \{1, 2\}$ such that $\mathrm{d_{TV}}(\boldsymbol{S}, \boldsymbol{H}_i) \le 6\epsilon$. We call the distribution $\boldsymbol{H}_i$ the winner of the competition between $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$. The overall algorithm Hypothesis-Testing$(\{\boldsymbol{H}_i\}_{i=1}^{\ell}, \epsilon', \delta')$ proceeds by running Choose-Hypothesis$(\boldsymbol{H}_i, \boldsymbol{H}_j, \epsilon', \delta'/(2\ell))$ for all pairs $(i, j) \in [\ell]^2$, $i \neq j$, and it outputs a distribution $\boldsymbol{H}_i$ that was never a loser (or "failure" if no such distribution exists). (See Lemma 11 of [DDS12b].) The bound on the running time for Hypothesis-Testing follows from the corresponding bound for Choose-Hypothesis so it suffices to establish the latter.

The routine Choose-Hypothesis works as follows (see Lemma 5 of [DDS12b]): It starts by computing the set $\mathcal{W}_1 = \{x \mid \boldsymbol{H}_1(x) > \boldsymbol{H}_2(x)\}$ and the corresponding probabilities $p_i = \boldsymbol{H}_i(\mathcal{W}_1)$ for $i = 1, 2$. It then draws $m = O((1/\epsilon^2)\log(1/\delta))$ samples from $\boldsymbol{S}$ and calculates the fraction τ of these samples that land in $\mathcal{W}_1$. If $\tau > p_1 - \epsilon$, it returns $\boldsymbol{H}_1$; if $\tau < p_2 + \epsilon$, it returns $\boldsymbol{H}_2$; otherwise, it declares a draw (and returns either of $\boldsymbol{H}_1$, $\boldsymbol{H}_2$).

By exploiting the fact that $\boldsymbol{H}_1, \boldsymbol{H}_2$ are t-flat distributions over the domain $\{0, 1, \ldots, N-1\}$ where $N = \mathrm{poly}(n)$ we can perform these calculations very efficiently. Indeed, let $\mathcal{I}^{(i)} = \{I_1^{(i)}, \ldots, I_t^{(i)}\}$ be the t-flat decomposition for $\boldsymbol{H}_i$, $i = 1, 2$, and $p_1^{(i)}, \ldots, p_t^{(i)}$ be the corresponding probabilities. Consider the common refinement $\mathcal{I}' = \{I_1', \ldots, I_{t'}'\}$ of $\mathcal{I}^{(1)}$ and $\mathcal{I}^{(2)}$ where $t' \le 2t$ and denote by $p_1', \ldots, p_{t'}'$ the corresponding probabilities. (The common refinement is obtained by taking all possible nonempty intervals of the form $I_i^{(1)} \cap I_j^{(2)}$.) By definition, for any $y, z \in I_j'$ either $y, z \in \mathcal{W}_1$ or $y, z \notin \mathcal{W}_1$. Hence $\mathcal{W}_1$ can be succinctly described by the sets $I_j' \in \mathcal{I}'$ it contains. The aforementioned computation can be performed with $O(t)$ comparisons. We then have that $p_i = \sum_{I_j' \in \mathcal{W}_1} \boldsymbol{H}_i(I_j') = \sum_{I_j' \in \mathcal{W}_1} |I_j'| p_j'$ which can be computed in time $O(t)$ in the RAM model. It is also easy to see that the fraction τ can be computed in time $O(mt)$. Hence, the overall running time of the routine is $O((t/\epsilon^2)\log(1/\delta))$ as desired.

5.4 Proof of Theorem 1.1

Now we are ready to prove the main learning result, Theorem 1.1. The algorithm first runs Learn-Sparse with size parameter set to $L = O(k^9/\epsilon^4)$, accuracy parameter $\epsilon' = \epsilon$, and confidence parameter $\delta = \frac{1}{20k}$, to obtain a hypothesis $\boldsymbol{H}_0$. Then for each $c = 1, \ldots, k-1$ the algorithm runs Learn-Heavy with parameter c, variance parameter set to $\Omega(\frac{k^{18}}{\epsilon^6}\log^2(1/\epsilon))$, accuracy parameter $\epsilon' = \epsilon$, and confidence parameter $\delta = \frac{1}{20k}$, to obtain a hypothesis $\boldsymbol{H}_c$. Finally, it runs Hypothesis-Testing using the explicit descriptions of $\boldsymbol{H}_0, \boldsymbol{H}_1, \ldots, \boldsymbol{H}_k$, accuracy parameter $\epsilon' = \epsilon$, and confidence parameter $\delta = \frac{1}{20k}$, and outputs the hypothesis $\boldsymbol{H}_i$ that Choose-Hypothesis returns.

The claimed sample complexity and running time of the algorithm follows easily from the lemmas proved earlier in this section. The correctness of the algorithm follows from those lemmas together with Corollary 4.8. This concludes the proof of Theorem 1.1.

References

[AHK12] A. Anandkumar, D. Hsu, and S. Kakade. A method of moments for mixture models and Hidden Markov Models. *Journal of Machine Learning Research - Proceedings Track*, 23:33.1–33.34, 2012. 1

[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

[BGK04] T. Batu, S. Guha, and S. Kannan. Inferring mixtures of Markov chains. In *COLT*, pages 186–199, 2004. 1

[BHJ92] A.D. Barbour, L. Holst, and S. Janson. *Poisson Approximation*. Oxford University Press, New York, NY, 1992. 1.2

[Bru12] F. Brunault (mathoverflow.net/users/6506). Estimates for Bézout coefficients. MathOverflow, October 2012. http://mathoverflow.net/questions/108723 (version: 2012-10-05). 3.1

[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. 1.2, 3.3

[CDSS13] S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Learning mixtures of structured distributions over discrete domains. In *SODA*, 2013. 1.1, 1.2, 5, 5.2, 5.2

[CGG02] M. Cryan, L. Goldberg, and P. Goldberg. Evolutionary trees can be learned in polynomial time in the two state general Markov model. *SIAM Journal on Computing*, 31(2):375–397, 2002. 1

[CGS11] L. Chen, L. Goldstein, and Q.-M. Shao. *Normal Approximation by Stein's Method*. Springer, 2011. 1.2, 1.2, 3.1, 5.2, 5.2

[CL10] L. H. Y. Chen and Y. K. Leong. From zero-bias to discretized normal approximation. 2010. 1.2

[Das99] S. Dasgupta. Learning mixtures of Gaussians. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science*, pages 634–644, 1999. 1

[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 *STOC*, pages 709–728, 2012. (document), 1, 1.1, 1.2, 1.3, 1.3, 5, 5.1, 5.1, 5.3, 5.3

[DG85] L. Devroye and L. Györfi. *Nonparametric Density Estimation: The L_1 View*. John Wiley & Sons, 1985. 1

[DHKS05] A. Dasgupta, J. Hopcroft, J. Kleinberg, and M. Sandler. On learning mixtures of heavy-tailed distributions. In *FOCS*, pages 491–500, 2005. 1

[DL01] L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. Springer Series in Statistics, Springer, 2001. 1, B.9, B.10

[DP11] C. Daskalakis and C. Papadimitriou. On Oblivious PTAS's for Nash Equilibrium. *STOC* 2009, pp. 75–84. Full version available as ArXiV report, 2011. 1.2

[DS00] S. Dasgupta and L. Schulman. A two-round variant of EM for Gaussian mixtures. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 143–151, 2000. 1

[Fan12] X. Fang. Discretized normal approximation by Stein's method. arXiv:1111.3162v3, 2012. 1.2

[FM99] Y. Freund and Y. Mansour. Estimating a mixture of two product distributions. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, pages 183–192, 1999. 1

[FOS05] J. Feldman, R. O'Donnell, and R. Servedio. Learning mixtures of product distributions over discrete domains. In *Proc. 46th Symposium on Foundations of Computer Science (FOCS)*, pages 501–510, 2005. 1

[FOS06] J. Feldman, R. O'Donnell, and R. Servedio. PAC learning mixtures of Gaussians with no separation assumption. In *Proc. 19th Annual Conference on Learning Theory (COLT)*, pages 20–34, 2006. 1

[GMRZ11] P. Gopalan, R. Meka, O. Reingold, and D. Zuckerman. Pseudorandom generators for combinatorial shapes. In *STOC*, pages 253–262, 2011. 1.2

[KG71] J. Keilson and H. Gerber. Some results for discrete unimodality. *J. American Statistical Association*, 66(334):386–389, 1971. 5.2

[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

[Kru86] J. Kruopis. Precision of approximation of the generalized binomial distribution by convolutions of poisson measures. *Lithuanian Mathematical Journal*, 26(1):37–49, 1986. 1.2

[MR05] E. Mossel and S. Roch. Learning nonsingular phylogenies and Hidden Markov Models. In *To appear in Proceedings of the 37th Annual Symposium on Theory of Computing (STOC)*, 2005. 1

[MR07] L. Mattner and B. Roos. A shorter proof of Kanter's Bessel function concentration bound. *Probability Theory and Related Fields*, 139(1-2):191–205, 2007. 3.3

[MV10] A. Moitra and G. Valiant. Settling the polynomial learnability of mixtures of Gaussians. In *FOCS*, pages 93–102, 2010. 1

[Pre83] E. L. Presman. Approximation of binomial distributions by infinitely divisible ones. *Theory Probab. Appl.*, 28:393–403, 1983. 1.2

[Rö07] A. Röllin. Translated Poisson Approximation Using Exchangeable Pair Couplings. *Annals of Applied Probability*, 17(5/6):1596–1614, 2007. 2.1

[Rey11] L. Reyzin. Extractors and the leftover hash lemma. http://www.cs.bu.edu/ reyzin/teaching/s11cs937/notes-leo-1.pdf, March 2011. B

[RR12] A. Röllin and N. Ross. Local limit theorems via Landau-Kolmogorov inequalities. arXiv:1011.3100v2, 2012. 2.1

[RSS12] Y. Rabani, L. Schulman, and C. Swamy. Learning mixtures of arbitrary distributions over large discrete domains. at http://arxiv.org/abs/1212.1527, 2012. 1

[Sco92] D.W. Scott. *Multivariate Density Estimation: Theory, Practice and Visualization*. Wiley, New York, 1992. 1

[Sil86] B. W. Silverman. *Density Estimation*. Chapman and Hall, London, 1986. 1

[Val84] L. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984. 1

[VV11] G. Valiant and P. Valiant. Estimating the unseen: an $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs. In *STOC*, pages 685–694, 2011. 1.2

[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

[Wik] Wikipedia. Kullback-Leibler Divergence between Normal Distributions. http://en.wikipedia.org/wiki/Multivariate_normal_distribution#Kullback.E2.80.93Leibler_divergence. B

A Proof of Observation 1.3

Recall Observation 1.3:

Observation 1.3. *Fix any integer $c \geq 1$. Let $\boldsymbol{S} = \boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n$ be a sum of n $(c, 2)$-moment IRVs. Let L be any algorithm which, given n and access to independent samples from $\boldsymbol{S}$, with probability at least $e^{-o(n^{1/c})}$ outputs a hypothesis random variable $\tilde{\boldsymbol{S}}$ such that $\mathrm{d_{TV}}(\boldsymbol{S}, \tilde{\boldsymbol{S}}) < 1/41$. Then L must use at least $n^{1/c}/10$ samples.*

Proof. We assume w.l.o.g. in the argument below that n is the c-th power of some integer and a multiple of 8. Let us define a probability distribution over possible target random variables $\boldsymbol{S} = \boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n$ as follows. First, a sequence of values $v_1, \ldots, v_{n^{1/c}/4}$ is chosen as follows: for each i, the value v_i is chosen independently and uniformly from $\{n^{1/c}/2 + 2i - 1, n^{1/c}/2 + 2i\}$. Given the outcome of $v_1, \ldots, v_{n^{1/c}/4}$, the $(c, 1)$-moment IRVs $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n/4}$ are defined as follows: for $1 \leq i \leq n^{1/c}/4$ and $1 \leq k \leq n^{1-1/c}$, the random variable $\boldsymbol{X}_{n^{1-1/c}(i-1)+k}$ takes value v_i with probability $1/n$ and takes value 0 with probability $1 - 1/n$ (so observe that the $n^{1-1/c}$ random variables $\boldsymbol{X}_{n^{1-1/c}(i-1)+1}, \ldots, \boldsymbol{X}_{n^{1-1/c} \cdot i}$ are identically distributed). For $n/4 < j \leq n$ the random variable $\boldsymbol{X}_j$ is identically 0.

We have the following easy lemmas:

Lemma A.1. *Fix any sequence of values $v_1, \ldots, v_{n^{1/c}/4}$ and corresponding sequence of $(c, 1)$-moment IRVs $\boldsymbol{X}_i$ as described above. For any value $r \in \{n^{1/c}/2 + 1, \ldots, n^{1/c}\}$ we have that $\mathbf{Pr}[\boldsymbol{S} = r] > 0$ if and only if $v_{\lceil (r - n^{1/c}/2)/2 \rceil} = r$. For each of the $n^{1/c}/4$ values of $r \in \{n^{1/c}/2 + 1, \ldots, n^{1/c}\}$ such that $\mathbf{Pr}[\boldsymbol{S} = r] > 0$, the value $\mathbf{Pr}[\boldsymbol{S} = r]$ is exactly $\frac{1}{n^{1/c}}(1 - \frac{1}{n})^{n/4-1} > \frac{1}{2n^{1/c}}$.*

The first claim of the lemma holds because any set of at least two numbers from $\{n^{1/c}/2 + 1, \ldots, n^{1/c}\}$ must sum to more than $n^{1/c}$. Hence the only way that $\boldsymbol{S}$ can equal r is if exactly one of the $n^{1-1/c}$ random variables $\boldsymbol{X}_{n^{1-1/c}\left(\lceil (r - n^{1/c}/2)/2 \rceil - 1\right)+1}, \ldots, \boldsymbol{X}_{n^{1-1/c}\left(\lceil (r - n^{1/c}/2)/2 \rceil\right)}$ takes value r and all other variables are 0. Since variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n/4}$ are non-zero with probability $1/n$ and the rest are identically 0, the probability of this is $\frac{1}{n^{1/c}}(1 - \frac{1}{n})^{n/4-1} > \frac{1}{2n^{1/c}}$.

The next lemma is an easy consquence of Chernoff bounds:

Lemma A.2. *Fix any sequence of $(c, 1)$-moment IRVs $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n/4}$ as described above. Consider a sequence of $n^{1/c}/10$ independent draws of $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n/4})$. With probability $1 - e^{-\Omega(n^{1/c})}$, the total number of indices $i \in \{1, \ldots, n/4\}$ such that $\boldsymbol{X}_i$ is ever nonzero in any of the $n^{1/c}/10$ draws is at most $n^{1/c}/20$.*

We are now ready to prove Observation 1.3. Let L be a learning algorithm that receives $n^{1/c}/10$ samples. Let $\boldsymbol{S}$ be drawn from the distribution over possible target random variables described above.

We consider an augmented learner L' that is given "extra information:" for each point in the sample instead of receiving just the value of $\boldsymbol{S} = \boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n$ the augmented learner receives the entire vector of outcomes of $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$. By Lemmas A.1 and A.2, with probability at least $1 - e^{-\Omega(n^{1/c})}$, the augmented learner receives (complete) information about the outcome of at most $n^{1/c}/20$ of the $n^{1/c}/4$ values $v_1, \ldots, v_{n^{1/c}/4}$ (we condition on this going forward). Since the choices of the v_i's are independent, for at least $n^{1/c}/5$ of the values v_i, the learning algorithm has no information about whether v_i is $n^{1/c}/2 + 2i - 1$

or $n^{1/c}/2 + 2i$, and hence it is equally likely that $\mathbf{Pr}[\boldsymbol{S} = r] > 0$ for r being either of these two values. Lemma A.1 implies that each of these $n^{1/c}/5$ values contributes (at least) $1/(4n^{1/c})$ of error in expectation (with respect to the randomness in the learning algorithm and the choice of v_i's) in the hypothesis output by the learning algorithm. Hence the expected L1 error of the hypothesis output is at least $1/20$. The proof of Observation 1.3 is now concluded with a Chernoff bound. □

B Tools from Probability

We begin by recalling some basic facts concerning total variation distance, starting with the "data processing inequality for total variation distance" (see part (iv) of Lemma 2 of [Rey11] for the proof):

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

$$\mathrm{d_{TV}}(F(\boldsymbol{X}), F(\boldsymbol{X}')) \le \mathrm{d_{TV}}(\boldsymbol{X}, \boldsymbol{X}').$$

Next we recall the subadditivity of total variation distance for independent random variables:

Proposition B.2. *Let $\boldsymbol{A}, \boldsymbol{A}', \boldsymbol{B}, \boldsymbol{B}'$ be integer random variables such that $(\boldsymbol{A}, \boldsymbol{A}')$ is independent of $(\boldsymbol{B}, \boldsymbol{B}')$. Then $\mathrm{d_{TV}}(\boldsymbol{A} + \boldsymbol{B}, \boldsymbol{A}' + \boldsymbol{B}') \le \mathrm{d_{TV}}(\boldsymbol{A}, \boldsymbol{A}') + \mathrm{d_{TV}}(\boldsymbol{B}, \boldsymbol{B}')$.*

Proof. By definition there is a coupling of $\boldsymbol{A}, \boldsymbol{A}'$ such that $\boldsymbol{A} = \boldsymbol{A}'$ except with probability at most $\mathrm{d_{TV}}(\boldsymbol{A}, \boldsymbol{A}')$, and similarly for $\boldsymbol{B}, \boldsymbol{B}'$. Taking these couplings independently we get that $\boldsymbol{A} + \boldsymbol{B} = \boldsymbol{A}' + \boldsymbol{B}'$ except with probability at most $\mathrm{d_{TV}}(\boldsymbol{A}, \boldsymbol{A}') + \mathrm{d_{TV}}(\boldsymbol{B}, \boldsymbol{B}')$, by the union bound. □

Proposition B.3. *Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be integer random variables which are independent conditioned on the outcome of a third discrete random variable $\boldsymbol{L}$. Further, let $\boldsymbol{Z}$ be an integer random variable independent of $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{L}$. Finally, let G be a set of "good" outcomes for $\boldsymbol{L}$ such that:*

- $\mathbf{Pr}[\boldsymbol{L} \notin G] \le \eta$*;*
- *for each $L \in G$ we have $\mathrm{d_{TV}}((\boldsymbol{Y} \mid \boldsymbol{L} = L), \boldsymbol{Z}) \le \epsilon$.*

Then $\mathrm{d_{TV}}(\boldsymbol{X} + \boldsymbol{Y}, \boldsymbol{X} \cdot \mathbf{1}_{\boldsymbol{L} \in G} + \boldsymbol{Z}) \le (1 - \eta)\epsilon + \eta \le \epsilon + \eta$.

Proof. Let $A \subseteq \mathbb{Z}$. Fix any good outcome $L \in G$ for $\boldsymbol{L}$ and any outcome x for $\boldsymbol{X}$. Since $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent conditioned on $\boldsymbol{L} = L$, and since $\boldsymbol{Z}$ is independent of $\boldsymbol{X}$ and $\boldsymbol{L}$, we have

$$|\mathbf{Pr}[\boldsymbol{Y} \in A - x \mid \boldsymbol{L} = L, \boldsymbol{X} = x] - \mathbf{Pr}[\boldsymbol{Z} \in A - x \mid \boldsymbol{L} = L, \boldsymbol{X} = x]| \le \mathrm{d_{TV}}((\boldsymbol{Y} \mid \boldsymbol{L} = L), \boldsymbol{Z}) \le \epsilon.$$

We may also replace $\mathbf{Pr}[\boldsymbol{Z} \in A - x \mid \boldsymbol{L} = L, \boldsymbol{X} = x]$ above with $\mathbf{Pr}[\boldsymbol{Z} \in A - x \cdot \mathbf{1}_{\boldsymbol{L} \in G} \mid \boldsymbol{L} = L, \boldsymbol{X} = x]$. Then multiplying the inequality by $\mathbf{Pr}[\boldsymbol{X} = x]$ and summing over all x yields

$$|\mathbf{Pr}[\boldsymbol{X} + \boldsymbol{Y} \in A \mid \boldsymbol{L} = L] - \mathbf{Pr}[\boldsymbol{X} \cdot \mathbf{1}_{\boldsymbol{L} \in G} + \boldsymbol{Z} \in A \mid \boldsymbol{L} = L]| \le \epsilon.$$

Multiplying the above by $\mathbf{Pr}[\boldsymbol{L} = L]$ and summing over all good $L \in G$ yields

$$|\mathbf{Pr}[\boldsymbol{X} + \boldsymbol{Y} \in A \wedge \boldsymbol{L} \in G] - \mathbf{Pr}[\boldsymbol{X} \cdot \mathbf{1}_{\boldsymbol{L} \in G} + \boldsymbol{Z} \in A \wedge \boldsymbol{L} \in G]| \le (1 - \eta)\epsilon.$$

Thus by a union bound, $|\mathbf{Pr}[\boldsymbol{X} + \boldsymbol{Y} \in A] - \mathbf{Pr}[\boldsymbol{X} \cdot \mathbf{1}_{\boldsymbol{L} \in G} + \boldsymbol{Z} \in A]| \le (1 - \eta)\epsilon + \eta$, completing the proof. □

We will use the following standard result which bounds the variation distance between two normal distributions in terms of their means and variances:

Proposition B.4. *Let $\mu_1, \mu_2 \in \mathbb{R}$ and $0 < \sigma_1 \le \sigma_2$. Then $\mathrm{d_{TV}}(\mathrm{N}(\mu_1, \sigma_1^2), \mathrm{N}(\mu_2, \sigma_2^2)) \le \frac{1}{2}\left(\frac{|\mu_1 - \mu_2|}{\sigma_1} + \frac{\sigma_2^2 - \sigma_1^2}{\sigma_1^2}\right)$.*

Proof. From [Wik], the Kullback-Leibler divergence from $\mathrm{N}(\mu_2, \sigma_2^2)$ to $\mathrm{N}(\mu_1, \sigma_1^2)$ is

$$\frac{1}{2}\left(\frac{\sigma_2^2}{\sigma_1^2} + \frac{(\mu_1-\mu_2)^2}{\sigma_1^2} - \ln\frac{\sigma_2^2}{\sigma_1^2} - 1\right) \le \frac{1}{2}\left(\left(\frac{\sigma_2^2-\sigma_1^2}{\sigma_1^2}\right)^2 + \frac{(\mu_1-\mu_2)^2}{\sigma_1^2}\right) \le \frac{1}{2}\left(\frac{\sigma_2^2-\sigma_1^2}{\sigma_1^2} + \frac{|\mu_1-\mu_2|}{\sigma_1}\right)^2,$$

where for the first inequality we used that $x - \ln x - 1 \le (1-x)^2$, for $x \ge 1$. The proof is concluded by applying Pinsker's inequality. □

We will also require:

Proposition B.5. *Let $\boldsymbol{G} \sim \mathrm{N}(\mu, \sigma^2)$ and $\lambda \in \mathbb{R}$. Then $\mathrm{d_{TV}}(\lfloor \boldsymbol{G} + \lambda\rceil, \lfloor \boldsymbol{G}\rceil + \lfloor\lambda\rceil) \le \frac{1}{2\sigma}$. Consequently, for $\lambda \in \mathbb{Z}$, if $\boldsymbol{Z} \sim Z(\mu, \sigma^2)$ and $\boldsymbol{Z}' \sim Z(\mu - \lambda, \sigma^2)$ then $\mathrm{d_{TV}}(\boldsymbol{Z}, \lambda + \boldsymbol{Z}') \le \frac{1}{2\sigma}$.*

Proof of Proposition B.5: For all $i \in \mathbb{Z}$ and being lax about measure zero events, we have:

- $\lfloor \boldsymbol{G} + \lambda\rceil = i \iff \boldsymbol{G} \in [i - \lambda - 0.5, i - \lambda + 0.5]$;
- $\lfloor \boldsymbol{G}\rceil + \lfloor\lambda\rceil = i \iff \boldsymbol{G} \in [i - \lfloor\lambda\rceil - 0.5, i - \lfloor\lambda\rceil + 0.5]$.

Hence, if f represents the probability density function of $\boldsymbol{G}$ we have:

$$\begin{aligned}
\mathrm{d_{TV}}(\lfloor \boldsymbol{G}+\lambda\rceil, \lfloor \boldsymbol{G}\rceil + \lfloor\lambda\rceil) &= \frac{1}{2}\sum_{i=-\infty}^{\infty}\left|\mathbf{Pr}[\lfloor \boldsymbol{G}+\lambda\rceil = i] - \mathbf{Pr}[\lfloor \boldsymbol{G}\rceil + \lfloor\lambda\rceil = i]\right| \\
&= \frac{1}{2}\sum_{i=-\infty}^{\infty}\left|\int_{i-\lambda-0.5}^{i-\lambda+0.5} f(x)dx - \int_{i-\lfloor\lambda\rceil-0.5}^{i-\lfloor\lambda\rceil+0.5} f(x)dx\right| \\
&= \frac{1}{2}\sum_{i=-\infty}^{\infty}\left|\int_{i-0.5}^{i+0.5} f(x-\lambda)dx - \int_{i-0.5}^{i+0.5} f(x-\lfloor\lambda\rceil)dx\right| \\
&= \frac{1}{2}\sum_{i=-\infty}^{\infty}\left|\int_{i-0.5}^{i+0.5} (f(x-\lambda) - f(x-\lfloor\lambda\rceil))dx\right| \\
&\le \frac{1}{2}\sum_{i=-\infty}^{\infty}\int_{i-0.5}^{i+0.5} |f(x-\lambda) - f(x-\lfloor\lambda\rceil)|dx \\
&= \frac{1}{2}\int_{-\infty}^{+\infty} |f(x-\lambda) - f(x-\lfloor\lambda\rceil)|dx \\
&= \mathrm{d_{TV}}(G+\lambda, G+\lfloor\lambda\rceil) \le \frac{|\lambda - \lfloor\lambda\rceil|}{\sigma} \le \frac{1}{2\sigma},
\end{aligned}$$

where in the second to last inequality of the above derivation we used Proposition B.4, and for the last equality we used that $f(x-\lambda)$ is the probability density function of $G+\lambda$, while $f(x-\lfloor\lambda\rceil)$ is the probability density function of $G + \lfloor\lambda\rceil$. □

Proposition B.6. *Let $\boldsymbol{G} \sim \mathrm{N}(\mu, \sigma^2)$. Then $\mathrm{d_{shift}}(\boldsymbol{G}) \le \frac{1}{2\sigma}$. Consequently, if $\boldsymbol{Z} \sim Z(\mu, \sigma^2)$ then $\mathrm{d_{shift}}(\boldsymbol{Z}) \le \frac{1}{2\sigma}$.*

Proof of Proposition B.6: Notice that $\boldsymbol{G}$ and $\boldsymbol{G}+1$ have the same variance and means differing by 1. The proof follows immediately by an application of Proposition B.4. □

The following is a simple consequence of the definition of shift-distance:

Fact B.7. *Let $\boldsymbol{X}$ an IRV and let $c \in \mathbb{Z}$. Then $\mathrm{d_{TV}}(\boldsymbol{X}, \boldsymbol{X}+c) \le |c| \cdot \mathrm{d_{shift}}(\boldsymbol{X})$.*

Proposition B.8. *Let $\boldsymbol{S}$ be an integer random variable and let $\boldsymbol{C}$ be a discrete random variable. Then $\mathrm{d_{shift}}(\boldsymbol{S}) \le \mathbf{E}_{\boldsymbol{C}}[\mathrm{d_{shift}}(\boldsymbol{S} \mid \boldsymbol{C})]$.*

Proof. Let $A \subseteq \mathbb{Z}$. Then

$$|\mathbf{Pr}[\boldsymbol{S} \in A] - \mathbf{Pr}[\boldsymbol{S}+1 \in A]| = \left|\mathop{\mathbf{E}}_{\boldsymbol{C}}[\mathbf{Pr}[\boldsymbol{S} \in A \mid \boldsymbol{C}]] - \mathop{\mathbf{E}}_{\boldsymbol{C}}[\mathbf{Pr}[\boldsymbol{S}+1 \in A \mid \boldsymbol{C}]]\right|$$

$$= \left|\mathop{\mathbf{E}}_{\boldsymbol{C}}[\mathbf{Pr}[\boldsymbol{S} \in A \mid \boldsymbol{C}] - \mathbf{Pr}[\boldsymbol{S}+1 \in A \mid \boldsymbol{C}]]\right| \le \mathop{\mathbf{E}}_{\boldsymbol{C}}[|\mathbf{Pr}[\boldsymbol{S} \in A \mid \boldsymbol{C}] - \mathbf{Pr}[\boldsymbol{S}+1 \in A \mid \boldsymbol{C}]|] = \mathop{\mathbf{E}}_{\boldsymbol{C}}[\mathrm{d}_{\mathrm{shift}}(\boldsymbol{S} \mid \boldsymbol{C})],$$

where the last step used the triangle inequality. □

We will also required the following classical inequalities:

The VC inequality. Given a family of subsets $\mathcal{A}$ over $[n]$, define $\|p\|_{\mathcal{A}} = \sup_{A \in \mathcal{A}} |p(A)|$. The *VC–dimension* of $\mathcal{A}$ is the maximum size of a subset $X \subseteq [n]$ that is shattered by $\mathcal{A}$ (a set X is shattered by $\mathcal{A}$ if for every $Y \subseteq X$ some $A \in \mathcal{A}$ satisfies $A \cap X = Y$).

Theorem B.9 (VC inequality, [DL01, p.31])**.** *Let $\widehat{p}_m$ be an empirical distribution of m samples from p. Let $\mathcal{A}$ be a family of subsets of VC–dimension d. Then*

$$\mathbf{E}\left[\|p - \widehat{p}_m\|_{\mathcal{A}}\right] \le O(\sqrt{d/m}).$$

Uniform convergence. We will also use the following uniform convergence bound:

Theorem B.10 ([DL01, p17])**.** *Let $\mathcal{A}$ be a family of subsets over $[n]$, and $\widehat{p}_m$ be an empirical distribution of m samples from p. Let X be the random variable $\|p - \hat{p}\|_{\mathcal{A}}$. Then we have*

$$\mathbf{Pr}\left[X - \mathbf{E}[X] > \eta\right] \le e^{-2m\eta^2}.$$