Learning Poisson Binomial Distributions

Constantinos Daskalakis*
MIT
costis@csail.mit.edu

Ilias Diakonikolas†
UC Berkeley
ilias@cs.berkeley.edu

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

July 18, 2011

Abstract

We consider a basic problem in unsupervised learning: learning an unknown *Poisson Binomial Distribution* over $\{0, 1, \ldots, n\}$. A Poisson Binomial Distribution (PBD) is a sum $X = X_1 + \cdots + X_n$ of n independent Bernoulli random variables which may have arbitrary expectations. We work in a framework where the learner is given access to independent draws from the distribution and must (with high probability) output a hypothesis distribution which has total variation distance at most ϵ from the unknown target PBD.

As our main result we give a highly efficient algorithm which learns to ϵ-accuracy using $\tilde{O}(1/\epsilon^3)$ samples independent of n. The running time of the algorithm is *quasilinear* in the size of its input data, i.e. $\tilde{O}(\log(n)/\epsilon^3)$ bit-operations (observe that each draw from the distribution is a $\log(n)$-bit string). This is nearly optimal since any algorithm must use $\Omega(1/\epsilon^2)$ samples. We also give positive and negative results for some extensions of this learning problem.

1 Introduction

We begin by considering a somewhat fanciful scenario: You are the manager of an independent weekly newspaper in a city of n people. Each week the i-th inhabitant of the city independently picks up a copy of your paper with probability p_i. Of course you do not know the values $p_1, \ldots, p_n$; each week you only see the total number of papers that have been picked up. For many reasons (advertising, production, revenue analysis, etc.) you would like to have a detailed "snapshot" of the probability distribution (pdf) describing how many readers you have each week. *Is there an efficient algorithm to construct a high-accuracy approximation of the pdf from a number of observations that is* independent *of the population n?* We show that the answer is "yes."

A *Poisson Binomial Distribution* (henceforth PBD) over the domain $[n] = \{0, 1, \ldots, n\}$ is the familiar distribution of a sum $X = \sum_{i=1}^{n} X_i$, where $X_1, \ldots, X_n$ are independent Bernoulli (0/1) random variables with $\mathbf{E}[X_i] = p_i$. The p_i's do not need to be all the same, and thus PBDs generalize the Binomial distribution $B(n, p)$ and, indeed, comprise a much richer class of distributions. (See Section 1.2.)

As PBDs are one of the most basic classes of discrete distributions they have been intensely-studied in probability and statistics (see Section 1.2); we note here that tail bounds on PBDs form an important special case of Chernoff/Hoeffding bounds [Che52, Hoe63, DP09]. In application domains, PBDs have many uses in research areas such as survey sampling, case-control studies, and survival analysis, see e.g. [CL97] for a survey of the many uses of these distributions in applications. It is thus natural to study the problem of learning/estimating an unknown PBD given access to independent samples drawn from the distribution; this is the problem we consider, and essentially settle in this paper.

We work in a natural PAC-style model of learning an unknown discrete probability distribution which is essentially the model of [KMR$^+$94]. In this learning framework for our problem, the learner is provided with independent samples drawn from an unknown PBD X. Using these samples, the learner must with probability $1 - \delta$ output a hypothesis distribution $\hat{X}$ such that the total variation distance $d_{\mathrm{TV}}(X, \hat{X})$ is at most ϵ, where $\epsilon, \delta > 0$ are accuracy and confidence parameters that are provided to the learner.[1] A *proper* learning algorithm in this framework outputs a distribution that is itself a Poisson Binomial Distribution, i.e. a vector $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_n)$ which describes the hypothesis PBD $\hat{X} = \sum_{i=1}^{n} \hat{X}_i$ where $\mathbf{E}[\hat{X}_i] = \hat{p}_i$.

1.1 Our results

Our main result is a highly efficient algorithm for learning PBDs from *constantly* many samples, i.e. quite surprisingly, the *sample complexity of learning PBDs over $[n]$ is independent of n*. We prove the following:

Theorem 1 (Main Theorem) *Let $X = \sum_{i=1}^{n} X_i$ be an unknown PBD.*

1. **[Learning PBDs from constantly many samples]** *There is an algorithm with the following properties: given n and access to independent draws from X, the algorithm uses $\tilde{O}(1/\epsilon^3) \cdot \log(1/\delta)$ samples from X, performs $\tilde{O}(\frac{1}{\epsilon^3} \log n \log \frac{1}{\delta})$ bit operations, [2] and with probability $1 - \delta$ outputs a (succinct description of a) distribution $\hat{X}$ over $[n]$ which is such that $d_{\mathrm{TV}}(\hat{X}, X) \leq \epsilon$.*

2. **[Properly learning PBDs from constantly many samples]** *There is an algorithm with the following properties: given n and access to independent draws from X, the algorithm uses $\tilde{O}(1/\epsilon^3) \cdot \log(1/\delta)$ samples from X, performs $(1/\epsilon)^{O(\log^2(1/\epsilon))} \cdot \tilde{O}(\log n \log \frac{1}{\delta})$ bit operations, and with probability $1 - \delta$ outputs a (succinct description of a) vector $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_n)$ defining a PBD $\hat{X}$ such that $d_{\mathrm{TV}}(\hat{X}, X) \leq \epsilon$.*

We note that since each sample drawn from X is a $\log(n)$-bit string, the number of bit-operations performed by our first algorithm is *quasilinear* in the length of its input. The sample complexity of both our algorithms is

---

[1][KMR$^+$94] used the Kullback-Leibler divergence as their distance measure but we find it more natural to use variation distance.

[2]We write $\tilde{O}(\cdot)$ to hide factors which are polylogarithmic in the argument to $\tilde{O}(\cdot)$; thus for example $\tilde{O}(a \log b)$ denotes a quantity which is $O(a \log b \cdot \log^c(a \log b))$ for some absolute constant c.

not far from optimal, since $\Omega(1/\epsilon^2)$ samples are required even to distinguish the (simpler) Binomial distributions $B(n, 1/2)$ and $B(n, 1/2 + \epsilon/\sqrt{n})$, which have variation distance $\Omega(\epsilon)$.

Motivated by these strong learning results for PBDs, we also consider learning a more general class of distributions, namely distributions of the form $X = \sum_{i=1}^{n} w_i X_i$ which are *weighted* sums of independent Bernoulli random variables. We give an algorithm which uses $O(\log n)$ samples and runs in $\mathrm{poly}(n)$ time if there are only constantly many different weights in the sum:

Theorem 2 (Learning sums of weighted independent Bernoulli random variables) *Let $X = \sum_{i=1}^{n} a_i X_i$ be a weighted sum of unknown independent Bernoullis such that there are at most k different values among $a_1, \ldots, a_n$. Then there is an algorithm with the following properties: given n, $a_1, \ldots, a_n$ and access to independent draws from X, it uses $k \log(n) \cdot \widetilde{O}(1/\epsilon^2) \cdot \log(1/\delta)$ samples from the target distribution X, runs in time $\mathrm{poly}(n^k \cdot \epsilon^{-k \log^2(1/\epsilon)}) \cdot \log(1/\delta)$, and with probability $1 - \delta$ outputs a hypothesis vector $\hat{p} \in [0, 1]^n$ defining independent Bernoulli random variables $\hat{X}_i$ with $\mathbf{E}[\hat{X}_i] = \hat{p}_i$ such that $d_{\mathrm{TV}}(\hat{X}, X) \le \epsilon$, where $\hat{X} = \sum_{i=1}^{n} a_i \hat{X}_i$.*

Note that setting all a_i's to 1 in Theorem 2 gives a weaker result than Theorem 1 in terms of running time and sample complexity. To complement Theorem 2, we also show that if there are many distinct weights in the sum, then even for weights with a very simple structure any learning algorithm must use many samples:

Theorem 3 (Sample complexity lower bound for learning sums of weighted independent Bernoullis) *Let $X = \sum_{i=1}^{n} i \cdot X_i$ be a weighted sum of unknown independent Bernoullis (where the i-th weight is simply i). Let L be any learning algorithm which, given n and access to independent draws from X, outputs a hypothesis distribution $\hat{X}$ such that $d_{\mathrm{TV}}(\hat{X}, X) \le 1/25$ with probability at least $e^{-o(n)}$. Then L must use $\Omega(n)$ samples.*

1.2 Related work

Many results in probability theory study approximations to the Poisson Binomial distribution via simpler distributions. In a well-known result, Le Cam [Cam60] shows that for any PBD $X = \sum_{i=1}^{n} X_i$ with $\mathbf{E}[X_i] = p_i$

$$d_{\mathrm{TV}}(X, \mathrm{Poi}(p_1 + \cdots + p_n)) \le 2 \sum_{i=1}^{n} p_i^2,$$

where $\mathrm{Poi}(\lambda)$ denotes the Poisson distribution with parameter λ. Subsequently many other proofs of this result and similar ones were given using a range of different techniques; [HC60, Che74, DP86, BHJ92] is a sampling of work along these lines, and Steele [Ste94] gives an extensive list of relevant references. Significant work has also been done on approximating PBDs by normal distributions (see e.g. [Ber41, Ess42, Mik93, Vol95]) and by Binomial distributions (see e.g. [Ehm91, Soo96, Roo00]). These results provide structural information about PBDs that can be well-approximated via simpler distributions, but fall short of our goal of obtaining approximations of a general, unknown PBD up to an *arbitrary accuracy*. Indeed, the approximations obtained in the probability literature (such as, the Poisson, Normal and Binomial approximations) typically depend on the first few moments of the target PBD, while higher moments are crucial for arbitrary approximation [Roo00].

Taking a different perspective, it is easy to show (see Section 2 of [KG71]) that every PBD is a unimodal distribution over $[n]$. The learnability of general unimodal distributions over $[n]$ is well understood: Birgé [Bir87a, Bir97] has given a computationally efficient algorithm that can learn any unimodal distribution over $[n]$ to variation distance ϵ from $O(\log(n)/\epsilon^3)$ samples, and has shown that any algorithm must use $\Omega(\log(n)/\epsilon^3)$ samples. (The [Bir87a] lower bound is stated for continuous unimodal distributions, but the arguments are easily adapted to the discrete case.) Our main result, Theorem 1, shows that the additional PBD assumption can be leveraged to obtain sample complexity *independent of n* with a computationally highly efficient algorithm.

So, how might one leverage the structure of PBDs to remove n from the sample complexity? The first property one might try to exploit is that a PBD assigns $1 - \epsilon$ of its mass to $O_\epsilon(\sqrt{n})$ points. So one could draw samples from the distribution to (approximately) identify these points and then try to estimate the probability

2

assigned to each such point to within high enough accuracy so that the overall estimation error is ϵ. Clearly, such an approach, if followed naïvely, would give $\text{poly}(n)$ sample complexity. Alternatively, one could run Birgé's algorithm on the restricted support of size $O_\epsilon(\sqrt{n})$, but that will not improve the asymptotic sample complexity. A different approach would be to construct a small ϵ-cover (under the total variation distance) of the space of all PBDs on n variables. Indeed, if such a cover has size N, it can be shown (see Lemma 11 of the full paper, or Chapter 7 of [DL01])) that a target PBD can be learned from $O(\log(N)/\epsilon^2)$ samples. Still it is easy to argue that any cover needs to have size $\Omega(n)$, so this approach too gives a $\log(n)$ dependence in the sample complexity.

Our approach, which removes n completely from the sample complexity, requires a refined understanding of the structure of the set of all PBDs on n variables, in fact one that is more refined than the understanding provided by the aforementioned results (approximating a PBD by a Poisson, Normal, or Binomial distribution). We give an outline of the approach in the next section.

1.3   Our approach

The starting point of our algorithm for learning PBDs is a theorem of [DP11, Das08] that gives detailed information about the structure of a small ϵ-cover (under the total variation distance) of the space of all PBDs on n variables (see Theorem 4). Roughly speaking, this result says that every PBD is either close to a PBD whose support is sparse, or is close to a translated "heavy" Binomial distribution. Our learning algorithm exploits the structure of the cover to close in on the information that is absolutely necessary to approximate an unknown PBD. In particular, the algorithm has two subroutines corresponding to the (aforementioned) different types of distributions that the cover maintains. First, assuming that the target PBD is close to a sparsely supported distribution, it runs Birgé's unimodal distribution learner over a carefully selected subinterval of $[n]$ to construct a hypothesis H_S; the (purported) sparsity of the distribution makes it possible for this algorithm to use $\tilde{O}(1/\epsilon^3)$ samples independent of n. Then, assuming that the target PBD is close to a translated "heavy" Binomial distribution, the algorithm constructs a hypothesis Translated Poisson Distribution H_P [Rö7] whose mean and variance match the estimated mean and variance of the target PBD; we show that H_P is close to the target PBD if the latter is not close to any sparse distribution in the cover. At this point the algorithm has two hypothesis distributions, H_S and H_P, one of which should be good; it remains to select one as the final output hypothesis. This is achieved using a form of "hypothesis testing" for probability distributions. The above sketch captures the main ingredients of Part (1) of Theorem 1, but additional work needs to be done to get the proper learning algorithm of Part (2), since neither the sparse hypothesis H_S output by Birgé's algorithm nor the Translated Poisson hypothesis H_S is a PBD. Via a sequence of transformations we are able to show that the Translated Poisson hypothesis H_P can be converted to a Binomial distribution $\text{Bin}(n', p)$ for some $n' \leq n$. For the sparse hypothesis, we obtain a PBD by searching a (carefully selected) subset of the ϵ-cover to find a PBD that is close to our hypothesis H_S (this search accounts for the increased running time in Part (2) versus Part (1)). We stress that for both the non-proper and proper learning algorithms sketched above, many technical subtleties and challenges arise in implementing the high-level plan given above, requiring a careful and detailed analysis which we give in full below. After all, eliminating n from the sample complexity is surprising and warrants some non-trivial technical effort.

To prove Theorem 2 we take a more general approach and then specialize it to weighted sums of independent Bernoullis with constantly many distinct weights. We show that for any class $\mathcal{S}$ of target distributions, if $\mathcal{S}$ has an ϵ-cover of size N then there is a generic algorithm for learning an unknown distribution from $\mathcal{S}$ to accuracy ϵ that uses $O((\log N)/\epsilon^2)$ samples. Our approach is rather similar to the algorithm of [DL01] for choosing a density estimate (but different in some details); it works by carrying out a tournament that matches every pair of distributions in the cover against each other. Our analysis shows that with high probability some ϵ-accurate distribution in the cover will survive the tournament undefeated, and that any undefeated tournament will with high probability be $O(\epsilon)$-accurate. We then specialize this general result to show how the tournament can be implemented efficiently for the class $\mathcal{S}$ of weighted sums of independent Bernoullis with constantly many distinct weights. Finally, the lower bound of Theorem 3 is proved by a direct information-theoretic argument.

1.4 Preliminaries

For a distribution X supported on $[n] = \{0, 1, \ldots, n\}$ we write $X(i)$ to denote the value $\Pr[X = i]$ of the pdf, and $X(\leq i)$ to denote the value $\Pr[X \leq i]$ of the cdf. For $S \subseteq [n]$ we write $X(S)$ to denote $\sum_{i \in S} X(i)$ and X_S to denote the conditional distribution of X restricted to S.

Recall that the *total variation distance* between two distributions X and Y over a finite domain D is

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

Fix a finite domain D, and let $\mathcal{P}$ denote some set of distributions over D. Given $\delta > 0$, a subset $\mathcal{Q} \subseteq \mathcal{P}$ is said to be a *δ-cover of $\mathcal{P}$* (w.r.t. total variation distance) if for every distribution P in $\mathcal{P}$ there exists some distribution Q in $\mathcal{Q}$ such that $d_{\mathrm{TV}}(P, Q) \leq \delta$.

We write $\mathcal{S} = \mathcal{S}_n$ to denote the set of all PBDs $X = \sum_{i=1}^{n} X_i$. We sometimes write $\{X_i\}$ to denote the PBD $X = \sum_{i=1}^{n} X_i$.

We also define the *Translated Poisson distribution* as follows.

Definition 1 ([RÖ7]) *We say that an integer random variable Y has a* translated Poisson distribution *with parameters μ and σ^2, written $Y = TP(\mu, \sigma^2)$, if $Y = \lfloor \mu - \sigma^2 \rfloor + Poisson(\sigma^2 + \{\mu - \sigma^2\})$, where $\{\mu - \sigma^2\}$ represents the fractional part of $\mu - \sigma^2$.*

Translated Poisson distributions are useful to us because known results bound how far they are from PBDs and from each other. We will use the following results:

Lemma 1 (see (3.4) of [RÖ7]) *Let $J_1, \ldots, J_n$ be a sequence of independent random indicators with $\mathbf{E}[J_i] = p_i$. Then*

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

where $\mu = \sum_{i=1}^{n} p_i$ and $\sigma^2 = \sum_{i=1}^{n} p_i(1 - p_i)$.

Lemma 2 (Lemma 2.1 of [BL06]) *Let $\mu_1, \mu_2 \in \mathbb{R}$ and $\sigma_1^2, \sigma_2^2 \in \mathbb{R}_+ \setminus \{0\}$ be such that $\lfloor \mu_1 - \sigma_1^2 \rfloor \leq \lfloor \mu_2 - \sigma_2^2 \rfloor$. Then*

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

2 Learning an unknown sum of Bernoullis from $\mathrm{poly}(1/\epsilon)$ samples

In this section we prove our main result, Theorem 1, by giving a sample- and time-efficient algorithm for learning an unknown PBD $X = \sum_{i=1}^{n} X_i$.

A cover for PBDs. An important ingredient in our analysis is the following theorem, which is an extension of Theorem 9 of the full version of [DP11]. It defines a cover (in total variation distance) of the space $\mathcal{S} = \mathcal{S}_n$ of all order-n PBDs:

Theorem 4 (Cover for PBDs) *For all $\epsilon > 0$, there exists an ϵ-cover $\mathcal{S}_\epsilon \subseteq \mathcal{S}$ of $\mathcal{S}$ such that*

1. *$|\mathcal{S}_\epsilon| \leq n^3 \cdot O(1/\epsilon) + n \cdot \left(\frac{1}{\epsilon}\right)^{O(\log^2 1/\epsilon)}$; and*

2. *The set $\mathcal{S}_\epsilon$ can be constructed in time linear in its representation size, i.e. $\tilde{O}(n^3/\epsilon) + \tilde{O}(n) \cdot \left(\frac{1}{\epsilon}\right)^{O(\log^2 1/\epsilon)}$.*

Moreover, if $\{Y_i\} \in \mathcal{S}_\epsilon$, then the collection $\{Y_i\}$ has one of the following forms, where $k = k(\epsilon) \leq C/\epsilon$ is a positive integer, for some absolute constant $C > 0$:

(i) *(Sparse Form) There is a value $\ell \leq k^3 = O(1/\epsilon^3)$ such that for all $i \leq \ell$ we have $\mathbf{E}[Y_i] \in \left\{\frac{1}{k^2}, \frac{2}{k^2}, \ldots, \frac{k^2-1}{k^2}\right\}$, and for all $i > \ell$ we have $\mathbf{E}[Y_i] \in \{0, 1\}$.*

(ii) *(k-heavy Binomial Form) There is a value $\ell \in \{0, 1, \ldots, n\}$ and a value $q \in \left\{\frac{1}{kn}, \frac{2}{kn}, \ldots, \frac{kn-1}{kn}\right\}$ such that for all $i \leq \ell$ we have $\mathbf{E}[Y_i] = q$; for all $i > \ell$ we have $\mathbf{E}[Y_i] \in \{0, 1\}$; and ℓ, q satisfy the bounds $\ell q \geq k^2 - \frac{1}{k}$ and $\ell q(1-q) \geq k^2 - k - 1 - \frac{3}{k}$.*

Finally, for every $\{X_i\} \in \mathcal{S}$ for which there is no ϵ-neighbor in $\mathcal{S}_\epsilon$ that is in sparse form, there exists a collection $\{Y_i\} \in \mathcal{S}_\epsilon$ in k-heavy Binomial form such that

(iii) *$d_{\mathrm{TV}}(\sum_i X_i, \sum_i Y_i) \leq \epsilon$; and*

(iv) *if $\mu = \mathbf{E}[\sum_i X_i]$, $\mu' = \mathbf{E}[\sum_i Y_i]$, $\sigma^2 = \mathrm{Var}[\sum_i X_i]$ and $\sigma'^2 = \mathrm{Var}[\sum_i Y_i]$, then $|\mu - \mu'| = O(\epsilon)$ and $|\sigma^2 - \sigma'^2| = O(1 + \epsilon \cdot (1 + \sigma^2))$.*

We remark that [Das08] establishes the same theorem, except that the size of the cover is $n^3 \cdot O(1/\epsilon) + n \cdot \left(\frac{1}{\epsilon}\right)^{O(1/\epsilon^2)}$. Indeed, this weaker bound is obtained by including in the cover all possible collections $\{Y_i\} \in \mathcal{S}$ in sparse form and all possible collections in k-heavy Binomial form, for $k = O(1/\epsilon)$ specified by the theorem. [DP11] obtains a smaller cover by only selecting a subset of the collections in sparse form included in the cover of [Das08]. Finally, the cover theorem stated in [Das08, DP11] does not include the part of the above statement following "finally." We provide a proof of this extension in Section 4.1.

We remark also that our analysis in this paper in fact establishes a slightly stronger version of the above theorem, with an improved bound on the cover size (as a function of n) and stronger conditions on the Binomial Form distributions in the cover. We present this strengthened version of the Cover Theorem in Section 4.2.

The learning algorithm. Our algorithm `Learn-PBD` has the general structure shown below (a detailed version is given later).

`Learn-PBD`

    1. Run `Learn-Sparse`$^X(n, \epsilon, \delta/3)$ to get hypothesis distribution H_S.
    2. Run `Learn-Poisson`$^X(n, \epsilon, \delta/3)$ to get hypothesis distribution H_P.
    3. Return the distribution which is the output of `Choose-Hypothesis`$^X(H_S, H_P, \epsilon, \delta/3)$.

Figure 1: `Learn-PBD`

The subroutine `Learn-Sparse`^{X} is given sample access to X and is designed to find an ϵ-accurate hypothesis if the target PBD X is ϵ-close to some sparse form PBD inside the cover $\mathcal{S}_\epsilon$; similarly, `Learn-Poisson`^{X} is designed to find an ϵ-accurate hypothesis if X is not ϵ-close to a sparse form PBD (in this case, Theorem 4 implies that X must be ϵ-close to some $k(\epsilon)$-heavy Binomial form PBD). Finally, `Choose-Hypothesis`^{X} is designed to choose one of the two hypotheses H_S, H_P as being ϵ-close to X. The following subsections describe and prove correctness of these subroutines. We remark that the subroutines `Learn-Sparse` and `Learn-Poisson` do not return the distributions H_S and H_P as a list of probabilities for every point in $[n]$; rather, they return a succinct description of these distributions in order to keep the running time of the algorithm logarithmic in n.

2.1 Learning when X is close to a Sparse Form PBD

Our starting point here is the simple observation that any PBD is a unimodal distribution over the domain $\{0, 1, \ldots, n\}$ (there is a simple inductive proof of this, or see Section 2 of [KG71]). This will enable us to use the algorithm of Birgé [Bir97] for learning unimodal distributions. We recall Birgé's result, and refer the reader to Section 5 for an explanation of how Theorem 5 as stated below follows from [Bir97].

Theorem 5 ([Bir97]) *For all $n, \epsilon, \delta > 0$, there is an algorithm that draws $\frac{\log n}{\epsilon^3} \cdot \tilde{O}(\log \frac{1}{\delta})$ samples from an unknown unimodal distribution X over $[n]$, does $\tilde{O}\left(\frac{\log^2 n}{\epsilon^3} \log^2 \frac{1}{\delta}\right)$ bit-operations, and outputs a (succinct description of a) hypothesis distribution H over $[n]$ that has the following form: H is uniform over subintervals $[a_1, b_1], [a_2, b_2], \ldots, [a_k, b_k]$, whose union $\cup_{i=1}^k [a_i, b_i] = [n]$, where $k = O\left(\frac{\log n}{\epsilon^3}\right)$. In particular, the algorithm outputs the lists a_1 through a_k and b_1 through b_k, as well as the total probability mass that H assigns to each subinterval $[a_i, b_i]$, $i = 1, \ldots, k$. Finally, with probability at least $1 - \delta$, $d_{\mathrm{TV}}(X, H) \leq \epsilon$.*

In the rest of this subsection we prove the following:

Lemma 3 *For all $n, \epsilon', \delta' > 0$, there is an algorithm* Learn-Sparse$^X(n, \epsilon', \delta')$ *that draws $O(\frac{1}{\epsilon'^3} \log \frac{1}{\epsilon'} \log \frac{1}{\delta'})$ samples from a target PBD X over $[n]$, does $\log n \cdot \tilde{O}\left(\frac{1}{\epsilon'^3} \log \frac{1}{\delta'}\right)$-bit operations, and outputs a (succinct description of a) hypothesis distribution H_S over $[n]$ that has the following form: its support is contained in an explicitly specified interval $[a, b] \subset [n]$, where $|b - a| = O(1/\epsilon'^3)$, and for every point in $[a, b]$ the algorithm explicitly specifies the probability assigned to that point by H_S.* [3] *Moreover, the algorithm has the following guarantee: Suppose X is ϵ'-close to some sparse form PBD Y in the cover $\mathcal{S}_{\epsilon'}$ of Theorem 4. Then, with probability at least $1 - \delta'$, $d_{\mathrm{TV}}(X, H_S) \leq c_1 \epsilon'$, for some absolute constant $c_1 \geq 1$, and the support of H_S is a subset of the support of Y.*

Proof: The Algorithm Learn-Sparse$^X(n, \epsilon', \delta')$ works as follows: It first draws $M = 32 \log(8/\delta')/\epsilon'^2$ samples from X and sorts them to obtain a list of values $0 \leq s_1 \leq \cdots \leq s_M \leq n$. In terms of these samples, let us define $\hat{a} := s_{\lceil 2\epsilon' M \rceil}$ and $\hat{b} := s_{\lfloor (1-2\epsilon')M \rfloor}$. We claim the following:

Claim 4 *With probability at least $1 - \delta'/2$, we have $X(\leq \hat{a}) \in [3\epsilon'/2, 5\epsilon'/2]$ and $X(\leq \hat{b}) \in [1 - 5\epsilon'/2, 1 - 3\epsilon'/2]$.*

Proof of Claim 4: We only show that $X(\leq \hat{a}) \geq 3\epsilon'/2$ with probability at least $1 - \delta'/8$, since the arguments for $X(\leq \hat{a}) \leq 5\epsilon'/2$, $X(\leq \hat{b}) \leq 1 - 3\epsilon'/2$ and $X(\leq \hat{b}) \geq 1 - 5\epsilon'/2$ are identical. Given that each of these conditions is met with probability at least $1 - \delta'/8$, the union bound establishes our claim.

To show that $X(\leq \hat{a}) \geq 3\epsilon'/2$ is satisfied with probability at least $1 - \delta'/8$ we argue as follows: Let $\alpha' = \max\{i \mid X(\leq i) < 3\epsilon'/2\}$. Clearly, $X(\leq \alpha') < 3\epsilon'/2$ while $X(\leq \alpha' + 1) \geq 3\epsilon'/2$. Given this, of M samples drawn from X an expected number of at most $3\epsilon' M/2$ samples are $\leq \alpha'$. It follows then from the Chernoff bound that the probability that more than $\frac{7}{4}\epsilon' M$ samples are $\leq \alpha'$ is at most $e^{-(\epsilon'/4)^2 M/2} \leq \delta'/8$. Hence, $\hat{a} \geq \alpha' + 1$, which implies that $X(\leq \hat{a}) \geq 3\epsilon'/2$. ∎

If $\hat{b} - \hat{a} > (C/\epsilon')^3$, where C is the constant in the statement of Theorem 4, the algorithm outputs "fail", returning the trivial hypothesis which puts probability mass 1 on the point 0. Otherwise, the algorithm runs Birgé's unimodal distribution learner (Theorem 5) on the conditional distribution $X_{[\hat{a}, \hat{b}]}$, and outputs the result of Birgé's algorithm. Since X is unimodal, it follows that $X_{[\hat{a}, \hat{b}]}$ is also unimodal, hence Birgé's algorithm is appropriate for learning it. The way we apply Birgé's algorithm to learn $X_{[\hat{a}, \hat{b}]}$ given samples from the original distribution X is the obvious one: we draw samples from X, ignoring all samples that fall outside of $[\hat{a}, \hat{b}]$, until the right $O(\log(1/\delta')\log(1/\epsilon')/\epsilon'^3)$ number of samples fall inside $[\hat{a}, \hat{b}]$, as required by Birgé's algorithm for learning a distribution of support of size $(C/\epsilon')^3$ with probability $1 - \delta'/4$. Once we have the right number of samples in $[\hat{a}, \hat{b}]$, we run Birgé's algorithm to learn the conditional distribution $X_{[\hat{a}, \hat{b}]}$. Note that the number of samples we need to draw from X until the right $O(\log(1/\delta')\log(1/\epsilon')/\epsilon'^3)$ number of samples fall inside $[\hat{a}, \hat{b}]$ is still $O(\log(1/\delta')\log(1/\epsilon')/\epsilon'^3)$, with probability at least $1 - \delta'/4$. Indeed, since $X([\hat{a}, \hat{b}]) = 1 - O(\epsilon')$, it follows from the Chernoff bound that with probability at least $1 - \delta'/4$, if $K = \Theta(\log(1/\delta')\log(1/\epsilon')/\epsilon'^3)$ samples are drawn from X, at least $K(1 - O(\epsilon'))$ fall inside $[\hat{a}, \hat{b}]$.

---

[3]In particular, our algorithm will output a list of pointers, mapping every point in $[a, b]$ to some memory location where the probability assigned to that point by H_S is written.

Analysis: It is easy to see that the sample complexity of our algorithm is as promised. For the running time, notice that, if Birgé's algorithm is invoked, it will return two lists of numbers a_1 through a_k and b_1 through b_k, as well as a list of probability masses $q_1, \ldots, q_k$ assigned to each subinterval $[a_i, b_i]$, $i = 1, \ldots, k$, by the hypothesis distribution H_S, where $k = O(\log(1/\epsilon')/\epsilon')$. In linear time, we can compute a list of probabilities $\hat{q}_1, \ldots, \hat{q}_k$, representing the probability assigned by H_S to every point of subinterval $[a_i, b_i]$, for $i = 1, \ldots, k$. So we can represent our output hypothesis H_S via a data structure that maintains $O(1/\epsilon'^3)$ pointers, having one pointer per point inside $[a, b]$. The pointers map points to probabilities assigned by H_S to these points. Thus turning the output of Birgé's algorithm into an explicit distribution over $[a, b]$ incurs linear overhead in our running time, and hence the running time of our algorithm is also as promised. Moreover, we also note that the output distribution has the promised structure, since in one case it has a single atom at 0 and in the other case it is the output of Birgé's algorithm on a distribution of support of size $(C/\epsilon')^3$.

It only remains to justify the last part of the lemma. Let Y be the sparse-form PBD that X is close to; say that Y is supported on $\{a', \ldots, b'\}$ where $b' - a' \le (C/\epsilon')^3$. Since X is ϵ'-close to Y in total variation distance it must be the case that $X(\le a' - 1) \le \epsilon'$. Since $X(\le \hat{a}) \ge 3\epsilon'/2$ by Claim 4, it must be the case that $\hat{a} \ge a'$. Similar arguments give that $\hat{b} \le b'$. So the interval $[\hat{a}, \hat{b}]$ is contained in $[a', b']$ and has length at most $(C/\epsilon')^3$. This means that Birgé's algorithm is indeed used correctly by our algorithm to learn $X_{[\hat{a},\hat{b}]}$, with probability at least $1 - \delta'/2$ (that is, unless Claim 4 fails). Now it follows from the correctness of Birgé's algorithm (Theorem 5) and the discussion above, that the hypothesis H_S output when Birgé's algorithm is invoked satisfies $d_{\mathrm{TV}}(H_S, X_{[\hat{a},\hat{b}]}) \le \epsilon'$, with probability at least $1 - \delta'/2$, i.e. unless either Birgé's algorithm fails, or we fail to get the right number of samples landing inside $[\hat{a}, \hat{b}]$. To conclude the proof of the lemma we note that:

$$
\begin{aligned}
2 d_{\mathrm{TV}}(X, X_{[\hat{a},\hat{b}]}) &= \sum_{i \in [\hat{a},\hat{b}]} |X_{[\hat{a},\hat{b}]}(i) - X(i)| + \sum_{i \notin [\hat{a},\hat{b}]} |X_{[\hat{a},\hat{b}]}(i) - X(i)| \\
&= \sum_{i \in [\hat{a},\hat{b}]} \left| \frac{1}{X([\hat{a}, \hat{b}])} X(i) - X(i) \right| + \sum_{i \notin [\hat{a},\hat{b}]} X(i) \\
&= \sum_{i \in [\hat{a},\hat{b}]} \left| \frac{1}{1 - O(\epsilon')} X(i) - X(i) \right| + O(\epsilon') \\
&= \frac{O(\epsilon')}{1 - O(\epsilon')} \sum_{i \in [\hat{a},\hat{b}]} \left| X(i) \right| + O(\epsilon') = O(\epsilon').
\end{aligned}
$$

So the triangle inequality gives: $d_{\mathrm{TV}}(H_S, X) = O(\epsilon')$, and Lemma 3 is proved. ∎

2.2 Learning when X is close to a k-heavy Binomial Form PBD

Lemma 5 *For all $n, \epsilon', \delta' > 0$, there is an algorithm* Learn-Poisson$^X(n, \epsilon', \delta')$ *that draws $O(\log(1/\delta')/\epsilon'^2)$ samples from a target PBD X over $[n]$, runs in time $O(\log n \cdot \log(1/\delta')/\epsilon'^2)$, and returns two parameters $\hat{\mu}$ and $\hat{\sigma}^2$. Moreover, the algorithm has the following guarantee: Suppose X is not ϵ'-close to any Sparse Form PBD in the cover $\mathcal{S}_{\epsilon'}$ of Theorem 4. Let H_P be the translated Poisson distribution with parameters $\hat{\mu}$ and $\hat{\sigma}^2$, i.e. $H_P = TP(\hat{\mu}, \hat{\sigma}^2)$. Then with probability at least $1 - \delta'$ we have $d_{\mathrm{TV}}(X, H_P) \le c_2 \epsilon'$, for some absolute constant $c_2 \ge 1$.*

Our proof plan is to exploit the structure of the cover of Theorem 4. In particular, if X is not ϵ'-close to any Sparse Form PBD in the cover, it must be ϵ'-close to a PBD in Heavy Binomial Form with approximately the same mean and variance as X, as specified by the final part of the cover theorem. Now, given that a PBD in Heavy Binomial Form is just a translated Binomial distribution, a natural strategy is to estimate the mean and variance of the target PBD X and output as a hypothesis a translated Poisson distribution with these parameters. We show that this strategy is a successful one.

7

We start by showing that we can estimate the mean and variance of the target PBD X.

Lemma 6 *For all $n, \epsilon, \delta > 0$, there exists an algorithm $\mathcal{A}(n, \epsilon, \delta)$ with the following properties: given access to a PBD X over $[n]$, it produces estimates $\hat{\mu}$ and $\hat{\sigma}^2$ for $\mu = \mathbf{E}[X]$ and $\sigma^2 = \mathrm{Var}[X]$ respectively such that with probability at least $1 - \delta$:*

$$|\mu - \hat{\mu}| \leq \epsilon \cdot \sigma \qquad and \qquad |\sigma^2 - \hat{\sigma}^2| \leq \epsilon \cdot \sigma^2 \sqrt{4 + \frac{1}{\sigma^2}}.$$

The algorithm uses $O(\log(1/\delta)/\epsilon^2)$ samples and runs in time $O(\log n \log(1/\delta)/\epsilon^2)$.

Proof of Lemma 6: We treat the estimation of μ and σ^2 separately. For both estimation problems we show how to use $O(1/\epsilon^2)$ samples to obtain estimates $\hat{\mu}$ and $\hat{\sigma}^2$ achieving the required guarantees with probability at least $2/3$. Then a routine procedure allows us to boost the success probability to $1 - \delta$ at the expense of a multiplicative factor $O(\log 1/\delta)$ on the number of samples. While we omit the details of the routine boosting argument, we remind the reader that it involves running the weak estimator $O(\log 1/\delta)$ times to obtain estimates $\hat{\mu}_1, \ldots, \hat{\mu}_{O(\log 1/\delta)}$ and outputting the median of these estimates, and similarly for estimating σ^2.

We proceed to specify and analyze the weak estimators for μ and σ^2 separately:

- *Weak estimator for μ:* Let $Z_1, \ldots, Z_m$ be independent samples from X, and let $\hat{\mu} = \frac{\sum_i Z_i}{m}$. Then

$$\mathbf{E}[\hat{\mu}] = \mu \quad \text{and} \quad \mathrm{Var}[\hat{\mu}] = \frac{1}{m}\mathrm{Var}[X] = \frac{1}{m}\sigma^2.$$

So Chebyshev's inequality implies that

$$\Pr[|\hat{\mu} - \mu| \geq t\sigma/\sqrt{m}] \leq \frac{1}{t^2}.$$

Choosing $t = \sqrt{3}$ and $m = 3/\epsilon^2$, the above imply that $|\hat{\mu} - \mu| \leq \epsilon\sigma$ with probability at least $2/3$.

- *Weak estimator for σ^2:* Let $Z_1, \ldots, Z_m$ be independent samples from X, and let $\hat{\sigma}^2 = \frac{\sum_i (Z_i - \frac{1}{m}\sum_i Z_i)^2}{m-1}$ be the unbiased sample variance (note the use of Bessel's correction). Then it can be checked [Joh03] that

$$\mathbf{E}[\hat{\sigma}^2] = \sigma^2 \quad \text{and} \quad \mathrm{Var}[\hat{\sigma}^2] = \sigma^4\left(\frac{2}{m-1} + \frac{\kappa}{m}\right),$$

where κ is the kurtosis of the distribution of X. To bound κ in terms of σ^2 suppose that $X = \sum_{i=1}^n X_i$, where $\mathbf{E}[X_i] = p_i$ for all i. Then

$$\kappa = \frac{1}{\sigma^4}\sum_i (1 - 6p_i(1 - p_i))(1 - p_i)p_i \qquad \text{(see [NJ05])}$$

$$\leq \frac{1}{\sigma^4}\sum_i |1 - 6p_i(1 - p_i)|(1 - p_i)p_i$$

$$\leq \frac{1}{\sigma^4}\sum_i (1 - p_i)p_i = \frac{1}{\sigma^2}.$$

So $\mathrm{Var}[\hat{\sigma}^2] = \sigma^4\left(\frac{2}{m-1} + \frac{\kappa}{m}\right) \leq \frac{\sigma^4}{m}(4 + \frac{1}{\sigma^2})$. So Chebyshev's inequality implies that

$$\Pr\left[|\hat{\sigma}^2 - \sigma^2| \geq t\frac{\sigma^2}{\sqrt{m}}\sqrt{4 + \frac{1}{\sigma^2}}\right] \leq \frac{1}{t^2}.$$

Choosing $t = \sqrt{3}$ and $m = 3/\epsilon^2$, the above imply that $|\hat{\sigma}^2 - \sigma^2| \leq \epsilon\sigma^2\sqrt{4 + \frac{1}{\sigma^2}}$ with probability at least $2/3$.

8

■

Proof of Lemma 5: Suppose now that X is not ϵ'-close to any PBD in sparse form inside the cover $\mathcal{S}_{\epsilon'}$ of Theorem 4. Then there exists a PBD Z in $k = k(\epsilon')$-heavy Binomial form inside $\mathcal{S}_{\epsilon'}$ that is within total variation distance ϵ' from X. We use the existence of such a Z to obtain lower bounds on the mean and variance of X. Indeed, suppose that the distribution of Z is $\mathrm{Bin}(\ell, q) + t$, i.e. a Binomial with parameters ℓ, q that is translated by t. Then Theorem 4 certifies that the following conditions are satisfied by the parameters $\ell, q, t, \mu = \mathbf{E}[X]$ and $\sigma^2 = \mathrm{Var}[X]$:

(a) $\ell q \geq k^2 - \frac{1}{k}$;

(b) $\ell q(1 - q) \geq k^2 - k - 1 - \frac{3}{k}$;

(c) $|t + \ell q - \mu| = O(\epsilon')$; and

(d) $|\ell q(1 - q) - \sigma^2| = O(1 + \epsilon \cdot (1 + \sigma^2))$.

In particular, conditions (b) and (d) above imply that

$$\sigma^2 = \Omega(k^2) = \Omega(1/\epsilon'^2) \geq \theta^2 \tag{1}$$

for some universal constant θ. Hence we can apply Lemma 6 with $\epsilon = \epsilon'/\sqrt{4 + \frac{1}{\theta^2}}$ and $\delta = \delta'$ to obtain—from $O(\log(1/\delta')/\epsilon'^2)$ samples and with probability at least $1 - \delta'$—estimates $\hat{\mu}$ and $\hat{\sigma}^2$ of μ and σ^2 respectively that satisfy

$$|\mu - \hat{\mu}| \leq \epsilon' \cdot \sigma \quad \text{and} \quad |\sigma^2 - \hat{\sigma}^2| \leq \epsilon' \cdot \sigma^2. \tag{2}$$

Now let Y be a random variable distributed according to the translated Poisson distribution $TP(\hat{\mu}, \hat{\sigma}^2)$. We conclude the proof of Lemma 5 by showing that Y and X are within $O(\epsilon')$ in total variation distance.

Claim 7 *If X and Y are as above, then $d_{\mathrm{TV}}(X, Y) \leq O(\epsilon')$.*

Proof of Claim 7: We make use of Lemma 1. Suppose that $X = \sum_{i=1}^n X_i$, where $\mathbf{E}[X_i] = p_i$ for all i. Lemma 1 implies that

$$
\begin{aligned}
d_{\mathrm{TV}}(X, TP(\mu, \sigma^2)) &\leq \frac{\sqrt{\sum_i p_i^3(1 - p_i)} + 2}{\sum_i p_i(1 - p_i)} \leq \frac{\sqrt{\sum_i p_i(1 - p_i)} + 2}{\sum_i p_i(1 - p_i)} \\
&\leq \frac{1}{\sqrt{\sum_i p_i(1 - p_i)}} + \frac{2}{\sum_i p_i(1 - p_i)} = \frac{1}{\sigma} + \frac{2}{\sigma^2} = O(\epsilon').
\end{aligned} \tag{3}
$$

It remains to bound the total variation distance between the translated Poisson distributions $TP(\mu, \sigma^2)$ and $TP(\hat{\mu}, \hat{\sigma}^2)$. For this we use Lemma 2. Lemma 2 implies

$$
\begin{aligned}
d_{\mathrm{TV}}(TP(\mu, \sigma^2), TP(\hat{\mu}, \hat{\sigma}^2)) &\leq \frac{|\mu - \hat{\mu}|}{\min(\sigma, \hat{\sigma})} + \frac{|\sigma^2 - \hat{\sigma}^2| + 1}{\min(\sigma^2, \hat{\sigma}^2)} \\
&\leq \frac{\epsilon'\sigma}{\min(\sigma, \hat{\sigma})} + \frac{\epsilon' \cdot \sigma^2 + 1}{\min(\sigma^2, \hat{\sigma}^2)} \\
&\leq \frac{\epsilon'\sigma}{\sigma/\sqrt{1 - \epsilon'}} + \frac{\epsilon' \cdot \sigma^2 + 1}{\sigma^2/(1 - \epsilon')} \\
&= O(\epsilon') + \frac{1 - \epsilon'}{\sigma^2} \\
&= O(\epsilon') + O(\epsilon'^2) \\
&= O(\epsilon').
\end{aligned} \tag{4}
$$

9

The claim follows from (3), (4) and the triangle inequality. This concludes the proof of Lemma 5 as well. ∎

As a final remark, we note that the algorithm described above does not need to know a priori whether or not X is ϵ'-close to a PBD in sparse form inside the cover $\mathcal{S}_{\epsilon'}$ of Theorem 4. The algorithm simply runs the estimator of Lemma 6 with $\epsilon = \epsilon'/\sqrt{4 + \frac{1}{\theta^2}}$ and $\delta' = \delta$ and outputs whatever estimates $\hat{\mu}$ and $\hat{\sigma}^2$ the algorithm of Lemma 6 produces.

2.3 Hypothesis testing

Our hypothesis testing routine `Choose-Hypothesis`^{X} runs a simple "competition" to choose a winner between two candidate hypothesis distributions H_1 and H_2 over $[n]$ that it is given in the input either explicitly, or in some succinct way. We show that if at least one of the two candidate hypotheses is close to the target distribution X, then with high probability over the samples drawn from X the routine selects as winner a candidate that is close to X. This basic approach of running a competition between candidate hypotheses is quite similar to the "Scheffé estimate" proposed by Devroye and Lugosi (see [DL96b, DL96a] and Chapter 6 of [DL01]), which in turn built closely on the work of [Yat85], but there are some small differences between our approach and theirs; the [DL01] approach uses a notion of the "competition" between two hypotheses which is not symmetric under swapping the two competing hypotheses, whereas our competition is symmetric. We obtain the following lemma, postponing all running-time analysis to the next section.

Lemma 8 *There is an algorithm* `Choose-Hypothesis`$^X(H_1, H_2, \epsilon', \delta')$ *which is given oracle access to X, two hypothesis distributions H_1, H_2 for X, an accuracy parameter ϵ', and a confidence parameter δ'. It makes $m = O(\log(1/\delta')/\epsilon'^2)$ draws from X and returns some $H \in \{H_1, H_2\}$. If one of H_1, H_2 has $d_{\mathrm{TV}}(H_i, X) \leq \epsilon'$ then with probability $1 - \delta'$ the H that* `Choose-Hypothesis` *returns has $d_{\mathrm{TV}}(H, X) \leq 6\epsilon'$.*

Proof: Let $\mathcal{W}$ be the support of X. To set up the competition between H_1 and H_2, we define the following subset of $\mathcal{W}$:

$$\mathcal{W}_1 = \mathcal{W}_1(H_1, H_2) := \{w \in \mathcal{W} \mid H_1(w) > H_2(w)\}. \tag{5}$$

Let then $p_1 = H_1(\mathcal{W}_1)$ and $q_1 = H_2(\mathcal{W}_1)$. Clearly, $p_1 > q_1$ and $d_{\mathrm{TV}}(H_1, H_2) = p_1 - q_1$.
   The competition between H_1 and H_2 is carried out as follows:

1. If $p_1 - q_1 \leq 5\epsilon'$, declare a draw and return either H_i. Otherwise:

2. Draw $m = O\left(\frac{\log(1/\delta')}{\epsilon'^2}\right)$ samples $s_1, \ldots, s_m$ from X, and let $\tau = \frac{1}{m}|\{i \mid s_i \in \mathcal{W}_1\}|$ be the fraction of samples that fall inside $\mathcal{W}_1$.

3. If $\tau > p_1 - \frac{3}{2}\epsilon'$, declare H_1 as winner and return H_1; otherwise,

4. if $\tau < q_1 + \frac{3}{2}\epsilon'$, declare H_2 as winner and return H_2; otherwise,

5. declare a draw and return either H_i.

It is not hard to check that the outcome of the competition does not depend on the ordering of the pair of distributions provided in the input; that is, on inputs (H_1, H_2) and (H_2, H_1) the competition outputs the same result for a fixed sequence of samples $s_1, \ldots, s_m$ drawn from X.
   The correctness of `Choose-Hypothesis` is an immediate consequence of the following claim. (In fact for Lemma 8 we only need item (i) below, but item (ii) will be handy later in the proof of Lemma 11.)

Claim 9 *Suppose that $d_{\mathrm{TV}}(X, H_1) \leq \epsilon'$. Then:*

(i) If $d_{\mathrm{TV}}(X, H_2) > 6\epsilon'$, then the probability that the competition between H_1 and H_2 does not declare H_1 as the winner is at most $e^{-m\epsilon'^2/2}$. (Intuitively, if H_2 is very bad then it is very likely that H_1 will be declared winner.)

(ii) If $d_{\mathrm{TV}}(X, H_2) > 4\epsilon'$, the probability that the competition between H_1 and H_2 declares H_2 as the winner is at most $e^{-m\epsilon'^2/2}$. (Intuitively, if H_2 is only moderately bad then a draw is possible but it is very unlikely that H_2 will be declared winner.)

Proof: Let $r = X(\mathcal{W}_1)$. The definition of the total variation distance implies that $|r - p_1| \leq \epsilon'$. Let us define the 0/1 (indicator) random variables $\{Z_j\}_{j=1}^m$ as $Z_j = 1$ iff $s_j \in \mathcal{W}_1$. Clearly, $\tau = \frac{1}{m}\sum_{j=1}^m Z_j$ and $\mathbb{E}[\tau] = \mathbb{E}[Z_j] = r$. Since the Z_j's are mutually independent, it follows from the Chernoff bound that $\Pr[\tau \leq r - \epsilon'/2] \leq e^{-m\epsilon'^2/2}$. Using $|r - p_1| \leq \epsilon'$ we get that $\Pr[\tau \leq p_1 - 3\epsilon'/2] \leq e^{-m\epsilon'^2/2}$. Hence:

- For part (i): If $d_{\mathrm{TV}}(X, H_2) > 6\epsilon'$, from the triangle inequality we get that $p_1 - q_1 = d_{\mathrm{TV}}(H_1, H_2) > 5\epsilon'$. Hence, the algorithm will go beyond step 1, and with probability at least $1 - e^{-m\epsilon'^2/2}$, it will stop at step 3, declaring H_1 as the winner of the competition between H_1 and H_2.

- For part (ii): If $p_1 - q_1 \leq 5\epsilon'$ then the competition declares a draw, hence H_2 is not the winner. Otherwise we have $p_1 - q_1 > 5\epsilon'$ and the above arguments imply that the competition between H_1 and H_2 will declare H_2 as the winner with probability at most $e^{-m\epsilon'^2/2}$.

This concludes the proof of Claim 9 and of Lemma 8. ∎

2.4 Proof of Theorem 1

We first treat Part (1) of the theorem, where the learning algorithm may output any distribution over $[n]$ and not necessarily a PBD. Our algorithm has the structure outlined in Figure 1 with the following modifications: (a) if the target total variation distance is ϵ, the second argument of both `Learn-Sparse` and `Learn-Poisson` is set to $\frac{\epsilon}{12\max\{c_1, c_2\}}$, where c_1 and c_2 are respectively the constants from Lemmas 3 and 5; (b) we replace the third step with `Choose-Hypothesis`$^X(H_S, \widehat{H_P}, \epsilon/8, \delta/3)$, where $\widehat{H_P}$ is defined in terms of H_P as described below. If `Choose-Hypothesis` returns H_S, then `Learn-PBD` also returns H_S, while if `Choose-Hypothesis` returns $\widehat{H_P}$, then `Learn-PBD` returns H_P. We proceed to the definition of $\widehat{H_P}$.

Definition of $\widehat{H_P}$: For every point i where $H_S(i) = 0$, we let $\widehat{H_P}(i) = H_P(i)$. For the points i where $H_S(i) \neq 0$, in Theorem 7 of Section 6 we describe an efficient deterministic algorithm that numerically approximates $H_P(i)$ to within an additive $\pm\epsilon/24s$, where $s = O(1/\epsilon^3)$ is the cardinality of the support of H_S. We define $\widehat{H_P}(i)$ to equal the approximation to $H_P(i)$ that is output by the algorithm of Theorem 7. Observe that $\widehat{H_P}$ satisfies $d_{\mathrm{TV}}(\widehat{H_P}, H_P) \leq \epsilon/24$, and therefore $|d_{\mathrm{TV}}(\widehat{H_P}, X) - d_{\mathrm{TV}}(X, H_P)| \leq \epsilon/24$. In particular, if $d_{\mathrm{TV}}(X, H_P) \leq \frac{\epsilon}{12}$, then $d_{\mathrm{TV}}(X, \widehat{H_P}) \leq \frac{\epsilon}{8}$, and if $d_{\mathrm{TV}}(X, \widehat{H_P}) \leq \frac{6\epsilon}{8}$, then $d_{\mathrm{TV}}(X, H_P) \leq \epsilon$.

We do not use H_P directly in `Choose-Hypothesis` because of computational considerations. Since H_P is a translated Poisson distribution, we cannot compute its values $H_P(i)$ exactly, but using approximate values may cause `Choose-Hypothesis` to make a mistake. So we use $\widehat{H_P}$ instead of H_P in `Choose-Hypothesis`; $\widehat{H_P}$ is carefully designed both to be close enough to H_P so that `Choose-Hypothesis` will select a probability distribution close to the target X, and to allow efficient computation of all probabilities that `Choose-Hypothesis` needs without much overhead. In particular, we remark that in running `Choose-Hypothesis` we do not a priori compute the value of $\widehat{H_P}$ at every point; we do instead a lazy evaluation of $\widehat{H_P}$, as explained in the running-time analysis below.

We proceed now to the analysis of our modified algorithm `Learn-PBD`. The sample complexity bound and correctness of our algorithm are immediate consequences of Lemmas 3, 5 and 8, taking into account the precise choice of constants and the distance between H_P and $\widehat{H_P}$. To bound the running time, Lemmas 3 and 5 bound the running time of Steps 1 and 2 of the algorithm, so it remains to bound the running time of

the `Choose-Hypothesis` step. Notice that $\mathcal{W}_1(H_S, \widehat{H_P})$ is a subset of the support of the distribution H_S. Hence to compute $\mathcal{W}_1(H_S, \widehat{H_P})$ it suffices to determine the probabilities $H_S(i)$ and $\widehat{H_P}(i)$ for every point i in the support of H_S. For every such i, $H_S(i)$ is explicitly given in the output of `Learn-Sparse`, so we only need to compute $\widehat{H_P}(i)$. Theorem 7 implies that the time needed to compute $\widehat{H_P}(i)$ is $\tilde{O}(\log^3(1/\epsilon) + \log n + |\hat{\mu}| + |\hat{\sigma}^2|)$, where $|\hat{\mu}|$ and $|\hat{\sigma}^2|$ are respectively the description complexities (bit lengths) of $\hat{\mu}$ and $\hat{\sigma}^2$. Since these parameters are output by `Learn-Poisson`, by inspection of that algorithm it is easy to see that they are each at most $O(\log n + \log\log(1/\delta) + \log(1/\epsilon))$. Hence, given that the support of H_S has cardinality $O(1/\epsilon^3)$, the overall time spent computing all probabilities under $\widehat{H_P}$ is $\tilde{O}(\frac{1}{\epsilon^3}\log n \log\frac{1}{\delta})$. After $\mathcal{W}_1$ is computed, the computation of the values $p_1 = H_S(\mathcal{W}_1)$, $q_1 = \widehat{H_P}(\mathcal{W}_1)$ and $p_1 - q_1$ takes time linear in the data produced by the algorithm so far, as these computations merely involve adding and subtracting probabilities that have already been explicitly computed by the algorithm. Computing the fraction of samples from X that fall inside $\mathcal{W}_1$ takes time $O\left(\log n \cdot \log(1/\delta)/\epsilon^2\right)$ and the rest of `Choose-Hypothesis` takes time linear in the size of the data that have been written down so far. Hence the overall running time of our algorithm is $\tilde{O}(\frac{1}{\epsilon^3}\log n \log\frac{1}{\delta})$. This gives Part (1) of Theorem 1.

Next we turn to Part (2) of Theorem 1, the proper learning result. We explain how to modify the algorithm of Part (1) to produce a PBD that is within $O(\epsilon)$ of the target X. We only need to add two post-processing steps converting H_S and H_P to PBDs; we describe and analyze these two steps below. For convenience we write c to denote $\max\{c_1, c_2\} \geq 1$ in the following discussion.

1. `Locate-Sparse`$(H_S, \frac{\epsilon}{12c})$: This routine searches the sparse-form PBDs inside the cover $\mathcal{S}_{\frac{\epsilon}{12c}}$ to identify a sparse-form PBD that is within distance $\frac{\epsilon}{6}$ from H_S, or outputs "fail" if it cannot find one. Note that if there is a sparse-form PBD Y that is $\frac{\epsilon}{12c}$-close to X and `Learn-Sparse` succeeds, then Y must be $\frac{\epsilon}{6}$-close to H_S, since by Lemma 3 whenever `Learn-Sparse` succeeds the output distribution satisfies $d_{TV}(X, H_S) \leq \frac{\epsilon}{12}$. We show that if there is a sparse-form PBD Y that is $\frac{\epsilon}{12c}$-close to X and `Learn-Sparse` succeeds (an event that occurs with probability $1 - \delta/3$, see Lemma 3), our `Locate-Sparse` search routine, described below, will output a sparse-form PBD that is $\frac{\epsilon}{6}$-close to H_S. Indeed, given the preceding discussion, if we searched over all sparse-form PBDs inside the cover, it would be trivial to meet this guarantee. To save on computation time, we prune the set of sparse-form PBDs we search over, completing the entire search in time $\left(\frac{1}{\epsilon}\right)^{O(\log^2 1/\epsilon)} \log n \log 1/\delta$.

   Here is a detailed explanation and run-time analysis of the improved search: First, note that the description complexity of H_S is $\text{poly}(1/\epsilon) \cdot \tilde{O}(\log n \log(1/\delta))$ as H_S is output by an algorithm with this running time. Moreover, given a sparse-form PBD in $\mathcal{S}_{\frac{\epsilon}{12c}}$, we can compute all probabilities in the support of the distribution in time $\text{poly}(1/\epsilon) \log n$. Indeed, by part (i) of Theorem 4 a sparse-form PBD has $O(1/\epsilon^3)$ non-trivial Bernoulli random variables and those each use probabilities p_i that are integer multiples of some value which is $\Omega(\epsilon^2)$. So an easy dynamic programming algorithm can compute all probabilities in the support of the distribution in time $\text{poly}(1/\epsilon) \log n$, where the $\log n$ overhead is due to the fact that the support of the distribution is some interval in $[n]$. Finally, we argue that we can restrict our search to only a small subset of the sparse-form PBDs in $\mathcal{S}_{\frac{\epsilon}{12c}}$. For this, we note that we can restrict our search to sparse-form PBDs whose support is a superset of the support of H_S. Indeed, the final statement of Lemma 3 implies that, if Y is an arbitrary sparse-form PBD that is $\frac{\epsilon}{12c}$-close to X, then with probability $1 - \delta/3$ the output H_S of `Learn-Sparse` will have support that is a subset of the support of Y. Given this, we only need to try $\left(\frac{1}{\epsilon}\right)^{O(\log^2 1/\epsilon)}$ sparse-form PBDs in the cover to find one that is close to H_S. Hence, the overall running time of our search is $\left(\frac{1}{\epsilon}\right)^{O(\log^2 1/\epsilon)} \tilde{O}(\log n \log 1/\delta)$.

2. `Locate-Binomial`$(\hat{\mu}, \hat{\sigma}^2, n)$: This routine tries to compute a Binomial distribution that is $O(\epsilon)$-close to H_P (recall that $H_P \equiv TP(\hat{\mu}, \hat{\sigma}^2)$). Analogous to `Locate-Sparse`, we will show that if X is not $\frac{\epsilon}{12c}$-close to any sparse-form distribution inside $\mathcal{S}_{\frac{\epsilon}{12c}}$ and `Learn-Poisson` succeeds (for convenience

we call these conditions our "working assumptions" in the following discussion), then the Binomial distribution output by our routine will be $O(\epsilon)$-close to H_P and thus $O(\epsilon)$-close to X.

Let $\hat{\mu}$ and $\hat{\sigma}^2$ be the parameters output by Learn-Poisson, and let μ and σ^2 be the (unknown) mean and variance of the target X. Our routine has several steps. The first two steps eliminate corner-cases in the values $\hat{\mu}$ and $\hat{\sigma}^2$ computed by Learn-Poisson, while the last step defines a Binomial distribution $B(\hat{n}, \hat{p})$ with $\hat{n} \le n$ that is close to $H_P \equiv TP(\hat{\mu}, \hat{\sigma}^2)$ under our working assumptions. (We note that a significant portion of the work below is to ensure that $\hat{n} \le n$, which does not seem to follow from a more direct approach. Getting $\hat{n} \le n$ is necessary in order for our learning algorithm for order-n PBDs to truly be proper.) Throughout (a), (b) and (c) below we assume that our working assumptions hold (note that this assumption is being used every time we employ results such as (1) or (2) from Section 2.2).

(a) Tweaking $\hat{\sigma}^2$: If $\hat{\sigma}^2 \le \frac{n}{4}$ then set $\sigma_1^2 = \hat{\sigma}^2$, and otherwise set $\sigma_1^2 = \frac{n}{4}$. We note for future reference that in both cases Equation (2) gives

$$\sigma_1^2 \le (1 + O(\epsilon))\sigma^2. \tag{6}$$

We claim that this setting of σ_1^2 results in $d_{\mathrm{TV}}(TP(\hat{\mu}, \hat{\sigma}^2), TP(\hat{\mu}, \sigma_1^2)) \le O(\epsilon)$. If $\hat{\sigma}^2 \le \frac{n}{4}$ then this variation distance is zero and the claim certainly holds. Otherwise we have the following (see Equation (2)):

$$\left(1 + \frac{\epsilon}{12c}\right)\sigma^2 \ge \hat{\sigma}^2 > \sigma_1^2 = \frac{n}{4} \ge \sum_{i=1}^{n} p_i(1 - p_i) = \sigma^2.$$

Hence, by Lemma 2 we get:

$$d_{\mathrm{TV}}(TP(\hat{\mu}, \hat{\sigma}^2), TP(\hat{\mu}, \sigma_1^2)) \le \frac{|\hat{\sigma}^2 - \sigma_1^2| + 1}{\sigma_1^2} \le \frac{O(\epsilon)\sigma^2 + 1}{\sigma^2} = O(\epsilon), \tag{7}$$

where we used the fact that $\sigma^2 = \Omega(1/\epsilon^2)$ (see (1)).

(b) Tweaking σ_1^2: If $\hat{\mu}^2 \le n(\hat{\mu} - \sigma_1^2)$ then set $\sigma_2^2 = \sigma_1^2$, and otherwise set $\sigma_2^2 = \frac{n\hat{\mu} - \hat{\mu}^2}{n}$. We claim that this results in $d_{\mathrm{TV}}(TP(\hat{\mu}, \sigma_1^2), TP(\hat{\mu}, \sigma_2^2)) \le O(\epsilon)$. If $\hat{\mu}^2 \le n(\hat{\mu} - \sigma_1^2)$ then as before the variation distance is zero and the claim holds. Otherwise, we observe that $\sigma_1^2 > \sigma_2^2$ and $\sigma_2^2 \ge 0$ (the last assertion follows from the fact that $\hat{\mu}$ must be at most n). So we have (see (2)) that

$$|\mu - \hat{\mu}| \le O(\epsilon)\sigma \le O(\epsilon)\mu, \tag{8}$$

which implies

$$n - \hat{\mu} \ge n - \mu - O(\epsilon)\sigma. \tag{9}$$

We now observe that

$$\mu^2 = \left(\sum_{i=1}^{n} p_i\right)^2 \le n\left(\sum_{i=1}^{n} p_i^2\right) = n(\mu - \sigma^2)$$

where the inequality is Cauchy-Schwarz. Rearranging this yields

$$\frac{\mu(n - \mu)}{n} \ge \sigma^2. \tag{10}$$

We now have that

$$\sigma_2^2 = \frac{\hat{\mu}(n - \hat{\mu})}{n} \ge \frac{(1 - O(\epsilon))\mu(n - \mu - O(\epsilon)\sigma)}{n} \ge (1 - O(\epsilon))\left(\sigma^2 - O(\epsilon)\sigma\right), \tag{11}$$

13

where the first inequality follows from (8) and (9) and the second follows from (10) and the fact that any PBD over n variables satisfies $\mu \le n$. Hence, by Lemma 2 we get:

$$d_{\mathrm{TV}}(TP(\hat{\mu}, \sigma_1^2), TP(\hat{\mu}, \sigma_2^2)) \le \frac{\sigma_1^2 - \sigma_2^2 + 1}{\sigma_2^2} \le \frac{(1 + O(\epsilon))\sigma^2 - (1 - O(\epsilon))\sigma^2 + O(\epsilon)\sigma + 1}{(1 - O(\epsilon))\sigma^2 - O(\epsilon)\sigma}$$

$$\le \frac{O(\epsilon)\sigma^2}{(1 - O(\epsilon))\sigma^2} = O(\epsilon), \tag{12}$$

where we used the bound $\sigma^2 = \Omega(1/\epsilon^2)$ (see (1)).

(c) Constructing a Binomial Distribution: We construct a Binomial distribution H_B that is $O(\epsilon)$-close to $TP(\hat{\mu}, \sigma_2^2)$. If we do this then we have $d_{\mathrm{TV}}(H_B, H_P) = O(\epsilon)$ by (7), (12) and the triangle inequality. The Binomial distribution H_B we construct is $\mathrm{Bin}(\hat{n}, \hat{p})$, where:

$$\hat{n} = \left\lfloor \frac{\hat{\mu}^2}{\hat{\mu} - \sigma_2^2} \right\rfloor \quad \text{and} \quad \hat{p} = \frac{\hat{\mu} - \sigma_2^2}{\hat{\mu}}.$$

Note that by the way σ_2^2 is set in step (b) above we indeed have $\hat{n} \le n$ as claimed in Part 2 of Theorem 1.

Let us bound the total variation distance between $\mathrm{Bin}(\hat{n}, \hat{p})$ and $TP(\hat{\mu}, \sigma_2^2)$. Using Lemma 1 we have:

$$d_{\mathrm{TV}}(\mathrm{Bin}(\hat{n}, \hat{p}), TP(\hat{n}\hat{p}, \hat{n}\hat{p}(1 - \hat{p}))) \le \frac{1}{\sqrt{\hat{n}\hat{p}(1 - \hat{p})}} + \frac{2}{\hat{n}\hat{p}(1 - \hat{p})}. \tag{13}$$

Notice that

$$\hat{n}\hat{p}(1 - \hat{p}) \ge \left(\frac{\hat{\mu}^2}{\hat{\mu} - \sigma_2^2} - 1 \right) \left(\frac{\hat{\mu} - \sigma_2^2}{\hat{\mu}} \right) \left(\frac{\sigma_2^2}{\hat{\mu}} \right) = \sigma_2^2 - \hat{p}(1 - \hat{p}) \ge (1 - O(\epsilon))\sigma^2 - 1 = \Omega(1/\epsilon^2),$$

where the next-to-last step used (11) and the last used the fact that $\sigma^2 = \Omega(1/\epsilon^2)$ (see (1). So plugging this into (13) we get:

$$d_{\mathrm{TV}}(\mathrm{Bin}(\hat{n}, \hat{p}), TP(\hat{n}\hat{p}, \hat{n}\hat{p}(1 - \hat{p}))) = O(\epsilon).$$

The next step is to compare $TP(\hat{n}\hat{p}, \hat{n}\hat{p}(1 - \hat{p}))$ and $TP(\hat{\mu}, \sigma_2^2)$. Lemma 2 gives:

$$d_{\mathrm{TV}}(TP(\hat{n}\hat{p}, \hat{n}\hat{p}(1 - \hat{p})), TP(\hat{\mu}, \sigma_2^2)) \le \frac{|\hat{n}\hat{p} - \hat{\mu}|}{\min(\sqrt{\hat{n}\hat{p}(1 - \hat{p})}, \sigma_2)} + \frac{|\hat{n}\hat{p}(1 - \hat{p}) - \sigma_2^2| + 1}{\min(\hat{n}\hat{p}(1 - \hat{p}), \sigma_2^2)}$$

$$\le \frac{1}{\sqrt{\hat{n}\hat{p}(1 - \hat{p})}} + \frac{2}{\hat{n}\hat{p}(1 - \hat{p})} = O(\epsilon).$$

By the triangle inequality we get $d_{\mathrm{TV}}(\mathrm{Bin}(\hat{n}, \hat{p}), TP(\hat{\mu}, \sigma_2^2) = O(\epsilon)$, which was our ultimate goal.

Given the above `Locate-Sparse` and `Locate-Binomial` routines, the algorithm `Proper-Learn-PBD` has the following structure: It first runs `Learn-PBD` with accuracy parameters ϵ, δ. If `Learn-PBD` returns the distribution H_S computed by subroutine `Learn-Sparse`, then `Proper-Learn-PBD` outputs the result of `Locate-Sparse`$(H_S, \frac{\epsilon}{12c})$. If, on the other hand, `Learn-PBD` returns the translated Poisson distribution $H_P = TP(\hat{\mu}, \hat{\sigma}^2)$ computed by subroutine `Learn-Poisson`, then `Proper-Learn-PBD` returns the Binomial distribution constructed by the routine `Locate-Binomial`$(\hat{\mu}, \hat{\sigma}^2, n)$. It follows from the correctness of `Learn-PBD` and the above discussion that, with probability $1 - \delta$, the output of `Proper-Learn-PBD` is within total variation distance $O(\epsilon)$ of the target X. The number of samples is the same as in `Learn-PBD`, and the running time is $\left(\frac{1}{\epsilon} \right)^{O(\log^2 1/\epsilon)} \cdot \tilde{O}(\log n \log 1/\delta)$.

This concludes the proof of Part 2 of Theorem 1, and thus of the entire theorem. ∎

3   Learning weighted sums of independent Bernoullis

In this section we consider a generalization of the problem of learning an unknown PBD, by studying the learnability of weighted sums of independent Bernoulli random variables $X = \sum_{i=1}^{n} w_i X_i$. (Throughout this section we assume for simplicity that the weights are "known" to the learning algorithm.) In Section 3.1 we show that if there are only constantly many different weights then such distributions can be learned by an algorithm that uses $O(\log n)$ samples and runs in time $\mathrm{poly}(n)$. In Section 3.2 we show that if there are n distinct weights then even if those weights have an extremely simple structure – the i-th weight is simply i – any algorithm must use $\Omega(n)$ samples.

3.1   Learning sums of weighted independent Bernoulli random variables with few distinct weights

Recall Theorem 2:

Theorem 2 *Let $X = \sum_{i=1}^{n} a_i X_i$ be a weighted sum of unknown independent Bernoulli random variables such that there are at most k different values in the set $\{a_1, \ldots, a_n\}$. Then there is an algorithm with the following properties: given $n, a_1, \ldots, a_n$ and access to independent draws from X, it uses $\log(n) \cdot \widetilde{O}(k \cdot \epsilon^{-2}) \cdot \log(1/\delta)$ samples from the target distribution X, runs in time $\mathrm{poly}(n^k \cdot (k/\epsilon)^{k \log^2(k/\epsilon)}) \cdot \log(1/\delta)$, and with probability $1 - \delta$ outputs a hypothesis vector $\hat{p} \in [0, 1]^n$ defining independent Bernoulli random variables $\hat{X}_i$ with $\mathbf{E}[\hat{X}_i] = p_i$ such that $d_{\mathrm{TV}}(\hat{X}, X) \leq \epsilon$, where $\hat{X} = \sum_{i=1}^{n} a_i \hat{X}_i$.*

Given a vector $\overline{a} = (a_1, \ldots, a_n)$ of weights, we refer to a distribution $X = \sum_{i=1}^{n} a_i X_i$ (where $X_1, \ldots, X_n$ are independent Bernoullis which may have arbitrary means) as an $\overline{a}$-*weighted sum of Bernoullis*, and we write $\mathcal{S}_{\overline{a}}$ to denote the space of all such distributions.

To prove Theorem 2 we first show that $\mathcal{S}_{\overline{a}}$ has an ϵ-cover that is not too large. We then show that by running a "tournament" between all pairs of distributions in the cover, using the hypothesis testing subroutine from Section 2.3, it is possible to identify a distribution in the cover that is close to the target $\overline{a}$-weighted sum of Bernoullis.

Lemma 10 *There is an ϵ-cover $\mathcal{S}_{\overline{a}, \epsilon} \subset \mathcal{S}_{\overline{a}}$ of size $|\mathcal{S}_{\overline{a}, \epsilon}| \leq (n/k)^{3k} \cdot (k/\epsilon)^{k \cdot O(\log^2(k/\epsilon))}$ that can be constructed in time $\mathrm{poly}(|\mathcal{S}_{\overline{a}, \epsilon}|)$.*

Proof: Let $\{b_j\}_{j=1}^{k}$ denote the set of distinct weights in $a_1, \ldots, a_n$, and let $n_j = \left|\{i \in [n] \mid a_i = b_j\}\right|$. With this notation, we can write $X = \sum_{j=1}^{k} b_j S_j = g(S)$, where $S = (S_1, \ldots, S_k)$ with each S_j a sum of n_j many independent Bernoulli random variables and $g(y_1, \ldots, y_k) = \sum_{j=1}^{k} b_j y_j$. Clearly we have $\sum_{j=1}^{k} n_j = n$. By Theorem 4, for each $j \in \{1, \ldots, k\}$ the space of all possible S_j's has an explicit (ϵ/k)-cover $\mathcal{S}_{\epsilon/k}^{j}$ of size $|\mathcal{S}_{\epsilon/k}^{j}| \leq n_j^3 \cdot O(k/\epsilon) + n \cdot (k/\epsilon)^{O(\log^2(k/\epsilon))}$. By independence across S_j's, the product $\mathcal{Q} = \prod_{j=1}^{k} \mathcal{S}_{\epsilon/k}^{j}$ is an ϵ-cover for the space of all possible S's, and hence the set

$$\{Q = \sum_{j=1}^{k} b_j S_j \; : \; (S_1, \ldots, S_k) \in \mathcal{Q}\}$$

is an ϵ-cover for $\mathcal{S}_{\overline{a}}$. So $\mathcal{S}_{\overline{a}}$ has an explicit ϵ-cover of size $|\mathcal{Q}| = \prod_{j=1}^{k} |\mathcal{S}_{\epsilon/k}^{j}| \leq (n/k)^{3k} \cdot (k/\epsilon)^{k \cdot O(\log^2(k/\epsilon))}$.

(We note that a slightly stronger quantitative bound on the cover size can be obtained using Theorem 6 instead of Theorem 4, but the improvement is negligible for our ultimate purposes.) ■

Lemma 11 *Let $\mathcal{S}$ be any collection of distributions over a finite set. Suppose that $\mathcal{S}_\epsilon \subset \mathcal{S}$ is an ϵ-cover of $\mathcal{S}$ of size N. Then there is an algorithm that uses $O(\epsilon^{-2} \log N \log(1/\delta))$ samples from an unknown target distribution $X \in \mathcal{S}$ and with probability $1 - \delta$ outputs a distribution $Z \in \mathcal{S}_\epsilon$ that satisfies $d_{\mathrm{TV}}(X, Z) \leq 6\epsilon$.*

15

Devroye and Lugosi (Chapter 7 of [DL01]) prove a similar result by having all pairs of distributions in the cover compete against each other using their notion of a competition, but again there are some small differences: their approach chooses a distribution in the cover which wins the maximum number of competitions, whereas our algorithm chooses a distribution that is never defeated (i.e. won or achieved a draw against all other distributions in the cover).

Proof: The algorithm performs a tournament by running the competition $\texttt{Choose-Hypothesis}^X(H_i, H_j, \epsilon, \delta/(2N))$ for every pair of distinct distributions H_i, H_j in the cover $\mathcal{S}_\epsilon$. It outputs a distribution $Y^\star \in \mathcal{S}_\epsilon$ that was never a loser (i.e. won or achieved a draw in all its competitions). If no such distribution exists in $\mathcal{S}_\epsilon$ then the algorithm outputs "failure."

Since $\mathcal{S}_\epsilon$ is an ϵ-cover of $\mathcal{S}$, there exists some $Y \in \mathcal{S}_\epsilon$ such that $d_{TV}(X, Y) \le \epsilon$. We first argue that with high probability this distribution Y never loses a competition against any other $Y' \in \mathcal{S}_\epsilon$ (so the algorithm does not output "failure"). Consider any $Y' \in \mathcal{S}_\epsilon$. If $d_{TV}(X, Y') > 4\epsilon$, by Lemma 9(ii) the probability that Y loses to Y' is at most $2e^{-m\epsilon^2/2} = O(1/N)$. On the other hand, if $d_{TV}(X, Y') \le 4\delta$, the triangle inequality gives that $d_{TV}(Y, Y') \le 5\epsilon$ and thus Y draws against Y'. A union bound over all N distributions in $\mathcal{S}_\epsilon$ shows that with probability $1 - \delta/2$, the distribution Y never loses a competition.

We next argue that with probability at least $1 - \delta/2$, every distribution $Y' \in \mathcal{S}_\epsilon$ that never loses has Y' close to X. Fix a distribution Y' such that $d_{TV}(Y', X) > 6\epsilon$; Lemma 9(i) implies that Y' loses to Y with probability $1 - 2e^{-m\epsilon^2/2} \ge 1 - \delta/(2N)$. A union bound gives that with probability $1 - \delta/2$, every distribution Y' that has $d_{TV}(Y', X) > 6\epsilon$ loses some competition.

Thus, with overall probability at least $1 - \delta$, the tournament does not output "failure" and outputs some distribution $Y^\star$ such that $d_{TV}(X, Y^\star)$ is at most 6ϵ. This proves the lemma. ∎

Proof of Theorem 2: We claim that the algorithm of Lemma 11 has the desired sample complexity and can be implemented to run in the claimed time bound. The sample complexity bound follows directly from Lemma 11. It remains to argue about the time complexity. Note that the running time of the algorithm is $\mathrm{poly}(|\mathcal{S}_{\overline{a},\epsilon}|)$ times the running time of a competition. We will show that a competition between $H_1, H_2 \in \mathcal{S}_{\overline{a},\epsilon}$ can be carried out by an efficient algorithm. This amounts to efficiently computing the probabilities $p_1 = H_1(\mathcal{W}_1)$ and $q_1 = H_2(\mathcal{W}_1)$. Note that $\mathcal{W} = \sum_{j=1}^k b_i \cdot \{0, 1, \dots, n_j\}$. Clearly, $|\mathcal{W}| \le \prod_{j=1}^k (n_j + 1) = O((n/k)^k)$. It is thus easy to see that p_1, q_1 can be efficiently computed as long as there is an efficient algorithm for the following problem: given $H = \sum_{j=1}^k b_j S_j \in \mathcal{S}_{\overline{a},\epsilon}$ and $w \in \mathcal{W}$, compute $H(w)$. Indeed, fix any such H, w. We have that

$$H(w) = \sum_{m_1, \dots, m_k} \prod_{j=1}^k \Pr_H[S_j = m_j],$$

where the sum is over all k-tuples $(m_1, \dots, m_k)$ such that $0 \le m_j \le n_j$ for all j and $b_1 m_1 + \dots + b_k m_k = w$ (as noted above there are at most $O((n/k)^k)$ such k-tuples). To complete the proof of Theorem 2 we note that $\Pr_H[S_j = m_j]$ can be computed in $O(n_j^2)$ time by standard dynamic programming. ∎

We close this subsection with the following remark: In recent work [DDS11] the authors have given a $\mathrm{poly}(\ell, \log(n), 1/\epsilon)$-time algorithm that learns any ℓ-modal distribution over $[n]$ (i.e. a distribution whose pdf has at most ℓ "peaks" and "valleys") using $O(\ell \log(n)/\epsilon^3 + (\ell/\epsilon)^3 \log(\ell/\epsilon) \log\log(\ell/\epsilon))$ samples. It is natural to wonder whether this algorithm could be used to efficiently learn a sum of n weighted independent Bernoulli random variables with k distinct weights, and thus give an alternate algorithm for Theorem 2, perhaps with better asymptotic guarantees. However, it is easy to construct a sum $X = \sum_{i=1}^n a_i X_i$ of n weighted independent Bernoulli random variables with k distinct weights such that X is 2^k-modal. Thus a naive application of the [DDS11] result would only give an algorithm with sample complexity exponential in k, rather than the quasi-linear sample complexity of our current algorithm. If the 2^k-modality of the above-mentioned example is the worst case (which we do not know), then the [DDS11] algorithm would give a $\mathrm{poly}(2^k, \log(n), 1/\epsilon)$-time algorithm for our problem that uses $O(2^k \log(n)/\epsilon^3) + 2^{O(k)} \cdot \tilde{O}(1/\epsilon^3)$ examples (so comparing with Theorem 2, exponentially worse sample complexity as a function of k, but exponentially better running time as a function of

n). Finally, in the context of this question (how many modes can there be for a sum of n weighted independent Bernoulli random variables with k distinct weights), it is interesting to recall the result of K.-I. Sato [Sat93] which shows that for any N there are two unimodal distributions X, Y such that $X + Y$ has at least N modes.

3.2 Sample complexity lower bound for learning sums of weighted independent Bernoullis

Recall Theorem 3:

Theorem 3 *Let $X = \sum_{i=1}^{n} i \cdot X_i$ be a weighted sum of unknown independent Bernoulli random variables (where the i-th weight is simply i). Let L be any learning algorithm which, given n and access to independent draws from X, outputs a hypothesis distribution $\hat{X}$ such that $d_{\mathrm{TV}}(\hat{X}, X) \le 1/25$ with probability at least $e^{-o(n)}$. Then L must use $\Omega(n)$ samples.*

Proof of Theorem 3: We define a probability distribution over possible target probability distributions X as follows: A subset $S \subset \{n/2 + 1, \dots, n\}$ of size $|S| = n/100$ is drawn uniformly at random from all $\binom{n/2}{n/100}$ possible outcomes.. The vector $\overline{p} = (p_1, \dots, p_n)$ is defined as follows: for each $i \in S$ the value p_i equals $100/n = 1/|S|$, and for all other i the value p_i equals 0. The i-th Bernoulli random variable X_i has $\mathbf{E}[X_i] = p_i$, and the target distribution is $X = X_{\overline{p}} = \sum_{i=1}^{n} iX_i$.

We will need two easy lemmas:

Lemma 12 *Fix any $S, \overline{p}$ as described above. For any $j \in \{n/2 + 1, \dots, n\}$ we have $X_{\overline{p}}(j) \ne 0$ if and only if $j \in S$. For any $j \in S$ the value $X_{\overline{p}}(j)$ is exactly $(100/n)(1 - 100/n)^{n/100 - 1} > 35/n$ (for n sufficiently large), and hence $X_{\overline{p}}(\{n/2 + 1, \dots, n\}) > 0.35$ (again for n sufficiently large).*

The first claim of the lemma holds because any set of $c \ge 2$ numbers from $\{n/2 + 1, \dots, n\}$ must sum to more than n. The second claim holds because the only way a draw x from $X_{\overline{p}}$ can have $x = j$ is if $X_j = 1$ and all other X_i are 0 (here we are using $\lim_{x \to \infty}(1 - 1/x)^x = 1/e$).

The next lemma is an easy consequence of Chernoff bounds:

Lemma 13 *Fix any $\overline{p}$ as defined above, and consider a sequence of $n/2000$ independent draws from $X_{\overline{p}} = \sum_i iX_i$. With probability $1 - e^{-\Omega(n)}$ the total number of indices $j \in [n]$ such that X_j is ever 1 in any of the $n/2000$ draws is at most $n/1000$.*

We are now ready to prove Theorem 3. Let L be a learning algorithm that receives $n/2000$ samples. Let $S \subset \{n/2 + 1, \dots, n\}$ and $\overline{p}$ be chosen randomly as defined above, and set the target to $X = X_{\overline{p}}$.

We consider an augmented learner L' that is given "extra information." For each point in the sample, instead of receiving the value of that draw from X the learner L' is given the entire vector $(X_1, \dots, X_n) \in \{0, 1\}^n$. Let T denote the set of elements $j \in \{n/2 + 1, \dots, n\}$ for which the learner is ever given a vector $(X_1, \dots, X_n)$ that has $X_j = 1$. By Lemma 13 we have $|T| \le n/1000$ with probability at least $1 - e^{-\Omega(n)}$; we condition on the event $|T| \le n/1000$ going forth.

Fix any value $\ell \le n/1000$. Conditioned on $|T| = \ell$, the set T is equally likely to be any ℓ-element subset of S, and all possible "completions" of T with an additional $n/100 - \ell \ge 9n/1000$ elements of $\{n/2 + 1, \dots, n\} \setminus T$ are equally likely to be the true set S.

Let H denote the hypothesis distribution over $[n]$ that algorithm L outputs. Let R denote the set $\{n/2 + 1, \dots, n\} \setminus T$; note that since $|T| = \ell \le n/1000$, we have $|R| \ge 499n/1000$. Let U denote the set $\{i \in R : H(i) \ge 30/n\}$. Since H is a distribution we must have $|U| \le n/30$. Each element in $S \setminus U$ "costs" at least $5/n$ in variation distance between X and H. Since S is a uniform random extension of T with at most $n/100 - \ell \in [9n/1000, n/100]$ unknown elements of R and $|R| \ge 499n/1000$, an easy calculation shows that $\Pr[|S \setminus U| > 8n/1000]$ is $1 - e^{-\Omega(n)}$. This means that with probability $1 - e^{-\Omega(n)}$ we have $d_{\mathrm{TV}}(X, H) \ge \frac{8n}{1000} \cdot \frac{5}{n} = 1/25$, and the theorem is proved. ∎

4 Extensions of the Cover Theorem of [DP11]

4.1 Proof of Theorem 4

We only need to argue that the ϵ-covers constructed in [Das08] and [DP11] satisfy the part of the theorem following "finally;" we will refer to this part of the theorem as the *last part* in the following discussion. Moreover, in order to avoid reproducing here the involved constructions of [Das08] and [DP11], we will assume that the reader has some familiarity with these constructions. Nevertheless, we will try to make our proof self-contained.

First, we claim that we only need to establish the last part of Theorem 4 for the cover obtained in [Das08]. Indeed, the ϵ-cover of [DP11] is just a subset of the $\epsilon/2$-cover of [Das08], which includes only a subset of the sparse form distributions in the $\epsilon/2$-cover of [Das08]. Moreover, for every sparse form distribution in the $\epsilon/2$-cover of [Das08], the ϵ-cover of [DP11] includes at least one sparse form distribution that is $\epsilon/2$-close in total variation distance. Hence, if the $\epsilon/2$-cover of [Das08] satisfies the last part of Theorem 4, it follows that the ϵ-cover of [DP11] also satisfies the last part of Theorem 4.

We proceed to argue that the cover of [Das08] satisfies the last part of Theorem 4. The construction of the ϵ-cover in [Das08] works roughly as follows: Given an arbitrary collection of indicators $\{X_i\}_{i=1}^n$ with expectations $\mathbf{E}[X_i] = p_i$ for all i, the collection is subjected to two filters, called the *Stage 1* and the *Stage 2* filters (see respectively Sections 5 and 6 of [Das08]). Using the same notation as [Das08] let us denote by $\{Z_i\}_i$ the collection output by the Stage 1 filter and by $\{Y_i\}_i$ the collection output by the Stage 2 filter. The collection output by the Stage 2 filter is included in the ϵ-cover of [Das08], satisfies that $d_{\mathrm{TV}}(\sum_i X_i, \sum_i Y_i) \leq \epsilon$, and is in either the heavy Binomial or the sparse form.

Let (μ_Z, σ_Z^2) and (μ_Y, σ_Y^2) denote respectively the (mean, variance) pairs of the variables $Z = \sum_i Z_i$ and $Y = \sum_i Y_i$. We argue first that the pair (μ_Z, σ_Z^2) satisfies $|\mu - \mu_Z| = O(\epsilon)$ and $|\sigma^2 - \sigma_Z^2| = O(\epsilon \cdot (1 + \sigma^2))$, where μ and σ^2 are respectively the mean and variance of $X = \sum_i X_i$. Next we argue that, if the collection $\{Y_i\}_i$ output by the Stage 2 filter is in heavy Binomial form, then (μ_Y, σ_Y^2) also satisfies $|\mu - \mu_Y| = O(\epsilon)$ and $|\sigma^2 - \sigma_Y^2| = O(1 + \epsilon \cdot (1 + \sigma^2))$.

- Proof for (μ_Z, σ_Z^2): The Stage 1 filter only modifies the indicators X_i with $p_i \in (0, 1/k) \cup (1 - 1/k, 1)$, for some well-chosen $k = O(1/\epsilon)$ (as in the statement of Theorem 4). For convenience let us define $\mathcal{L} = \{i \,|\, p_i \in (0, 1/k)\}$ and $\mathcal{H} = \{i \,|\, p_i \in (1 - 1/k, 1)\}$ as in [Das08]. The filter of Stage 1 rounds the expectations of the indicators indexed by $\mathcal{L}$ to some value in $\{0, 1/k\}$ so that no expectation is altered by more than an additive $1/k$, and the sum of these expectations is not modified by more than an additive $1/k$. Similarly, the expectations of the indicators indexed by $\mathcal{H}$ are rounded to some value in $\{1 - 1/k, 1\}$. See the details of how the rounding is performed in Section 5 of [Das08]. Let us then denote by $\{p_i'\}_i$ the expectations of the indicators $\{Z_i\}_i$ resulting from the rounding. We argue that the mean and variance of $Z = \sum_i Z_i$ is close to the mean and variance of X. Indeed,

$$|\mu - \mu_Z| = \left| \sum_i p_i - \sum_i p_i' \right| = \left| \sum_{i \in \mathcal{L} \cup \mathcal{H}} p_i - \sum_{i \in \mathcal{L} \cup \mathcal{H}} p_i' \right| \leq O(1/k) = O(\epsilon). \tag{14}$$

Similarly,

$$|\sigma^2 - \sigma_Z^2| = \left| \sum_i p_i(1 - p_i) - \sum_i p_i'(1 - p_i') \right|$$

$$= \left| \sum_{i \in \mathcal{L}} p_i(1 - p_i) - \sum_{i \in \mathcal{L}} p_i'(1 - p_i') \right| + \left| \sum_{i \in \mathcal{H}} p_i(1 - p_i) - \sum_{i \in \mathcal{H}} p_i'(1 - p_i') \right|.$$

We proceed to bound the two terms of the above summation separately. Since the argument is symmetric

18

for $\mathcal{L}$ and $\mathcal{H}$ we only do $\mathcal{L}$. We have

$$\left| \sum_{i \in \mathcal{L}} p_i(1 - p_i) - \sum_{i \in \mathcal{L}} p_i'(1 - p_i') \right| = \left| \sum_{i \in \mathcal{L}} (p_i - p_i')(1 - (p_i + p_i')) \right|$$

$$= \left| \sum_{i \in \mathcal{L}} (p_i - p_i') - \sum_{i \in \mathcal{L}} (p_i - p_i')(p_i + p_i') \right|$$

$$\leq \left| \sum_{i \in \mathcal{L}} (p_i - p_i') \right| + \left| \sum_{i \in \mathcal{L}} (p_i - p_i')(p_i + p_i') \right|$$

$$\leq \frac{1}{k} + \sum_{i \in \mathcal{L}} |p_i - p_i'|(p_i + p_i')$$

$$\leq \frac{1}{k} + \frac{1}{k} \sum_{i \in \mathcal{L}} (p_i + p_i')$$

$$\leq \frac{1}{k} + \frac{1}{k} \left(2 \sum_{i \in \mathcal{L}} p_i + 1/k \right)$$

$$\leq \frac{1}{k} + \frac{1}{k} \left(\frac{2}{1 - 1/k} \sum_{i \in \mathcal{L}} p_i(1 - 1/k) + 1/k \right)$$

$$\leq \frac{1}{k} + \frac{1}{k} \left(\frac{2}{1 - 1/k} \sum_{i \in \mathcal{L}} p_i(1 - p_i) + 1/k \right)$$

$$\leq \frac{1}{k} + \frac{1}{k^2} + \frac{2}{k - 1} \sum_{i \in \mathcal{L}} p_i(1 - p_i).$$

Using the above (and a symmetric argument for index set $\mathcal{H}$) we obtain:

$$|\sigma^2 - \sigma_Z^2| \leq \frac{2}{k} + \frac{2}{k^2} + \frac{2}{k - 1} \sigma^2 = O(\epsilon)(1 + \sigma^2). \tag{15}$$

- Proof for (μ_Y, σ_Y^2): After the Stage 1 filter is applied to the collection $\{X_i\}$, the resulting collection of random variables $\{Z_i\}$ has expectations $p_i' \in \{0, 1\} \cup [1/k, 1 - 1/k]$, for all i. The Stage 2 filter has different form depending on the cardinality of the set $\mathcal{M} = \{i \mid p_i' \in [1/k, 1 - 1/k]\}$. In particular, if $|\mathcal{M}| \geq k^3$ the output of the Stage 2 filter is in heavy Binomial form, while if if $|\mathcal{M}| < k^3$ the output of the Stage 2 filter is in sparse form. As we are only looking to provide a guarantee for the distributions in heavy Binomial form, it suffices to only consider the former case next.

  - $|\mathcal{M}| \geq k^3$: Let $\{Y_i\}$ be the collection produced by Stage 2 and let $Y = \sum_i Y_i$. Then Lemma 6.1 in [Das08] implies that

  $$|\mu_Z - \mu_Y| = O(\epsilon) \text{ and } |\sigma_Z^2 - \sigma_Y^2| = O(1).$$

  Combining this with (14) and (15) gives

  $$|\mu - \mu_Y| = O(\epsilon) \text{ and } |\sigma^2 - \sigma_Y^2| = O(1 + \epsilon \cdot (1 + \sigma^2)).$$

This concludes the proof of Theorem 4. ∎

4.2 Improved Version of Theorem 4

In our new improved version of the Cover Theorem, the k-heavy Binomial Form distributions in the cover are actually Binomial distributions $\text{Bin}(\ell, q)$ (rather than translated Binomial distributions as in the original version) for some $\ell \leq n$ and some q which is of the form (integer)$/\ell$ (rather than q of the form (integer)$/(kn)$ as in the original version). This gives an improved bound on the cover size. For clarity we state in full the improved version of Theorem 4 below:

Theorem 6 (Cover for PBDs, stronger version) *For all $\epsilon > 0$, there exists an ϵ-cover $\mathcal{S}_\epsilon \subseteq \mathcal{S}$ of $\mathcal{S}$ such that*

1. *$|\mathcal{S}_\epsilon| \leq n^2 + n \cdot \left(\frac{1}{\epsilon}\right)^{O(\log^2 1/\epsilon)}$; and*

2. *The set $\mathcal{S}_\epsilon$ can be constructed in time linear in its representation size, i.e. $\tilde{O}(n^2) + \tilde{O}(n) \cdot \left(\frac{1}{\epsilon}\right)^{O(\log^2 1/\epsilon)}$.*

Moreover, if $\{Z_i\} \in \mathcal{S}_\epsilon$, then the collection $\{Z_i\}$ has one of the following forms, where $k = k(\epsilon) \leq C/\epsilon$ is a positive integer, for some absolute constant $C > 0$:

(i) *(Sparse Form) There is a value $\ell \leq k^3 = O(1/\epsilon^3)$ such that for all $i \leq \ell$ we have $\mathbf{E}[Z_i] \in \left\{\frac{1}{k^2}, \frac{2}{k^2}, \ldots, \frac{k^2-1}{k^2}\right\}$, and for all $i > \ell$ we have $\mathbf{E}[Z_i] \in \{0, 1\}$.*

(ii) *(Binomial Form) There is a value $\bar{\ell} \in \{0, 1, \ldots, n\}$ and a value $\bar{q} \in \left\{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}\right\}$ such that for all $i \leq \bar{\ell}$ we have $\mathbf{E}[Z_i] = \bar{q}$; for all $i > \bar{\ell}$ we have $\mathbf{E}[Z_i] = 0$; and $\bar{\ell}, \bar{q}$ satisfy the bounds $\bar{\ell}\bar{q} \geq k^2 - 2 - \frac{1}{k}$ and $\bar{\ell}\bar{q}(1 - \bar{q}) \geq k^2 - k - 3 - \frac{3}{k}$.*

Finally, for every $\{X_i\} \in \mathcal{S}$ for which there is no ϵ-neighbor in $\mathcal{S}_\epsilon$ that is in sparse form, there exists a collection $\{Z_i\} \in \mathcal{S}_\epsilon$ in Binomial form such that

(iii) *$d_{\text{TV}}\left(\sum_i X_i, \sum_i Z_i\right) \leq \epsilon$; and*

(iv) *if $\mu = \mathbf{E}[\sum_i X_i]$, $\bar{\mu} = \mathbf{E}[\sum_i Z_i]$, $\sigma^2 = \text{Var}[\sum_i X_i]$ and $\bar{\sigma}^2 = \text{Var}[\sum_i Z_i]$, then $|\mu - \bar{\mu}| = 2 + O(\epsilon)$ and $|\sigma^2 - \bar{\sigma}^2| = O(1 + \epsilon \cdot (1 + \sigma^2))$.*

Proof: Suppose that $X = \{X_i\} \in \mathcal{S}$ is a PBD that is not ϵ_1-close to any Sparse Form PBD in the cover $\mathcal{S}_{\epsilon_1}$ of Theorem 4, where $\epsilon_1 = \Theta(\epsilon)$ is a suitable (small) constant multiple of ϵ (more on this below). Let μ, σ^2 denote the mean and variance of $\sum_i X_i$. Parts (iii) and (iv) of Theorem 4 imply that there is a collection $\{Y_i\} \in \mathcal{S}_{\epsilon_1}$ in k-heavy Binomial Form that is close to $\sum_i X_i$ both in variation distance and in its mean μ' and variance σ'^2. More precisely, let ℓ, q be the parameters defining $\{Y_i\}$ as in part (ii) of Theorem 4 and let μ', σ'^2 be the mean and variance of $\sum_i Y_i$; so we have $\mu' = \ell q + t$ for some integer $0 \leq t \leq n - \ell$ and $\sigma'^2 = \ell q(1 - q) \geq \Omega(1/\epsilon_1^2)$ from part (ii). This implies that the bounds $|\mu - \mu'| = O(\epsilon_1)$ and $|\sigma^2 - \sigma'^2| = O(1 + \epsilon_1 \cdot (1 + \sigma^2))$ of (iv) are at least as strong as the bounds given by Equation (2) (here we have used the fact that ϵ_1 is a suitably small constant multiple of ϵ), so we may use the analysis of Section 2.2. The analysis of Section 2.2 (Claim 7 and Lemma 2) gives that $d_{\text{TV}}(X, TP(\mu', \sigma'^2)) \leq O(\epsilon_1)$.

Now the analysis of `Locate-Binomial` (from Section 2.4) implies that $TP(\mu', \sigma'^2)$ is $O(\epsilon_1)$-close to a Binomial distribution $\text{Bin}(\hat{n}, \hat{p})$. We first observe that in Step 2.a of Section 2.4, the variance $\sigma'^2 = \ell q(1 - q)$ is at most $n/4$ and so the σ_1^2 that is defined in Step 2.a equals σ'^2. We next observe that by the Cauchy-Schwarz inequality we have $\mu'^2 \leq n(\mu' - \sigma'^2)$, and thus the value σ_2^2 defined in Step 2.b of Section 2.4 also equals σ'^2. Thus we have that the distribution $\text{Bin}(\hat{n}, \hat{p})$ resulting from `Locate-Binomial` is defined by

$$\hat{n} = \left\lfloor \frac{(\ell q + t)^2}{\ell q^2 + t} \right\rfloor \quad \text{and} \quad \hat{p} = \frac{\ell q^2 + t}{\ell q + t}.$$

So we have established that X is $O(\epsilon_1)$-close to the Binomial distribution $\text{Bin}(\hat{n}, \hat{p})$. We first establish that the parameters $\hat{n}, \hat{p}$ and the corresponding mean and variance $\hat{\mu} = \hat{n}\hat{p}$, $\hat{\sigma}^2 = \hat{n}\hat{p}(1 - \hat{p})$ satisfy the bounds

claimed in parts (ii) and (iv) of Theorem 6. To finally prove the theorem we will take $\bar{\ell} = \hat{n}$ and $\bar{q}$ to be $\hat{p}$ rounded to the nearest integer multiple of $1/n$, and we will show that the Binomial distribution $\mathrm{Bin}(\bar{\ell}, \bar{q})$ satisfies all the claimed bounds.

If $t = 0$ then it is easy to see that $\hat{n} = \ell$ and $\hat{p} = q$ and all the claimed bounds in parts (ii) and (iv) of Theorem 6 hold as desired for $\hat{n}, \hat{p}, \hat{\mu}$ and $\hat{\sigma}^2$. Otherwise $t \geq 1$ and we have

$$\hat{\mu} = \hat{n}\hat{p} \geq \left(\frac{(\ell q + t)^2}{\ell q^2 + t} - 1 \right) \cdot \left(\frac{\ell q^2 + t}{\ell q + t} \right) \geq \ell q + t - 1 \geq \ell q \geq k^2 - 1/k,$$

and similarly

$$\hat{\sigma}^2 = \hat{n}\hat{p}(1 - \hat{p}) \geq \left(\frac{(\ell q + t)^2}{\ell q^2 + t} - 1 \right) \cdot \left(\frac{\ell q^2 + t}{\ell q + t} \right) \cdot \left(\frac{\ell q - \ell q^2}{\ell q + t} \right) = \ell q(1 - q) - \hat{p}(1 - \hat{p}) \geq k^2 - k - 2 - \frac{3}{k},$$

so we have the bounds claimed in (ii). Similarly, we have

$$\mu' = \ell q + t = \left(\frac{(\ell q + t)^2}{\ell q^2 + t} \right) \cdot \left(\frac{\ell q^2 + t}{\ell q + t} \right) \geq \hat{\mu} = \hat{n}\hat{p} \geq \mu' - 1$$

so from part (iv) of Theorem 4 we get the desired bound $|\mu - \hat{\mu}| \leq 1 + O(\epsilon)$ of Theorem 6. Recalling that $\sigma'^2 = \ell q(1 - q)$, we have shown above that $\hat{\sigma}^2 \geq \sigma'^2 - 1$; we now observe that

$$\sigma'^2 = \left(\frac{(\ell q + t)^2}{\ell q^2 + t} \right) \cdot \left(\frac{\ell q^2 + t}{\ell q + t} \right) \cdot \left(\frac{\ell q - \ell q^2}{\ell q + t} \right) \geq \hat{n}\hat{p}(1 - \hat{p}) = \hat{\sigma}^2,$$

so from part (iv) of Theorem 4 we get the desired bound $|\sigma^2 - \hat{\sigma}^2| \leq O(1 + \epsilon(1 + \sigma^2))$ of Theorem 6.

Finally, we take $\bar{\ell} = \hat{n}$ and $\bar{q}$ to be $\hat{p}$ rounded to the nearest multiple of $1/n$ as described above; $Z = \mathrm{Bin}(\bar{\ell}, \bar{q})$ is the desired Binomial distribution whose existence is claimed by the theorem, and the parameters $\bar{\mu}, \bar{\sigma}^2$ of the theorem are $\bar{\mu} = \bar{\ell}\bar{q}$, $\bar{\sigma}^2 = \bar{\ell}\bar{q}(1 - \bar{q})$. Passing from $\mathrm{Bin}(\hat{n}, \hat{p})$ to $\mathrm{Bin}(\bar{\ell}, \bar{q})$ changes the mean and variance of the Binomial distribution by at most 1, so all the claimed bounds from parts (ii) and (iv) of Theorem 6 indeed hold. To finish the proof of the theorem it remains only to show that $d_{\mathrm{TV}}(\mathrm{Bin}(\bar{\ell}, \hat{p}), \mathrm{Bin}(\bar{\ell}, \bar{q})) \leq O(\epsilon)$. Similar to Section 2.2 this is done by passing through Translated Poisson distributions. We show that

$$d_{\mathrm{TV}}(\mathrm{Bin}(\bar{\ell}, \hat{p}), TP(\bar{\ell}\hat{p}, \bar{\ell}\hat{p}(1 - \hat{p}))), \quad d_{\mathrm{TV}}(TP(\bar{\ell}\hat{p}, \bar{\ell}\hat{p}(1 - \hat{p})), TP(\bar{\ell}\bar{q}, \bar{\ell}\bar{q}(1 - \bar{q}))), \quad \text{and}$$
$$d_{\mathrm{TV}}(TP(\bar{\ell}\bar{q}, \bar{\ell}\bar{q}(1 - \bar{q})), \mathrm{Bin}(\bar{\ell}, \bar{q}))$$

are each at most $O(\epsilon)$, and invoke the triangle inequality.

1. Bounding $d_{\mathrm{TV}}(\mathrm{Bin}(\bar{\ell}, \hat{p}), TP(\bar{\ell}\hat{p}, \bar{\ell}\hat{p}(1 - \hat{p})))$: Using Lemma 1, we get

$$d_{\mathrm{TV}}(\mathrm{Bin}(\bar{\ell}, \hat{p}), TP(\bar{\ell}\hat{p}, \bar{\ell}\hat{p}(1 - \hat{p}))) \leq \frac{1}{\sqrt{\bar{\ell}\hat{p}(1 - \hat{p})}} + \frac{2}{\bar{\ell}\hat{p}(1 - \hat{p})}.$$

   Since $\bar{\ell}\hat{p} = \hat{n}\hat{p} \geq k^2 - 1/k = \Omega(1/\epsilon^2)$ we have that the RHS above is at most $O(\epsilon)$.

2. Bounding $d_{\mathrm{TV}}(TP(\bar{\ell}\hat{p}, \bar{\ell}\hat{p}(1 - \hat{p})), TP(\bar{\ell}\bar{q}, \bar{\ell}\bar{q}(1 - \bar{q})))$: Let $\tilde{\sigma}^2$ denote $\min\{\bar{\ell}\hat{p}(1 - \hat{p}), \bar{\ell}\bar{q}(1 - \bar{q})\}$. Since $|\bar{q} - \hat{p}| \leq 1/n$, we have that $\bar{\ell}\bar{q}(1 - \bar{q}) = \bar{\ell}\hat{p}(1 - \hat{p}) \pm O(1) = \Omega(1/\epsilon^2)$ so $\tilde{\sigma} = \Omega(1/\epsilon)$. We use Lemma 2, which tells us that

$$d_{\mathrm{TV}}(TP(\bar{\ell}\hat{p}, \bar{\ell}\hat{p}(1 - \hat{p})), TP(\bar{\ell}\bar{q}, \bar{\ell}\bar{q}(1 - \bar{q}))) \leq \frac{|\bar{\ell}\hat{p} - \bar{\ell}\bar{q}|}{\tilde{\sigma}} + \frac{|\bar{\ell}\hat{p}(1 - \hat{p}) - \bar{\ell}\bar{q}(1 - \bar{q})| + 1}{\tilde{\sigma}^2}. \quad (16)$$

   Since $|\hat{p} - \bar{q}| \leq 1/n$, we have that $|\bar{\ell}\hat{p} - \bar{\ell}\bar{q}| = \bar{\ell}|\hat{p} - \hat{q}| \leq \bar{\ell}/n \leq 1$, so the first fraction on the RHS of (16 is $O(\epsilon)$. The second fraction is at most $(O(1) + 1)/\tilde{\sigma}^2 = O(\epsilon^2)$, so we get $d_{\mathrm{TV}}(TP(\bar{\ell}\hat{p}, \bar{\ell}\hat{p}(1 - \hat{p})), TP(\bar{\ell}\bar{q}, \bar{\ell}\bar{q}(1 - \bar{q}))) \leq O(\epsilon)$ as desired.

3. Bounding $d_{\mathrm{TV}}(TP(\bar{\ell}\bar{q}, \bar{\ell}\bar{q}(1 - \bar{q})), \mathrm{Bin}(\bar{\ell}, \bar{q}))$: We use Lemma 1 similar to the first case above, together with the lower bound $\tilde{\sigma} = \Omega(1/\epsilon)$, to get the desired $O(\epsilon)$ upper bound.

This concludes the proof of Theorem 6. ∎

5 Birgé's theorem: Learning unimodal distributions

Here we briefly explain how Theorem 5 follows from [Bir97]. We assume that the reader is moderately familiar with the paper [Bir97].

Birgé (see his Theorem 1 and Corollary 1) upper bounds the expected variation distance between the target distribution (which he denotes f) and the hypothesis distribution that is constructed by his algorithm (which he denotes $\hat{f}_n$; it should be noted, though, that his "n" parameter denotes the number of samples used by the algorithm, while we will denote this by "m", reserving "n" for the domain $\{1, \ldots, n\}$ of the distribution). More precisely, [Bir97] shows that this expected variation distance is at most that of the Grenander estimator (applied to learn a unimodal distribution when the mode is known) plus a lower-order term. For our Theorem 5 we take Birgé's "η" parameter to be ϵ. With this choice of η, by the results of [Bir87a, Bir87b] bounding the expected error of the Grenander estimator, if $m = O(\log(n)/\epsilon^3)$ samples are used in Birgé's algorithm then the expected variation distance between the target distribution and his hypothesis distribution is at most $O(\epsilon)$. To go from expected error ϵ to an ϵ-accurate hypothesis with probability $1 - \delta$, we run the above-described algorithm $O(\log(1/\delta))$ times so that with probability at least $1 - \delta$ some hypothesis obtained is ϵ-accurate. Then we use our hypothesis testing procedure of Lemma 8, or, more precisely, the extension provided in Lemma 11, to identify an $O(\epsilon)$-accurate hypothesis. (The use of Lemma 11 is why the running time of Theorem 5 depends quadratically on $\log(1/\delta)$.)

It remains only to argue that a single run of Birgé's algorithm on a sample of size $m = O(\log(n)/\epsilon^3)$ can be carried out in $\tilde{O}(\log^2(n)/\epsilon^3)$ bit operations (recall that each sample is a $\log(n)$-bit string). His algorithm begins by locating an $r \in [n]$ that approximately minimizes the value of his function $d(r)$ (see Section 3 of [Bir97]) to within an additive $\eta = \epsilon$ (see Definition 3 of his paper); intuitively this r represents his algorithm's "guess" at the true mode of the distribution. To locate such an r, following Birgé's suggestion in Section 3 of his paper, we begin by identifying two consecutive points in the sample such that r lies between those two sample points. This can be done using $\log m$ stages of binary search over the (sorted) points in the sample, where at each stage of the binary search we compute the two functions d^- and d^+ and proceed in the appropriate direction. To compute the function $d^-(j)$ at a given point j (the computation of d^+ is analogous), we recall that $d^-(j)$ is defined as the maximum difference over $[1, j]$ between the empirical cdf and its convex minorant over $[1, j]$. The convex minorant of the empirical cdf (over m points) can be computed in $\tilde{O}((\log n)m)$ bit-operations (where the $\log n$ comes from the fact that each sample point is an element of $[n]$), and then by enumerating over all points in the sample that lie in $[1, j]$ (in time $O((\log n)m)$) we can compute $d^-(j)$. Thus it is possible to identify two adjacent points in the sample such that r lies between them in time $\tilde{O}((\log n)m)$. Finally, as Birgé explains in the last paragraph of Section 3 of his paper, once two such points have been identified it is possible to again use binary search to find a point r in that interval where $d(r)$ is minimized to within an additive η. Since the maximum difference between d^- and d_+ can never exceed 1, at most $\log(1/\eta) = \log(1/\epsilon)$ stages of binary search are required here to find the desired r.

Finally, once the desired r has been obtained, it is straightforward to output the final hypothesis (which Birgé denotes $\hat{f}_n$. As explained in Definition 3, this hypothesis is the derivative of $\tilde{F}_n^r$, which is essentially the convex minorant of the empirical cdf to the left of r and the convex majorant of the empirical cdf to the right of r. As described above, given a value of r these convex majorants and minorants can be computed in $\tilde{O}((\log n)m)$ time, and the derivative is simply a collection of uniform distributions as claimed. This concludes our sketch of how Theorem 5 follows from [Bir97].

6 Efficient Evaluation of the Poisson Distribution

In this section we provide an efficient algorithm to compute an additive approximation to the Poisson probability mass function. This seems like a basic operation in numerical analysis, but we were not able to find it explicitly in the literature.

Before we state our theorem we need some notation. For a positive integer n, denote by $|n|$ its description

complexity (bit complexity), i.e. $|n| = \lceil \log_2 n \rceil$. We represent a positive rational number q as $\frac{q_1}{q_2}$, where q_1, q_2 are relatively prime positive integers. The description complexity of q is defined to be $|q| = |q_1| + |q_2|$. We are now ready to state our theorem for this section:

Theorem 7 *There is an algorithm that, on input a rational number $\lambda > 0$, and integers $k \geq 0$ and $t > 0$, produces an estimate $\widehat{p_k}$ such that*

$$|\widehat{p_k} - p_k| \leq \frac{1}{t},$$

where $p_k = \frac{\lambda^k e^{-\lambda}}{k!}$ is the probability that the Poisson distribution of parameter λ assigns to integer k. The running time of the algorithm is $\tilde{O}(|t|^3 + |k| \cdot |t| + |\lambda| \cdot |t|)$.

Proof: Clearly we cannot just compute $e^{-\lambda}$, λ^k and $k!$ separately, as this will take time exponential in the description complexity of k and λ. We follow instead an indirect approach. We start by rewriting the target probability as follows

$$p_k = e^{-\lambda + k \ln(\lambda) - \ln(k!)}.$$

Motivated by this formula, let

$$E_k := -\lambda + k \ln(\lambda) - \ln(k!).$$

Note that $E_k \leq 0$. Our goal is to approximate E_k to within high enough accuracy and then use this approximation to approximate p_k.

In particular, the main part of the argument involves an efficient algorithm to compute an approximation $\widehat{\widehat{E_k}}$ to E_k satisfying

$$\left| \widehat{\widehat{E_k}} - E_k \right| \leq \frac{1}{4t} \leq \frac{1}{2t} - \frac{1}{8t^2}. \tag{17}$$

This approximation has bit complexity $\tilde{O}(|k| + |\lambda| + |t|)$ and can be computed in time $\tilde{O}(|k| \cdot |t| + |\lambda| + |t|^3)$.

We first show how to use such an approximation to complete the proof. We claim that it suffices to approximate $e^{\widehat{\widehat{E_k}}}$ to within an additive error $\frac{1}{2t}$. Indeed, if $\widehat{p_k}$ is the result of this approximation, then:

$$\widehat{p_k} \leq e^{\widehat{\widehat{E_k}}} + \frac{1}{2t} \leq e^{E_k + \frac{1}{2t} - \frac{1}{8t^2}} + \frac{1}{2t} \leq e^{E_k + \ln(1 + \frac{1}{2t})} + \frac{1}{2t}$$

$$\leq e^{E_k} \left(1 + \frac{1}{2t} \right) + \frac{1}{2t} \leq p_k + \frac{1}{t};$$

and similarly

$$\widehat{p_k} \geq e^{\widehat{\widehat{E_k}}} - \frac{1}{2t} \geq e^{E_k - (\frac{1}{2t} - \frac{1}{8t^2})} - \frac{1}{2t} \geq e^{E_k - \ln(1 + \frac{1}{2t})} - \frac{1}{2t}$$

$$\geq e^{E_k} \left/ \left(1 + \frac{1}{2t} \right) \right. - \frac{1}{2t} \geq e^{E_k} \left(1 - \frac{1}{2t} \right) - \frac{1}{2t} \geq p_k - \frac{1}{t}.$$

We will need the following lemma:

Lemma 14 *Let $\alpha \leq 0$ be a rational number. There is an algorithm that computes an estimate $\widehat{e^{\alpha}}$ such that*

$$\left| \widehat{e^{\alpha}} - e^{\alpha} \right| \leq \frac{1}{2t}$$

and has running time $\tilde{O}(|\alpha| \cdot |t| + |t|^2)$.

Proof: Since $e^\alpha \in [0,1]$, the point of the additive grid $\{\frac{i}{4t}\}_{i=1}^{4t}$ closest to e^α achieves error at most $1/(4t)$. Equivalently, in a logarithmic scale, consider the grid $\{\ln \frac{i}{4t}\}_{i=1}^{4t}$ and let $j^* := \arg\min_j \left\{ \left| \alpha - \ln(\frac{j}{4t}) \right| \right\}$. Then, we have that

$$\left| \frac{j^*}{(4t)} - e^\alpha \right| \le \frac{1}{4t}.$$

The idea of the algorithm is to approximately identify the point j^*, by computing approximations to the points of the logarithmic grid combined with a binary search procedure. Indeed, consider the "rounded" grid $\{\widehat{\ln \frac{i}{4t}}\}_{i=1}^{4t}$ where each $\widehat{\ln(\frac{i}{4t})}$ is an approximation to $\ln(\frac{i}{4t})$ that is accurate to within an additive $\frac{1}{16t}$. Notice that, for $i = 1, \ldots, 4t$:

$$\ln\left(\frac{i+1}{4t}\right) - \ln\left(\frac{i}{4t}\right) = \ln\left(1 + \frac{1}{i}\right) \ge \ln\left(1 + \frac{1}{4t}\right) > 1/8t.$$

Given that our approximations are accurate to within an additive $1/16t$, it follows that the rounded grid $\{\widehat{\ln \frac{i}{4t}}\}_{i=1}^{4t}$ is monotonic in i.

The algorithm does not construct the points of this grid explicitly, but adaptively as it needs them. In particular, it performs a binary search in the set $\{1, \ldots, 4t\}$ to find the point $i^* := \arg\min_i \left\{ \left| \alpha - \widehat{\ln(\frac{i}{4t})} \right| \right\}$. In every iteration of the search, when the algorithm examines the point j, it needs to compute the approximation $g_j = \widehat{\ln(\frac{j}{4t})}$ and evaluate the distance $|\alpha - g_j|$. It is known that the logarithm of a number x with a binary fraction of L bits and an exponent of $o(L)$ bits can be computed to within a relative error $O(2^{-L})$ in time $\tilde{O}(L)$ [Bre75]. It follows from this that g_j has $O(|t|)$ bits and can be computed in time $\tilde{O}(|t|)$. The subtraction takes linear time, i.e. it uses $O(|\alpha| + |t|)$ bit operations. Therefore, each step of the binary search can be done in time $O(|\alpha|) + \tilde{O}(|t|)$ and thus the overall algorithm has $O(|\alpha| \cdot |t|) + \tilde{O}(|t|^2)$ running time.

The algorithm outputs $\frac{i^*}{4t}$ as its final approximation to e^α. We argue next that the achieved error is at most an additive $\frac{1}{2t}$. Since the distance between two consecutive points of the grid $\{\ln \frac{i}{4t}\}_{i=1}^{4t}$ is more than $1/(8t)$ and our approximations are accurate to within an additive $1/16t$, a little thought reveals that $i^* \in \{j^*-1, j^*, j^*+1\}$. This implies that $\frac{i^*}{4t}$ is within an additive $1/2t$ of e^α as desired, and the proof of the lemma is complete. ∎

We now proceed to describe how to approximate $e^{\widehat{\widehat{E_k}}}$. Recall that we want to output an estimate $\widehat{p_k}$ such that $|\widehat{p_k} - e^{\widehat{\widehat{E_k}}}| \le 1/(2t)$. We distinguish the following cases:

- If $\widehat{\widehat{E_k}} \ge 0$, we output $\widehat{p_k} := 1$. Indeed, given that $\left| \widehat{\widehat{E_k}} - E_k \right| \le \frac{1}{4t}$ and $E_k \le 0$, if $\widehat{\widehat{E_k}} \ge 0$ then $\widehat{\widehat{E_k}} \in [0, \frac{1}{4t}]$. Hence, because $t \ge 1$, $e^{\widehat{\widehat{E_k}}} \in [1, 1+1/2t]$, so 1 is within an additive $1/2t$ of the right answer.

- Otherwise, $\widehat{p_k}$ is defined to be the estimate obtained by applying Lemma 14 for $\alpha := \widehat{\widehat{E_k}}$. Given the bit complexity of $\widehat{\widehat{E_k}}$, the running time of this procedure will be $\tilde{O}(|k| \cdot |t| + |\lambda| \cdot |t| + |t|^3)$.

Hence, the overall running time is $\tilde{O}(|k| \cdot |t| + |\lambda| \cdot |t| + |t|^3)$.

We now show how to compute $\widehat{\widehat{E_k}}$. There are several steps to our approximation:

1. (Stirling's Asymptotic Approximation): Recall Stirling's asymptotic approximation (see e.g. [Whi80] p.193):

$$\ln k! = k \ln(k) - k + (1/2) \cdot \ln(2\pi) + \sum_{j=2}^{m} \frac{B_j \cdot (-1)^j}{j(j-1) \cdot k^{j-1}} + O(1/k^m).$$

where B_k are the Bernoulli numbers. We define an approximation of $\ln k!$ as follows:

$$\widehat{\ln k!} := k \ln(k) - k + (1/2) \cdot \ln(2\pi) + \sum_{j=2}^{m_0} \frac{B_j \cdot (-1)^j}{j(j-1) \cdot k^{j-1}}$$

for $m_0 := \lceil \frac{|t|}{|k|} \rceil + 1$.

2. (Definition of an approximate exponent $\widehat{E_k}$): Define $\widehat{E_k} := -\lambda + k \ln(\lambda) - \widehat{\ln(k!)}$. Given the above discussion, we can calculate the distance of $\widehat{E_k}$ to the true exponent E_k as follows:

$$|E_k - \widehat{E_k}| \le |\ln(k!) - \widehat{\ln(k!)}| \le O(1/k^{m_0}) \tag{18}$$

$$\le \frac{1}{10t}. \tag{19}$$

So we can focus our attention to approximating $\widehat{E_k}$. Note that $\widehat{E_k}$ is the sum of $m_0 + 2 = O(\frac{\log t}{\log k})$ terms. To approximate it within error $1/(10t)$, it suffices to approximate each summand within an additive error of $O(1/(t \cdot \log t))$. Indeed, we so approximate each summand and our final approximation $\widehat{\widehat{E_k}}$ will be the sum of these approximations. We proceed with the analysis:

3. (Estimating 2π): Since 2π shows up in the above expression, we should try to approximate it. It is known that the first ℓ digits of π can be computed exactly in time $O(\log \ell \cdot M(\ell))$, where $M(\ell)$ is the time to multiply two ℓ-bit integers [Sal76, Bre76]. For example, if we use the Schönhage-Strassen algorithm for multiplication [SS71], we get $M(\ell) = O(\ell \cdot \log \ell \cdot \log \log \ell)$. Hence, choosing $\ell := \lceil \log_2(12t \cdot \log t) \rceil$, we can obtain in time $\tilde{O}(|t|)$ an approximation $\widehat{2\pi}$ of 2π that has a binary fraction of ℓ bits and satisfies:

$$|\widehat{2\pi} - 2\pi| \le 2^{-\ell} \implies (1 - 2^{-\ell})2\pi \le \widehat{2\pi} \le (1 + 2^{-\ell})2\pi.$$

Note that, with this approximation, we have

$$\left| \ln(2\pi) - \ln(\widehat{2\pi}) \right| \le \ln(1 - 2^{-\ell}) \le 2^{-\ell} \le 1/(12t \cdot \log t).$$

4. (Floating-Point Representation): We will also need accurate approximations to $\ln \widehat{2\pi}$, $\ln k$ and $\ln \lambda$. We think of $\widehat{2\pi}$ and k as multiple-precision floating point numbers base 2. In particular,

   - $\widehat{2\pi}$ can be described with a binary fraction of $\ell + 3$ bits and a constant size exponent; and
   - $k \equiv 2^{\lceil \log k \rceil} \cdot \frac{k}{2^{\lceil \log k \rceil}}$ can be described with a binary fraction of $\lceil \log k \rceil$, i.e. $|k|$, bits and an exponent of length $O(\log \log k)$, i.e. $O(\log |k|)$.

Also, since λ is a positive rational number, $\lambda = \frac{\lambda_1}{\lambda_2}$, where λ_1 and λ_2 are positive integers of at most $|\lambda|$ bits. Hence, for $i = 1, 2$, we can think of λ_i as a multiple-precision floating point number base 2 with a binary fraction of $|\lambda|$ bits and an exponent of length $O(\log |\lambda|)$. Hence, if we choose $L = \lceil \log_2(12(3k+1)t^2 \cdot k \cdot \lambda_1 \cdot \lambda_2) \rceil = O(|k| + |\lambda| + |t|)$, we can represent all numbers $\widehat{2\pi}, \lambda_1, \lambda_2, k$ as multiple precision floating point numbers with a binary fraction of L bits and an exponent of $O(\log L)$ bits.

5. (Estimating the logs): It is known that the logarithm of a number x with a binary fraction of L bits and an exponent of $o(L)$ bits can be computed to within a relative error $O(2^{-L})$ in time $\tilde{O}(L)$ [Bre75]. Hence, in time $\tilde{O}(L)$ we can obtain approximations $\widehat{\ln \widehat{2\pi}}, \widehat{\ln k}, \widehat{\ln \lambda_1}, \widehat{\ln \lambda_2}$ such that:

   - $|\widehat{\ln k} - \ln k| \le 2^{-L} \ln k \le \frac{1}{12(3k+1)t^2}$; and similarly
   - $|\widehat{\ln \lambda_i} - \ln \lambda_i| \le \frac{1}{12(3k+1)t^2}$, for $i = 1, 2$;
   - $|\widehat{\ln \widehat{2\pi}} - \ln \widehat{2\pi}| \le \frac{1}{12(3k+1)t^2}$.

25

6. (Estimating the terms of the series): To complete the analysis, we also need to approximate each term of the form $c_j = \frac{B_j}{j(j-1) \cdot k^{j-1}}$ up to an additive error of $O(1/(t \cdot \log t))$. We do this as follows: We compute the numbers B_j and k^{j-1} exactly, and we perform the division approximately.

Clearly, the positive integer k^{j-1} has description complexity $j \cdot |k| = O(m_0 \cdot |k|) = O(|t| + |k|)$, since $j = O(m_0)$. We compute k^{j-1} exactly using repeated squaring in time $\tilde{O}(j \cdot |k|) = \tilde{O}(|t| + |k|)$. It is known [Fil92] that the rational number B_j has $\tilde{O}(j)$ bits and can be computed in $\tilde{O}(j^2) = \tilde{O}(|t|^2)$ time. Hence, the approximate evaluation of the term c_j (up to the desired additive error of $1/(t \log t)$) can be done in $\tilde{O}(|t|^2 + |k|)$, by a rational division operation (see e.g. [Knu81]). The sum of all the approximate terms takes linear time, hence the approximate evaluation of the entire truncated series (comprising at most $m_0 \le |t|$ terms) can be done in $\tilde{O}(|t|^3 + |k| \cdot |t|)$ time overall.

Let $\widehat{\widehat{E_k}}$ be the approximation arising if we use all the aforementioned approximations. It follows from the above computations that

$$\left| \widehat{\widehat{E_k}} - \widehat{E_k} \right| \le \frac{1}{10t}.$$

7. (Overall Error): Combining the above computations we get:

$$\left| \widehat{\widehat{E_k}} - E_k \right| \le \frac{1}{4t}.$$

The overall time needed to obtain $\widehat{\widehat{E_k}}$ was $\tilde{O}(|k| \cdot |t| + |\lambda| + |t|^3)$ and the proof of the theorem is complete. ∎

7 Conclusion and open problems

While we have essentially settled the sample and time complexity of learning an unknown Poisson Binomial Distribution to high accuracy, several natural goals remain for future work. One goal is to obtain a proper learning algorithm which is as computationally efficient as our non-proper algorithm. Another goal is to understand the sample complexity of learning *log-concave* distributions over $[n]$ (a distribution X over $[n]$ is log-concave if $p_i^2 \ge p_{i+1}p_{i-1}$ for every i, where p_j denotes $\Pr[X = j]$). Every PBD over $[n]$ is log-concave (see Section 2 of [KG71]), and every log-concave distribution over $[n]$ is unimodal; thus this class lies between the class of PBDs (now known to be learnable from $\tilde{O}(1/\epsilon^3)$ samples) and the class of unimodal distributions (for which $\Omega(\log(n)/\epsilon^3)$ samples are necessary). Can log-concave distributions over $[n]$ be learned from $\mathrm{poly}(1/\epsilon)$ samples independent of n? If not, what is the dependence of the sample complexity on n?

References

[Ber41]   Andrew C. Berry. The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the American Mathematical Society*, 49(1):122–136, 1941.

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

[Bir87a]   L. Birgé. Estimating a density under order restrictions: Nonasymptotic minimax risk. *Annals of Statistics*, 15(3):995–1012, 1987.

[Bir87b]   L. Birgé. On the risk of histograms for estimating decreasing densities. *Annals of Statistics*, 15(3):1013–1022, 1987.

[Bir97] L. Birgé. Estimation of unimodal densities without smoothness assumptions. *Annals of Statistics*, 25(3):970–981, 1997.

[BL06] A. D. Barbour and T. Lindvall. Translated poisson approximation for markov chains. *Journal of Theoretical Probability*, 19, 2006.

[Bre75] R. P. Brent. Multiple-precision zero-finding methods and the complexity of elementary function evaluation. *Analytic Computational Complexity (J. F. Traub ed.)*, pages 151–176, 1975. Academic Press, New York.

[Bre76] R. P. Brent. Fast multiple-precision evaluation of elementary functions. *Journal of the ACM*, 23(2):242–251, 1976.

[Cam60] L. Le Cam. An approximation theorem for the Poisson binomial distribution. *Pacific J. Math*, 10:1181–1197, 1960.

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

[Che74] L.H.Y. Chen. On the convergence of Poisson binomial to Poisson distributions. *Ann. Probab.*, 2:178–180, 1974.

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

[Das08] Constantinos Daskalakis. An Efficient PTAS for Two-Strategy Anonymous Games. *WINE* 2008, pp. 186-197. Full version available as ArXiV report, 2008.

[DDS11] C. Daskalakis, I. Diakonikolas, and R.A. Servedio. Learning k-modal distributions via testing. Arxiv report, 2011.

[DL96a] L. Devroye and G. Lugosi. Nonasymptotic universal smoothing factors, kernel complexity and Yatracos classes. *Annals of Statistics*, 25:2626–2637, 1996.

[DL96b] L. Devroye and G. Lugosi. A universally acceptable smoothing factor for kernel density estimation. *Annals of Statistics*, 24:2499–2512, 1996.

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

[DP86] P. Deheuvels and D. Pfeifer. A semigroup approach to Poisson approximation. *Ann. Probab.*, 14:663–676, 1986.

[DP09] D. Dubhashi and A. Panconesi. *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press, Cambridge, 2009.

[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.

[Ehm91] Werner Ehm. Binomial approximation to the Poisson binomial distribution. *Statistics and Probability Letters*, 11:7–16, 1991.

[Ess42] Carl-Gustav Esseen. On the liapunoff limit of error in the theory of probability. *Arkiv för matematik, astronomi och fysik*, A:1–19, 1942.

[Fil92]   Sandra Fillebrown. Faster computation of bernoulli numbers. *Journal of Algorithms*, 13(3):431 – 445, 1992.

[HC60]   S.L. Hodges and L. Le Cam. The Poisson approximation to the binomial distribution. *Ann. Math. Statist.*, 31:747–740, 1960.

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

[Joh03]   J.L. Johnson. *Probability and Statistics for Computer Science*. John Wiley & Sons, Inc., New York, NY, USA, 2003.

[KG71]   J. Keilson and H. Gerber. Some Results for Discrete Unimodality. *J. American Statistical Association*, 66(334):386–389, 1971.

[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.

[Knu81]   Donald E. Knuth. *The Art of Computer Programming, Volume II: Seminumerical Algorithms, 2nd Edition*. Addison-Wesley, 1981.

[Mik93]   V.G. Mikhailov. On a refinement of the central limit theorem for sums of independent random indicators. *Theory Probab. Appl.*, 38:479–489, 1993.

[NJ05]   S. Kotz N.L. Johnson, A.W. Kemp. *Univariate discrete distributions*. John Wiley & Sons, Inc., New York, NY, USA, 2005.

[RÖ7]   A. Röllin. Translated Poisson Approximation Using Exchangeable Pair Couplings. *Annals of Applied Probability*, 17(5/6):1596–1614, 2007.

[Roo00]   B. Roos. Binomial approximation to the Poisson binomial distribution: The Krawtchouk expansion. *Theory Probab. Appl.*, 45:328–344, 2000.

[Sal76]   Eugene Salamin. Computation of pi using arithmetic-geometric mean. *Mathematics of Computation*, 30(135):565–570, 1976.

[Sat93]   Ken-Iti Sato. Convolution of unimodal distributions can produce any number of modes. *Annals of Probability*, 21(3):1543–1549, 1993.

[Soo96]   S.Y.T. Soon. Binomial approximation for dependent indicators. *Statist. Sinica*, 6:703–714, 1996.

[SS71]   A. Schönhage and V. Strassen. Schnelle multiplikation grosser zahlen. *Computing*, 7:281–292, 1971.

[Ste94]   J.M. Steele. Le Cam's Inequality and Poisson Approximation. *Amer. Math. Monthly*, 101:48–54, 1994.

[Vol95]   A. Yu. Volkova. A refinement of the central limit theorem for sums of independent random indicators. *Theory Probab. Appl.*, 40:791–794, 1995.

[Whi80]   E.T. Whittaker. *A course of modern analysis*. Cambridge University Press, 1980.

[Yat85]   Y. G. Yatracos. Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *Annals of Statistics*, 13:768–774, 1985.