Testing Poisson Binomial Distributions

Jayadev Acharya*

EECS, MIT
jayadev@csail.mit.edu

Constantinos Daskalakis†

EECS, MIT
costis@mit.edu

Abstract

A Poisson Binomial distribution over n variables is the distribution of the sum of n independent Bernoullis. We provide a sample near-optimal algorithm for testing whether a distribution P supported on $\{0, \ldots, n\}$ to which we have sample access is a Poisson Binomial distribution, or far from all Poisson Binomial distributions. The sample complexity of our algorithm is $O(n^{1/4})$ to which we provide a matching lower bound. We note that our sample complexity improves quadratically upon that of the naive "learn followed by tolerant-test" approach, while instance optimal identity testing [VV14] is not applicable since we are looking to simultaneously test against a whole family of distributions.

1 Introduction

Given independent samples from an unknown probability distribution P over $\{0, \ldots, n\}$, can you explain P as the distribution of the sum of n independent Bernoullis? For example, P may be the number of faculty attending the weekly faculty meeting, and you may be looking to test whether your observations are consistent with the different faculty decisions being independent. It is a problem of testing against a *family* of distributions:

PbdTesting: Given $\epsilon > 0$ and sample access to an unknown distribution P over $\{0, \ldots, n\}$, test whether $P \in \mathcal{PBD}_n$, or $d_{TV}(P, \mathcal{PBD}_n) > \epsilon$, where $\mathcal{PBD}_n$ is the set of Poisson Binomial distributions over n variables.

Besides any practical applications, the theoretical interest in studying PbdTesting, and for that matter testing membership to other classes of distributions, stems from the fact that "being a Poisson Binomial distribution" is not a symmetric property of a distribution; hence the results of [VV11] cannot be brought to bear. At the same time, "being a Poisson Binomial Distribution" does not fall into the shape restrictions to a distribution, such as uniformity [GR00, BFF+01, Pan08] or monotonicity [BKR04], for which (near-)optimal testing algorithms have been obtained. While there has been a lot of work on learning distributions from a class of distributions [CDSS13, FOS05, MV10, BS10], there is still a large gap in our current knowledge about the complexity of testing against general families of distributions, unless both the unknown distribution and the family have been restricted a priori [DDS+13, DKN14].

An obvious approach to PbdTesting is to learn a candidate Poisson Binomial distribution Q that is $\epsilon/2$-close to P, if P truly is a Poisson Binomial distribution. This is known to be quite cheap, only requiring $\tilde{O}(1/\epsilon^2)$ samples from P [DDS12]. We can then use a tolerant tester to test $d_{TV}(P, Q) \le \epsilon/2$ vs $d_{TV}(P, Q) > \epsilon$. Such a tester would allow us to distinguish $P \in \mathcal{PBD}_n$ vs $d_{TV}(P, \mathcal{PBD}_n) > \epsilon$, as $d_{TV}(P, Q) \le \epsilon/2 \Leftrightarrow P \in \mathcal{PBD}_n$.

Given that any $Q \in \mathcal{PBD}_n$ has effective support $O(\sqrt{n \log 1/\epsilon})$,[1] we can easily construct a tolerant tester that uses $\tilde{O}(\sqrt{n}/\epsilon^2)$ samples, resulting in overall sampling complexity of $\tilde{O}(\sqrt{n}/\epsilon^2)$. On the other hand, we do not see how to substantially improve this

*Supported by grant from MITEI-Shell program.

†Supported by a Sloan Foundation Fellowship, a Microsoft Research Faculty Fellowship and NSF Award CCF-0953960 (CAREER) and CCF-1101491.

[1] *Effective support* is the smallest set of contiguous integers where the distribution places all but ϵ of its probability mass.

approach, given the lower bound of $\Omega(m/\log m)$ for tolerant identity testing distributions of support size m [VV11].

A somewhat different approach would circumvent the use of a tolerant identity tester, by exploiting the small amount of tolerance accommodated by known (non-tolerant) identity testers. For instance, [BFF$^+$01] show that, given a distribution Q of support m and $\tilde{O}(\sqrt{m}) \cdot \mathrm{poly}(1/\epsilon)$ samples from an unknown distribution P over the same support, one can distinguish $d_{TV}(P,Q) \le \frac{\epsilon^3}{4\sqrt{m}\log m}$ vs $d_{TV}(P,Q) > \epsilon$. Hence, we can try to first find a candidate $Q \in \mathcal{PBD}_n$ that is $\frac{\epsilon^3}{4\sqrt{m}\log m}$-close to P, if $P \in \mathcal{PBD}_n$, and then do (non-tolerant) identity testing against Q. In doing so, we can use $m = O(\sqrt{n \log 1/\epsilon})$, since that is the worst case effective support of P, if $P \in \mathcal{PBD}_n$.

The testing step of this approach is cheaper, namely $\tilde{O}(n^{1/4}) \cdot \mathrm{poly}(1/\epsilon)$ samples, but now the learning step becomes more expensive, namely $\tilde{\Omega}(\sqrt{n}) \cdot \mathrm{poly}(1/\epsilon)$ samples, as the required learning accuracy is more extravagant than before.

Is there then a fundamental barrier, imposing a sample complexity of $\tilde{\Omega}(\sqrt{n})$? We show that the answer is "no," namely

THEOREM 1. *For $n, \epsilon, \delta > 0$, there exists an algorithm,* `Testing PBDs`*, that uses*

$$O\left(\frac{n^{1/4}\sqrt{\log(1/\epsilon)}}{\epsilon^2} + \frac{\log^{2.5}(1/\epsilon)}{\epsilon^6}\right) \cdot \log(1/\delta)$$

independent samples from an unknown distribution P over $\{0, \dots, n\}$ and, with probability $\ge 1 - \delta$, outputs **Yes PBD***, if $P \in \mathcal{PBD}_n$, and* **No PBD***, if $d_{TV}(P, \mathcal{PBD}_n) > \epsilon$. The time complexity of the algorithm is*

$$O\Big(n^{1/4}\sqrt{\log(1/\epsilon)}/\epsilon^2 + (1/\epsilon)^{O(\log^2 1/\epsilon)}\Big) \cdot \log(1/\delta).$$

The proof of Theorem 1 can be found in Section 3. We also show that the dependence of our sample complexity on n cannot be improved, by providing a matching lower bound in Section 4 as follows.

THEOREM 2. *Any algorithm for* PBDTESTING *requires $\Omega(n^{1/4}/\epsilon^2)$ samples.*

One might be tempted to deduce Theorem 2 from the lower bound for identity testing against Binomial$(n, 1/2)$, which has been shown to require $\Omega(n^{1/4}/\epsilon^2)$ samples [Pan08, VV14]. However, testing against a class of distributions may very well be easier than testing against a specific member of the class. (As a trivial example consider the class of all distributions over $\{0, \dots, n\}$, which are trivial to test.) Still, for the class $\mathcal{PBD}_n$, we establish the same lower bound as for Binomial$(n, 1/2)$, deducing that the dependence of our sample complexity on n is tight up to constant factors, while the dependence on ϵ of the leading term in our sample complexity is tight up to a logarithmic factor.

1.1 Related work and our approach

Our testing problem is intimately related to the following fundamental problems:

IDENTITYTESTING: Given a known distribution Q and independent samples from an unknown distribution P, which are both supported on $[m] := \{0, \dots, m\}$, determine whether $P = Q$ OR $d_{TV}(P,Q) > \epsilon$. If $d_{TV}(P,Q) \in (0, \epsilon]$, then any answer is allowed.

TOLERANT-IDENTITYTESTING: Given a known distribution Q and independent samples from an unknown distribution P, which are both supported on $[m]$, determine whether $d_{TV}(P,Q) \le \epsilon/2$ OR $d_{TV}(P,Q) > \epsilon$. If $d_{TV}(P,Q) \in (\epsilon/2, \epsilon]$, then any answer is allowed.

It is known that IDENTITYTESTING can be solved from a near-optimal $\tilde{O}(\sqrt{m}/\epsilon^2)$ number of samples [BFF$^+$01, Pan08]. The guarantee is obviously probabilistic: with probability $\ge 2/3$, the algorithm outputs "equal," if $P = Q$, and "different," if $d_{TV}(P,Q) > \epsilon$. On the other hand, even testing whether P equals the uniform distribution over $[m]$ requires $\Omega(\sqrt{m}/\epsilon^2)$ samples.

While the identity tester of [BFF$^+$01] allows in fact a little bit of tolerance (namely distinguishing $d_{TV}(P,Q) \le \frac{\epsilon^3}{4\sqrt{m}\log m}$ vs $d_{TV}(P,Q) > \epsilon$), it does not accommodate a tolerance of $\epsilon/2$. Indeed, [VV11] show that there is a gap in the sample complexity of tolerant vs non-tolerant testing, showing that the tolerant version requires $\Omega(m/\log m)$ samples.

As discussed earlier, these results on identity testing in conjunction with the algorithm of [DDS12] for learning Poisson Binomial distributions can be readily used to solve PBDTESTING, albeit with suboptimal sample complexity. Moreover, recent work of Valiant and Valiant [VV14] pins down the optimal sampling complexity of IDENTITYTESTING up to constant factors for any distribution Q, allowing sample-optimal testing on an instance to instance basis. However, their algorithm is not applicable to PBDTESTING, as it allows testing whether an unknown distribution P equals a specific distribution Q vs being ϵ-far from Q, but not testing whether P belongs to a class of distributions vs being ϵ-far from all distributions in the class.

Our Approach. What we find quite interesting is that our "learning followed by tolerant testing" approach seems optimal: The learning algorithm of [DDS12] is optimal up to logarithmic factors, and there is strong evidence that tolerant identity testing a Poisson Binomial distribution requires $\tilde{\Omega}(\sqrt{n})$ samples. So where are we losing?

We observe that, even though PbdTesting can be reduced to tolerant identity testing of the unknown distribution P to a single Poisson Binomial distribution Q, we cannot consider the latter problem out of context, shooting at optimal testers for it. Instead, we are really trying to solve the following problem:

Tolerant-(Identity+PBD)-Testing: Given a known Q and independent samples from an unknown distribution P, which are both supported on $[m]$, determine whether $(d_{TV}(P,Q) \leq \epsilon_1$ And $P \in \mathcal{PBD}_n)$ Or $(d_{TV}(P,Q) > \epsilon_2)$. In all other cases, any answer is allowed.

The subtle difference between Tolerant-(Identity+PBD)-Testing and Tolerant-IdentityTesting is the added clause "And $P \in \mathcal{PBD}_n$," which, it turns out, makes a big difference for certain Q's. In particular, we would hope that, when Q and P are Poisson Binomial distributions with about the same variance, then the ℓ_1 bound $d_{TV}(P,Q) \leq \epsilon_1$ implies a good enough ℓ_2 bound, so that Tolerant-(Identity+PBD)-Testing can be reduced to tolerant identity testing in the ℓ_2 norm. We proceed to sketch the steps of our tester in more detail.

The first step is to run the learning algorithm of [DDS12] on $\tilde{O}(1/\epsilon^2)$ samples from P. The result is some $P_{\rm pbd} \in \mathcal{PBD}_n$ such that, if $P \in \mathcal{PBD}_n$ then, with probability $\geq .99$, $d_{TV}(P, P_{\rm pbd}) < \epsilon/10$. We then bifurcate depending on the variance of the learned $P_{\rm pbd}$. For some constant C to be decided, we consider the following cases.

- **Case 1:** $\sigma^2(P_{\rm pbd}) < \frac{C\cdot\log^4 1/\epsilon}{\epsilon^8}$

  In this case, $P_{\rm pbd}$ assigns probability mass of $\geq 1-\epsilon/5$ to an interval $\mathcal{I}$ of size $O(\log^{2.5}(1/\epsilon)/\epsilon^4)$. If $P \in \mathcal{PBD}_n$, then the ℓ_1 distance between P and $P_{\rm pbd}$ over $\mathcal{I}$ is at most $\epsilon/5$ (with probability at least 0.99). If $d_{TV}(P, \mathcal{PBD}_n) > \epsilon$, then over the same interval, the ℓ_1 distance is at least $4\epsilon/5$. We can therefore do tolerant identity testing restricted to support $\mathcal{I}$, with $O(|\mathcal{I}|/\epsilon^2) = O(\log^{2.5}(1/\epsilon)/\epsilon^6)$ samples. To this end, we use a simple tolerant identity test whose sample complexity is tight up to a logarithm in the support size, and very easy to analyze. Its use here does not affect the dependence of the overall sample complexity on n.

- **Case 2:** $\sigma^2(P_{\rm pbd}) \geq \frac{C\cdot\log^4 1/\epsilon}{\epsilon^8}$

  We are looking to reduce this case to a Tolerant-(Identity+PBD)-Testing task for an appropriate distribution Q that will make the reduction to tolerant identity testing in the ℓ_2 norm feasible. First, it follows from [DDS12] that in this case we can actually assume that $P_{\rm pbd}$ is a Binomial distribution. Next, we use $O(n^{1/4}/\epsilon^2)$ samples to obtain estimates $\hat{\mu}$ and $\hat{\sigma}^2$ for the mean and variance of P, and consider the Translated Poisson distribution $P_{tp}(\hat{\mu}, \hat{\sigma}^2)$ with parameters $\hat{\mu}$ and $\hat{\sigma}^2$; see Definition 5. If $P \in \mathcal{PBD}_n$, then with good probability (i) $\hat{\mu}$ and $\hat{\sigma}^2$ are extremely accurate as characterized by Lemma 1, and (ii) using the Translated Poisson approximation to the Poisson Binomial distribution, Lemma 4, we can argue that $d_{TV}(P, P_{tp}(\hat{\mu}, \hat{\sigma}^2)) \leq \frac{\epsilon^2}{10}$.

  Before getting to the heart of our test, we perform one last check. We calculate an estimate of $d_{TV}(P_{tp}(\hat{\mu}, \hat{\sigma}^2), P_{\rm pbd})$ that is accurate to within $\pm\epsilon/5$. If our estimate $\hat{d}_{TV}(P_b, P_{\rm pbd}) > \epsilon/2$, we can safely deduce $d_{TV}(P, \mathcal{PBD}_n) > \epsilon$. Indeed, if $P \in \mathcal{PBD}_n$, we would have seen $\hat{d}_{TV}(P_{tp}(\hat{\mu}, \hat{\sigma}^2), P_{\rm pbd}) \leq d_{TV}(P_{tp}(\hat{\mu}, \hat{\sigma}^2), P_{\rm pbd}) + \epsilon/5 \leq d_{TV}(P_{tp}(\hat{\mu}, \hat{\sigma}^2), P) + d_{TV}(P, P_{\rm pbd}) + \epsilon/5 \leq \frac{\epsilon^2}{10} + \frac{\epsilon}{10} + \frac{\epsilon}{5} < 2\epsilon/5$.

  If $\hat{d}_{TV}(P_{tp}(\hat{\mu}, \hat{\sigma}^2), P_{\rm pbd}) \leq \epsilon/2$, then

  $$\begin{aligned} & d_{TV}(P, P_{tp}(\hat{\mu}, \hat{\sigma}^2)) \\ \geq & d_{TV}(P, P_{\rm pbd}) - d_{TV}(P_{\rm pbd}, P_{tp}(\hat{\mu}, \hat{\sigma}^2)) \\ \geq & d_{TV}(P, P_{\rm pbd}) - \frac{7\epsilon}{10}. \end{aligned}$$

  At this point, there are two possibilities we need to distinguish between: *either* $P \in \mathcal{PBD}_n$ and $d_{TV}(P, P_{tp}(\hat{\mu}, \hat{\sigma}^2)) \leq \frac{\epsilon^2}{10}$, *or* $d_{TV}(P, \mathcal{PBD}_n) > \epsilon$ and $d_{TV}(P, P_{tp}(\hat{\mu}, \hat{\sigma}^2)) \geq \frac{3\epsilon}{10}$.[2] We argue that we can use an ℓ_2 test to solve this instance of Tolerant-(Identity+PBD)-Testing.

Clearly, we can boost the probability of error to any $\delta > 0$ by repeating the above procedure $\log(1/\delta)$ times and outputting the majority. Our algorithm is provided as Algorithm 2.

[2] The two cases identified here correspond to Cases 3a and 3b of Section 3, except that we include some logarithmic factors in the total variation distance bound in Case 3a for minute technical reasons.

2 Preliminaries

We provide some basic definitions, and state results that will be useful in our analysis.

Definition 1. *The* truncated logarithm *function* tlog *is defined as* $\mathrm{logt}(x) = \max\{1, \log x\}$*, for all* $x \in (0, +\infty)$*, where* $\log x$ *represents the natural logarithm of* x*.*

Definition 2. *Let* P *be a distribution over* $[n] = \{0, \ldots, n\}$*. The* ϵ-effective support *of* P *is the length of the smallest interval where the distribution places all but at most* ϵ *of its probability mass.*

Definition 3. *A* Poisson Binomial Distribution (PBD) *over* $[n]$ *is the distribution of* $X = \sum_{i=1}^{n} X_i$*, where the* X_i*'s are (mutually) independent Bernoulli random variables.* $\mathcal{PBD}_n$ *is the set of all Poisson Binomial distributions over* $[n]$*.*

Definition 4. *The* total variation distance *between two distributions* P *and* Q *over a finite set* A *is* $d_{TV}(P, Q) \overset{\text{def}}{=} \frac{1}{2}\sum_{i\in A}|P(i) - Q(i)|$*. The total variation distance between two sets of distributions* $\mathcal{P}$ *and* $\mathcal{Q}$ *is* $d_{TV}(\mathcal{P}, \mathcal{Q}) \overset{\text{def}}{=} \inf_{P\in\mathcal{P}, Q\in\mathcal{Q}} d_{TV}(P, Q)$*.*

We make use of the following learning algorithm, allowing us to learn an unknown Poisson Binomial distribution over $[n]$ from $\tilde{O}(1/\epsilon^2)$ samples (independent of n). Namely,

Theorem 3. ([DDS12]) *For all* $n, \epsilon > 0$*, there is an algorithm that uses* $\tilde{O}(1/\epsilon^2)$ *samples from an unknown* $P \in \mathcal{PBD}_n$ *and outputs some* $P_{\text{pbd}} \in \mathcal{PBD}_n$ *such that*

- P_{pbd} *is supported on an interval of length* $O(1/\epsilon^3)$*,*
- *or* P_{pbd} *is a Binomial distribution.*

Moreover, with probability ≥ 0.99*,* $d_{TV}(P_{\text{pbd}}, P) < \epsilon$ *and, if the algorithm outputs a Binomial distribution, the standard deviation of the output distribution* P_{pbd} *is within a factor of* 2 *of the standard deviation of the unknown distribution* P*. Furthermore, the running time of the algorithm is* $\tilde{O}(\log n) \cdot (1/\epsilon)^{O(\log^2 1/\epsilon)}$*.* ■

In the same paper the authors show that the mean and variance of a Poisson Binomial distribution can be estimated using a few samples. They use the empirical mean and variance estimates and bound the means and variances of these estimates.

Lemma 1. ([DDS12]) *For all* $\epsilon' > 0$*, there is an algorithm that, using* $O(1/\epsilon'^2)$ *samples from an unknown PBD with mean* μ *and variance* σ^2*, produces estimates* μ'*,* σ' *such that*

$$|\mu - \mu'| < \epsilon' \cdot \sigma \quad \text{and} \quad |\sigma^2 - \sigma'^2| < \epsilon' \cdot \sigma^2\sqrt{4 + \frac{1}{\sigma^2}},$$

with probability $\geq .99$*.* ■

Since Poisson Binomial variables are sums of indicators, the following bound is also helpful.

Lemma 2. (Chernoff bound for sums of Indicators [Lev]) *Let* $X_1, \ldots, X_n$ *be independent Bernoulli random variables,* $X = X_1 + \ldots + X_n$*, and* $\sigma^2 = \mathrm{Var}(X)$*. Then, for* $0 < \lambda < 2\sigma$*,*

$$\mathrm{Prob}\left(|X - \mathbb{E}[X]| > \lambda\sigma\right) < 2e^{-\lambda^2/4}.$$

The following result is a tail bound on Poisson random variables obtained from the Chernoff Bounds [MU05].

Lemma 3. ([ADJ+12]) *If* X *is a Poisson* λ *random variable, then for* $x \geq \lambda$*,*

$$\Pr(X \geq x) \leq \exp\left(-\frac{(x-\lambda)^2}{2x}\right),$$

and for $x \leq \lambda$*,*

$$\Pr(X \leq x) \leq \exp\left(-\frac{(x-\lambda)^2}{2\lambda}\right).$$

Poisson Binomial distributions are specified by n parameters, and consequently there has been a great deal of interest in approximating them via distributions with only a few parameters. One such class of distributions are *Translated Poisson distributions*, defined next.

Definition 5. ([Rol07]) *A* Translated Poisson distribution*, denoted* $P_{tp}(\mu, \sigma^2)$*, is the distribution of a random variable* $Y = \lfloor \mu - \sigma^2 \rfloor + Z$*, where* Z *is a random variable distributed according to the Poisson distribution* $\mathrm{poi}\left(\sigma^2 + \{\mu - \sigma^2\}\right)$*. Here* $\lfloor x \rfloor$ *and* $\{x\}$ *denote respectively the integral and fractional parts of* x *respectively.*

The following lemma bounds the total variation and ℓ_∞ distance between a Poisson Binomial and a Translated Poisson distribution with the same mean and variance. The bound on total variation distance is taken directly from [Rol07], while the ℓ_∞ bound is obtained via simple substitutions in their ℓ_∞ bound.

Lemma 4. ([Rol07]) *Let* $X = \sum_i X_i$*, where the* X_i*'s are independent Bernoulli random variables,*

$X_i \sim B(p_i)$. Also, let $q_{\max} = \max_k \Pr(X = k)$, $\mu = \sum p_i$ and $\sigma^2 = \sum p_i(1-p_i)$. The following hold:

$$d_{TV}(X, P_{tp}(\mu,\sigma^2)) \le \frac{2+\sqrt{\sum p_i^3(1-p_i)}}{\sum p_i(1-p_i)};$$
$$\ell_\infty\left(X, P_{tp}(\mu,\sigma^2)\right) \le \frac{2+2\sqrt{q_{\max}\sum p_i^3(1-p_i)}}{\sum p_i(1-p_i)};$$
$$q_{\max} \le d_{TV}(X, P_{tp}(\mu,\sigma^2)) + \frac{1}{2.3\sigma}.$$

Finally, the total variation distance between two Translated Poisson distributions can be bounded as follows.

Lemma 5. ([BL07]) *Let P_{tp1} and P_{tp2} be Translated Poisson distributions with parameters $(\mu_1,\ \sigma_1^2)$ and $(\mu_2,\ \sigma_2^2)$ respectively. Then,*

$$d_{TV}(P_{tp1}, P_{tp2}) \le \frac{|\mu_1-\mu_2|}{\min\{\sigma_1,\sigma_2\}} + \frac{|\sigma_1^2-\sigma_2^2|+1}{\min\{\sigma_1^2,\sigma_2^2\}}.$$

3 Testing PBD's

We fill in the details of the outline provided in Section 1.1. Our algorithm is given in Algorithm 2 in the appendix.

Our algorithm starts by running the algorithm of Theorem 3 with accuracy $\epsilon/10$ to find $P_{\text{pbd}} \in \mathcal{PBD}_n$. If the unknown distribution $P \in \mathcal{PBD}_n$, then with probability ≥ 0.99, $d_{TV}(P, P_{\text{pbd}}) \le \epsilon/10$.

As in the outline, we next consider two cases, depending on the variance of P_{pbd}:[3]

- **Sparse Case:** when $Var(P_{\text{pbd}}) < \frac{C\cdot \text{logt}^4 1/\epsilon}{\epsilon^8}$.

- **Heavy case:** when $Var(P_{\text{pbd}}) \ge \frac{C\cdot \text{logt}^4 1/\epsilon}{\epsilon^8}$.

Clearly, if the distribution P_{pbd} given by Theorem 3 is supported on an interval of length $O(1/\epsilon^3)$, then we must be in the sparse case. Hence, the only way we can be in the heavy case is when P_{pbd} is a Binomial distribution with variance larger than our threshold. We treat the two cases separately next.

3.1 Sparse case Our goal is to perform a simple tolerant identity test to decide whether $d_{TV}(P, P_{\text{pbd}}) \le \epsilon/10$ or $d_{TV}(P, P_{\text{pbd}}) > \epsilon$. We first develop the tolerant identity test.

Simple Tolerant Identity Test: The test is given in the appendix as Algorithm 1 and is based on the folklore result described as Lemma 6.

Lemma 6. *Let $\epsilon > 0$, and P be an arbitrary distribution over a finite set A of size $|A| = m$. With $O(m/\epsilon^2)$ independent samples from P, we can compute a distribution Q over A such that $d_{TV}(P,Q) \le \epsilon$, with probability at least .99. In fact, the empirical distribution achieves this bound.*

Lemma 6 enables the simple tolerant identity tester, whose pseudocode is given in Algorithm 1, which takes $O(m/\epsilon^2)$ samples from a distribution P over m elements and outputs whether it is $\le \epsilon/10$ close or $> 2\epsilon/5$ far from a known distribution Q. The simple idea is that with sufficiently many samples, the empirical distribution $\hat{P}$ satisfies $d_{TV}(P,\hat{P}) < \epsilon/10$ (by Lemma 6), which allows us to distinguish between $d_{TV}(P,Q) \le \epsilon/10$ and $d_{TV}(P,Q) > 2\epsilon/5$.

Finishing the Sparse Case: Lemma 2 implies that there exists an interval $\mathcal{I}$ of length $O(\frac{1}{\epsilon^4}\cdot \text{logt}^{2.5}\frac{1}{\epsilon})$ such that $P_{\text{pbd}}(\mathcal{I}) \ge 1-\epsilon/5$. Let us find such an interval $\mathcal{I}$, and consider the distribution P' that equals P on $\mathcal{I}$ and places all remaining probability on -1. Similarly, let us define P'_{pbd} from P_{pbd}. It is easy to check that:

- if $P \in \mathcal{PBD}_n$, then $d_{TV}(P', P'_{\text{pbd}}) \le \epsilon/10$, since $d_{TV}(P, P_{\text{pbd}}) \le \epsilon/10$ and P', P'_{pbd} are coarsenings of P and P_{pbd} respectively.

- if $d_{TV}(P, \mathcal{PBD}_n) > \epsilon$, then $d_{TV}(P', P'_{\text{pbd}}) > 2\epsilon/5$. (This follows easily from the fact that P_{pbd} places less than $\epsilon/5$ mass outside of $\mathcal{I}$.)

Hence, we can use our simple tolerant identity tester (Algorithm 1) to distinguish between these cases from $O(|\mathcal{I}|/\epsilon^2) = O(\frac{1}{\epsilon^6}\cdot \text{logt}^{2.5}\frac{1}{\epsilon})$ samples.

3.2 Heavy case In this case, it must be that $P_{\text{pbd}} = \text{Binomial}(n', p)$ and $n'p(1-p) \ge \frac{C\cdot \text{logt}^4 \frac{1}{\epsilon}}{\epsilon^8}$. The high level plan for this case is the following:

1. First, using Theorem 3, we argue that, if $P \in \mathcal{PBD}_n$, then its variance is also $\Omega\left(\frac{\text{logt}^4\frac{1}{\epsilon}}{\epsilon^8}\right)$ large.

2. Next, we apply Lemma 1 with $O(n^{1/4}/\epsilon^2)$ samples to get estimates $\hat{\mu}$ and $\hat{\sigma}^2$ of the mean and variance of P. If $P \in \mathcal{PBD}_n$, then these estimates are very accurate, with probability at least 0.99, and, by Lemmas 4 and 5, the corresponding Translated Poisson distribution $P_{tp}(\hat{\mu}, \hat{\sigma}^2)$ is $\frac{\epsilon^2}{C'\sqrt{\text{logt}\frac{1}{\epsilon}}}$-close to P, for our choice of C' (that we can tune by choosing C large enough).

[3] Notice that in defining our two cases we use the truncated logarithm instead of the logarithm function in our threshold variance. This choice is made for trivial technical reasons. Namely, this logarithmic factor will appear in denominators later on, and it is useful to truncate it to avoid singularities.

3. Then, with a little preprocessing, we can get to a state where we need to distinguish between the following, for any $C' \geq 10$ of our choice:
   (a) $P \in \mathcal{PBD}_n$ and P is $\frac{\epsilon^2}{C'\sqrt{\text{logt}\frac{1}{\epsilon}}}$-*close to* $P_{tp}(\hat{\mu}, \hat{\sigma}^2)$ OR
   (b) $d_{TV}(P, \mathcal{PBD}_n) > \epsilon$ and P is $3\epsilon/10$-*far from* $P_{tp}(\hat{\mu}, \hat{\sigma}^2)$.
4. Finally, using the ℓ_∞ bound in Lemma 4, we show that, if the first case holds, then P is close to $P_{tp}(\hat{\mu}, \hat{\sigma}^2)$ even in the ℓ_2 distance. Using this, we show that it suffices to design an ℓ_2 test against $P_{tp}(\hat{\mu}, \hat{\sigma}^2)$. The computations of this algorithm are similar to the $\chi-$squared statistic used for testing closeness of distributions in [ADJ+12, CDVV14].

We proceed to flesh out these steps:

Step 1: By Theorem 3, if $P \in \mathcal{PBD}_n$ and a Binomial(n', p) is output by the algorithm of [DDS12] then P's variance is within a factor of 4 from $n'p(1-p)$, with probability at least 0.99. Hence, if $P \in \mathcal{PBD}_n$ and we are in the heavy case, then we know that, with probability ≥ 0.99:

$$Var(P) > \frac{C \cdot \text{logt}^4 \frac{1}{\epsilon}}{4\epsilon^8}. \tag{3.1}$$

Going forward, if $P \in \mathcal{PBD}_n$, we condition on (3.1), which happens with good probability.

Step 2: Let us denote by μ and σ^2 the mean and variance of the unknown P. If $P \in \mathcal{PBD}_n$, then clearly $\sigma^2 \leq n/4$. So let us use $\epsilon' = \frac{\epsilon}{(n/4)^{1/8}}$ in Lemma 1 to compute estimates $\hat{\mu}$ and $\hat{\sigma}^2$ of μ and σ^2 respectively. Given that $\sigma > 1$ for a choice of $C \geq 4$ in (3.1), we get:

Claim 1. *If $P \in \mathcal{PBD}_n$ and $C \geq 4$, then the outputs $\hat{\mu}$ and $\hat{\sigma}^2$ of the algorithm of Lemma 1 computed from $O(n^{1/4}/\epsilon^2)$ samples from P satisfy the following with probability ≥ 0.99:*

$$|\mu - \hat{\mu}| < \frac{\epsilon}{\sigma^{1/4}}\sigma \quad \text{and} \quad |\sigma^2 - \hat{\sigma}^2| < 3\frac{\epsilon}{\sigma^{1/4}}\sigma^2. \tag{3.2}$$

Using these bounds, we show next that $P_{tp}(\hat{\mu}, \hat{\sigma}^2)$ is a good approximator of P, if $P \in \mathcal{PBD}_n$.

Claim 2. *If $P \in \mathcal{PBD}_n$ and (3.1), (3.2) hold, then for any constant C' there exists large enough C:*

$$d_{TV}(P, P_{tp}(\hat{\mu}, \hat{\sigma}^2)) < \frac{3}{\sigma} + \frac{14\epsilon}{\sigma^{1/4}} \leq \frac{\epsilon^2}{C'\sqrt{\text{logt}\frac{1}{\epsilon}}}. \tag{3.3}$$

Proof. We first note that for large enough C we have $\sigma > 256$ so (3.2) implies that $\frac{\sqrt{7}}{2}\sigma > \hat{\sigma} > \hat{\sigma} > \frac{\sigma}{2}$. By the first bound of Lemma 4 we have that:

$$d_{TV}(P, P_{tp}(\mu, \sigma^2)) < \frac{2+\sigma}{\sigma^2} < \frac{3}{\sigma}. \tag{3.4}$$

Using (3.2) and $\hat{\sigma} > \sigma/2$ in Lemma 5 gives

$$d_{TV}(P_{tp}(\mu, \sigma^2), P_{tp}(\hat{\mu}, \hat{\sigma}^2)) < 14\frac{\epsilon}{\sigma^{1/4}}.$$

So, from triangle inequality, $d_{TV}(P, P_{tp}(\hat{\mu}, \hat{\sigma}^2)) < \frac{3}{\sigma} + \frac{14\epsilon}{\sigma^{1/4}}$. Plugging (3.1) into this bound gives $d_{TV}(P, P_{tp}(\hat{\mu}, \hat{\sigma}^2)) \leq \frac{\epsilon^2}{C'\sqrt{\text{logt}\frac{1}{\epsilon}}}$, when C is large enough.

Going forward, for any C' of our choice (to be determined), we choose C to be large enough as required by Claims 1 and 2. In particular, this choice ensures $\sigma > 256$, and $\frac{\sqrt{7}}{2}\sigma > \hat{\sigma} > \frac{\sigma}{2}$. Moreover, if $P \in \mathcal{PBD}_n$, we condition on (3.2) and (3.3), which hold with good probability.

Step 3: We do some pre-processing that allows us to reduce our problem to distinguishing between cases 3a and 3b. Given our work in Steps 1 and 2, if $P \in \mathcal{PBD}_n$, then with good probability,

$$d_{TV}(P_{\text{pbd}}, P_{tp}(\hat{\mu}, \hat{\sigma}^2)) \leq \frac{\epsilon}{10} + \frac{\epsilon^2}{C'\sqrt{\text{logt}\frac{1}{\epsilon}}} < \epsilon/5, \tag{3.5}$$

for $C' \geq 10$. Given that $P_{tp}(\hat{\mu}, \hat{\sigma}^2)$ and P_{pbd} are explicit distributions we can compute (without any samples) an estimate $\hat{d}_{TV}(P_{tp}(\hat{\mu}, \hat{\sigma}^2), P_{\text{pbd}}) = d_{TV}(P_{tp}(\hat{\mu}, \hat{\sigma}^2), P_{\text{pbd}}) \pm \epsilon/5$. Based on this estimate we distinguish the following cases:

- If $\hat{d}_{TV}(P_{tp}(\hat{\mu}, \hat{\sigma}^2), P_{\text{pbd}}) > \epsilon/2$, we can safely deduce $d_{TV}(P, \mathcal{PBD}_n) > \epsilon$. Indeed, if $P \in \mathcal{PBD}_n$, then by (3.5) we would have seen
$$\begin{aligned}&\hat{d}_{TV}(P_{tp}(\hat{\mu}, \hat{\sigma}^2), P_{\text{pbd}})\\ \leq &d_{TV}(P_{tp}(\hat{\mu}, \hat{\sigma}^2), P_{\text{pbd}}) + \epsilon/5 \leq 2\epsilon/5.\end{aligned}$$
- If $\hat{\sigma}^2 > n/2$, we can also safely deduce that $d_{TV}(P, \mathcal{PBD}_n) > \epsilon$. Indeed, if $P \in \mathcal{PBD}_n$, then $\sigma^2 \leq n/4$, hence $\hat{\sigma}^2 \leq n/2$ by our assumption that $\frac{\sqrt{7}}{2}\sigma > \hat{\sigma}$.
- So it remains to consider the case $\hat{d}_{TV}(P_{tp}(\hat{\mu}, \hat{\sigma}^2), P_{\text{pbd}}) \leq \epsilon/2$. This implies
$$\begin{aligned}&d_{TV}(P, P_{tp}(\hat{\mu}, \hat{\sigma}^2))\\ \geq &d_{TV}(P, P_{\text{pbd}}) - d_{TV}(P_{\text{pbd}}, P_{tp}(\hat{\mu}, \hat{\sigma}^2))\\ \geq &d_{TV}(P, P_{\text{pbd}}) - \frac{7\epsilon}{10}. \qquad (3.6)\end{aligned}$$

Now, if $d_{TV}(P, \mathcal{PBD}_n) > \epsilon$, then (3.6) implies $d_{TV}(P, P_{tp}(\hat{\mu}, \hat{\sigma}^2)) > \frac{3\epsilon}{10}$. On the other hand, if $P \in \mathcal{PBD}_n$, then (3.3) implies $d_{TV}(P, P_{tp}(\hat{\mu}, \hat{\sigma}^2)) \leq \frac{\epsilon^2}{C'\sqrt{\mathrm{logt}\frac{1}{\epsilon}}}$, for any $C' \geq 10$ of our choice. So, it suffices to be able to distinguish between cases 3a and 3b.

Step 4: We now show that an ℓ_2 test suffices for distinguishing between Cases 3a and 3b. We start by bounding the ℓ_2 distance between P and $P_{tp}(\hat{\mu}, \hat{\sigma}^2)$ in the two cases of interest. We start with the easy bound, corresponding to Case 3b.

Case 3b: Using Cauchy-Schwarz Inequality, we can lower bound the ℓ_2 distance between P and $P_{tp}(\hat{\mu}, \hat{\sigma}^2)$ in this case.

CLAIM 3. *$d_{TV}(P, P_{tp}(\hat{\mu}, \hat{\sigma}^2)) > 3\epsilon/10$ implies:*

$$\ell_2^2(P, P_{tp}(\hat{\mu}, \hat{\sigma}^2)) > \frac{c\epsilon^2}{\hat{\sigma}\sqrt{\mathrm{logt}(1/\epsilon)}},$$

for some absolute constant c.

Proof. By Lemma 3, $P_{tp}(\hat{\mu}, \hat{\sigma}^2)$ assigns $\geq 1 - \epsilon/10$ of its mass to an interval $\mathcal{I}$ of length $O(\hat{\sigma}\sqrt{\log(1/\epsilon)})$. Therefore, the ℓ_1 distance between P and $P_{tp}(\hat{\mu}, \hat{\sigma}^2)$ over $\mathcal{I}$ is at least $d_{TV}(P, P_{tp}(\hat{\mu}, \hat{\sigma}^2)) - \epsilon/10 > 3\epsilon/10 - \epsilon/10 > 0.2\epsilon$. Applying the Cauchy-Schwarz Inequality, *over this interval*:

$$\ell_2^2(P, P_{tp}(\hat{\mu}, \hat{\sigma}^2))_{\mathcal{I}} \geq \frac{\ell_1^2(P, P_{tp}(\hat{\mu}, \hat{\sigma}^2))_{\mathcal{I}}}{|\mathcal{I}|} \geq \frac{c\epsilon^2}{\hat{\sigma}\sqrt{\log(1/\epsilon)}},$$

for some constant $c > 0$. In the above inequality we denote by $\ell_1(P, P_{tp}(\hat{\mu}, \hat{\sigma}^2))_{\mathcal{I}}$ and $\ell_2(P, P_{tp}(\hat{\mu}, \hat{\sigma}^2))_{\mathcal{I}}$ the ℓ_1 and ℓ_2 norms respectively of the vectors obtained by listing the probabilities assigned by P and $P_{tp}(\hat{\mu}, \hat{\sigma}^2)$ on all points in set $\mathcal{I}$.

Case 3a: Rollin's result stated in Section 2 provides bounds on the ℓ_1 and ℓ_∞ distance between a PBD and its corresponding translated Poisson distribution. Using these bounds we can show:

CLAIM 4. *In Case 3a:*

$$\ell_2^2(P, P_{tp}(\hat{\mu}, \hat{\sigma}^2)) \leq \frac{5.3}{\hat{\sigma}} \cdot \frac{2\epsilon^2}{C'\sqrt{\mathrm{logt}\frac{1}{\epsilon}}}.$$

Proof. Recall that, by our choice of C, $\sigma > 256$, hence (3.2) implies that $\frac{\sqrt{7}}{2}\sigma > \hat{\sigma} > \frac{\sigma}{2} > 128$. Next, we recall the following bound on the ℓ_2 norm of any two distributions P and Q:

$$\ell_2^2(P, Q) \leq \ell_\infty(P, Q)\ell_1(P, Q). \tag{3.7}$$

Claim 2 takes care of the ℓ_1 term when we substitute $Q = P_{tp}(\hat{\mu}, \hat{\sigma}^2)$. We now bound the ℓ_∞ term. For any distributions P and Q it trivially holds that $\ell_\infty(P, Q) < \max\{\max_i\{P(i)\}, \max_i\{Q(i)\}\}$. By the third part of Lemma 4 and Equation (3.4):

$$\max_i P(i) \leq d_{TV}(P, P_{tp}(\mu, \sigma^2)) + \frac{1}{2.3\sigma} \leq \frac{4}{\sigma} \leq \frac{5.3}{\hat{\sigma}}.$$

To bound $\max_i Q(i)$, a standard Stirling approximation on the definition of Translated Poisson distribution shows that the largest probability of $P_{tp}(\hat{\mu}, \hat{\sigma}^2)$ is at most $1.5/\hat{\sigma}$. Hence, $\ell_\infty(P, P_{tp}(\hat{\mu}, \hat{\sigma}^2)) \leq \frac{5.3}{\hat{\sigma}}$. Plugging the above bounds into Equation (3.7) with $Q = P_{tp}(\hat{\mu}, \hat{\sigma}^2)$ shows the result.

Claims 4 and 3 show that the ratio of the squared ℓ_2 distance between P and $P_{tp}(\hat{\mu}, \hat{\sigma}^2)$ in Case 3b versus Case 3a can be made larger than any constant, by choosing C' large enough. To distinguish between the two cases we employ a tolerant ℓ_2 identity test, based on an unbiased estimator T_n of the squared ℓ_2 distance between P and $P_{tp}(\hat{\mu}, \hat{\sigma}^2)$ described in Section 3.2.1. We will see that distinguishing between Cases 3a and 3b boils down to showing that in the latter case:

$$Var(T_n) \ll \mathbb{E}[T_n]^2.$$

3.2.1 Unbiased ℓ_2^2 Estimator.

Throughout this section we assume that distributions are sampled in the Poisson sampling framework, where the number of samples K drawn from a distribution are distributed as a Poisson random variable of some mean k of our choice, instead of being a fixed number k of our choice. This simplifies the variance computations by inducing independence among the number of times each symbol appears, as we discuss next. Due to the sharp concentration of the Poisson distribution, the number of samples we draw satisfies $K \leq 2k$, with probability at least $1 - (\frac{e}{4})^k$.

Suppose $K \sim \mathrm{poi}\,(k)$ samples X_1^K are generated from a distribution P_1 over $[n]$. Let K_i be the random variable denoting the number of appearances of symbol i. Then K_i is distributed according to $\mathrm{poi}\,(\lambda_i)$, where $\lambda_i \stackrel{\mathrm{def}}{=} kP_1(i)$, independently of all other K_j's. Let also $\lambda_i' \stackrel{\mathrm{def}}{=} kP_2(i)$, and define:

(3.8)
$$T_n = T_n(X_1^K, P_2) = \frac{1}{k^2}\sum_{i\in[n]}\Big[(K_i - \lambda_i')^2 - K_i\Big].$$

A straightforward, albeit somewhat tedious, computation involving Poisson moments shows that

LEMMA 7.

$$\mathbb{E}[T_n] = \ell_2^2(P_1, P_2) = \frac{1}{k^2}\sum_{i=1}^{n}(\lambda_i - \lambda_i')^2$$

and

$$Var(T_n) = \frac{2}{k^4}\sum_{i=1}^{n}\left[\lambda_i^2 + 2\lambda_i(\lambda_i - \lambda_i')^2\right].$$

We use Lemma 7 with $P_1 = P$ and $P_2 = P_{tp}(\hat{\mu}, \hat{\sigma}^2)$ to bound the variance of T_n in terms of its squared expected value in Case 3b, where $d_{TV}(P, P_{tp}(\hat{\mu}, \hat{\sigma}^2)) > 0.3\epsilon$.

LEMMA 8. *There is an absolute constant C_1 such that if*

$$k \geq C_1\frac{\sqrt{\hat{\sigma}\cdot \mathrm{logt}(1/\epsilon)}}{\epsilon^2},$$

and $d_{TV}(P, P_{tp}(\hat{\mu}, \hat{\sigma}^2)) > 0.3\epsilon$, then

$$\mathrm{Var}(T_n) < \frac{1}{20}\mathbb{E}[T_n]^2.$$

Proof. From Lemma 7:

$$(3.9)\quad Var(T_n) = \frac{2}{k^4}\sum_{i=1}^{n}\lambda_i^2 + \frac{4}{k^4}\sum_{i=1}^{n}\lambda_i(\lambda_i - \lambda_i')^2.$$

We will show that each term on the right hand side is less than $\mathbb{E}[T_n]^2/40$:

- The second term can be bounded by using the following:

$$\sum_i \lambda_i(\lambda_i' - \lambda_i)^2 \overset{(a)}{\leq} \Big[\sum_i \lambda_i^2\Big]^{\frac{1}{2}}\Big[\sum_i(\lambda_i' - \lambda_i)^4\Big]^{\frac{1}{2}} \overset{(b)}{\leq} \Big[\sum_i \lambda_i^2\Big]^{\frac{1}{2}}\Big[\sum_i(\lambda_i' - \lambda_i)^2\Big],$$

  where (a) uses the Cauchy-Schwarz inequality and (b) follows from the fact that, for positive reals $a_1, \ldots a_n$, $(\sum_i a_i)^2 \geq \sum a_i^2$. Therefore, to bound the second term of the right hand side of (3.9) by $\mathbb{E}[T_n]^2/40$ it suffices to show that

$$4\Big[\sum_i \lambda_i^2\Big]^{\frac{1}{2}}\Big[\sum_i(\lambda_i' - \lambda_i)^2\Big] < \frac{1}{40}\Big[\sum_i(\lambda_i' - \lambda_i)^2\Big]^2,$$

  which holds if

$$(3.10)\qquad \Big[\sum_i \lambda_i^2\Big] < \frac{1}{160^2}\Big[\sum_i(\lambda_i' - \lambda_i)^2\Big]^2.$$

- To bound the first term of the right hand side of (3.9) by $\mathbb{E}[T_n]^2/40$ it suffices to show that

$$(3.11)\qquad \Big[\sum_i \lambda_i^2\Big] < \frac{1}{80}\Big[\sum_i(\lambda_i' - \lambda_i)^2\Big]^2.$$

Note that (3.10) is stronger than (3.11). Therefore, we only prove (3.10). Recall, from the proof of Claim 4, that $\max_i P_{tp}(\hat{\mu}, \hat{\sigma}^2)(i) \leq \frac{1.5}{\hat{\sigma}}$. Using $\lambda_i' = kP_{tp}(\hat{\mu}, \hat{\sigma}^2)(i)$ and $\sum \lambda_i = k$,

$$\begin{aligned}\sum_i(\lambda_i' - \lambda_i)^2 &> \sum_i \lambda_i^2 - 2\sum_i \lambda_i\lambda_i' \\ &\geq \sum_i \lambda_i^2 - \frac{3k}{\hat{\sigma}}\sum_i \lambda_i \\ &= \sum_i \lambda_i^2 - \frac{3k^2}{\hat{\sigma}},\end{aligned}$$

and hence

$$\sum_i(\lambda_i' - \lambda_i)^2 + \frac{3k^2}{\hat{\sigma}} > \sum_i \lambda_i^2.$$

Let $y \overset{\text{def}}{=} \sum_i(\lambda_i' - \lambda_i)^2$. It suffices to show that

$$\frac{1}{160^2}y^2 > y + \frac{3k^2}{\hat{\sigma}},$$

which holds if the following conditions are satisfied: $y > 2\cdot 160^2$ and $y^2 > 6\cdot 160^2\frac{k^2}{\hat{\sigma}}$. By Claim 3,

$$y = \sum_i(\lambda_i' - \lambda_i)^2 > \frac{k^2 c\epsilon^2}{\hat{\sigma}\sqrt{\mathrm{logt}(1/\epsilon)}},$$

so the conditions hold as long as:

$$k \geq C_1\frac{\sqrt{\hat{\sigma}\cdot \mathrm{logt}(1/\epsilon)}}{\epsilon^2},$$

and C_1 is a large enough constant.

3.2.2 Finishing the Heavy Case. Recall that the ratio of the squared ℓ_2 distance between P and $P_{tp}(\hat{\mu}, \hat{\sigma}^2)$ in Case 3b versus Case 3a can be made larger than any constant, by choosing C' large enough. This follows from Claims 3 and 4. Let us choose C' so that this ratio is > 100. Now let us draw $K \sim \mathrm{poi}\,(k)$ samples from the unknown distribution P, where $k \geq C_1\frac{\sqrt{\hat{\sigma}\cdot \mathrm{logt}(1/\epsilon)}}{\epsilon^2}$ and C_1 is determined by Lemma 8, and compute T_n using (3.8) with $P_1 = P$ and $P_2 = P_{tp}(\hat{\mu}, \hat{\sigma}^2)$.

By Lemma 7, $\mathbb{E}[T_n] = \ell_2^2(P, P_{tp}(\hat{\mu}, \hat{\sigma}^2))$. Moreover:

- In Case 3a, by Markov's Inequality, T_n does not exceed 10 times its expected value with probability at least 0.9.
- In Case 3b, from Chebychev's Inequality and Lemma 8 it follows that

$$\text{Prob}\left(|T_n - \mathbb{E}[T_n]| > \frac{\mathbb{E}[T_n]}{\sqrt{2}}\right) < \frac{1}{10}.$$

It follows that we can distinguish between the two cases with probability at least 0.9 by appropriately thresholding T_n. One possible value for the threshold is one quarter of the bound of Claim 3. This is the threshold used in Algorithm 2. This concludes the proof of correctness of the heavy case algorithm.

3.3 Correctness and Sample Complexity of Overall Algorithm.

We have argued the correctness of our algorithm conditioning on various events. The overall probability of correctness is at least $0.99^2 \cdot 0.9 \geq 0.75$. Indeed, one 0.99 factor accounts for the success of Theorem 3, if $P \in \mathcal{PBD}_n$. If the algorithm continues in the sparse case, the second 0.99 factor accounts for the success of the Simple Tolerant Identity Test, and we don't need to pay the factor of 0.9. If the algorithm continues in the heavy case, the second 0.99 factor accounts for the success of Lemma 1, if $P \in \mathcal{PBD}_n$, and the 0.9 factor accounts for the success of the ℓ_2 test. (In this analysis, we have assumed that we use fresh samples, each time our algorithm needs samples from P.) Clearly, running the algorithm $O(\log(1/\delta))$ times and outputting the majority of answers drives the probability of error down to any desired δ, at a cost of a factor of $O(\log(1/\delta))$ in the overall sample complexity.

Let us now bound the sample complexity of our algorithm. It is easy to see that the expected number of samples is as desired, namely: $O\left(\frac{n^{1/4}\sqrt{\text{logt}(1/\epsilon)}}{\epsilon^2} + \frac{\text{logt}^{2.5}(1/\epsilon)}{\epsilon^6}\right)$ times a factor of $O(\log 1/\delta)$ from repeating $O(\log 1/\delta)$ times. (It is easy to convert this to a worst-case bound on the sample complexity, by adding an extra check to our algorithm that aborts computation whenever $K \geq \Omega(k)$.)

4 Lower Bound

We now show that any algorithm for PbdTesting requires $\Omega(n^{1/4}/\epsilon^2)$ samples, in the spirit of [Pan04, ADJ+12].

Our lower bound will be based on constructing two classes of distributions $\mathcal{P}'$ and $\mathcal{Q}'$ such that

(a) $\mathcal{P}'$ consists of the single distribution $P_0 \overset{\text{def}}{=} \mathit{Binomial}(n, 1/2)$.

(b) a uniformly chosen Q from $\mathcal{Q}'$ satisfies $d_{TV}(Q, \mathcal{PBD}_n) > \epsilon$ with probability > 0.99.

(c) any algorithm that succeeds in distinguishing P_0 from a uniformly chosen distribution from $\mathcal{Q}'$ with probability > 0.6 requires $\geq c \cdot n^{1/4}/\epsilon^2$ samples, for an absolute constant $c > 0$.

Suppose k_{min} is the least number of samples required for PbdTesting with success probability $> 2/3$. We show that if the conditions above are satisfied then

$$k_{min} \geq c \cdot n^{1/4}/\epsilon^2.$$

The argument is straight-forward as we can use the PbdTesting algorithm with k_{min} samples to distinguish P_0 from a uniformly chosen $Q \in \mathcal{Q}'$, by just checking whether $Q \in \mathcal{PBD}_n$ or $d_{TV}(Q, \mathcal{PBD}_n) > \epsilon$. The success probability of the algorithm is at least $2/3 \cdot 0.99 > 0.6$. Indeed, by (b) a uniformly chosen distribution from $\mathcal{Q}'$ is at least ϵ away from $\mathcal{PBD}_n$ with probability > 0.99, and the PbdTesting algorithm succeeds with probability $> 2/3$ on those distributions. Along with (c) this proves the lower bound on k_{min}.

We now construct $\mathcal{Q}'$. Since P_0 is $\mathit{Binomial}(n, 1/2)$,

$$P_0(i) = \binom{n}{i}\left(\frac{1}{2}\right)^n,$$

and $P_0(i) = P_0(n - i)$.

Without loss of generality assume n is even. For each of the $2^{n/2}$ vectors $z_0 z_1 \ldots z_{n/2-1} \in \{-1, 1\}^{n/2}$, define a distribution Q over $\{0, 1, \ldots, n\}$ as follows, where c is an absolute constant specified later.

$$Q(i) = \begin{cases} (1 - c\epsilon z_i)P_0(i) & \mathit{if} \quad i < n/2, \\ P_0(n/2) & \mathit{if} \quad i = n/2, \\ (1 + c\epsilon z_{n-i})P_0(i) & \mathit{otherwise}. \end{cases}$$

The class $\mathcal{Q}'$ is the collection of these $2^{n/2}$ distributions. We proceed to prove (b) and (c).

Proof of Item (b): We need to prove that a uniformly picked distribution in $\mathcal{Q}'$ is ϵ−far from $\mathcal{PBD}_n$ with probability > 0.99. Since Poisson Binomials are log-concave, and hence unimodal, it will suffice to show that in fact distributions in $\mathcal{Q}'$ are ϵ−far from all unimodal distributions. The intuition for this is that when a distribution is picked at random from $\mathcal{Q}'$ it is equally likely to be above P_0 or under P_0 at any point i of its support. Since P_0 is a well behaved function, namely it varies smoothly around its mean, we expect then that typical distributions in $\mathcal{Q}'$ with have a lot of modes.

We say that a distribution $Q \in \mathcal{Q}'$ has a mode at i if

$$Q(i{-}1) < Q(i) > Q(i{+}1) \text{ or } Q(i{-}1) > Q(i) < Q(i{+}1).$$

We consider $i \in \mathcal{I}_l \overset{\text{def}}{=} [n/2 - 4\sqrt{n}, n/2 + 4\sqrt{n}]$. Then by Lemma 2 for P_0,

$$P_0(\mathcal{I}_l) \geq 1 - 2e^{-8} > 0.99.$$

Note that for $i \in \mathcal{I}_l$,

$$\frac{P_0(i+1)}{P_0(i)} = \frac{n-i}{i+1} \in \left[1 - \frac{18}{\sqrt{n}}, 1 + \frac{18}{\sqrt{n}}\right]. \tag{4.12}$$

We need the following simple result to show that most distributions in $\mathcal{Q}'$ are $\epsilon-$far from all unimodal distributions.

CLAIM 5. *Suppose $a_1 \geq a_2$ and $b_1 \leq b_2$, then*

$$|a_1 - b_1| + |a_2 - b_2| \geq |a_1 - a_2|$$

Proof. By the triangle inequality,

$$|a_1 - b_1| + |b_2 - a_2| \geq |a_1 - a_2 + b_2 - b_1| \geq a_1 - a_2.$$

Using $a_1 \geq a_2$ proves the result.

Consider any unimodal distribution R over $[n]$. Suppose its unique mode is at j. Suppose $j{\geq}n/2$ (the other possibility is treated symmetrically) and R is increasing until j. Then for $Q \in \mathcal{Q}'$, let $\mathcal{I}_l \ni i < j$ be such that $Q(i) = P_0(i) \cdot (1 + c \cdot \epsilon)$ and $Q(i+1) = P_0(i+1) \cdot (1 - c \cdot \epsilon)$. If $c > 200$ and $\epsilon > 100/\sqrt{n}$, then by (4.12), $Q(i+1) < Q(i)$ (for large enough n), and therefore

$$\begin{aligned}
&|Q(i+1) - R(i+1)| + |Q(i) - R(i)|\\
\geq& Q(i) - Q(i+1)\\
=& P_0(i) \cdot (1 + c \cdot \epsilon) - P_0(i+1) \cdot (1 - c \cdot \epsilon)\\
\geq& P_0(i) \cdot c\epsilon.
\end{aligned}$$

This can be used to lower bound the ℓ_1 distance from a typical distribution $Q \in \mathcal{Q}'$ to any unimodal distribution. Simple Chernoff bounds show that a randomly chosen string of length $\Theta(\sqrt{n})$ over $\{+1, -1\}$ has $\Theta(\sqrt{n})$ occurrences of +1-1 and -1+1 in consecutive locations with high probability. Moreover, note that in the interval $\mathcal{I}_l$, $P_0(i) = \Theta(1/\sqrt{n})$. Using this along with the bound above shows that taking c large enough proves that a random distribution in $\mathcal{Q}'$ is $\epsilon-$far from all unimodal distributions with high probability.

Proof of Item (c): We consider the distribution obtained by picking a distribution uniformly from $\mathcal{Q}'$ and generating $K = \text{poi}\,(k)$ samples from it. (By the concentration of the Poisson distribution, it suffices to prove a lower bound w.r.t. the mean k of the Poisson.) Let $\bar{Q}^k$ denote the distribution over poi (k) length samples thus generated. Since a distribution is chosen at random, the z_i's are independent of each other. Therefore, K_i, the number of occurrences of symbol i is independent of all K_j's except $j = n - i$. Using this we get the following decomposition

$$\begin{aligned}
&\bar{Q}^k(K_0 = k_0, \ldots, K_n = k_n)\\
=&\prod_{i=0}^{n/2} \bar{Q}^k(K_i = k_i, K_{n-i} = k_{n-i}).
\end{aligned}$$

Now K_i and K_{n-i} are generated either by poi (λ_i^-), where $\lambda_i^- \overset{\text{def}}{=} k(1 - c\epsilon)P_0(i)$, or by poi (λ_i^+), where $\lambda_i^+ \overset{\text{def}}{=} k(1 + c\epsilon)P_0(i)$ with equal probability. Therefore:

$$\begin{aligned}
&\bar{Q}^k(K_i = k_i, K_{n-i} = k_{n-i})\\
=&\frac{1}{2}[\text{poi}(\lambda_i^+, k_i)\text{poi}(\lambda_i^-, k_{n-i}) + \text{poi}(\lambda_i^-, k_i)\text{poi}(\lambda_i^+, k_{n-i})]
\end{aligned} \tag{4.13}$$

$$\begin{aligned}
=&\frac{1}{2}\frac{e^{-2kP_0(i)}}{k_i!k_{n-i}!}(kP_0(i))^{k_i+k_{n-i}}\cdot\\
&\left[(1+c\epsilon)^{k_i}(1-c\epsilon)^{k_{n-i}} + (1-c\epsilon)^{k_i}(1+c\epsilon)^{k_{n-i}}\right].
\end{aligned}$$

Let P_0^k denote distribution over poi (k) samples from the Binomial P_0. By independence of multiplicities,

$$\begin{aligned}
&P_0^k(K_1 = k_1, \ldots, K_n = k_n)\\
=&\prod_{i=1}^{n} \text{poi}(kP_0(i), k_i)\\
=&\frac{e^{-kP_0(i)}}{k_i!}(kP_0(i))^{k_i}.
\end{aligned}$$

Our objective is to bound $d_{TV}(P_0^k, \bar{Q}^k)$. We use the following.

LEMMA 9. ([DL01]) *For any distributions P and Q*

$$2d_{TV}(P, Q)^2 \leq \log \mathbb{E}_Q\left[\frac{Q}{P}\right].$$

Proof. By Pinsker's Inequality [CT06], and concavity of logarithms,

$$2d_{TV}(P, Q)^2 \leq KL(Q, P) = \mathbb{E}_Q\left[\log\frac{Q}{P}\right] \leq \log\left[\mathbb{E}_Q\frac{Q}{P}\right].$$

We consider the ratio of $\bar{Q}^k$ to P_0^k, and obtain

$$\frac{\bar{Q}^k(K_0=k_0,\dots,K_n=k_n)}{P_0^k(K_0=k_0,\dots,K_n=k_n)}$$
$$=\prod_{i=0}^{\frac{n}{2}-1}\frac{\bar{Q}^k(K_i=k_i,K_{n-i}=k_{n-i})}{P_0^k(K_i=k_i)P_0^k(K_{n-i}=k_{n-i})}$$
$$=\prod_{i=0}^{\frac{n}{2}-1}\frac{(1+c\epsilon)^{k_i}(1-c\epsilon)^{k_{n-i}}+(1-c\epsilon)^{k_i}(1+c\epsilon)^{k_{n-i}}}{2}$$

where we used (4.13). We can use this now to calculate the following expectation

$$\mathbb{E}_{\bar{Q}^k}\left[\frac{\bar{Q}^k}{P_0^k}\right]$$
$$=\prod_{i=0}^{n/2-1}\left[\sum_{k_i\geq 0,k_{n-i}\geq 0}\bar{Q}^k(K_i=k_i,K_{n-i}=k_{n-i})\cdot\right.$$
$$\left.\frac{1}{2}\left((1+c\epsilon)^{k_i}(1-c\epsilon)^{k_{n-i}}+(1-c\epsilon)^{k_i}(1+c\epsilon)^{k_{n-i}}\right)\right].$$

For $X\sim\text{poi}\,(\lambda)$, elementary calculus shows that

$$\mathbb{E}[a^X]=e^{\lambda(a-1)}.$$

Combining with (4.13), and using $P_0(i)=P_0(n-i)$, the above expression simplifies to

$$\mathbb{E}_{\bar{Q}^k}\left[\frac{\bar{Q}^k}{P_0^k}\right]$$
$$=\prod_{i=0}^{n/2-1}\frac{1}{2}\left[e^{c\epsilon k(1+c\epsilon)P_0(i)}e^{-c\epsilon k(1-c\epsilon)P_0(i)}\right.$$
$$\left.+e^{-c\epsilon k(1+c\epsilon)P_0(i)}e^{c\epsilon k(1-c\epsilon)P_0(i)}\right]$$
$$=\prod_{i=0}^{n/2-1}\frac{1}{2}\left[e^{2c^2\epsilon^2kP_0(i)}+e^{-2c^2\epsilon^2kP_0(i)}\right]$$
$$\leq e^{2c^4\epsilon^4k^2\sum_{i=0}^{n/2-1}P_0(i)^2},$$

where the last step uses, $e^x+e^{-x}\leq 2e^{x^2/2}$. Using Stirling's approximation, we get:

$$\sum_{i=0}^{n/2-1}P_0(i)^2\leq\max_i P_0(i)\leq P_0\left(\frac{n}{2}\right)=\binom{n}{\frac{n}{2}}\frac{1}{2^n}\leq\frac{1}{\sqrt{n}}.$$

Therefore,

$$d_{TV}(P_0^k,\bar{Q}^k)^2\leq\frac{c^4\epsilon^4k^2}{\sqrt{n}}.$$

Unless $k=\Omega(n^{1/4}/\epsilon^2)$, there is no test to distinguish a distribution picked uniformly from $\mathcal{Q}'$ versus P_0. This proves item (c).

Acknowledgements

The authors thank Piotr Indyk, Gautam Kamath, and Ronitt Rubinfeld for helpful discussions and thoughts. Jayadev Acharya thanks Ananda Theertha Suresh for valuable suggestions.

References

[ADJ+12] Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Shengjun Pan, and Ananda Theertha Suresh. Competitive classification and closeness testing. In *COLT*, pages 22.1–22.18, 2012. 3, 4, 4

[BFF+01] Tugkan Batu, Lance Fortnow, Eldar Fischer, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. Testing random variables for independence and identity. In *FOCS*, pages 442–451, 2001. 1, 1.1

[BKR04] Tugkan Batu, Ravi Kumar, and Ronitt Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *STOC*, pages 381–390. ACM, 2004. 1

[BL07] A. Barbour and Torgny Lindvall. Translated poisson approximation for markov chains. *Journal of Theoretical Probability*, 2007. 5

[BS10] Mikhail Belkin and Kaushik Sinha. Polynomial learning of distribution families. In *FOCS*, pages 103–112, 2010. 1

[CDSS13] Siu On Chan, Ilias Diakonikolas, Rocco A. Servedio, and Xiaorui Sun. Learning mixtures of structured distributions over discrete domains. In *SODA*, pages 1380–1394, 2013. 1

[CDVV14] Siu On Chan, Ilias Diakonikolas, Paul Valiant, and Gregory Valiant. Optimal algorithms for testing closeness of discrete distributions. In *SODA*, pages 1193–1203, 2014. 4

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements of information theory (2. ed.)*. Wiley, 2006. 4

[DDS12] Constantinos Daskalakis, Ilias Diakonikolas, and Rocco A. Servedio. Learning poisson binomial distributions. In *STOC*, pages 709–728, 2012. 1, 1.1, 1.1, 3, 1, 3.2, 2

[DDS+13] Constantinos Daskalakis, Ilias Diakonikolas, Rocco A. Servedio, Gregory Valiant, and Paul Valiant. Testing k-modal distributions: Optimal algorithms via reductions. pages 1833–1852, 2013. 1

[DKN14] Ilias Diakonikolas, Daniel M. Kane, and Vladimir Nikishkin. Testing identity of structured distributions. abs/1410.2266, 2014. 1

[DL01] Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer, 2001. 9

[FOS05] Jon Feldman, Ryan O'Donnell, and Rocco A. Servedio. Learning mixtures of product distributions over discrete domains. In *FOCS*, pages 501–510, 2005. 1

[GR00] Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. *Electronic Colloquium on Computational Complexity (ECCC)*, 7(20), 2000. 1

[Lev] Kirill Levchenko. Chernoff bound. http://cseweb.ucsd.edu/klevchen/techniques/chernoff.pdf. 2

[MU05] Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge University Press, 2005. 2

[MV10] Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *FOCS*, pages 93–102, 2010. 1

[Pan04] Liam Paninski. Variational minimax estimation of discrete distributions under kl loss. In *NIPS*, 2004. 4

[Pan08] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008. 1, 1, 1.1

[Rol07] Adrian Rollin. Translated poisson approximation using exchangeable pair couplings. *The Annals of Applied Probability*, 17(5/6):1596–1614, 10 2007. 5, 2, 4

[VV11] Gregory Valiant and Paul Valiant. Estimating the unseen: an n/log(n)-sample estimator for entropy and support size, shown optimal via new clts. In *STOC*. ACM, 2011. 1, 1.1

[VV14] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *FOCS*, 2014. (document), 1, 1.1

```
Algorithm 1: Simple Tolerant Identity Test
Input: Known $Q$ over finite set $A$ of cardinality $|A| = m$, $\epsilon > 0$, and
       independent samples $X_1, , \ldots, X_k$ from unknown $P$
Output: Close, if $d_{TV}(P, Q) \leq \epsilon/10$, and far, if $d_{TV}(P, Q) > 2\epsilon/5$
$\hat{P} \leftarrow$ empirical distribution of $X_1^k$;
if $d_{TV}(\hat{P}, Q) < 2.5\epsilon/10$ then
    output close;
else
    output far;
end
```

```
Algorithm 2: Testing PBDs
Input: Independent samples from unknown $P$ over $[n]$, $\epsilon$, $\delta > 0$
Output: With probability $\geq 0.75$, output Yes PBD, if $P \in \mathcal{PBD}_n$, and No PBD,
        if $d_{TV}(P, \mathcal{PBD}_n) > \epsilon$.
Using $\tilde{O}(1/\epsilon^2)$ samples, run the algorithm of [DDS12] with accuracy $\epsilon/10$ to obtain
$P_{\text{pbd}} \in \mathcal{PBD}_n$;
if $\sigma(P_{\text{pbd}})^2 < \frac{C \cdot \text{logt}^4 1/\epsilon}{\epsilon^8}$ then
    Find an interval $\mathcal{I}$ of length $O(\text{logt}^{2.5}(1/\epsilon)/\epsilon^4)$ such that
    $P_{\text{pbd}}(\mathcal{I}) \geq 1 - \epsilon/5$;
    Run Simple Tolerant Identity Test to compare $P$ and $P_{\text{pbd}}$ on $\mathcal{I}$, using
    $k = O\left(\frac{\text{logt}^{2.5}(1/\epsilon)}{\epsilon^6}\right)$ samples;
    if Simple Tolerant Identity test outputs close then
        output Yes PBD;
    else
        output No PBD;
    end
else
    Use $O(n^{1/4}/\epsilon^2)$ samples to estimate the mean, $\hat{\mu}$, and variance, $\hat{\sigma}^2$, of $P$;
    Calculate an estimate $\hat{d}_{TV}(P_{tp}(\hat{\mu}, \hat{\sigma}^2), P_{\text{pbd}})$ of
    $d_{TV}(P_{tp}(\hat{\mu}, \hat{\sigma}^2), P_{\text{pbd}})$ that is accurate to within $\pm\epsilon/5$;
    if $\hat{d}_{TV}(P_{tp}(\hat{\mu}, \hat{\sigma}^2), P_{\text{pbd}}) > \epsilon/2$ OR $\hat{\sigma}^2 > n/2$ then
        output No PBD;
    end
    Draw $K \sim \text{poi}(k)$ samples, where $k \geq C_1 \frac{\sqrt{\hat{\sigma} \cdot \text{logt}(1/\epsilon)}}{\epsilon^2}$ and $C_1$ is as
    determined by Lemma 8;
    Let $K_i$ be the number of samples that equal $i$;
    if $\frac{1}{k^2} \sum \left[(K_i - kP_{tp}(\hat{\mu}, \hat{\sigma}^2)(i))^2 - K_i\right] <$
    $0.25 \cdot \frac{c\epsilon^2}{\hat{\sigma}\sqrt{\text{logt}(1/\epsilon)}}$, where c is the constant from Claim 3, then
        output Yes PBD
    else
        output No PBD
    end
end
```