Testing Mixtures of Discrete Distributions

Maryam Aliakbarpour MARYAMA@MIT.EDU
CSAIL, MIT

Ravi Kumar RAVI.K53@GMAIL.COM
Google

Ronitt Rubinfeld RONITT@CSAIL.MIT.EDU
CSAIL, MIT, TAU

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

There has been significant study on the sample complexity of testing properties of distributions over large domains. For many properties, it is known that the sample complexity can be substantially smaller than the domain size. For example, over a domain of size n, distinguishing the uniform distribution from distributions that are far from uniform in ℓ_1-distance uses only $O(\sqrt{n})$ samples.

However, the picture is very different in the presence of arbitrary noise, even when the amount of noise is quite small. In this case, one must distinguish if samples are coming from a distribution that is ϵ-close to uniform from the case where the distribution is $(1-\epsilon)$-far from uniform. The latter task requires nearly linear in n samples (Valiant, 2008; Valiant and Valiant, 2017a).

In this work, we present a noise model that on one hand is more tractable for the testing problem, and on the other hand represents a rich class of noise families. In our model, the noisy distribution is a mixture of the original distribution and noise, where the latter is known to the tester either explicitly or via sample access; the form of the noise is also known *a priori*. Focusing on the identity and closeness testing problems leads to the following mixture testing question: Given samples of distributions p, q_1, q_2, can we test if p is a mixture of q_1 and q_2? We consider this general question in various scenarios that differ in terms of how the tester can access the distributions, and show that indeed this problem is more tractable. Our results show that the sample complexity of our testers are exactly the same as for the classical non-mixture case.

1. Introduction

Distribution testing (Batu et al., 2013) has been studied extensively for the past many years (see Canonne (2015) for a survey). In the vanilla version, the problem is to quickly test if a discrete distribution has a certain property or is statistically far from any distribution with that property. The tester has access to samples from the distribution and strives to be as frugal as possible in the number of samples it uses. Many statistical properties, including various distances between distributions, are well understood in this model. There have been several relaxations to the basic testing model including tolerant testing (where the tester should also accept if the distribution is close to having a property), the conditional samples model (where the tester can access the distribution conditioned on a specified subset), making stylized assumptions about the distribution (monotone, sparse support, high-dimensional, etc), and so on. In each of these works, the aim has been to push the boundaries of our understanding: when do sample-efficient testers exist? Here, by sample-efficient, we mean the number of samples should be sub-linear in the domain size.

There are many scenarios in which a distribution is observed along with noise; in some cases, even the form of the noise is known *a priori*. One such scenario is the so-called *identity testing* problem in which the tester has a known (explicitly specified) distribution and its goal is to check if a given distribution, available as samples, is close to the known distribution. For example, assume that the distribution of the top million queries to a web search engine is known in advance. Then, identity testing would be a quick way to check how close the daily query distribution is to this known distribution. However, in reality, there are natural minor variations to the daily query distribution, which may cause the identity tester to fail. This is clearly undesirable.

An option to tackle the noise would be to use testers that are *tolerant* to noise. Unfortunately, even simple versions of tolerant testers are faced with near-linear lower bounds on the sample complexity, making this option uninteresting. For example, one can distinguish if a distribution on a domain of size n is uniform or far from uniform in ℓ_1-distance using $\Theta(\sqrt{n})$ samples (Paninski, 2008). However, an algorithm that distinguishes between near-uniform distributions and distributions that are far from uniform requires $\Omega(n/\log n)$ samples (Valiant, 2008; Valiant and Valiant, 2017b). Hence, to achieve sub-linear sample complexity, we need more judicious, stylized assumptions about the noise—how it is available to the tester and if it is adversarial.

A different yet natural way to model the above scenario is to view it as a mixture of distributions. In the above example, one of the components of the mixture can be interpreted as the signal and the other component can be thought of as the noise. More generally, the tester is given the components of a mixture of two distributions. However, it does not know the mixing parameter, i.e., the magnitude of the contribution of each component to the mixture. The mixture testing problem is then to test if a distribution is close to a mixture of two given distributions or is far from any manner in which the two distributions can be mixed. As we will see, by making reasonable assumptions on the form of the noise and how it is available to the tester, the tolerant testing lower bounds can be circumvented and one can obtain testers with sub-linear sample complexity.

Main contributions. In this work, we consider distribution testing of mixtures of two distributions q_1 and q_2. For ease of exposition, let us call the first component q_1 the *original distribution* and the second component q_2 the *noise*, and let $[n] = \{1, \ldots, n\}$ be the domain of both q_1 and q_2. The simplest version of our problem is: given sample access to distribution p, and for known distributions q_1, q_2, is $p = \alpha q_1 + (1-\alpha)q_2$ for some α, or is p far from $\alpha q_1 + (1-\alpha)q_2$ for every $\alpha \in [0,1]$? Note that the tester is not given the mixture parameter α. We further study the case when q_1, q_2 are not given explicitly to the algorithm, as well as other generalizations.

We mainly focus on identity and closeness testing, which are two basic instances of hypothesis testing that have received much attention in the theory, machine learning, and statistics communities; see the works of Goldreich et al. (1998); Batu et al. (2001, 2013); Batu (2001); Batu et al. (2002, 2004); Paninski (2008); Valiant (2008); Goldreich and Ron (2011); Indyk et al. (2012); Levi et al. (2013); Daskalakis et al. (2013); Acharya et al. (2014); Chan et al. (2014); Falahatgar et al. (2015); Diakonikolas et al. (2015a,b); Acharya et al. (2015); Aliakbarpour et al. (2016); Canonne et al. (2016, 2017); Daskalakis and Pan (2017); Valiant and Valiant (2017b); Diakonikolas et al. (2018); Stewart et al. (2018); Blum and Hu (2018) and the surveys of Rubinfeld (2012) and Canonne (2015).

The mixture testing problem has a more constrained model compared to the tolerant testing problem, so one might hope to bypass the existing lower bounds. However, the mixture testing problem can also run into near-linear sample complexity lower bounds if one does not provide the tester with sufficient access to the mixture components. Indeed, if the tester does not have access to the noise, we show the mixture testing problem becomes as hard as tolerant testing, necessitating $\Omega(n/\log n)$ samples (Theorem 19). Hence, to show nontrivial positive results, the tester must have access to some kind of information about the noise. We consider the following three cases for the noise, namely, (i) when the noise is given as an explicitly specified distribution, (ii) when the tester does not explicitly know the noise distribution, but does have sample access to it, and (iii) when there is no explicit description or access to samples from the noise distribution, but it is known that the noise distribution comes from a *class* of distributions, e.g., the set of k-histogram distributions. For the first, we obtain a tester with sample complexity $\Theta(\sqrt{n}/\epsilon^2)$ and for the second, we obtain a tester with sample complexity $\Theta(\sqrt{n}/\epsilon^2 + n^{2/3}/\epsilon^{4/3})$ where ϵ is a given proximity parameter; these show that the complexity of our testers is exactly the same as for the classical non-mixture case. For the third, when the noise is assumed to come from the set of k-histogram distributions, we obtain an identity tester that uses $\tilde{O}(\sqrt{kn})$ samples.

2. Preliminaries

For the rest of the paper, we use the following notation. For a distribution p over $[n]$, we use $p(i)$ to denote the probability of element $i \in [n]$ and for a subset $S \subseteq [n]$, let $p(S) = \sum_{i \in S} p(i)$. We use $\|.\|_p$ to indicate the ℓ_p-norm of a vector. We typically use the ℓ_1-distance and say p and q are *ϵ-close* if $\|p - q\|_1 < \epsilon$ and *ϵ-far* otherwise. Let $\mathcal{U}_n$ denote the uniform distribution on $[n]$; we drop the subscript when the domain is clear from the context. Distribution p is a *mixture* of q_1 and q_2 if there exists $\alpha \in [0, 1]$ such that $p = (1 - \alpha)\, q_1 + \alpha q_2$. We call α the *mixture parameter*. We use q_α to denote the mixture $(1 - \alpha) q_1 + \alpha q_2$ when the components q_1 and q_2 are clear from the context.

Background. Through this paper we consider several *distribution testing* problems: For a given *property* of distributions, we use Π to denote a set of distributions that satisfy the property. The distance of distribution p to Π is the ℓ_1-distance between p and the closest distribution q in Π. In a *distribution testing problem*, the goal is to distinguish whether p is in Π or is ϵ-far from Π. We say an algorithm is a *tester for property* Π if the following is true with probability $2/3$. [1]

- Completeness: If p is in Π, then the algorithm outputs accept.
- Soundness: If p is ϵ-far from Π, the algorithm outputs reject.

The algorithm is an *(ϵ', ϵ)-tolerant tester*, if it also satisfies the stronger completeness property that when p is ϵ'-close to some distribution in Π, then the algorithm outputs accept (with probability at least $2/3$). These definitions can be extended to the case of properties of collections of more than one distribution. Although in the standard setting we receive samples from at least one distribution in the collection, the testing problems may be defined with respect to other methods of access.

We make one of the three following assumptions regarding the algorithm's view of the distributions: (i) The distribution is *explicitly given* or *known* if the algorithm knows the probability of each domain element under the distribution. (ii) The distribution is *given by samples* if the algorithm has access to an oracle that provides samples from the distribution. (iii) The distribution is not known nor given by samples but is a member of a given class of distributions.

The term *identity testing* is used to refer to the setting in which we test if a distribution, which we have sample access to, is equal to a known one. Note that this is equivalent to testing property

[1] The success probability of $2/3$ is arbitrary here. Given such tester, we can achieve a success probability of $1 - \delta$, via standard amplification methods, at the cost of a $\log(1/\delta)$ multiplicative increase in the sample complexity.

$\Pi = \{q\}$. The term *closeness testing* refers to the setting in which we test if two distributions, both available via samples, are equal or not; in this case, Π is the set of pairs of equal distributions.

Mixture testing problems. Suppose p, q_1, and q_2 are distributions over $[n]$. Let $\Pi_{q_1,q_2} := \{(1-\alpha)\, q_1 + \alpha\, q_2 \mid \alpha \in [0,1]\}$ (we usually drop the subscripts q_1, q_2 when they are clear from context). In a *mixture testing problem*, the goal is to distinguish whether a distribution p given via samples is in Π_{q_1,q_2} or ϵ-far from any distribution in Π_{q_1,q_2} with probability at least 2/3. We investigate the following problems, which differ in the way that mixture testing algorithm can access q_1, q_2. Note however that the mixture parameter α is *not* given to the tester. (i) An algorithm is an *identity tester in the presence of known noise* if it solves the mixture testing problem when q_1, q_2 are known to the tester. (ii) An algorithm is a *closeness tester in the presence of noise that is accessible via samples* if it solves the mixture testing problem when q_1, q_2 are not explicitly given, but samples of each are provided to the tester. (iii) An algorithm is an *identity tester in the presence of class $\mathcal{C}$-noise* if it can distinguish whether p is a mixture of a known distribution q_1 and some $q_2 \in \mathcal{C}$. Note that such an algorithm is a property tester for $\Pi := \{(1-\alpha)\, q_1 + \alpha\, q_2 \mid q_2 \in \mathcal{C}, \text{ and } \alpha \in [0,1]\}$.

Note that one can also define “closeness testing in the presence of known noise”, and “identity testing in the presence of noise that is given via samples”, but our lower bounds will show that the sample complexity of these tasks is the same as the sample complexity of closeness testing in the presence of noise that is given via samples.

3. An overview of our results and techniques

3.1. Testing identity in the presence of known noise

We first consider the problem of testing if distribution p, given via samples, can be expressed as mixture of known distributions q_1 and q_2. We show the following.

Theorem 1 *Given two known distributions q_1, q_2, and $\epsilon > 0$, there is an identity tester in the presence of known noise that uses $O(\sqrt{n}/\epsilon^2)$ samples. Furthermore, $\Omega(\sqrt{n}/\epsilon^2)$ samples are required.*

At a high level, we take the following steps to test if p is a mixture or ϵ-far from it. First, we develop an algorithm (*learner*) to learn mixture distributions. The learner receives samples from p and outputs a mixture distribution q_α. If p is a mixture, then we show that the learner finds a mixture distribution q_α that is ϵ'-close to p for some proximity parameter $\epsilon' := \Theta(\epsilon)$; and if p is not a mixture, the learner outputs q_α with no specific guarantee. Second, we use the distance between p and q_α as a measure to decide about p: if p is ϵ'-close to q_α, we accept p; and if p is ϵ-far from q_α, we reject it. This approach results in a tester for p. In fact, if p is a mixture, then we show that the learner finds a q_α that is ϵ'-close and if p is ϵ-far from being a mixture, then we show that p has to be ϵ-far from any mixture distribution, including q_α.

The challenge in this approach is to distinguish whether p is close to q_α or far from it. In general, testing whether two distributions are ϵ'-close or ϵ-far from each other requires $\Omega(n/\log n)$ samples. However, we show that we can exploit the structural properties of mixture distributions to achieve a sample-efficient algorithm. Below we provide a more detailed description of the steps.

The learner. The algorithm begins by assuming that the given distribution p is indeed a mixture, and attempts to learn the mixture parameter: If p is a mixture, then we show that it can be learned to error $\epsilon' = \Theta(\epsilon)$ using $O(1/\epsilon^2)$ samples given q_1 and q_2. The algorithm picks a subset S of elements such that it contains every element i for which $q_1(i) \geq q_2(i)$ and estimates the weight of these elements according to p, i.e., $p(S)$. S satisfies that $q_1(S) - q_2(S)$ is exactly the total variation distance between q_1 and q_2. Comparing $p(S)$ with the weight of these elements according to q_1 and q_2 guides us to choose a mixture parameter α, and allows us to bound the distance between p and $q_\alpha \coloneqq (1-\alpha)q_1 + \alpha q_2$. (Instead of learning α, one might do a grid search on $\alpha \in [0, 1]$; however the granularity required could make the resulting algorithm sub-optimal.)

Assessing the distance between p and q_α. After obtaining q_α, the task of distinguishing whether distribution p is a mixture or ϵ-far from a mixture boils down to testing if p is ϵ'-close to q_α or is ϵ-far from it. We propose a scheme to *reshape* the distributions p and q_α and get two new distributions p' and q'_α such that for p that is a mixture, the ℓ_2-distance between p' and q'_α is at most $O(\epsilon/\sqrt{n})$. Furthermore, in the case where p is ϵ-far from being a mixture, p' is ϵ-far from q'_α. It is known that one can efficiently distinguish the case that $\|p' - q'_\alpha\|_2 \leq O(\epsilon/\sqrt{n})$ versus $\|p' - q'_\alpha\|_1 \geq \epsilon$ using $O(\sqrt{n}/\epsilon^2)$ samples (Diakonikolas and Kane, 2016; Chan et al., 2014)).

Here, we elaborate further on how we reshape the distributions. Similar techniques have been used previously to reduce the ℓ_2-norm, e.g., in Diakonikolas and Kane (2016). Here, we use it to bound the ℓ_2-distance between p' and q'_α. The reshaping process is as follows. Define p', the reshaped distribution of p with a new domain which is larger than the domain of p. For each element i, we determine an integer a_i solely based on q_1, q_2, and q_α. Then we add (i, j) for all j in $[a_i]$ to the domain of p'. We set the probability of element (i, j) to be $p(i)/a_i$. Also, we reshape q_α according to the same process and get q'_α.

But how can reshaping reduce the ℓ_2-distance? Given that p is a mixture, for each element i in the domain, the discrepancy between the probability of i according to p' and q'_α, $|p(i) - q_\alpha(i)|/a_i$, is proportional to $|q_1(i) - q_2(i)|/a_i$. With this observation, we set the a_i's such that they make the discrepancy $O(\epsilon/n)$ for each element. This ensures the ℓ_2-distance between p' and q'_α is $O(\epsilon/\sqrt{n})$.

The arguments described above are formalized in Theorem 5. In addition, in the case where q_1 and q_2 are uniform, this problem is as hard as testing if a distribution is uniform, which needs at least $\Omega(\sqrt{n}/\epsilon^2)$ samples (Paninski (2008)), showing that the sample complexity of our algorithm is tight. Furthermore, we match the sample complexity of the standard identity tester where there is no noise involved.

3.2. Testing closeness in the presence of noise that is accessible via samples

We next investigate the problem of testing closeness of distributions in the presence of noise that is accessible via samples. Suppose we have sample access to three distributions, p, q_1, and q_2, over $[n]$. The goal is to test if there is a mixture parameter α^* such that $p = (1-\alpha^*)\, q_1 + \alpha^* q_2$, or p is ϵ-far from any distribution in this form.

Similarly to the identity testing algorithm explained earlier, our approach is first attempt to learn p. That is, we design an algorithm that finds a candidate mixture distribution, $q_\alpha \coloneqq (1-\alpha)q_1 + \alpha\, q_2$, such that if p is a mixture of q_1 and q_2, then p and q_α will be $(\epsilon/\sqrt{n})$-close to p in ℓ_2-distance; and if p is not a mixture, the algorithm finds a distribution q_α with no specific guarantees. Then, we test to see if p is $(\epsilon/(2\sqrt{n}))$-close to q_α in ℓ_2-distance, or $(\epsilon/\sqrt{n})$-far from it. The answer of the test dictates if we should accept or reject p. Indeed, if p is a mixture distribution, q_α is very close to p, and the test will accept p. If p is ϵ-far from being a mixture, then p is ϵ-far from q_α, and furthermore p and q_α are $(\epsilon/\sqrt{n})$-far from each other in ℓ_2-distance, so that the test will reject p.

But how do we learn p? Since we are looking for q_α, which is close to p in ℓ_2-distance, we study the problem of estimating the ℓ_2-distance between p and a mixture distribution of q_1 and q_2. Inspired by the ℓ_2-distance estimator proposed by Chan et al. (2014), we propose a statistic such that given α it estimates the ℓ_2-distance between p and q_α: $f(\alpha) := \sum_{i=1}^{n}(X_i - (1-\alpha)Y_i - \alpha Z_i)^2 - X_i - (1-\alpha)^2 Y_i - \alpha^2 Z_i$, where X_i, Y_i, and Z_i are the number of instances of element i among samples from p, q_1, and q_2 respectively. The statistic is designed such that it is equal to $s^2\|p - q(\alpha)\|_2^2$ in expectation where s is the number of samples from each distribution p, q_1, and q_2.

Given the sample sets, the goal is to use the quadratic function f to find a candidate α. For now, assume p is a mixture of q_1 and q_2 with parameter α^*. We make two observations about $f(\alpha^*)$: (i) the expectation of $f(\alpha)$ is minimum, in fact zero, when $\alpha = \alpha^*$, and (ii) we provide a threshold T for which $|f(\alpha^*)|$ is at most T with high probability. Although α^* is not given to the algorithm, we wish to pick a candidate α that is very close to α^*. We use the above two observations as a guide to take the following strategy: find α that minimizes $f(\alpha)$ while $|f(\alpha)|$ is at most T. This method apparently finds several candidate α's. We establish that if p is a mixture, then one of the candidate α's will result in a mixture distribution q_α that is $(\epsilon/2\sqrt{n})$-close to p in ℓ_2 distance. (Once again, a grid search on $\alpha \in [0,1]$ will not yield an optimal sample complexity.)

From there on, we only need to test if any of the candidates we found are $(\epsilon/2\sqrt{n})$-close to p or not. If p is a mixture we are promised that one of the candidates will pass the test. Otherwise we show that all candidates have to give distributions that are ϵ-far (implying $(\epsilon/\sqrt{n})$-far in ℓ_2-distance) from p by definition, so all of them will fail. Our approach yields the following result:

Theorem 2 *Assume we have sample access to three distributions p, q_1, and q_2 over $[n]$. There exists a closeness tester in the presence of noise that uses $O(\sqrt{n}/\epsilon^2 + n^{2/3}/\epsilon^{4/3})$ samples. Furthermore $\Omega(\sqrt{n}/\epsilon^2 + n^{2/3}/\epsilon^{4/3})$ samples are required.*

See Theorem 10 for the formal statement of the result. For the lower bound of sample complexity, we establish that the lower bound for standard closeness testers holds in the mixture setting as well, even in the case where q_1 or q_2 is known. In particular, we show given sample access to p and q_1, testing whether p is a mixture of q_1 and the uniform distribution requires $\Omega(\sqrt{n}/\epsilon^2 + n^{2/3}/\epsilon^{4/3})$ samples (Proposition 20). Hence, one cannot hope to achieve a better sample complexity.

3.3. Testing identity in the presence of k-flat noise

On the one hand, the sample complexity of distribution testing under arbitrary noise is significantly worse than that of noise-free distribution testing. On the other hand, we have seen that the sample complexity of distribution testing with noise (either known or given via sample access) is very similar to the sample complexity of noise-free distribution testing. This raises the question of whether one can relax the requirement of the access to the noise by the tester and still achieve better sample complexity. The next problem we consider is the identity testing problem when there is no direct access to the noise (either via samples, or an explicit description) except for the promise that the noise comes from a class, in particular, the class of k-flat distributions.

We say a distribution is *k-flat* if the probability mass function of the distribution is a piece-wise constant function with k pieces. We investigate the following problem: given a known distribution q and having sample access to p, can we distinguish if p is a mixture of q and some k-flat distribution, or p is ϵ-far from any such distribution? We provide an algorithm that uses $\tilde{O}(\sqrt{kn})$ samples.

Inspired by the identity tester proposed in Batu et al. (2001), we propose the following approach. First, we guess the k intervals on which the noise is constant. Then, we take the elements of each interval and further partition them into subsets (not necessarily contiguous) such that in each subset the probability of the elements according to q are very similar to each other (similar enough so that we can show that q is nearly-uniform on each subset). For a mixture distribution p, if we have guessed the intervals correctly, p is almost uniform within each subset since it is a mixture of an almost uniform q and a constant function (noise). Hence to see if p is a mixture, we first test each of these subsets and see if p is close to uniform on them. We then estimate the total weights that p assigns to each of these subsets and determine if the weights are consistent with a mixture of q and some k-flat distribution. One challenge is to find a sampling method that guarantees good results for all initial guesses of the k intervals describing the noise. See Appendix A for details.

3.4. Lower bounds

We show that testing identity with respect to the uniform distribution when the noise component can be an arbitrary distribution, requires near-linear samples, i.e., $\Omega(n/\log n)$. More specifically,

Theorem 19 *Assume p is a distribution on $[n]$. There exists a constant parameter ϵ such that distinguishing the following cases with probability at least $2/3$ requires $\Omega(n/\log n)$ samples.*

- *There exists a noise distribution on $[n]$, namely η, and an $\alpha \leq \epsilon_1$ such that p is a mixture of uniform and η with parameter α, i.e., $p = (1-\alpha)\,\mathcal{U} + \alpha\,\eta$.*
- *There is no noise distribution η such that $p = (1-\alpha)\,\mathcal{U} + \alpha\,\eta$ unless $\alpha = 1$.*

The main idea is to reduce this problem to the that of testing the T*-bigness property* (Aliakbarpour et al., 2019 (this proceedings)), which holds if all probabilities are above a given threshold T.

4. Identity testing of mixtures in the presence of known noise

In this section, we give an algorithm which tests if p is close to a mixture where both the components are explicitly known. As before, we assume the mixture parameter α is unknown.

The main idea is attempt to learn a mixture distribution q_α that is close to p. Using q_1, q_2, and q_α, we then reshape the distribution p to another distribution p' and use the same reshaping to transform q_α to q'_α. The reshaping has the property that in the case that p is indeed a mixture, then p' and q'_α will be extremely close to each other in ℓ_2-distance and if p is not a mixture, then p' and q'_α will be quite far from each other. Thus we can use a (non-tolerant) identity tester on p' and q'_α.

In the rest of this section, we present the three main steps of the testing algorithm. The first step is a learner algorithm that finds q_α (Section 4.1). The second step is a reshaping process that transforms the distributions p and q_α into p' and q'_α respectively (Section 4.2). The third step is to put these pieces together to get the identity tester (Section 4.3).

4.1. The learner

At a high level, the leaner proceeds as follows. Observe that if p is a mixture of q_1 and q_2, then there is a parameter $\alpha^* \in [0,1]$ such that $p = (1-\alpha^*)q_1 + \alpha^* q_2$, so to learn p it is sufficient to learn α^*. Let S be the set of all domain elements, x, where $q_1(x)$ is at least $q_2(x)$. By definition, for $S \subseteq [n]$, we have $p(S) = (1-\alpha^*)q_1(S) + \alpha^* q_2(S)$, which leads to $\alpha^* = (q_1(S) - p(S))\,/\,(q_1(S) - q_2(S))$.

The idea then is to replace $p(S)$ with its estimate, say, w_S to get an estimate α of α^*. We formally describe the procedure in Algorithm 1 and prove its correctness in Lemma 3.

Lemma 3 *Suppose $p = (1-\alpha^*)q_1 + \alpha^* q_2$. Using $O(1/\epsilon^2)$ samples, Algorithm 1 outputs a mixture parameter $\alpha \leq \alpha^*$ such that $\mathbf{Pr}[\, \|q_\alpha - p\|_1 < \epsilon] \geq 5/6$.*

The proof is presented in Appendix C.

4.2. Reshaping the distributions

Algorithm 1: Learning mixture of two known distributions.

```
Procedure: MIXTURE-LEARNER($q_1$, $q_2$, $n$, $\epsilon$, sample access to $p$)
if $\|q_1 - q_2\|_1 \leq \epsilon$ then
    return 0
end
$S \leftarrow \{i \in [n] \mid q_1(i) > q_2(i)\}$
$m \leftarrow O(1/\epsilon^2)$
$w_S \leftarrow$ (# samples in $S$)$/m$
$\alpha \leftarrow \dfrac{q_1(S) - w_S - \epsilon/4}{q_1(S) - q_2(S)}$
return $\alpha$
```

Using Algorithm 1, given p, we can obtain a mixture parameter α and a mixture distribution q_α for which (i) if p is the mixture of q_1 and q_2 with parameter α^*, then q_α is ϵ'-close to p for a proximity parameter $\epsilon' = \epsilon/6$, and $\alpha \leq \alpha^*$ and (ii) if p is ϵ-far from being a mixture, then p is ϵ-far from q_α. Ideally, we wish to use an identity tester to see if q_α and p are roughly the same or far from each other. Unfortunately, this is not possible in general, unless p and q_α are very close on *every* domain element. To resolve this issue, the goal in this section is to introduce two distributions p' and q'_α such that (i) when p and q_α are close, p' and q'_α are very close to each other on every domain element and (ii) when p and q_α are far, p' and q'_α are far. Our reshaping process is inspired by the method of Diakonikolas and Kane (2016): For each element $i \in [n]$, using q_α, we define:

$$a_i := \lfloor n q_\alpha(i) \rfloor + \left\lfloor \frac{n|q_\alpha(i) - q_2(i)|}{\|q_\alpha - q_2\|_1} \right\rfloor + 1.$$

Note that the process in Diakonikolas and Kane (2016) uses only the first and third terms of the above sum in defining a_i. We start the reshaping process by associating $a_i \geq 1$ buckets to each domain element $i \in [n]$ to form a new domain $D = \{(i,j) \mid i \in [n] \text{ and } j \in [a_i]\}$. To draw a sample from p', we first draw a sample i from p, then we sample j from $[a_i]$, and return the pair (i,j) as the sample from p'. We say p' is a *reshaping of* p with respect to q_α. Clearly $p'(i,j) = p(i)/a_i$. In a similar manner, we define the reshaping q'_α of q_α and once again, we have $q'_\alpha(i,j) = q_\alpha(i)/a_i$.

We next prove several crucial properties of the reshaped distributions.

Lemma 4 *Let p' and q'_α be the result of the reshaping of p and q_α with respect to q_α as described above. Then, the following hold:*

(i) The ℓ_1-distance after reshaping does not change: $\|p - q_\alpha\|_1 = \|p' - q'_\alpha\|_1$.
(ii) The domain size of p' and q'_α, $|D| \leq 3n$.
(iii) The ℓ_2-norm of q'_α, $\|q'_\alpha\|_2 \leq \sqrt{3/n}$.
(iv) If p is a mixture distribution, q_α is ϵ'-close to p, and α is at most α^*, then $|p'(i,j) - q'_\alpha(i,j)| \leq \epsilon'/n$ for all $(i,j) \in D$.

Appendix C contains the proofs.

4.3. The mixture testing algorithm

In this section, we use the learner and the reshaped distributions to obtain an identity tester for mixtures of two known distributions.

Theorem 5 *Given a proximity parameter ϵ, Algorithm 2 is identity tester in the presence of known noise that uses $O(\sqrt{n}/\epsilon^2)$ samples.*

Algorithm 2: Identity tester in the presence of known noise.

```
Procedure: MIXTURE-IDENTITY-TESTER($q_1, q_2, n, \epsilon$, sample access to $p$)
$\epsilon' \leftarrow \epsilon/6$
$\alpha \leftarrow$ MIXTURE-LEARNER($q_1, q_2, n, \epsilon'$)
for $i = 1$ to $n$ do
    $a_i \leftarrow \lfloor n q_\alpha(i) \rfloor + \left\lfloor \frac{n|q_\alpha(i) - q_2(i)|}{\|q_\alpha - q_2\|_1} \right\rfloor + 1$
    $q'_\alpha(i) = \frac{p(i)}{a_i}$
end
$x_1, \ldots, x_s \leftarrow \Theta\left(\frac{\sqrt{n}}{\epsilon^2}\right)$ samples from $p$
for $i = 1$ to $s$ do
    $r \leftarrow$ uniform at random in $[a_i]$
    $y_i \leftarrow (x_i, r)$
end
return IDENTITY-TESTER($q'_\alpha$, $\epsilon$, $\{y_1, \ldots, y_s\}$ as samples from $p'$)
```

Proof Let $\epsilon' = \epsilon/6$. In the completeness case, p is a mixture distribution with parameter α^*. Therefore, with probability at least 5/6, q_α is a mixture distribution with parameter α for which q_α is ϵ'-close to p. Let p' and q'_α be the reshaped distributions described in Section 4.2. By Lemma 4(ii), $|D| \leq 3n$. Moreover, for any $(i, j) \in D$, $|p'(i, j) - q'_\alpha(i, j)| \leq \epsilon'/n$, which implies that

$$\|p' - q'_\alpha\|_2 \leq \sqrt{\frac{|D| \cdot \epsilon'^2}{n^2}} \leq \sqrt{\frac{\epsilon^2}{12\,n}} \leq \frac{\epsilon}{2\sqrt{|D|}}\,.$$

Conversely, if p is ϵ-far from being a mixture distribution, then it has to be ϵ-far from q_α. By Lemma 4, p' and q'_α are ϵ-far from each other. Therefore, $\|p' - q'_\alpha\|_2^2 \geq \epsilon/\sqrt{|D|}$. Using the identity tester (IDENTITY-TESTER) provided in Diakonikolas and Kane (2016) (see Remark 2.7 and Remark 2.8), there exists an algorithm that can distinguish the above cases with probability $5/6$ using $O(\sqrt{|D|}/\epsilon^2)$ samples. Thus, with probability $2/3$ both the invoked learner and the tester returns the right answer. Also, the sample complexity is $O(\sqrt{n}/\epsilon^2 + 1/\epsilon^2) = O(\sqrt{n}/\epsilon^2)$. Hence the proof is complete. ∎

5. Testing mixtures in the presence of noise that is accessible via samples

In this section, we provide an algorithm for the testing closeness of distributions in the presence of noise that is accessible via samples. We assume we have sample access to three distributions p, q_1, and q_2, over $[n]$ and the goal is to test if p is a mixture of q_1 and q_2. Our approach is first to learn p in an indirect manner. Specifically, we design an algorithm that finds a candidate mixture distribution $q_\alpha := (1 - \alpha)q_1 + \alpha\, q_2$ such that with high probability if p is a mixture of q_1 and q_2, then q_α will be close to p. We claim that the answer to the test "is p close to q_α" can be used to test if p is close to a mixture of q_1 and q_2. Indeed, if p is a mixture distribution, by the property of the learning algorithm, q_α is close to p and hence the test will accept. Conversely, if p is far from being a mixture, then p is far from any mixture distribution including q_α, and hence the test will reject.

In particular, the candidate q_α will be such that (i) if p is a mixture, then $\|p - q_\alpha\|_2 \leq c\epsilon/\sqrt{n}$ for a sufficiently small constant c and (ii) if p is ϵ-far from being a mixture, then $\|p - q_\alpha\|_2 \geq \epsilon/\sqrt{n}$. As we will see, the robust ℓ_2-distance tester of Chan et al. (2014) can efficiently distinguish these two cases. Since we are looking for q_α that is close to p in ℓ_2-distance, we study how we can estimate the ℓ_2-distance between p and a mixture distribution q_1 and q_2. Let s be the expected number of samples we draw; s will be specified later. Assume we draw *Poi*(s) samples[2] from p, q_1, and q_2. Let X, Y, and Z denote the (multi)set of samples from p, q_1, and q_2 respectively. Let X_i, Y_i, and Z_i be the numbers of instances of element $i \in [n]$ in each sample set. Consider the following statistic:[3]

$$f(\alpha) := \sum_{i=1}^{n} (X_i - (1-\alpha)Y_i - \alpha Z_i)^2 - X_i - (1-\alpha)^2 Y_i - \alpha^2 Z_i. \tag{1}$$

Note that if we fix the sample sets X, Y, and Z, f is a quadratic function of α. We show that the above statistic has the expected value of $s^2\|p - q_\alpha\|_2^2$.

If p is a mixture distribution with parameter α^*, then $\mathbf{E}[\,f(\alpha^*)\,] = 0$, where the expectation is taken over the randomness of the samples. Hence, a natural candidate to approximate α^* is some α where f achieves its (near-)minimum. To do this, we first show that if p is a mixture, then we can choose a threshold parameter T such that $|f(\alpha^*)| \leq T$ with high probability. Then we pick $\alpha \in [0,1]$ that minimizes $f(\alpha)$ with the constraint that $|f(\alpha)| \leq T$. Since f is a quadratic, let $\hat{\alpha}_1$ and $\hat{\alpha}_2$ be the solutions. We then show that if p is a mixture of q_1 and q_2, with high probability, at least one of $\|p - q_{\hat{\alpha}_1}\|_2$ or $\|p - q_{\hat{\alpha}_2}\|_2$ is small (Section 5.1).

For the rest of the section, let b be a parameter specified later for which $\|p\|_2^2$, $\|q_1\|_2^2/2$, and $\|q_2\|_2^2/2$ are bounded by b. Thus for any mixture distribution q_α, we have $\|q_\alpha\|_2^2 \leq b$. Also, let $\gamma = \epsilon^2/(10n)$, $s = c_s \cdot (\sqrt{b}/(\gamma/2))$ for a sufficiently large constant c_s, and let $T = s^2\gamma$. (All missing proofs are in Appendix D.)

5.1. Finding candidates

In this section, we aim to learn a mixture distribution. More precisely, we are looking for α's such that if p is a mixture distribution, then with high probability, $\|p - q_\alpha\|_2 \leq \epsilon/(2\sqrt{n})$.

Theorem 6 *Suppose p is a mixture distribution. Given X, Y, and Z, with probability* 0.94, *one can compute a candidate set $\mathcal{M}$, $|\mathcal{M}| \leq 5$, for which there exists $\alpha \in \mathcal{M}$ such that* $\|p - q_\alpha\|_2^2 \leq \frac{\epsilon^2}{4n}$.

Proof We consider two cases based on the ℓ_2-distance of q_1 and q_2. Suppose $\|q_1 - q_2\|_2^2 \leq \epsilon^2/(4\,n)$. If p is a mixture distribution with parameter α^*, then we have:

$$\|q_1 - p\|_2^2 = \sum_{i=1}^{n} (q_1(i) - (1-\alpha^*)q_1(i) - \alpha^*\, q_2(i))^2 = \sum_{i=1}^{n} \alpha^{*2} \cdot (q_1(i) - q_2(i))^2$$
$$= \alpha^{*2} \cdot \|q_1 - q_2\|_2^2 \leq \|q_1 - q_2\|_2^2 \leq \frac{\epsilon^2}{4\,n}\,.$$

Hence, q_1, which is a mixture distribution with parameter $\alpha = 0$, is a candidate.

We now focus on the case $\|q_1 - q_2\|_2^2 \geq \epsilon^2/(4\,n)$. Without loss of generality, $\alpha^* > 1/2$ (otherwise swap q_1 and q_2). Fixing X, Y, Z, we write (1) as $f(\alpha) = A\alpha^2 + B\alpha + C$, where

$$A := \sum_{i=1}^{n} (Y_i - Z_i)^2 - Z_i - Y_i, \quad B := 2\sum_{i=1}^{n} Y_i + X_iY_i + Y_iZ_i - Y_i^2 - X_iZ_i, \quad C := \sum_{i=1}^{n} (X_i - Y_i)^2 - X_i - Y_i. \tag{2}$$

[2] *Poi*(s) is a Poisson random variable with parameter s.

[3] This is motivated by the ℓ_2-distance estimator proposed in Chan et al. (2014) in which they draw a set of samples from p and q and use the statistic $\sum_{i=1}^{n}(X_i - Y_i)^2 - X_i - Y_i$, where X_i (resp., Y_i) is the number of times $i \in [n]$ occurs in the samples from p (resp., q).

As explained earlier, the idea is to use $f(\alpha)$ to find a proper candidate α.

We now study the properties of A and B. It turns out that A is the same as the statistic for testing closeness where *Poi*(s) samples are drawn from q_1 and q_2. From Chan et al. (2014),

$$\mathbf{E}_{X,Y,Z}[A] = s^2\|q_1 - q_2\|_2^2 \quad \text{and} \quad \mathbf{Var}_{X,Y,Z}[A] \leq 8s^3 \, \|q_1 - q_2\|_4^2 \sqrt{b} + 8\, s^2\, b\,.$$

We show (Lemma 21) that with probability 0.99, there is a constant $c_A \in [0.9, 1.1]$ such that

$$A = c_A \cdot s^2\|q_1 - q_2\|_2^2. \tag{3}$$

B might not have a nice closed-form expression when p is an arbitrary distribution, but when p is a mixture, it has the following property.

Lemma 7 *Suppose p is a mixture of q_1 and q_2 with parameter $\alpha^* \geq 1/2$. Let B be a function of the sample sets X, Y, and Z as defined in (2) and let $\gamma < \|q_1 - q_2\|_2^2$. If the sample sets, X, Y, and Z, each have $\Theta(\sqrt{b}/\gamma)$ samples, then with probability 0.99, there exists $c_B \in [0, 1]$ such that*

$$B = -2\, c_B \cdot \alpha^* \|q_1 - q_2\|_2^2\,. \tag{4}$$

We now analyze $f(\alpha)$ for a fixed α. (The following in fact holds for any distribution p over $[n]$.)

Lemma 8 *For a fixed α,*

$$\mathbf{E}_{X,Y,Z}[f(\alpha)] = s^2 \cdot \|p - q_\alpha\|_2^2 \quad \textit{and} \quad \mathbf{Var}_{X,Y,Z}[f(\alpha)] \leq 8\, s^3 \cdot \sqrt{b} \cdot \|p - q_\alpha\|_4^2 + 8\, s^2 \cdot b\,.$$

By Lemma 8 and Lemma 21, with probability 0.99, if $p = (1 - \alpha^*)q_1 + \alpha^* q_2$, then

$$|f(\alpha^*)| \leq s^2\gamma = T. \tag{5}$$

With probability 0.97, all of (3), (4), and (5) hold; we condition on this from now on.

Since $f(\alpha)$ is a quadratic and since $A > 0$ from (3), let $\alpha_{\min} = -B/(2A)$ where f achieves its minimum. We define $\hat{\alpha}_1$ and $\hat{\alpha}_2$ as follows:

$$\hat{\alpha}_1 = \mathop{\arg\min}_{\alpha \in [\alpha_{\min}, 1], f(\alpha) \leq T} f(\alpha), \quad \text{and} \quad \hat{\alpha}_2 = \mathop{\arg\min}_{\alpha \in [0, \alpha_{\min}], f(\alpha) \leq T} f(\alpha)\,. \tag{6}$$

Note that (5) guarantees that either $\hat{\alpha}_1$ or $\hat{\alpha}_2$ exists depending on if $\alpha^* > \alpha_{\min}$ or not; they can also be found very efficiently by binary search. It remains to show that one of $q_{\hat{\alpha}_1}$ and $q_{\hat{\alpha}_2}$ is very close to p in ℓ_2-distance.

Lemma 9 *We have either $\|p - q_{\hat{\alpha}_1}\|_2 \leq \frac{2\,T}{0.9\, s^2}$ or $\|p - q_{\hat{\alpha}_2}\|_2 \leq \frac{2\,T}{0.9\, s^2}$.*

Note that by choice of our parameters, we have $2\,T/(0.9s^2) < \epsilon^2/(4\,n)$. Hence, either $q_{\hat{\alpha}_1}$ or $q_{\hat{\alpha}_2}$ is a candidate. Thus our potential candidates so far are $\alpha = 0$, $q_{\hat{\alpha}_1}$, and $q_{\hat{\alpha}_1}$. In addition, given our assumption for $\alpha^* > 1/2$, we need to compute the corresponding $\hat{\alpha}_1$ and $\hat{\alpha}_2$ when q_1 and q_2 are swapped. Hence, we have at most five candidates for α. ■

5.2. Mixture closeness tester

In this section, we provide our algorithm and prove its correctness in the following theorem.

Theorem 10 *Given a proximity parameter ϵ, Algorithm 3 is an closeness tester in the presence of noise that is accessible via samples and it uses $\Theta(\sqrt{n}/\epsilon^2 + n^{2/3}/\epsilon^{4/3})$ samples.*

Algorithm 3: Closeness tester in the presence of noise that is accessible via samples.

```
Procedure: MIXTURE-CLOSENESS-TESTER ($n$, $\epsilon$, sample access to $p$, $q_1$, $q_2$)
$k \leftarrow \Theta\left(\min\left(n, n^{2/3}/\epsilon^{4/3}\right)\right)$ samples to reduce the $\ell_2$-norm of the distribution.
$p'$, $q'_1$, $q'_2 \leftarrow p$, $q_1$, and $q_2$ after flattening.
$b \leftarrow \Theta\left(1/k\right)$
$s \leftarrow \Theta(n \cdot \sqrt{b}/\epsilon^2)$
$X, Y, Z \leftarrow$ Poi($s$) samples from each distribution $p'$, $q'_1$, and $q'_2$.
$\mathcal{C} \leftarrow$ set of candidates
for $\alpha \in \mathcal{C}$ do
    if $\ell_2^2$-DIST-ESTIMATOR$\left(b, \frac{\epsilon^2}{4n}, q'_\alpha, p'\right) \leq \frac{\epsilon^2}{2n}$ then
        return accept
    end
end
return reject
```

Proof We reduce the ℓ_2-norm of the three input distributions via the reshaping technique proposed in Diakonikolas and Kane (2016). Let S be a multi-set consisting of $3k$ samples, where k samples are chosen from each distribution p, q_1, and q_2. For $i \in [n]$, we assign b_i buckets to element i where b_i is the number of instances of element i in set S plus one. For a distribution d over $[n]$, we define d' to be a distribution over all the buckets, $D := \{(i, j) \mid i \in [n] \text{ and } j \in [b_i]\}$. We generate a sample from d' via the following process: (i) draw a sample $i \sim d$, (ii) pick $j \in [b_i]$ uniformly at random, and (iii) output (i, j). The probability of any element (i, j) according to d' is $d(i)/b_i$. It is known that flattening does not change the ℓ_1-distance between two distributions. Let p', q'_1, and q'_2 be the distributions p, q_1, and q_2 after flattening. We show that a mixture distribution will remain a mixture after flattening. More precisely, if p is a mixture of q_1 and q_2 with parameter α^*, then it is easy to see that p' is a mixture of distributions q'_1 and q'_2 with the same parameter α^*. Thus, it suffices to test if p' is a mixture of q'_1 and q'_2.

By setting $k = \Theta(\min(n, n^{2/3}/\epsilon^{4/3}))$, according to (Diakonikolas and Kane, 2016, Lemma II.6) and Markov's inequality, we can assume the ℓ_2-norms of all three distributions p', q'_1, and q'_2 are at most b with probability at least 0.99, where we set $b = 1/\min(n, n^{2/3}/\epsilon^{4/3}) = \Theta(1/k)$. Also, note tht $|D| = \Theta(n)$.

Given Theorem 6, one can find a set $\mathcal{M}$ of at most five candidates. If p' is a mixture of q'_1 and q'_2, then there is an $\alpha \in \mathcal{M}$ such that $\|q'_\alpha - p'\|_2 \leq \epsilon/(2\sqrt{|D|})$. On the other hand, if p' is ϵ-far from being a mixture, it is also ϵ-far from all $\alpha \in \mathcal{M}$; using the Cauchy–Schwarz inequality, we have $\|q'_\alpha - p'|_2 \geq \epsilon/\sqrt{|D|}$. Note that Chan et al. (2014) showed one can estimate the ℓ_2-distance accurately using $\Theta(|D| \cdot \sqrt{b}/\epsilon^2)$ samples and with probability 0.99 (see Lemma 22 in Appendix D.)

By a union bound, the probability that $\mathcal{M}$ does not contain the right α, the probability that the ℓ_2 estimation fails, and the probability that *Poi*$(\lambda) < 100\lambda$ sum up to below $1/3$. Hence, with probability 2/3, the algorithm outputs the right answer and the total number of samples is $\Theta(k + n\sqrt{b}/\epsilon^2) = \Theta(\sqrt{n}/\epsilon^2 + n^{2/3}/\epsilon^{4/3})$. ■

Acknowledgments

MA is supported by funds from the MIT-IBM Watson AI Lab (Agreement No. W1771646), the NSF grants IIS-1741137, and CCF-1733808. RR is supported by by funds from the MIT-IBM Watson AI Lab (Agreement No. W1771646) the NSF grants CCF-1650733, CCF-1733808, IIS-1741137 and CCF-1740751.

References

Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda T. Suresh. Sublinear algorithms for outlier detection and generalized closeness testing. In *IEEE ISIT*, pages 3200–3204, 2014.

Jayadev Acharya, Costantinoss Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. In *NIPS*, pages 3591–3599, 2015.

Maryam Aliakbarpour, Eric Blais, and Ronitt Rubinfeld. Learning and testing junta distributions. In *COLT*, pages 19–46, 2016.

Maryam Aliakbarpour, Themis Gouleakis, John Peebles, Ronitt Rubinfeld, and Anak Yodpinyanee. Towards testing monotonicity of distributions over general posets. In *COLT*, 2019 (this proceedings).

Tugkan Batu. *Testing Properties of Distributions*. PhD thesis, Cornell University, 2001.

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

Tugkan Batu, Sanjoy Dasgupta, Ravi Kumar, and Ronitt Rubinfeld. The complexity of approximating entropy. In *STOC*, pages 678–687, 2002.

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

Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing closeness of discrete distributions. *JACM*, 60(1):4:1–4:25, 2013.

Avrim Blum and Lunjia Hu. Active tolerant testing. In *COLT*, pages 474–497, 2018.

Clément L Canonne. A survey on distribution testing: Your data is big. but is it blue? *ECCC*, 22: 63, 2015.

Clément L. Canonne, Ilias Diakonikolas, Themis Gouleakis, and Ronitt Rubinfeld. Testing shape restrictions of discrete distributions. In *STACS*, pages 25:1–25:14, 2016.

Clément L. Canonne, Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Testing Bayesian networks. In *COLT*, pages 370–448, 2017.

Siu-on Chan, Ilias Diakonikolas, Paul Valiant, and Gregory Valiant. Optimal algorithms for testing closeness of discrete distributions. In *SODA*, pages 1193–1203, 2014.

Constantinos Daskalakis and Qinxuan Pan. Square Hellinger subadditivity for bayesian networks and its applications to identity testing. In *COLT*, pages 697–703, 2017.

Constantinos Daskalakis, Ilias Diakonikolas, Rocco A. Servedio, Gregory Valiant, and Paul Valiant. Testing k-modal distributions: Optimal algorithms via reductions. In *SODA*, pages 1833–1852, 2013.

Ilias Diakonikolas and Daniel M. Kane. A new approach for testing properties of discrete distributions. In *FOCS*, pages 685–694, 2016.

Ilias Diakonikolas, Daniel M. Kane, and Vladimir Nikishkin. Testing identity of structured distributions. In *SODA*, pages 1841–1854, 2015a.

Ilias Diakonikolas, Daniel M. Kane, and Vladimir Nikishkin. Optimal algorithms and lower bounds for testing closeness of structured distributions. In *FOCS*, pages 1183–1202, 2015b.

Ilias Diakonikolas, Themis Gouleakis, J. Peebles, and Eric Price. Collision-based testers are optimal for uniformity and closeness. *ECCC*, 23:178, 2016.

Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Sharp bounds for generalized uniformity testing. In *NeurIPS*, pages 6204–6213, 2018.

Moein Falahatgar, Ashkan Jafarpour, Alon Orlitsky, Venkatadheeraj Pichapati, and Ananda Theertha Suresh. Faster algorithms for testing under conditional sampling. In *COLT*, pages 607–636, 2015.

Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. In *Studies in Complexity and Cryptography. Miscellanea on the Interplay between Randomness and Computation*, pages 68–75. Springer, 2011.

Oded Goldreich, Shafi Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. *JACM*, 45:653–750, 1998.

Piotr Indyk, Reut Levi, and Ronitt Rubinfeld. Approximating and testing k-histogram distributions in sub-linear time. In *PODS*, pages 15–22, 2012.

Reut Levi, Dana Ron, and Ronitt Rubinfeld. Testing properties of collections of distributions. *Theory of Computing*, 9(8):295–347, 2013.

Liam Paninski. A coincidence-based test for uniformity given very sparsely-sampled discrete data. *IEEE TOIT*, 54:4750–4755, 2008.

Ronitt Rubinfeld. Taming big probability distributions. *XRDS*, 19(1):24–28, 2012.

Alistair Stewart, Ilias Diakonikolas, and Clément L. Canonne. Testing for families of distributions via the Fourier transform. In *NeurIPS*, pages 10084–10095, 2018.

Gregory Valiant and Paul Valiant. Estimating the unseen: Improved estimators for entropy and other properties. *JACM*, 64(6):37:1–37:41, 2017a.

Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SICOMP*, 46(1):429–455, 2017b.

Paul Valiant. Testing symmetric properties of distributions. In *STOC*, pages 383–392, 2008.

Appendix A. Testing under k-flat noise

We have so far considered the problems of identity testing and closeness testing in the presence of the noise that is directly accessible and proved these problems have the same sample complexity as their respective noise-free versions. These results raise the question of whether one can replace the requirement of access to the noise by an assumption that restricts the noise to be in a *class* of distributions and still achieve improved sample complexity compared to the near-linear lower bound we mentioned earlier. In this section we develop a tester for identity testing when the noise distribution belongs to the class of *k-flat distributions* without any further information. This assumption means that the noise can be *any* k-flat distribution, while the parameters of the k-flat distribution are not known to the tester, nor given via samples.

A.1. Preliminaries

We begin by formally defining k-flat distributions: We say $\mathcal{I} = \{I_1, \ldots, I_k\}$ is a *k-segmentation* of $[n]$ if and only if $I_1, \ldots, I_k$ are k disjoint intervals that cover $[n]$. Also, we say a function $f : [n] \to \mathbb{R}$ is a *k-flat function* if and only if there is a k-segmentation of $[n]$, namely $\mathcal{I} = \{I_1, \ldots, I_k\}$, such that for any two elements, x and y, in the same interval in $\mathcal{I}$, $f(x)$ is equal to $f(y)$. A distribution is a *k-flat distribution* if and only if its probability mass function is a k-flat function.

We next define concepts that will be necessary for describing our algorithms. For any distribution p and a partition $\mathcal{D} = \{D_1, \ldots D_t\}$ of its domain, the *coarsening* of p over $\mathcal{D}$, denoted by $p_{\langle D\rangle}$, is a distribution over the sets in $\mathcal{D}$ where the probability of each set D_i is $\sum_{x\in D_i} p(x)$. For a subset $D \subseteq [n]$, we define the *restriction* of p to D, denoted by $p_{|D}$, to be a distribution over D for which the probability of $x \in D$ is equal to $p(x \mid x \in D)$. Although the restriction is well-defined only when $p(D)$ is not zero, abusing notation, we define $\|p_{|D} - q_{|D}\|_1$ to be zero if $p(D)$ or $q(D)$ is zero.

Also, throughout this section, we study different schemes for partitioning the domain. In addition to k-segmentation, which is defined earlier, two other schemes are defined as follows: Given a known distribution q, Batu et al. (2001) provide a partitioning scheme, called *bucketing*, which places elements with similar probability in the same bucket. Note that, in contrast with k-segmentation, this scheme does not necessarily place consecutive elements in the same bucket.

Definition 11 (Similar to Batu et al. (2001)) *Assume we have a known distribution q over $[n]$. Given a parameter ϵ, we define the* bucketing *of the domain,* $\textsc{Bucket}(q, n, \epsilon)$*, to be a set of v subsets of the domain, $\mathcal{B} = \{B_1, \ldots, B_v\}$, where each subset is defined as below:*

$$B_1 \stackrel{\text{def}}{=} \left\{ x \in [n] \,\middle|\, q(x) \le \frac{\epsilon^2}{n} \right\}, \text{ and}$$

$$B_i \stackrel{\text{def}}{=} \left\{ x \in [n] \,\middle|\, \frac{(1+\epsilon)^i \epsilon^2}{n} < q(x) \le \frac{(1+\epsilon)^{i+1}\epsilon^2}{n} \right\} \quad \text{for } i = 2, \ldots, v\,.$$

We define the last partitioning scheme below. This partition is a refinement of the bucketing with respect to a k-segmentation $\mathcal{I}$.

Definition 12 *Assume $\mathcal{I}$ is a k-segmentation of $[n]$, and $\mathcal{B}$ is a bucketing of $[n]$ containing v disjoint subsets. We define $\mathcal{D}(\mathcal{I}, \mathcal{B}) = \{D_{i,j,\ell}\}_{(i,j)\in[k]\times[v]}$ to be* a division of the domain *for which the $D_{i,j,\ell}$'s are the intersection of the ith interval and the jth bucket. Formally, $D_{i,j,\ell}$ is defined as:*

$$D_{i,j,\ell} \coloneqq \{x \in [n] \,|\, x \in I_i \cap B_j\} \, .$$

The problem of testing identity in the presence of k-flat noise. Suppose we are given a known distribution q, and sample access to a distribution p both over the domain $[n]$. Let $\mathcal{C}$ denote the class of all k-flat distributions over $[n]$. The problem of testing identity in the presence of k-flat noise boils down to distinguishing the following cases with probability at least 2/3:

- There exists a mixture parameter α^* and a k-flat distribution r^* over $[n]$ such that p is a mixture of q and r^* with parameter α^*, i.e., $p = (1-\alpha^*)q + \alpha^* \, r^*$.
- p is ϵ-far from any distribution of the form $(1-\alpha)q + \alpha \, r$ where $r \in \mathcal{C}$ and $\alpha \in [0, 1]$.

A.2. The algorithm

We start by explaining the properties of the partitioning schemes we defined earlier. Let $\mathcal{B} = \textsc{Bucket}(q, n, \epsilon')$ be the bucketing of the domain elements for a parameter $\epsilon' \coloneqq \epsilon/14$. The algorithm can obtain this bucketing since q and ϵ is known to the algorithm. The bucketing scheme is designed such that the probabilities of the elements in a bucket are within a $(1+\epsilon)$-factor of each other (except for B_1). This property implies that the restriction of q to any bucket is extremely close to the uniform distribution.

Now, assume that p is in fact a mixture of q and a k-flat distribution r^*. We denote the k-segmentation of r^* by $\mathcal{I}^* = \{I_1^*, \ldots, I_k^*\}$ (which is not known to the algorithm). By definition, the restriction of r^* on any $I_i^* \in \mathcal{I}^*$ is a uniform distribution. Consider the division $\mathcal{D}(\mathcal{I}^*, \mathcal{B})$, described in Definition 12. Observe that $D_{i,j,\ell} \in \mathcal{D}$ is a subset of both I_i and B_j. One can show that the restriction of r^* is uniform on $D_{i,j,\ell}$, and the restriction of q to $D_{i,j,\ell}$ is very close to the uniform distribution as well. Thus, p, which is assumed to be the mixture of q and r^*, must be very close to the uniform distribution on $D_{i,j,\ell}$. We formally prove this claim in Lemma 15.

Based on the above observation, our tester looks for two qualities in p to assert that it is a mixture distribution: Given a division $\mathcal{D}(\mathcal{I}, \mathcal{B})$, (i) are the restrictions of p to the $D_{i,j,\ell}$'s almost uniform and (ii) is the overall shape of p over $D_{i,j,\ell}$'s (i.e., $p_{\langle \mathcal{D} \rangle}$) consistent with a mixture of q and a k-flat noise distribution? More specifically, our tester follows these steps. For every k-segmentation $\mathcal{I}$, the tester checks that the restriction of p to each $D_{i,j,\ell} \in \mathcal{D}(\mathcal{I}, \mathcal{B})$ is almost uniform. If it figures out that it is not the case, it abandons the current segmentation, and start over with another one. If at some point, the tester passes this step, it checks the overall shape of p. It draws enough samples from p and forms the empirical distribution $\hat{p}$ from the samples. Then it checks whether there exists a k-flat *function*, f, such that $\hat{p}_{\langle \mathcal{D} \rangle}$ is consistent with a mixture of q and f. If the tester finds a k-segmentation such that the distribution passes the two steps above, then it asserts that p is a mixture and outputs accept. Otherwise, it outputs reject.

Based on our first observation, one can expect the tester to accept a mixture distribution p. However, the main challenge is to show that the tester rejects when p is ϵ-far from being a mixture. To

Algorithm 4: Identity testing in the presence of k-flat noise

```
Procedure: IDENTITY-TESTER-K-FLAT-NOISE($q, n, \epsilon$, sample access to $p$)
$\epsilon' \leftarrow \epsilon/14$
$\mathcal{B} \leftarrow (q, n, , \epsilon')$
$M \leftarrow$ Multiset of $s = \widetilde{\Theta}_\epsilon(k\sqrt{n})$ i.i.d. samples from $p$.
for every possible $k$-segmentation $\mathcal{I}$ do
    $\mathcal{D} \leftarrow \mathcal{D}(\mathcal{I}, \mathcal{B})$
    for $i = 1$ to $k$ do
        for $j = 2$ to $v$ do
            $M_{i,j} \leftarrow M \cap D_{i,j,\ell}$
            if $|M_{i,j}| \geq \epsilon' s/4(k \cdot v)$ then
                if $\|p_{|D_{i,j,\ell}} - \mathcal{U}_{|D_{i,j,\ell}}\|_2 \geq 2\,\epsilon'/\sqrt{|D_{i,j,\ell}|}$ then
                    Continue with another segmentation.
                end
            end
        end
    end
    $\hat{p} \leftarrow$ Empirical distribution built by samples in $M$
    for $\alpha = 0, \epsilon'/2, \epsilon', \ldots, 1$ do
        $f \leftarrow$ find a $k$-flat function on $\mathcal{I}$ such that $\hat{p}_{\langle\mathcal{D}\rangle}$ is $2\,\epsilon'$-close to $(1-\alpha)q_{\langle\mathcal{D}\rangle} + \alpha\, f_{\langle\mathcal{D}\rangle}$ if $f$ exists
        then
            return accept
        end
    end
end
return reject
```

prove this fact, we also use the following observation. Suppose we have two distributions p and p'. Let $\mathcal{P}$ be a partition of their domain. We prove that if p and p' are ϵ-far from each other, there is a noticeable discrepancy between either their coarsening distributions over $\mathcal{P}$ or their restrictions to the subsets in $\mathcal{P}$ (Lemma 16). This observation implies that if p is ϵ-far from being a mixture distribution, then at least one the steps will fail. Hence, we distinguish both cases with high probability.

We describe our tester in Algorithm 4 and show its correctness in Theorem 13. Later, we also discuss how to avoid trying all $\mathcal{I}$'s and achieve a polynomial time algorithm.

Theorem 13 *Algorithm 4 is an identity tester in the presence of k-flat noise that uses $\widetilde{O}(\sqrt{nk}/\epsilon^{3.5})$ samples.*

Proof We set $\epsilon' = \epsilon/14$. We denote the number of buckets in $\mathcal{B} = \text{BUCKETS}(q, n, \epsilon')$ by v. Let t denote $k \cdot v$. Without loss of generality we assume $t \leq n$. Otherwise, one could learn the distribution p up to $\epsilon/2$ ℓ_1-distance error via $O(n/\epsilon^2)$ samples, and trivially check if it is $\epsilon/2$-close to a mixture of q and a k-flat distribution.

Consider a segmentation $\mathcal{I}$, and a division $\mathcal{D} := \mathcal{D}(\mathcal{I}, \mathcal{B})$. To obtain better sample complexity, we need to make sure that the size of each set in $\mathcal{D}$ is not greater than $\lceil n/t \rceil$. In the case that a large set of size $z > \lceil n/t \rceil$ exists, we split it into $D_{i,j,\ell} := \lfloor z \cdot t/n \rfloor + 1$ sets of roughly the same size and denote them by $D_{i,j,\ell}$ for $\ell \in [D_{i,j,\ell}]$. The new sets form a new partition of the domain. We call it a *refined division*, denoted $\widetilde{\mathcal{D}} := \widetilde{\mathcal{D}}(\mathcal{I}, \mathcal{B}, n, t)$. Note that this replacement will not asymptotically increase the total number of sets in the division, since $\widetilde{\mathcal{D}}$ has $\sum_{i,j} D_{i,j,\ell} \leq 2\,t$ many sets.

Now, we establish that for a sufficiently large number of samples, the three steps in the algorithm succeed with high probability. First, in the following lemma, we show that $O(t \cdot \log n/\epsilon^2)$ samples are enough to obtain an empirical distribution $\hat{p}$ such that for all the divisions $\mathcal{D}$ $\hat{p}_{\langle \mathcal{D} \rangle}$ and $p_{\langle \mathcal{D} \rangle}$ are ϵ'-close to each other with probability 0.9.

Lemma 14 *Assume p is a distribution over $[n]$. Let $\hat{p}$ be an empirical distribution formed by $\Theta(\min(n, kv \log n) \cdot (\log \delta^{-1})/\epsilon'^2)$ samples from p. Fix a bucketing of the domain $\mathcal{B}$ =*BUCKET*(q, n, ϵ'). For every k-segmentation $\mathcal{I}$, and the corresponding refined division of the domain $\mathcal{D} = \mathcal{D}(\mathcal{I}, \mathcal{B})$, the coarsening of p and the empirical distribution $\hat{p}$ over $\mathcal{D}$ is at most ϵ'-far from each other with probability at least $1 - \delta$.*

The proof is presented in Section A.3.

Second, we show that if $p(D_{i,j,\ell})$, for a fixed i, j, and ℓ, is at least $\epsilon'/|\widetilde{\mathcal{D}}| = \Theta(\epsilon'/t)$, then $M_{i,j}$ contains at least $\epsilon'/(4t)$ fraction of the samples with high probability. Note that there are at most $\Theta(n^2 \cdot v)$ set $D_{i,j,\ell}$ for a fixed $\mathcal{B}$. Using the Chernoff bound, the claim is true for all $D_{i,j,\ell}$'s with probability 0.9 if we draw more than $\Theta(\log(n^2 \cdot v)t/\epsilon')$ samples.

Third, we show if $M_{i,j}$ contains enough samples, then with high probability we can distinguish whether $\|p_{|D_{i,j,\ell}} - \mathcal{U}_{|D_{i,j,\ell}}\|_2^2$ is at most $\epsilon'^2/|D_{i,j,\ell}|$, or it is at least $2\epsilon'^2/|D_{i,j,\ell}|$: If we draw $\Theta(t/\epsilon' \cdot (\log(n^2 \cdot v) \cdot \sqrt{n/t}/\epsilon^2))$ samples, we receive $O((\log(n^2 \cdot v) \cdot \sqrt{n/t}/\epsilon^2) = O((\log(n^2 \cdot v) \cdot \sqrt{|D_{i,j,\ell}|}/\epsilon^2)$ samples from any set $D_{i,j,\ell}$ with $p(D_{i,j,\ell}) \geq \epsilon'/t$. Based on (Diakonikolas et al., 2016, Theorem 1), with probability $1 - 1/3$, we can distinguish whether $\|p_{|D_{i,j,\ell}} - U_{|D_{i,j,\ell}|}\|_2^2$ is at most $2\epsilon'^2/|D_{i,j,\ell}|$ or at least $\epsilon^2/|D_{i,j,\ell}|$ using $\Theta(\sqrt{|D_{i,j,\ell}|}/\epsilon^2)$ samples. By repeating this $\Theta(\log(n^2 \cdot v))$ times and taking the majority answer, we can be assured to obtain the correct answer for the test on all the $D_{i,j,\ell}$'s with probability at least 0.9. Thus, we need $O(\sqrt{n \cdot t} \cdot (\log n + \log v)/\epsilon^3)$ samples for this step.

In the above three steps, we need the following number of samples:

$$\begin{aligned} &O(t \cdot \log n/\epsilon^2 + \sqrt{n \cdot t} \cdot (\log n + \log v)/\epsilon^3 + \log(n^2 \cdot v)t/\epsilon') \\ &\qquad = O(\sqrt{n \cdot k}/\epsilon^{3.5}) \cdot \text{Polylog}(n, \epsilon^{-1}) = \widetilde{O}(\sqrt{n \cdot k}/\epsilon^{3.5})\,. \end{aligned}$$

By a union bound, the probability than any of the above steps goes wrong is at most 0.3. Hence, for the rest of the proof, we assume that the algorithm carries out the steps as expected with probability at least 2/3. Given this assumption, we show in both the completeness case and the soundness case, the algorithm outputs the correct answer.

Completeness: In this case, there exist a k-flat distribution over $\mathcal{I}$, r, and a parameter α^* such that $p = (1 - \alpha^*)q + \alpha^* r$. First, note that p in each $D_{i,j,\ell} \in \widetilde{\mathcal{D}}$ is close to the uniform distribution. In particular, we have the following lemma.

Lemma 15 *Suppose p is a mixture of q and r with parameter α. Let $\mathcal{I}^*$, $\mathcal{B}$, and $\widetilde{\mathcal{D}}$, be the partitions we defined earlier. For any non-empty set, $D_{i,j,\ell} \in \mathcal{D}$, if $j > 1$, then the restriction of p to the set, $p_{|D_{i,j,\ell}}$, is ϵ-close to the uniform distribution in ℓ_1-distance and $\epsilon/\sqrt{|D_{i,j,\ell}|}$-close to the uniform distribution in ℓ_2-distance.*

For the proof of the lemma, see Section A.3.

The lemma implies that for all the $D_{i,j,\ell}$, $p_{|D_{i,j,\ell}}$ is close to the uniform distribution. Hence, the algorithm while considering the segmentation $\mathcal{I}^*$, will not continue with another segmentation since $p_{|D_{i,j,\ell}}$ is being far from uniform, and the algorithm will move on to the next step.

Also, we show that a k-flat function, f, exists, because r is a solution itself. We have $\Omega(t/\epsilon'^2)$ samples which is enough to learn the coarsening of p over $\widetilde{\mathcal{D}}$. Thus, the coarsening of the empirical distribution, $\hat{p}$, is ϵ'-close to the coarsening of p over $\widetilde{\mathcal{D}}$. There exists an iteration in the algorithm in which we try a parameter α such that $\alpha - \alpha^*$ is at most $\epsilon'/2$. Therefore, r itself is a solution the algorithm is looking for:

$$\begin{aligned}\|\hat{p}_{\langle\widetilde{\mathcal{D}}\rangle} - ((1-\alpha)q_{\langle\widetilde{\mathcal{D}}\rangle} + \alpha\, r_{\langle\widetilde{\mathcal{D}}\rangle})\| &\leq \|\hat{p}_{\langle\widetilde{\mathcal{D}}\rangle} - p_{\langle\widetilde{\mathcal{D}}\rangle}\|_1 \\ &+ \|p_{\langle\widetilde{\mathcal{D}}\rangle} - ((1-\alpha^*)q_{\langle\widetilde{\mathcal{D}}\rangle} + \alpha^*\, r_{\langle\widetilde{\mathcal{D}}\rangle})\| \\ &+ \|((1-\alpha^*)q_{\langle\widetilde{\mathcal{D}}\rangle} + \alpha^*\, r_{\langle\widetilde{\mathcal{D}}\rangle}) - ((1-\alpha)q_{\langle\widetilde{\mathcal{D}}\rangle} + \alpha\, r_{\langle\widetilde{\mathcal{D}}\rangle})\| \\ &\leq \epsilon' + 0 + \frac{\epsilon'}{2} \cdot \|q_{\langle\widetilde{\mathcal{D}}\rangle} - r_{\langle\widetilde{\mathcal{D}}\rangle}\|_1 \leq 2\epsilon'.\end{aligned}$$

Hence, the algorithm will not output reject.

Soundness: In this case, p is ϵ-far from any mixture distribution $q_\alpha = ((1-\alpha)q_{\langle\widetilde{\mathcal{D}}\rangle} + \alpha\, r_{\langle\widetilde{\mathcal{D}}\rangle})$ for any k-flat distribution r and $\alpha \in [0,1]$. We have the following structural lemma (similar to Lemma 6 in Batu et al. (2001)) which bounds the distance between p and q_α from above:

Lemma 16 *Assume p and q are two distributions on $[n]$, and let $\widetilde{\mathcal{D}}$ be a refined division of the domain elements. Then, we have*

$$\|p - q\|_1 \leq \left\|p_{\langle\widetilde{\mathcal{D}}\rangle} - q_{\langle\widetilde{\mathcal{D}}\rangle}\right\|_1 + \sum_{D\in\widetilde{\mathcal{D}}} \left\|p_{|D} - q_{|D}\right\|_1 \cdot \min(p(D), q(D)).$$

The proof is in Section A.3.

Since the distance between p and q_α is at least ϵ, we can apply this lemma to obtain a lower bound for the two quantities in the right hand side of the equation above.

$$\begin{aligned}\epsilon = 14\epsilon' &< \|p - q_\alpha\|_1 \\ &\leq \left\|p_{\langle\widetilde{\mathcal{D}}\rangle} - (q_\alpha)_{\langle\widetilde{\mathcal{D}}\rangle}\right\|_1 + \sum_{D_{i,j,\ell}} \left\|p_{|D_{i,j,\ell}} - (q_\alpha)_{|D_{i,j,\ell}}\right\|_1 \cdot \min\left(p(D_{i,j,\ell}), q_\alpha(D_{i,j,\ell})\right). \end{aligned} \tag{7}$$

At least one of the two terms on the right hand side above is greater than $7\epsilon'$. Net, we show if the algorithm reaches to the point that forms the empirical distribution, then the second term is at most $7\epsilon'$. On the other hand, if the algorithm outputs accept, then the first term is at most $5\epsilon'$. Hence, these two events cannot happen at the same time while $|p - q| \geq \epsilon$.

Formally, if there is no $D_{i,j,\ell}$ such that causes the algorithm to move forward to the next segmentation, then for each $D_{i,j,\ell}$ either the weight of the set is not larger than $\epsilon'/|\widetilde{\mathcal{D}}|$, or the ℓ_2-distance between $P_{|D_{i,j,\ell}}$ and the uniform distribution is not more than $\sqrt{2\epsilon'^2/|D_{i,j,\ell}|}$. In the following lemma, we show that this situation implies that the second term in Equation 7 is at most $6.42\,\epsilon'$.

Lemma 17 *Suppose for every non-empty $D_{i,j,\ell}$ in the division $\widetilde{\mathcal{D}} = \widetilde{\mathcal{D}}(\mathcal{I}, \mathcal{B})$, either $p(D_{i,j,\ell})$ is at most $\epsilon'/|\widetilde{\mathcal{D}}|$, or $\|p_{|D_{i,j,\ell}} - \mathcal{U}_{|D_{i,j,\ell}}\|_2^2$ is at most $2\epsilon'^2/|D_{i,j,\ell}|$. Let q_α be a mixture of q and a k-flat distribution over $\mathcal{I}$ with an arbitrary α in $[0,1]$. Then, the following holds*

$$\sum_{i=1}^{k}\sum_{j=1}^{v} \|p_{|D_{i,j,\ell}} - (q_\alpha)_{|D_{i,j,\ell}}\|_1 \cdot \min\left(p(D_{i,j,\ell}), q_\alpha(D_{i,j,\ell})\right) \leq 6.42\,\epsilon' \,.$$

See Section A.3 for the proof.

On the other hand, if the algorithm outputs accept, it implies that there exists a function f and α, such that $\|\hat{p} - ((1-\alpha)\, q + \alpha\, f)\|_1$ is at most $2\epsilon'$. In the following lemma, we show it implies that there exists a q_α such that $\|p_{\langle \mathcal{D}\rangle} - (q_\alpha)_{\langle \mathcal{D}\rangle}\|_1$ is at most 5ϵ.

Lemma 18 *Assume p, $\hat{p}$, and q are three distributions over $[n]$, and $f : [n] \to R^+$ is a k-flat function over k-segmentation $\mathcal{I}$. For a division $\mathcal{D}$, suppose $\hat{p}_{\langle \mathcal{D}\rangle}$ is ϵ'-close to $p_{\langle \mathcal{D}\rangle}$, and there exists $\alpha \in [0,1]$ such that $\|\hat{p} - ((1-\alpha)\, q + \alpha\, f)\|_1$ is at most $2\epsilon'$. Then, there exists a k-flat distribution r, such that p is $5\epsilon'$-close to the mixture of r and q with parameter α.*

The proof of the lemma is in Section A.3.

Moreover, outputting accept means that the two terms in Equation 7 are at most $6.42\epsilon' + 5\epsilon' < 14\epsilon'$, which contradicts the fact that one of them has to be $7\epsilon'$. Hence, the proof is complete. ■

A faster algorithm. In the interest of a simpler exposition, the algorithm described above tries all possible k-segmentations. However, there are at most $O(n^2 \cdot v)$ possible subsets that could appear as $D_{i,j,\ell}$'s. Hence, one can test uniformity of p on each of them separately regardless of $\mathcal{I}$. Moreover, finding a k-flat function f for which the ℓ_1-distance between $\hat{p}$ and the mixture of q and f is minimized, can be done via dynamic programming: we define $d[i,j]$ to be the smallest ℓ_1-distance between $\hat{p}$ and mixture of q and any j-flat distribution when we consider only the first i elements of the domain. We compute $d[i,j]$ using the previously computed $d[i', j-1]$:

$$d[i,j] = \min_{1 \leq i' < i} d[i', j-1] + \text{cost}([i', i]),$$

where the cost$([i', i])$ is defined as follows: We set the cost of an interval to infinity if any subset of $[i', i]$ which would have appeared in the divisions (i.e, all subsets in such form $[i', i] \cap B_z$) for $z = 1, \ldots, v$) does not pass the uniformity test. Otherwise, cost$([i, i'])$ is the minimum ℓ_1-distance between $\hat{p}$ and a mixture of q and a constant function for the elements in $[i', i]$. Since we are only looking for k-flat functions rather than distributions, the updates can be computed locally and independently of the rest of segments.

A.3. Proofs

Proof of Lemma 14: Suppose we draw s samples from p. Let n_x indicate the number of occurrences of x among the samples. Let $\hat{p}$ be the empirical distribution formed by s samples which means that $\hat{p}(x) := n_x/s$. The goal is to show that for every segmentation $\mathcal{I}$, the coarsening of $\hat{p}$ and p over $\widetilde{\mathcal{D}}(\mathcal{I}, \mathcal{B}, n, t)$ are ϵ'-close with probability at least $1-\delta$. We build on the standard idea that is used to show that $O(n/\epsilon^2)$ samples is sufficient to learn a distribution over $[n]$ within ϵ error in ℓ_1-distance. Consider $\widetilde{\mathcal{D}}$ which contains $\Theta(t)$ disjoint subsets of $[n]$. The ℓ_1-distance between the coarsening of p and the empirical distribution is defined as follows:

$$\begin{aligned}\|p_{\langle\mathcal{D}\rangle} - \hat{p}_{\langle\mathcal{D}\rangle}\|_1 &= \sum_{i=1}^{k}\sum_{j=1}^{v}\sum_{\ell=1}^{D_{i,j,\ell}}\left|\sum_{x\in D_{i,j,\ell}} p(x) - \sum_{x\in D_{i,j,\ell}} \hat{p}(x)\right| \\ &= \sum_{x\in[n]} \operatorname{sign}\left(\sum_{x\in D_{i,j,\ell}} p(x) - \sum_{x\in D_{i,j,\ell}} \hat{p}(x)\right)\cdot(p(x)-\hat{p}(x)).\end{aligned}$$

We need to show that the above quantity is at most ϵ' for any segmentation $\mathcal{I}$ and its corresponding division $\mathcal{D}(\mathcal{I},\mathcal{B})$. However, we prove a stronger claim: Suppose we have a collection C of vectors of length n with entries in $\{+1,-1\}$ for which the following is true:

- For every refined division $\widetilde{\mathcal{D}}$, an every set $D_{i,j,\ell}\in\widetilde{\mathcal{D}}$, there exists a vector $c\in C$ such that if x is in $D_{i,j,\ell}$, then $c_x = \operatorname{sign}\left(\sum_{x\in D_{i,j,\ell}} p(x) - \sum_{x\in D_{i,j,\ell}} \hat{p}(x)\right)$.
- For all $c\in C$, $\sum_{x\in[n]} c_x\cdot(p(x)-\hat{p}(x))$ is at most ϵ' with probability at least $1-\delta$.

The proof is complete if we establish this claim, so now we focus on proving that the collection C exists. We first put a vector c corresponding to each refined division. Then we show there is an upper bound for the size of the collection. Next, we show since there are not too many vectors in the collection, with high probability, $\sum_{x\in[n]} c_x\cdot(p(x)-\hat{p}(x))$ is at most ϵ' for any $c\in C$.

Clearly, there are no more than 2^n possible vectors. However, we get a better bound for the cases when k is not arbitrarily large. We begin by considering a refined division $\widetilde{\mathcal{D}}$. Fix a set $B\in\mathcal{B}$. If two elements in B are in the same interval I_i for $i\in[k]$, then they will have the same c_x as well. Thus, if we sort elements in B, and then write the corresponding c_x's, then we get a sequence of $+1$ and -1 where the sign is changed in at most $\Theta(t)$ places. To uniquely represent the sequence, one can determine the indices where the sign changed and indicate whether the sequence starts with $+1$ or -1. Thus, the total number of such sequences is:

$$\sum_{i=0}^{\Theta(t)} 2\cdot\binom{|B|-1}{i} \le 2\cdot\left((|B|)^{\Theta(t)}+1\right) \le n^{\Theta(t)}\,.$$

Note that we have at most v such subsets of the domain $B\in\mathcal{B}$. Thus, the total number of vectors in C is at most $\min(2^n, n^{\Theta(t\cdot v)})$.

Next, we show that if we draw enough samples, the probability of $\sum_{x\in[n]} c_x\cdot(p(x)-\hat{p}(x)) \ge \epsilon'$ for any $c\in C$ is at most δ. Fix a vector $c=(c_1,\ldots,c_n)$ in $\{+1,-1\}^n$. Consider the following random process: we draw a sample from p, namely x; if c_x is one, output one and otherwise output zero. In other words upon receiving sample x, we output $(1+c_x)/2$. Assume $x_1,\ldots,x_s$ are s samples from p that form the empirical distribution. Suppose that we generate $b_1,\ldots,b_s$ according to the process using these samples from p, i.e., $b_j=(1+c_{x_j})/2$. The b_j's are s independent random variables with the following expected value.

$$\mathbf{E}[b_j] = \sum_{x=1}^{n} p(x)\cdot\frac{1+c_x}{2} = \frac{1+\sum_x c_x\cdot p(x)}{2}.$$

Clearly, the average of b_j's are close to its expectation with high probability, we use this fact to show that $\sum_x c_x \cdot (p(x) - \hat{p}(x))$ are close to zero as well. Recall n_x is the number of occurrences of element x in the sample set. Using the Hoeffding bounds, we achieve:

$$\begin{aligned}
\mathbf{Pr}\left[\left|\sum_x c_x \cdot (p(x) - \hat{p}(x))\right| \geq \epsilon'\right] &= \mathbf{Pr}\left[\left|\sum_x c_x \cdot p(x) - \sum_x c_x \cdot \frac{n_x}{s}\right| \geq \epsilon'\right] \\
&= \mathbf{Pr}\left[\left|\sum_x c_x \cdot p(x) - \frac{\sum_{j=1}^s c_{x_j}}{s}\right| \geq \epsilon'\right] \\
&= \mathbf{Pr}\left[\left|\frac{1 + \sum_x c_x \cdot p(x)}{2} - \frac{s + \sum_{j=1}^s c_{x_j}}{2\,s}\right| \geq \frac{\epsilon'}{2}\right] \\
&= \mathbf{Pr}\left[\left|\frac{1 + \sum_x c_x \cdot p(x)}{2} - \frac{\sum_{j=1}^s b_j}{s}\right| \geq \frac{\epsilon'}{2}\right] \\
&= \mathbf{Pr}\left[\left|\mathbf{E}[b_i] - \frac{\sum_{j=1}^s b_j}{s}\right| \geq \frac{\epsilon'}{2}\right] \leq 2\,\exp\left(-\,s\,{\epsilon'}^2/2\right).
\end{aligned} \tag{8}$$

Therefore, by setting $s = \Theta(\min(n, t \cdot v \cdot \log n) \cdot (\log \delta^{-1})/\epsilon'^2)$ and using Equation 8 and a union bound, for every $c \in C$, $\sum_{x\in[n]} c_x \cdot (p(x) - \hat{p}(x)) \geq \epsilon'$ is at most ϵ' with probability $1 - \delta$. This completes the proof. ■

Proof of Lemma 15: Fix a non-empty set $D_{i,j,\ell}$ in $\widetilde{\mathcal{D}}$ for some $j > 1$. To prove the lemma, we show that the ratio of the maximum and the minimum probability according to p in $D_{i,j,\ell}$ is at most $1 + \epsilon'$. Consider two elements in $D_{i,j,\ell}$ namely x and y (if there is only one element in $D_{i,j,\ell}$ the claim is apparent). Without loss of generality assume $q(x) \leq q(y)$. By definition of $D_{i,j,\ell}$, x and y are in the same interval of $\mathcal{I}$, so $r(x)$ and $r(y)$ are equal. Thus, we have:

$$1 \leq \frac{p(y)}{p(x)} = \frac{(1-\alpha)q(y) + \alpha\, r(y)}{(1-\alpha)q(x) + \alpha\, r(x)} \leq \frac{q(y)}{q(x)} \leq 1 + \epsilon\,, \tag{9}$$

the second to last inequality is true, because we have $q(y) \geq q(x) > 0$. Also, the last inequality is true since both x and y are in B_j. In the proof of Lemma 8 in Batu et al. (2001), it is show that if the ratio of the probabilities in a set, in our case $D_{i,j,\ell}$, is bounded by $(1+\epsilon)$, then for all $x \in D_{i,j,\ell}$, $|p(x) - (1/|D_{i,j,\ell}|)|$ is at most $\epsilon/|D_{i,j,\ell}|$. This completes the proof. ■

Proof of Lemma 16: Fix a set in $\widetilde{\mathcal{D}}$, namely D, which $p(D)$ and $q(D)$ are non-zero, we have the following:

$$\begin{aligned}
\left\|p_{|D} - q_{|D}\right\|_1 &= \sum_{x\in D} \left|p_{|D}(x) - q_{|D}(x)\right| = \sum_{x\in D}\left|\frac{p(x)}{p(D)} - \frac{q(x)}{q(D)}\right| \\
&= \sum_{x\in D}\left|\frac{p(x)}{p(D)} - \frac{q(x)}{p(D)} + \frac{q(x)}{p(D)} - \frac{q(x)}{q(D)}\right| \\
&= \sum_{x\in D}\left|\frac{p(x)-q(x)}{p(D)} - q(x)\cdot\frac{q(D)-p(D)}{p(D)\,q(D)}\right| \\
&\geq \frac{1}{p(D)}\sum_{x\in D}|p(x)-q(x)| - \frac{|p(D)-q(D)|}{p(D)}\cdot\sum_{x\in D}\frac{q(x)}{q(D)} \\
&= \frac{1}{p(D)}\left(\sum_{x\in D}|p(x)-q(x)| - |p(D)-q(D)|\right).
\end{aligned}$$

Therefore, we have:

$$\sum_{x\in D}|p(x)-q(x)| \leq |p(D)-q(D)| + \left\|p_{|D}-q_{|D}\right\|_1\cdot p(D)\,. \tag{10}$$

If we swap p and q in the above inequality, and replicate the equations, we have:

$$\sum_{x\in D}|p(x)-q(x)| \leq |p(D)-q(D)| + \left\|p_{|D}-q_{|D}\right\|_1\cdot q(D)\,. \tag{11}$$

Putting Equation 10 and Equation 11 together, we get:

$$\sum_{x\in D}|p(x)-q(x)| \leq |p(D)-q(D)| + \left\|p_{|D}-q_{|D}\right\|_1\cdot \min(p(D),q(D))\,.$$

If at least one of $p(D)$ and $q(D)$ is zero, it implies:

$$\sum_{x\in D}|p(x)-q(x)| = |p(D)-q(D)|\,.$$

Hence, we have:

$$\begin{aligned}
\|p-q\|_1 &\leq \sum_{D\in\widetilde{\mathcal{D}}}|p(D)-q(D)| + \sum_{D\in\widetilde{\mathcal{D}}}\left\|p_{|D}-q_{|D}\right\|_1\cdot\min(p(D),q(D)) \\
&\leq \left\|p_{\langle\widetilde{\mathcal{D}}\rangle} - q_{\langle\widetilde{\mathcal{D}}\rangle}\right\|_1 + \sum_{D\in\widetilde{\mathcal{D}}}\left\|p_{|D}-q_{|D}\right\|_1\cdot\min(p(D),q(D))\,.
\end{aligned}$$

■

Proof of Lemma 17: We first consider a non-empty $D_{i,j,\ell}$ when $j = 1$. Since $j = 1$, $D_{i,j,\ell}$ is a subset of B_1. For each $x \in D_{i,j,\ell}$, $q(x)$ is at most ϵ'^2/n. Also, r is a k-flat on $\mathcal{I}$, and since $D_{i,j,\ell}$ is a subset of I_i, for all $x \in D_{i,j,\ell}$, $r(x)$ is the same. We denote this quantity, $r(x)$, by b. Here, we prove that either $q_\alpha(D_{i,j,\ell})$ is small, or $q_\alpha(D_{i,j,\ell})$ has to be close to uniform.

We have two cases. First, suppose $\alpha \cdot b$ is at most ϵ'/n. In this case, $q_\alpha(D_{i,j,\ell})$ is at most $\epsilon'^2|D_{i,j,\ell}|/n$. Thus, the total weight of such sets, sum of $q_\alpha(D_{i,j,\ell})$'s, is at most ϵ'^2. Second, assume $\alpha \cdot b$ is greater that ϵ'/n. On the other hand, $q(x)$ is at most ϵ'^2/n. These two facts implies for each x in $D_{i,j,\ell}$:

$$\begin{aligned}
|D_{i,j,\ell}| \cdot q_\alpha(D_{i,j,\ell}) \left|(q_\alpha)_{|D_{i,j,\ell}}(x) - \frac{1}{|D_{i,j,\ell}|}\right| &= |D_{i,j,\ell}| \cdot q_\alpha(D_{i,j,\ell}) \cdot \left|\frac{q_\alpha(x)}{q_\alpha(D_{i,j,\ell})} - \frac{1}{|D_{i,j,\ell}|}\right| \\
&= \left| |D_{i,j,\ell}| \cdot (1-\alpha) q(x) - (1-\alpha) \sum_{x \in D_{i,j,\ell}} q(x) \right| \\
&\leq |D_{i,j,\ell}| \frac{\epsilon'^2}{n} \leq \epsilon' \, |D_{i,j,\ell}| \, \alpha \, b \leq \epsilon' \cdot q_\alpha(D_{i,j,\ell}) \,.
\end{aligned}$$

Therefore, the ℓ_2^2-distance between $(q_\alpha)_{|D_{i,j,\ell}}$ and the uniform distribution is bounded:

$$\|(q_\alpha)_{|D_{i,j,\ell}}(x) - \mathcal{U}_{|D_{i,j,\ell}}\|_2^2 = \sum_{x \in D_{i,j,\ell}} \left((q_\alpha)_{|D_{i,j,\ell}}(x) - \frac{1}{|D_{i,j,\ell}|} \right)^2 \leq \frac{\epsilon'^2}{|D_{i,j,\ell}|} \,.$$

Note that if j is greater than one, the ℓ_2-distance between $(q_\alpha)_{|D_{i,j,\ell}}$ and the uniform distribution is bounded by $\epsilon'/\sqrt{|D_{i,j,\ell}|}$ as well. Therefore, if $p_{|D_{i,j,\ell}}$ is close uniform distribution, it has to be close to $(q_\alpha)_{|D_{i,j,\ell}}$ as well. That is,

$$\begin{aligned}
\|p_{|D_{i,j,\ell}} - (q_\alpha)_{|D_{i,j,\ell}}\|_1 &\leq \|p_{|D_{i,j,\ell}} - \mathcal{U}_{|D_{i,j,\ell}}\|_1 + \|\mathcal{U}_{|D_{i,j,\ell}} - (q_\alpha)_{|D_{i,j,\ell}}\|_1 \\
&\leq \sqrt{|D_{i,j,\ell}|} \left(\|p_{|D_{i,j,\ell}} - \mathcal{U}_{|D_{i,j,\ell}}\|_2 + \|\mathcal{U}_{|D_{i,j,\ell}} - (q_\alpha)_{|D_{i,j,\ell}}\|_2 \right) \leq 2.42\epsilon' \,.
\end{aligned}$$

Hence, given the discussion above there are three possibilities for $D_{i,j,\ell}$. (i) $p(D_{i,j,\ell})$ is at most $\epsilon'/(k.v)$. Since ℓ_1-distance is at most 2, the total contribution of these sets in the sum below is at most $2\epsilon'$. (ii) $q_\alpha(D_{i,j,\ell})$ is at most $\epsilon'^2|D_{i,j,\ell}|/n$, so the total contribution of these sets is at most $2\epsilon'^2$. (iii) $\|p_{|D_{i,j,\ell}} - (q_\alpha)_{|D_{i,j,\ell}}\|_1$ is at most $2.42\,\epsilon'$.

$$\|p_{|D_{i,j,\ell}} - (q_\alpha)_{|D_{i,j,\ell}}\|_1 \cdot \min\left(p(D_{i,j,\ell}), q_\alpha(D_{i,j,\ell})\right) \leq 6.42\,\epsilon' \,.$$

Hence, the proof is complete. ■

Proof of Lemma 18: First, consider a degenerate case. If $\alpha = 0$, then the claim is trivially true by the triangle inequality: $\|p - q\|_1 \leq \|p - \hat{p}\|_1 + \|\hat{p} - q\|_1 \leq 3\epsilon'$. Thus, assume $\alpha > 0$.

For now, consider the case that there exists x such that $f(x)$ is not zero, so $\sum_x f(x)$ is greater than zero. First, we show that since $\hat{p}$ is close to the mixture of q and f, the sum of the $f(x)$'s has to be close to one. That is,

$$\begin{gathered}
-2\,\epsilon' \leq \sum_x \hat{p}(x) - ((1-\alpha)\,q(x) - \alpha\, f(x)) \leq 2\,\epsilon' \quad \Rightarrow \\
-2\,\epsilon' \leq 1 - (1-\alpha) - \alpha \cdot \sum_x f(x)) \leq 2\,\epsilon' \quad \Rightarrow \\
-\frac{2\,\epsilon'}{\alpha} \leq 1 - \sum_x f(x) \leq \frac{2\,\epsilon'}{\alpha} \quad \Rightarrow \quad \left|1 - \sum_x f(x)\right| \leq \frac{2\,\epsilon'}{\alpha} \,.
\end{gathered} \tag{12}$$

We define $r : [n] \to [0,1]$ to be the normalization of f for which $r(x) = f(x)/\sum_x f(x)$ for all x in the domain. If f is a k-flat function, then r will be a k-flat distribution. Now, we show that the mixture of q and f is close to the mixture of q and r with mixture parameter α.

$$\begin{aligned}\|(1-\alpha)\,q+\alpha\,f-(1-\alpha)\,q+\alpha\,r\|_1 &= \sum_x |(1-\alpha)\,q(x)+\alpha\,f(x)-(1-\alpha)\,q(x)+\alpha\,r(x)| \\ &= \alpha\cdot\sum_x |f(x)-r(x)| = \alpha\cdot\sum_x f(x)\cdot\left|1-\frac{1}{\sum_y f(y)}\right| \\ &= \alpha\cdot\left|\sum_x f(x)-1\right| \le 2\,\epsilon'\,,\end{aligned}$$

where the last inequality is due to Equation 12. Moreover, by the triangle inequality, we have:

$$\|\hat{p}-((1-\alpha)\,q+\alpha\,r)\|_1 \le \|\hat{p}-((1-\alpha)\,q+\alpha\,f)\|_1+\|(1-\alpha)\,q+\alpha\,f-(1-\alpha)\,q+\alpha\,r\|_1 \le 4\,\epsilon'\,.$$

Now, assume $f(x)$ is zero for all x in $[n]$. We show that if we set r to be the uniform distribution over $[n]$, the same result holds. First, observe that the uniform distribution is a k-flat distribution for any $k \ge 1$. Then we show that $(1-\alpha)\,q+\alpha r$ is $4\epsilon'$-close to p'. Since $r(x) = 1/n$ for all x in $[n]$,

$$\|\hat{p}-((1-\alpha)\,q+\alpha\,r)\|_1 \le \|\hat{p}-((1-\alpha)\,q)\|_1+\alpha\,.$$

On the other hand, since $\hat{p}$ is $2\epsilon'$-close to $(1-\alpha)\,q$, one can show that α is at most ϵ'.

$$2\,\epsilon' \ge \|\hat{p}-((1-\alpha)\,q+\alpha\,f)\|_1 \ge \sum_x \hat{p}(x)-(1-\alpha)\,q(x) = 1-1+\alpha = \alpha\,.$$

Therefore, whether $\sum_x f(x)$ is zero or not, there exists a k-flat distribution, r for which $\|\hat{p}-((1-\alpha)\,q+\alpha\,r)\|_1$ is at most $4\,\epsilon'$. Since p' is ϵ'-close to p, and by triangle inequality, we have:

$$\|p-((1-\alpha)\,q+\alpha\,r)\|_1 \le 5\epsilon'\,,$$

which concludes the proof. ■

Appendix B. Lower bounds

In this section, we present lower bounds for testing mixtures in different settings discussed earlier.

Theorem 19 *Assume p is a distribution on $[n]$. There exists a constant parameter ϵ such that distinguishing the following cases with probability at least $2/3$ requires $\Omega(n/\log n)$ samples.*

- *There exists a noise distribution on $[n]$, namely η, and an $\alpha \le \epsilon_1$ such that p is a mixture of uniform and η with parameter α, i.e., $p = (1-\alpha)\,\mathcal{U}+\alpha\,\eta$.*
- *There is no noise distribution η such that $p = (1-\alpha)\,\mathcal{U}+\alpha\,\eta$ unless $\alpha = 1$.*

Proof We prove by showing a reduction from mixture testing to testing bigness property of distributions. A distribution called T-*big* if the probability of any domain element is at least T (Aliakbarpour et al., 2019 (this proceedings)). In addition, they showed there exist two constant parameters ϵ and β and two family of distributions, namely $\mathcal{F}^+$ and $\mathcal{F}^-$, such that the following is true

- All distribution in $\mathcal{F}^+$ are $1/(\beta n)$-big.
- All distribution in $\mathcal{F}^-$ are ϵ-far from being $1/(\beta n)$-big. Moreover, all the probability of each element according to the distributions is either zero or at least $1/(\beta n)$.
- Using $o(n/\log n)$ samples from a distribution in the families, no algorithm can distinguish whether the distribution was from $\mathcal{F}^+$ or $\mathcal{F}^-$ with probability at least $2/3$.

Let $\epsilon = 1/\beta$. We show that any algorithm that can test mixtures as described in theorem, can distinguish $\mathcal{F}^+$ and $\mathcal{F}^-$ with high probability.

First, we show that for any $1/(\beta n)$-big distribution, denoted by p^+, there exists distribution η such that p^+ is a mixture of η and uniform distribution, meaning $p^+ = \alpha\,\eta + (1-\alpha)\,\mathcal{U}$ for $\alpha = 1/\beta$. Let η assign the following probability to the ith element of the domain:

$$\eta(i) = \frac{p(i) - \frac{1}{\beta n}}{1 - \frac{1}{\beta}}\,.$$

It is not hard to see that η as defined above is a probability distribution. Since p is $1/(\beta n)$-big distribution, all the $p(i)$ are at least $1/(\beta n)$, so all the $\eta(i)$'s are non-negative. Also, $\sum_i \eta(i) = 1$. Clearly, p^+ is a mixture in the form $\alpha\,\eta + (1-\alpha)\,\mathcal{U}$ for $\alpha = 1/\beta = \epsilon$.

Note that for any distribution p^- in $\mathcal{F}^-$ there is at least one element (in fact many elements) that has probability zero. Otherwise, all elements would have probability at least $1/(\beta n)$, and the distribution would be big. On the other hand, any distribution that is mixed with uniform with parameter $\alpha < 1$ cannot have any zero probability element. Thus, p^- is not a mixture of the form $\alpha\,\eta + (1-\alpha)\,\mathcal{U}$ when $\alpha \neq 1$.

Thus, any algorithm that can test mixture property as defined in the theorem has to accept p^+ and reject p^-. However, we know this is not possible unless the algorithms gets $\Omega(n/\log n)$ samples. This completes the proof. ■

Proposition 20 *When we have sample access to q and p, any closeness tester in the presence of uniform noise $\Omega\left(\max\left(n^{2/3}/\epsilon^{4/3}, \sqrt{n}/\epsilon^2\right)\right)$ samples.*

Proof First, note that one can reduce testing uniformity to this problem by setting q equal to the uniform distribution. Therefore, it requires at least $\Omega(\sqrt{n}/\epsilon^2)$ samples by the lower bound for uniformity testing shown in Paninski (2008).

Now, we establish that $\Omega(n^{2/3}/\epsilon^{4/3})$ many samples is also required. Without loss of generality, assume $\epsilon \geq 4^{3/4}/n^{1/4}$. Otherwise $\sqrt{n}/\epsilon^2$ would be the dominating term in the lower bound up to a constant factor. To prove the lower bound, we use two distributions (and any random relabeling of them) used in proving lower bounds for testing closeness of distributions Batu et al. (2013); Valiant and Valiant (2017a,b); Chan et al. (2014). More precisely, we define two distributions p^* and q^* such that distinguishing (p^*, q^*) and (q^*, q^*) (and any random relabeling of them) requires $\Omega(n^{2/3}/\epsilon^{4/3})$ samples. On the other hand, we show that any $\Omega(q^*, \mathcal{U}, \epsilon)$-mixture tester has to distinguish (p^*, q^*) and (q^*, q^*). Thus, the statement of the proposition is concluded.

Let $a = 4\epsilon/n$ and $b = \epsilon^{4/3}/n^{2/3}$. Consider three disjoint subset of domain elements $[n]$, namely A, B, and C each of size $(1-\epsilon)/b$, ϵ/a, and ϵ/a respectively. Let p and q be the following distributions:

$$p^* = b\mathbb{1}_A + a\mathbb{1}_B, \qquad q^* = b\mathbb{1}_A + a\mathbb{1}_C.$$

Note that p^* is ϵ-far from any mixture distribution of q^* and $\mathcal{U}$ with parameter $\alpha \in [0,1]$, since

$$\begin{aligned}
\|p^* - q^*_\alpha\|_1 &= \sum_i |p^*(i) - (1-\alpha)q^*(i) - \alpha/n| \\
&\geq \sum_{i \in B} |a - \alpha/n| + \sum_{i \in C} |-(1-\alpha)a - \alpha/n| \\
&\geq \frac{n}{4}\left(|a - \alpha/n| + |(1-\alpha)a + \alpha/n|\right) \\
&\geq \frac{n}{4} \cdot a \geq \epsilon.
\end{aligned}$$

Clearly, in the case where $p = q^*$ and $q = q^*$, p is a mixture of q and $\mathcal{U}$ with mixture parameter $\alpha = 0$, and in the case where $p = p^*$ and $q = q^*$, p is ϵ-far from any mixture distribution of q^* and $\mathcal{U}$. Thus, a $(q, \mathcal{U}, \epsilon)$-mixture tester has to distinguish between (q^*, q^*) and (p^*, q^*). By proposition 4.1 in Chan et al. (2014), we know that this task requires $\Omega(n^{2/3}/\epsilon^{4/3})$ samples. ■

Appendix C. Proofs for Section 4

Proof of Lemma 3: First, observe that if q_1 and q_2 are ϵ-close, then p is ϵ-close to q_1 as well, so distribution q_1 which is a mixture with parameter $\alpha = 0$ is a valid output. For the remainder of the proof, we assume q_1 and q_2 are ϵ-far from each other.

Let $S = \{i \in [n] \mid q_1(i) > q_2(i)\}$. We have $\alpha^* = \frac{q_1(S) - p(S)}{q_1(S) - q_2(S)}$. The only unknown value in the above expression is $p(S)$, which we estimate using $O(1/\epsilon^2)$ samples from p. We show by replacing $p(S)$, we get a viable estimate for α^*.

Let w_S be the estimate that is the ratio of the samples that are in S. By the Hoeffding bound, we have $\mathbf{Pr}[|p(S) - w_s| \leq \epsilon/4] \geq 5/6$. We define our estimate of α^* as: $\alpha := \frac{q_1(S) - (w_S + \epsilon/4)}{q_1(S) - q_2(S)}$. The reason that we add $\epsilon/4$ to w_S is to assure an overestimation of $p(S)$, so α becomes smaller than α^* with high probability. I.e., since with high probability $p(S) \leq w_S + \epsilon/4$, we get:

$$\alpha^* = \frac{q_1(S) - p(S)}{q_1(S) - q_2(S)} \geq \frac{q_1(S) - (w_S + \epsilon/4)}{q_1(S) - q_2(S)} = \alpha.$$

Below, we show q_α is close to p in ℓ_1-distance. Based on the definition of S, $q_1(S) - q_2(S)$ is equal to the total variation distance (i.e., half of the ℓ_1-distance) between q_1 and q_2. With probability $5/6$,

$$\begin{aligned}
\|p - q_\alpha\|_1 &= \sum_i |p(i) - q_\alpha(i)| \;=\; \sum_i |(1-\alpha^*)q_1(i) + \alpha^*\, q_2(i) - (1-\alpha)q_1(i) - \alpha\, q_2(i)| \\
&= \sum_i |(\alpha - \alpha^*)(q_1(i) - q_2(i))| \;=\; |(\alpha - \alpha^*)| \cdot \|q_1 - q_2\|_1 \\
&= 2\,|(\alpha - \alpha^*)| \cdot (q_1(S) - q_2(S)) \;=\; 2\left|p(S) - w_S - \tfrac{\epsilon}{4}\right| \;\leq\; \epsilon.
\end{aligned}$$

■

Proof of Lemma 4: To prove (i), note that

$$\|p' - q'_\alpha\|_1 = \sum_{i=1}^{n}\sum_{j=1}^{a_i} |p'(i,j) - q'_\alpha(i,j)| = \sum_{i=1}^{n}\sum_{j=1}^{a_i} \frac{|p(i) - q_\alpha(i)|}{a_i} = \sum_{i=1}^{n} |p(i) - q_\alpha(i)| = \|p - q_\alpha\|_1.$$

For (ii),

$$\begin{aligned}
|D| &= \sum_{i=1}^{n} a_i \leq \sum_{i=1}^{n}\left(nq_\alpha(i) + \tfrac{n|q_\alpha(i) - q_2(i)|}{\|q_\alpha - q_2\|_1} + 1\right) \\
&= n\left(\sum_{i=1}^{n} q_\alpha(i)\right) + \tfrac{n}{\|q_\alpha - q_2\|_1}\left(\sum_{i=1}^{n} |q_\alpha(i) - q_2(i)|\right) + n = 3n\,.
\end{aligned}$$

We now prove (iii). If $q_\alpha(i) < 1/n$, then $q'_\alpha(i,j) < 1/n$. If $q_\alpha(i) \geq 1/n$, then $a_i \geq n q_\alpha(i)$ and hence $q'_\alpha(i,j) \leq 1/n$. Therefore, $q'_\alpha(i,j) \leq 1/n$ for all i, j. Since the domain size of q'_α is at most $3n$, $\|q'_\alpha\|_2$ is at most $\sqrt{3/n}$.

Finally, we show (iv). Since p is a mixture distribution, there is an $\alpha^* \in [0,1]$ such that $p = (1-\alpha^*)q_1 + \alpha^* q_2$. Also, we have that q_α has a mixture parameter $\alpha \leq \alpha^*$. Furthermore, we also have $\|p - q_\alpha\|_1 \leq \epsilon'$. Let $\beta = (\alpha^* - \alpha)/(1-\alpha)$ which is in $[0,1]$. Observe that

$$(1-\beta)q_\alpha + \beta\, q_2 = (1-\beta)(1-\alpha)q_1 + (\alpha(1-\beta)+\beta)\, q_2 = (1-\alpha^*)q_1 + \alpha^*\, q_2 = p\,.$$

Thus, p is a mixture of q_α and q_2. For an element $(i,j) \in D$, we can bound the difference of $p'(i,j)$ and $q'_\alpha(i,j)$ as follows, which finishes the proof.

$$\begin{aligned} |p'(i,j) - q'_\alpha(i,j)| &= \tfrac{|p(i)-q_\alpha(i)|}{a_i} = \tfrac{\beta\cdot|q_\alpha(i)-q_2(i)|}{a_i} \leq \tfrac{\beta\cdot d\cdot|q_\alpha(i)-q_2(i)|}{n\cdot|q_\alpha(i)-q_2(i)|} = \tfrac{\beta d}{n} \\ &= \tfrac{1}{n}\sum_{i=1}^{n} \beta|q_\alpha(i) - q_2(i)| = \tfrac{1}{n}\sum_{i=1}^{n} |p(i) - q_\alpha(i)| = \tfrac{\|p-q_\alpha\|_1}{n} \leq \tfrac{\epsilon'}{n}. \end{aligned}$$ ■

Appendix D. Proofs for Section 5

Proof of Lemma 8: In this proof, we adapt the proof of Proposition 3.1 from Chan et al. (2014). Recall that

$$f(\alpha) = \sum_{i=1}^{n} (X_i - (1-\alpha)Y_i - \alpha Z_i)^2 - X_i - (1-\alpha)^2 Y_i - \alpha^2 Z_i.$$

Via the Poissonization method, we can assume X_i (similarly Y_i and Z_i) is a random variable from $\mathrm{Poi}(s\, p(i))$ (similarly $\mathrm{Poi}(s\, q_1(i))$ and $\mathrm{Poi}(s\, q_2(i))$), which is drawn independently from the rest of the random variables. Note that if x is a Poisson random variable with mean λ, then $\mathbf{E}\big[x^2\big]$ is $\lambda^2 + \lambda$. Using this equation and the independence of the random variables, for a fixed α, we have:

$$\mathbf{E}_{X,Y,Z}[f(\alpha)] = s^2 \cdot \|p - q_\alpha\|_2^2\,.$$

Now, we bound the variance of $f(X,Y,Z,\alpha)$ for a fixed α. Let W_i denote a single term in the summation:

$$W_i \coloneqq (X_i - (1-\alpha)Y_i - \alpha Z_i)^2 - X_i - (1-\alpha)^2 Y_i - \alpha^2 Z_i\,.$$

Using the moments of the Poisson distribution, we have

$$\begin{aligned} \mathbf{Var}_{X,Y,Z}[W_i] &= \mathbf{E}_{X,Y,Z}\big[W_i^2\big] - \mathbf{E}_{X,Y,Z}[W_i]^2 \\ &= 4\, s^3\, (p(i) - (1-\alpha)q_1(i) - \alpha\, q_2(i))^2 \cdot (p(i) + (1-\alpha)^2 q_1(i) + \alpha^2 q_2(i)) \\ &+ 2\, s^2\, (p(i) + (1-\alpha)^2 q_1(i) + \alpha^2 r(i))^2 \\ &\leq 4\, s^3\, (p(i) - (1-\alpha)q_1(i) - \alpha\, q_2(i))^2 \cdot (p(i) + (1-\alpha)q_1(i) + \alpha q_2(i)) \\ &+ 2\, s^2\, (p(i) + (1-\alpha)q_1(i) + \alpha q_2(i))^2 \\ &= 4\, s^3\, (p(i) - q_\alpha(i))^2 \cdot (p(i) + q_\alpha(i)) + 2\, s^2\, (p(i) + q_\alpha(i))^2. \end{aligned}$$

Now, we bound the variance of f which is the sum of n independent terms, W_i's. Using the Cauchy–Schwarz inequality, and the fact that $(p(i) + q_\alpha(i))^2$ is at most $2p(i)^2 + 2q_\alpha(i)^2$, we have

$$\begin{aligned}\mathbf{Var}_{X,Y,Z}[f(\alpha)] &= \sum_{i=1}^{n} \mathbf{Var}_{X,Y,Z}[W_i] \\ &\leq 4\, s^3\, (p(i) - q_\alpha(i))^2 \cdot (p(i) + q_\alpha(i)) + 2\, s^2\, (p(i) + q_\alpha(i))^2 \\ &\leq 4\, s^3 \sqrt{\left(\sum_{i=1}^{n} (p(i) - q_\alpha(i))^4\right) \cdot \left(\sum_{i=1}^{n} (p(i) + q_\alpha(i))^2\right)} + 4\, s^2 \left(\|p\|_2^2 + \|q_\alpha\|_2^2\right) \\ &\leq 8\, s^3 \cdot \|p - q_\alpha\|_4^2 \cdot \sqrt{b} + 8\, s^2\, b\,,\end{aligned}$$

where b is at least $\|p\|_2^2$ and $\|q_\alpha\|_2^2$ by the first condition of the theorem. ■

Proof of Lemma 7: Recall that B is defined to be $2\sum_{i=1}^{n} Y_i + X_iY_i + Y_iZ_i - Y_i^2 - X_iZ_i$. To analyze the expected value and the variance of B, we consider each terms in the sum. Let B_i denote a single term in the sum after ignoring constant 2:

$$B_i \coloneqq Y_i + X_iY_i + Y_iZ_i - Y_i^2 - X_iZ_i\,.$$

Note that via the Poissonization method, the X_i's, the Y_i's, the Z_i's, and consequently the B_i's are independent random variables. Note that if x is a Poisson random variable with mean λ, then $\mathbf{E}\left[x^2\right]$ is $\lambda^2 + \lambda$. Using this equation, we compute the expected value of B_i:

$$\begin{aligned}\mathbf{E}_{X,Y,Z}[B_i] &= -s^2 \left(q_1(i)^2 + p(i)q_1(i) + q_1(i)q_2(i) - p(i)q_2(i)\right) \\ &= -\alpha^* s^2 (q_1(i) - q_2(i))^2\,.\end{aligned}$$

Thus, the expected value of $-B$ is the following:

$$\mathbf{E}_{X,Y,Z}[-B] = -\sum_{i=1}^{n} 2B_i = \sum_{i=1}^{n} 2\alpha^* s^2 (q_1(i) - q_2(i))^2 = 2\alpha^* s^2 \|q_1 - q_2\|_2^2,$$

where $2\alpha^*$ is a constant between $[1, 2]$. Using the first four moments of the Poisson distribution and the fact that $\alpha \leq 1$, we have the following:

$$\begin{aligned}\mathbf{Var}_{X,Y,Z}[B_i] &= \mathbf{E}_{X,Y,Z}\left[B_i^2\right] - \mathbf{E}_{X,Y,Z}[B_i]^2 \\ &= s^3\, \alpha^* (1 + \alpha^*)\, (q_1(i) - q_2(i))^2 (q_1(i) + q_2(i)) + 2\, s^3\, (q_1(i) - q_2(i))^2\, q_1(i) \\ &\quad + s^2 \left(\alpha^* (q_2(i)^2 - q_1(i)^2) + q_1(i)\, (3q_1(i) + 2q_2(i))\right) \\ &\leq 4\, s^3\, (q_1(i) + q_2(i))(q_1(i) - q_2(i))^2 + s^2\, (q_1(i) + q_2(i))^2 + 3\, s^2\, q_1(i)^2\,.\end{aligned}$$

Using the bound above and the Cauchy–Schwarz inequality, we bound the variance of B as follows:

$$\begin{aligned}\mathbf{Var}_{X,Y,Z}[B] &= 4\sum_{i=1}^{n}\mathbf{Var}_{X,Y,Z}[B_i]\\ &\leq 16\,s^3\sum_{i=1}^{n}(q_1(i)+q_2(i))(q_1(i)-q_2(i))^2 + 8\|q_2\|_2^2 + 20\,s^2\|q_1\|_2^2\\ &\leq 16s^3\sqrt{\left(\sum_{i=1}^{n}(q_1(i)-q_2(i))^4\right)\cdot\left(\sum_{i=1}^{n}(q_1(i)+q_2(i))^2\right)} + 28\,s^2 b\\ &\leq 32\,s^3\|q_1-q_2\|_4^2\sqrt{b} + 28\,s^2\,b\,.\end{aligned}$$

Clearly, the variance of $-B$ is equal to the variance of B, and it is bounded the same as above. Note that γ is at most $\|q_1 - q_2\|_2^2$, and the sample sets, X, Y, and Z, each have at least $\Theta(\sqrt{b}/\gamma)$ samples. By Lemma 21, there exists $c_B \in [0,1]$ such that

$$-B = 2\,c_B\,\alpha^*\|q_1 - q_2\|_2^2,$$

with probability 0.99 which concludes the proof. ■

Proof of Lemma 9: Consider the statistic as a function of α: $f(\alpha) = A\alpha^2 + B\alpha + C$. Since A is positive, f takes its minimum at $\alpha_{\min} := (-B)/2A$. By Equation 3 and Equation 4, $\alpha_{\min}$ is $c_B\alpha^*/c_A$, and for any α, we have:

$$\begin{aligned}f(\alpha^*) - f(\alpha) &= A({\alpha^*}^2 - \alpha^2) + B(\alpha^* - \alpha)\\ &= c_A\,s^2\,\|q_1-q_2\|_2^2\,(\alpha^*-\alpha)\left(\alpha^* + \alpha - \frac{2\alpha^*(c_B)}{c_A}\right)\\ &= c_A\,s^2\,\|q_1-q_2\|_2^2\,(\alpha^*-\alpha)\,(\alpha^*+\alpha-2\alpha_{\min})\,.\end{aligned} \tag{13}$$

Depending on whether α^* is larger than $\alpha_{\min}$ or not, we consider the following cases.

Case 1: $\boldsymbol{\alpha^* \geq \alpha_{\min}}$. Let $\hat{\alpha}_1$ be the smallest number in $[\alpha_{\min}, 1]$ for which $|f(\hat{\alpha}_1)|$ is at most T. Clearly, $\hat{\alpha}_1$ exists since α^* is a potential solution, so the solution interval is not empty. Note that based on the way we pick $\hat{\alpha}_1$, the following are true: (i) $\hat{\alpha}_1$ is at most α^*, (ii) $f(\hat{\alpha}_1)$ is at least $-T$, and (ii) since A is positive, and f is increasing over $[\alpha_min, 1]$, then $f(\hat{\alpha}_1)$ is at most $f(\alpha^*)$. Hence, by Equation 13, we have:

$$0 \leq f(\alpha^*) - f(\hat{\alpha}_1) = c_A\,s^2\,\|q_1-q_2\|_2^2\,(\alpha^*-\hat{\alpha}_1)\,(\alpha^*+\hat{\alpha}_1-2\alpha_{\min}) \leq 2\,T\,.$$

If we replace $\alpha^* + \hat{\alpha}_1 - 2\alpha_{\min}$ by a smaller quantity ,$\alpha^* - \hat{\alpha}_1$, where both are positive then we have:

$$s^2\,\|q_1-q_2\|_2^2\,(\alpha^*-\hat{\alpha}_1)^2 \leq \frac{2\,T}{c_A} \leq \frac{2}{0.9}\,T\,. \tag{14}$$

Case 2: $\boldsymbol{\alpha^* \leq \alpha_{\min}}$. We replicate what we did in the previous case. Let $\hat{\alpha}_2$ be the largest number in $[0, \alpha_{\min}]$ for which $|f(\hat{\alpha}_2)|$ is at most T. Clearly, $\hat{\alpha}_2$ exists since α^* is a potential solution, so the solution interval is not empty. Note that based on the way we pick $\hat{\alpha}_2$, the following are true:

(i) $\hat{\alpha}_2$ is at least α^*, (ii) $f(\hat{\alpha}_2)$ is at least $-T$, and (iii) since A is positive, and f is decreasing over $[0, \alpha_m in]$, then $f(\hat{\alpha}_2)$ is at most $f(\alpha^*)$. Hence, by Equation 13, we have:

$$\begin{aligned} 0 \leq f(\alpha^*) - f(\hat{\alpha}_2) &= c_A\, s^2 \, \|q_1 - q_2\|_2^2 \, (\alpha^* - \hat{\alpha}_2)\,(\alpha^* + \hat{\alpha}_2 - 2\alpha_{\min}) \\ &= c_A\, s^2 \, \|q_1 - q_2\|_2^2 \, (\hat{\alpha}_2 - \alpha^*)\,(2\alpha_{\min} - \alpha^* - \hat{\alpha}_2) \leq 2\,T\,. \end{aligned}$$

If we replace $2\alpha_{\min}\alpha^* - \hat{\alpha}_2$ by a smaller quantity, $\hat{\alpha}_2 - \alpha^*$, where both are positive, then we have:

$$s^2 \, \|q_1 - q_2\|_2^2 \, (\alpha^* - \hat{\alpha}_2)^2 \leq \frac{2\,T}{c_A} \leq \frac{2}{0.9}\,T\,. \tag{15}$$

The left side of Equation 14 and Equation 15 are in the form of the ℓ_2-distance between two mixture distributions p and $q_{\hat{\alpha}}$ due to the following:

$$\|p - q_{\hat{\alpha}}\|_2^2 = \sum_{i=1}^{n} (p(i) - q_{\hat{\alpha}}(i))^2 = \sum_{i=1}^{n} ((1-\alpha^*)\,q_1(i) + \alpha^* q_2(i) - (1-\hat{\alpha})\,q_1(i) - \hat{\alpha}\,q_2(i))^2 \tag{16}$$

$$= (\alpha^* - \hat{\alpha})^2 \sum_{i=1}^{n} (q_1(i) - q_2(i))^2 = (\alpha^* - \hat{\alpha})^2 \|q_1 - q_2\|_2^2\,. \tag{17}$$

Note that we are either in case 1 or case 2. So, on of the two equations, Equation 14, Equation 15 has to be true. By Equation 16, of the following is true.

$$\|p - q_{\hat{\alpha}_1}\|_2 \leq \frac{2\,T}{0.9\,s^2} \qquad\qquad \|p - q_{\hat{\alpha}_2}\|_2 \leq \frac{2\,T}{0.9\,s^2},$$

which concludes the proof. ■

Lemma 21 *[Adapted from Chan et al. (2014)] Assume a random variable, namely R, has the following properties:*

$$\mathbf{E}[R] = c_1 s^2 \|q_1 - q_2\|_2^2, \qquad \mathbf{Var}[R] \leq c_2 s^3 \|q_1 - q_2\|_4^2 \sqrt{b} + c_3 s^2 b, \tag{18}$$

where c_1, c_2, and c_3 are three positive constants, s is an integer, q_1 and q_1 are two distributions over $[n]$, and b is a real number which is greater than $\|q_1\|_2^2$ and $\|q_2\|_2^2$. If s is at least $c \cdot \sqrt{b}/\tau$ for sufficiently large c, then with probability 0.99 the following is true:

- *If $\|q_1 - q_2\|_2^2$ is at most τ, then $|R|$ is at most $2\,c_1\,\tau\,s^2$.*
- *If $\|q_1 - q_2\|_2^2$ is at least τ, then R is between $0.9 \cdot \mathbf{E}[R]$ and $1.1 \cdot \mathbf{E}[R]$.*

Proof We use Chebyshev's inequality to prove the lemma. For the first case, by the ℓ_p-norms inequality, we have the following:

$$\begin{aligned} \mathbf{Pr}\left[|R - \mathbf{E}[R]| \geq c_1\,\tau s^2\right] &\leq \frac{\mathbf{Var}[R]}{c_1^2 \tau^2 s^4} \leq \frac{c_2 \|q_1 - q_2\|_4^2 \sqrt{b}}{c_1^2 \tau^2 s} + \frac{c_3\,b}{c_1^2 \tau^2 s^2} \\ &\leq \frac{c_2 \|q_1 - q_2\|_2^2 \sqrt{b}}{c_1^2 \tau^2 s} + \frac{c_3\,b}{c_1^2 \tau^2 s^2} \leq \frac{c_2 \sqrt{b}}{c_1^2 \tau\, s} + \frac{c_3\,b}{c_1^2 \tau^2 s^2} \leq 0.01, \end{aligned}$$

Algorithm 5: An algorithm for estimating the ℓ_2-distance squared Chan et al. (2014)

```
Procedure: $\ell_2^2$-ESTIMATOR($b, \sigma$, sample access to $r_1$ and $r_2$)
$s \leftarrow \Theta(\sqrt{b}/\sigma))$
$X \leftarrow$ Draw $s$ samples from $r_1$.
$Y \leftarrow$ Draw $s$ samples from $r_2$.
return $\frac{1}{s^2} \sum_{i=1}^{n} (X_i - Y_i)^2 - X_i - Y_i$
```

where the last inequality is true when $s \geq \max(200\, c_2/c_1^2, \sqrt{200\, c_3}/c_1)\sqrt{b}/\tau$.

For the second case, we have the following:

$$\begin{aligned}
\mathbf{Pr}[|R - \mathbf{E}[R]| \geq 0.1 \cdot \mathbf{E}[R]] &\leq \frac{100\, \mathbf{Var}[R]}{\mathbf{E}[R]^2} \leq \frac{100\, c_2 \|q_1 - q_2\|_4^2 \sqrt{b}}{c_1^2\, s\, \|q_1 - q_2\|_2^4} + \frac{100\, c_3\, b}{c_1^2\, s^2\, \|q_1 - q_2\|_2^4} \\
&\leq \frac{100\, c_2 \sqrt{b}}{c_1^2\, s\, \|q_1 - q_2\|_2^2} + \frac{100\, c_3\, b}{c_1^2\, s^2\, \|q_1 - q_2\|_2^4} \\
&\leq \frac{100\, c_2 \sqrt{b}}{c_1^2 \tau\, s} + \frac{100\, c_3\, b}{c_1^2 \tau^2 s^2} \leq 0.01,
\end{aligned}$$

where the last inequality is true when $s \geq \max(20000\, c_2/c_1^2, \sqrt{20000\, c_3}/c_1)\sqrt{b}/\tau$. This completes the proof. ■

Lemma 22 *[Restated from Chan et al. (2014)] The procedure ℓ_2^2-ESTIMATOR (b, σ, r_1, r_2) described in Algorithm 5, that uses $\Theta(\sqrt{b}/\sigma)$ samples, has the following property with probability 0.99:*

- *If $\|r_1 - r_2\|_2^2$ is at most σ, then $|R|$ is at most $2\, \sigma\, s^2$.*
- *If $\|r_1 - r_2\|_2^2$ is at least σ, then R is between $0.9 \cdot \mathbf{E}[R]$ and $1.1 \cdot \mathbf{E}[R]$.*

Proof We use the ℓ_2^2-distance estimator proposed in Chan et al. (2014). However, for the sake of completeness, we provide the process in Algorithm 5. X and Y are two sample sets each containing s samples from r_1 and r_2 respectively. Let X_i and Y_i indicate the numbers of samples in X and Y respectively. The authors showed that the expected value of the statistic $\sum_{i=1}^{n} (X_i - Y_i)^2 - X_i - Y_i$ is $s^2 \|r_1 - r_2\|_2^2$, and the variance is bounded by $8 s^3\, \|r_1 - r_2\|_4^2\, \sqrt{b} + 8\, s^2\, b$. By Lemma 21, if we draw $\Theta(\sqrt{b}/\gamma)$ samples, then the algorithm will have the desired property with probability 0.99. ■