

Exploring the Gap Between Tolerant and Non-Tolerant Distribution Testing

Sourav Chakraborty, Eldar Fischer, Arijit Ghosh, Gopinath Mishra, and Sayantan Sen

Abstract—The framework of distribution testing is currently ubiquitous in the field of property testing. In this model, the input is a probability distribution accessible via independently drawn samples from an oracle. The testing task is to distinguish a distribution that satisfies some property from a distribution that is far in some distance measure from satisfying it. The task of tolerant testing imposes a further restriction, that distributions close to satisfying the property are also accepted. This work focuses on the connection between the sample complexities of non-tolerant testing of distributions and their tolerant testing counterparts. When limiting our scope to label-invariant (symmetric) properties of distributions, we prove that the gap is at most quadratic, ignoring poly-logarithmic factors. Conversely, the property of being the uniform distribution is indeed known to have an almost-quadratic gap. When moving to general, non-tolerant testing of distributions and their tolerant testing focuses on the connection between the sample complexities of

I. INTRODUCTION

Let $D$ be a distribution over a finite set $\Omega$, and $\mathcal{P}$ be a property, that is, a set of distributions over $\Omega$. Given access to independent random samples from $\Omega$ according to the distribution $D$, we are interested in the problem of distinguishing whether the distribution $D$ is $\eta$-close to having the property $\mathcal{P}$, or is $\varepsilon$-far from having the property $\mathcal{P}$, where $\eta$ and $\varepsilon$ are two fixed proximity parameters such that $0 \leq \eta < \varepsilon \leq 2$. The distance of the distribution $D$ from the property $\mathcal{P}$ is defined as $\min_{P} ||D - \mathcal{P}||_1$, where $||D - \mathcal{P}||_1$ denotes the $\ell_1$-distance between the distributions $D$ and $\mathcal{P}$. A distribution $D$ is said to be $\eta$-close to $\mathcal{P}$ if the distance of $D$ from $\mathcal{P}$ is at most $\eta$. Similarly, $D$ is said to be $\varepsilon$-far from $\mathcal{P}$, if the distance of $D$ from $\mathcal{P}$ is at least $\varepsilon$. The goal is to design a tester that uses as few samples as possible. For $\eta > 0$, the problem of distinguishing the two cases is referred to as the tolerant distribution testing problem of $\mathcal{P}$, and the particular case where $\eta = 0$ is referred to as the non-tolerant distribution testing problem of $\mathcal{P}$. The sample complexity (tolerant and non-tolerant testing) is the number of samples required by the best algorithm that can distinguish with high probability (usually with probability at least 2/3) whether the distribution $D$ is $\eta$-close to having the property $\mathcal{P}$, or is $\varepsilon$-far from having the property $\mathcal{P}$.

While results and techniques from distribution testing are already interesting in their own right, they have also found numerous applications in central problems in Theoretical Computer Science, and in particular in property testing, e.g., graph isomorphism testing [1], [2] and function isomorphism testing [3], learning theory [4], [5], [6], and differential privacy [7], [8], [9], [10]. Thus, understanding the tolerant and non-tolerant sample complexity of distribution testing is a central problem in theoretical computer science.

There have been extensive studies of non-tolerant and tolerant testing of some specific distribution properties like uniformity, identity with a fixed distribution, equality of two distributions and independence of a joint distribution [11], [12], [13], [14], [15], [16]. Various other specific distribution properties have also been studied [17], [18]. Then, some works

0018-9448 © 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
investigated general tests for the large class of all shape-restricted properties of distributions, which contains properties like monotonicity, log-concavity, modality etc., [19] and [20]. This paper proves general results about the gap between tolerant and non-tolerant distribution testing for large classes of properties.

A. Our Results

We now informally present our results. The formal definitions are presented in Section II. We assume that the distributions are supported over a set \( \Omega = [n] = \{1, 2, \ldots, n\} \). We first prove a result about label-invariant distribution properties (properties that are invariant under all permutations of \( \Omega \)). We show that, for any label-invariant distribution property, there is at most a quadratic blowup in its tolerant sample complexity as compared to its non-tolerant counterpart, ignoring poly-logarithmic factors.

**Theorem 1 (Informal):** Any label-invariant distribution property that can be non-tolerantly tested using \( \Lambda \) samples, can also be tolerantly tested using \( O(\min\{\Lambda^2, n\}) \) samples, where \( n \) is the size of the support of the distribution.\(^2\)

This result gives a unified way for obtaining tolerant testers from their non-tolerant counterparts. The above result will be stated and proved formally in Section IV. Moreover, in Section IV-C, we give a constructive variant of the tolerant tester of Theorem 1, when the property can be expressed as the feasible solution to a set of linear inequalities.

**Theorem 2 (Informal):** Any label-invariant distribution property that can be non-tolerantly tested using \( \Lambda \) samples and can also be expressed as a feasible solution to \( m \) linear inequalities, can also be tolerantly tested using \( O(\min\{\Lambda^2, n\}) \) samples and in time polynomial in \( m \) and \( n \), where \( n \) is the size of the support of the distribution.

Note that any distribution can be learned by using \( O(n) \) samples, that is, given \( O(n/\varepsilon^2) \) samples from an unknown distribution \( D_1 \), we can construct an explicit new distribution \( D_2 \) such that \( |D_1 - D_2| \leq \varepsilon \), where \( \varepsilon \in (0, 2) \) is an input parameter.\(^3\) Note that if \( \Lambda = \Omega(\sqrt{n}) \), Theorem 1 is obvious. It is only interesting if \( \Lambda = o(\sqrt{n}) \). Now we present a property for which this connection is useful. Consider a natural distribution property: given a distribution \( D \) and a parameter \( k \), we want to decide whether the support size of \( D \) is at most \( k \) or \( \varepsilon \)-far from having support at most \( k \). If \( k = o(\sqrt{n}) \), the query complexity for testing this problem is \( O(k \log k) \) \(^{[21]}\).

We would like to point out that the quadratic gap of Theorem 1 and Theorem 2 are tight, in the sense that there can not be any general procedure to construct a tolerant tester from their non-tolerant counterpart in better than quadratic blowup. This is due to the fact that there exists a distribution property (testing uniformity of a distribution) that requires almost quadratic blowup. For example, testing non-tolerant uniformity requires \( \Theta(\sqrt{n}) \) samples and requires \( \Theta(n/\log n) \) samples for tolerant testing.

It is a natural question to investigate the extent to which the quadratic gap is necessary. Though we are not resolving this question completely, as a first step in the direction of extending the above result, we consider the notion of non-concentrated properties. By the notion of a non-concentrated distribution, intuitively, we mean that there is no significant portion of the base set of the distribution that carries only a negligible weight, making the probability mass of the distribution well distributed among its indices. Specifically, any subset \( X \subseteq [n] \), for which \( |X| \) is above some threshold (say \( \beta n \) with \( \beta \in (0, \frac{1}{2}) \)), has probability mass of at least another threshold (say \( \alpha \) with \( \alpha \in (0, \frac{1}{2}) \)). A property is said to be non-concentrated if only non-concentrated distributions can satisfy the property. We prove a lower bound on the testing of any non-concentrated property (not necessarily label-invariant).

**Theorem 3 (Informal):** In order to non-tolerantly test any non-concentrated distribution property, \( \Omega(\sqrt{n}) \) samples are required, where \( n \) is the size of the support of the distribution.

The sufficiency of the quadratic gap between tolerant testing and non-tolerant testing for any non-concentrated property follows from the above theorem, since by a folklore result, only \( O(n) \) many samples are required to learn any distribution approximately.

The proof of Theorem 3 for label-invariant non-concentrated properties is a generalization of the proof of the \( \Omega(\sqrt{n}) \) lower bound for classical uniformity testing, while for the whole theorem, that is, for the general (not label-invariant) non-concentrated properties, a more delicate argument is required. The formal proof is presented in Section V-B.

The next natural question is about the sample complexity of any tolerant tester for non-concentrated properties. We address this question for label-invariant non-concentrated properties by proving the following theorem in Section V. However, the question is left open for non-label-invariant properties.

**Theorem 4 (Informal):** The sample complexity for tolerantly testing any non-concentrated label-invariant distribution property is \( \Omega(n^{1-o(1)}) \), where \( n \) is the size of the support of the distribution.

Note that Theorem 3 and Theorem 4 imply that for the general class of non-concentrated label-invariant properties whose non-tolerant sample complexity is \( \Theta(\sqrt{n}) \), the quadratic gap is almost tight. Note that fundamental properties like uniformity falls under this class. As already noted, non-tolerant uniformity testing requires \( \Theta(\sqrt{n}) \) samples and tolerant uniformity testing requires \( \Theta(n/\log n) \) samples. The tightness of this quadratic gap for general properties remains open.

A natural question related to tolerant testing is:

**How many samples are required to learn a distribution?**

As pointed out earlier, any distribution can be learnt using \( O(n) \) samples. But what if the distribution happens to be very concentrated? We present an upper bound result for learning a distribution, in which the sample complexity depends on the minimum cardinality of any set \( S \subseteq [n] \) over which the unknown distribution is concentrated.

**Theorem 5 (Informal):** To learn a distribution approximately, \( O(|S|) \) samples are enough, where \( S \subseteq [n] \) is an unknown set of minimum cardinality whose mass is close...
to 1. Note that $|S|$ is also unknown, and the algorithm adapts to it.

Observe that we cannot learn a distribution supported on the set $S$ using $o(|S|)$ samples, so the above result is essentially tight.

**B. Related Works**

Several forms of distribution testing have been investigated for over a century in statistical theory [22], [23], while combinatorial properties of distributions have been explored over the last two decades in Algorithm Theory, Machine Learning and Information Theory [24], [25], [26], [27]. In Algorithm Theory, the investigation into testing properties of distributions started with the work of Goldreich and Ron [28], even though it was not directly stated there in these terms. Batu et al. [11] launched the intensive study of property testing of distributions with the problem of equivalence testing. Later, Batu et al. [12] studied the problems of identity and independence testing of distributions. Since then there has been a flurry of interesting works in this model. For example, Paninski [13] proved tight bounds on uniformity testing, Valiant and Valiant [15] resolved the tolerant sample complexity for a large class of label-invariant properties that includes uniformity testing, Acharya et al. [29] proved various optimal testing results under several distance measures, and Valiant and Valiant [16] studied the sample complexity of instance optimal identity testing. In [17], Batu and Canonne studied the problem of generalized uniformity testing, where the distribution is promised to be supported on an unknown set $S$, and proved a tight bound of $\tilde{\Theta}(|S|^{2/3})$ samples for non-tolerant uniformity testing. This is in contrast to the non-tolerant uniformity testing of a distribution supported over $[n]$, whose sample complexity is $\Theta(\sqrt{n})$, ignoring the dependence on the proximity parameter. Daskalakis et al. [30] studied the problem of tolerant testing under various distance measures. Very recently, Canonne et al. [31] revisited the problem of determining the sample complexity of tolerant identity testing, where they proved the optimal dependence on the proximity parameters. Going beyond studying specific properties, Canonne et al. [19] studied the class of shape-restricted properties of a distribution, a condition general enough to contain several interesting properties like monotonicity, log-concavity, $t$-modality etc. Their result was later improved by Fischer et al. [20]. See the surveys of Canonne [32], [33] for a more exhaustive list.

While the most studied works concentrate on non-tolerant testing of distributions, a natural extension is to test such properties tolerantly. Since the introduction of tolerant testing in the pioneering work of Parnas et al. [34], that defined this notion for classical (non-distribution) property testing, there have been several works in this framework. Note that it is nontrivial in many cases to construct tolerant testers from their non-tolerant counterparts, as in the case of tolerant junta testing [35] for example. In a series of works, it has been proved that tolerant testing of the most natural distribution properties, like uniformity, requires an almost linear number of samples [14], [15].

Now a natural question arises about how the sampling complexity of tolerant testing is related to non-tolerant testing of distributions in general. To the best of our knowledge, there is no known example with more than a quadratic gap.

It would also be interesting to bound the gap for sample-based testing as defined in the work of Goldreich and Ron [36]. This model was investigated further in the work of Fischer et al. [37], where a general upper bound for non-tolerant sample-based testing of strongly testable properties was provided.

1) **Organization of the Paper:** Section II contains the definitions used throughout the paper. The high-level overview of the proofs of all our theorems are presented in Section III. Section IV contains the formal statement and proof of Theorem 1. Theorem 2 has been formally stated and proved in Section IV-C. Theorem 3, Theorem 4 and Theorem 5 are formally stated and proved in Section V, Section V-B and Section VI respectively. Finally we conclude with some open problems in Section VII.

**II. Notation and Definitions**

A probability distribution $D$ over a universe $\Omega = [n]$ is a non-negative function $D : \Omega \rightarrow [0,1]$ such that $\sum_{i \in \Omega} D(i) = 1$. For $S \subseteq \Omega$, the mass of $S$ is defined as $D(S) = \sum_{i \in S} D(i)$, where $D(i)$ is the mass of $i$ in $D$. The support of a probability distribution $D$ on $\Omega$ is denoted by $\text{Supp}(D)$. For any distribution $D$, by the top $t$ elements of $D$, we refer to the first $t$ elements in the support of $D$ when the elements in the support are sorted according to a non-increasing order of their probability masses in $D$. When we write $\tilde{O}(\cdot)$, it sometimes suppresses a poly-logarithmic term in $n$ and the inverse of the proximity parameter(s), as well as the inverse of the difference of two proximity parameters. Although there are several other distance measures, in this work, we mainly focus on the $\ell_1$ distance.

**Definition 6 (Distribution Property):** Let $D$ denote the set of all distributions over $\Omega$. A distribution property $\mathcal{P}$ is a subset of $D$. A distribution $D \in \mathcal{P}$ is said to be in the property or to satisfy the property. Otherwise, $D$ is said to be not in the property or to not satisfy the property.

**Definition 7 (Label-Invariant Property):** Let us consider a property $\mathcal{P}$. For a distribution $D$ and a permutation $\sigma : \Omega \rightarrow \Omega$, consider the distribution $D_{\sigma}$ defined as $D_{\sigma}(\sigma(i)) = D(i)$ (equivalently, $D_{\sigma}(i) = D(\sigma^{-1}(i))$) for each $i \in \Omega$. If for every distribution $D$ in $\mathcal{P}$, $D_{\sigma}$ is also in $\mathcal{P}$ for every permutation $\sigma$, then the property $\mathcal{P}$ is said to be label-invariant.

**Definition 8 (Distance Between Two Distributions):** The distance between two distributions $D_1$ and $D_2$ over $\Omega$ is the standard $\ell_1$ distance between them, which is defined as $||D_1 - D_2||_1 := \sum_{i \in \Omega} |D_1(i) - D_2(i)|$. For $\eta \in [0,2]$, $D_1$ and $D_2$ are $\eta$-close if $||D_1 - D_2||_1 \leq \eta\sum_{i \in \Omega} D_2(i)$.

2Given two unknown probability distributions that can be accessed via samples from their respective oracles, equivalence testing refers to the problem of distinguishing whether they are same or far from each other.

3Given an unknown distribution accessible via samples, the problem of identity testing refers to the problem of distinguishing whether it is identical to a known distribution or far from it.

4To be precise, the exact lower bounds for non-tolerant uniformity testing is $\Omega(\sqrt{n})$, and for tolerant uniformity testing it is $\Omega(\frac{1}{\log \varepsilon})$, where $n$ is the support size of the distribution and the proximity parameter $\varepsilon$ is constant.
D_2 are said to be \( \eta \)-close to each other if \( ||D_1 - D_2||_1 \leq \eta \). Similarly, for \( \varepsilon \in [0, 2] \), D_1 and D_2 are said to be \( \varepsilon \)-far from each other if \( ||D_1 - D_2||_1 \geq \varepsilon \).

**Definition 9 (Distance of a Distribution From a Property):** The distance of a distribution \( D \) from a property \( \mathcal{P} \) is the minimum \( \ell_1 \)-distance between \( D \) and any distribution in \( \mathcal{P} \). For \( \eta \in [0, 2] \), a distribution \( D \) is said to be \( \eta \)-close to \( \mathcal{P} \) if the distance of \( D \) from \( \mathcal{P} \) is at most \( \eta \). Analogously, for \( \varepsilon \in [0, 2] \), a distribution \( D \) is said to be \( \varepsilon \)-far from \( \mathcal{P} \) if the distance of \( D \) from \( \mathcal{P} \) is at least \( \varepsilon \).

**Definition 10 ((\( \eta, \varepsilon \))-Tester):** An \((\eta, \varepsilon)\)-tester for a distribution property is a randomized algorithm that has sample access to the unknown distribution (upon query it can receive elements of \( \Omega \), each drawn according to the unknown distribution, independently of any previous query or the algorithm’s private coins), and distinguishes whether the distribution is \( \eta \)-close to the property or \( \varepsilon \)-far from the property, with probability at least \( \frac{1}{2} \), where \( \eta \) and \( \varepsilon \) are proximity parameters such that \( 0 \leq \eta < \varepsilon \leq 2 \). The tester is said to be tolerant when \( \eta > 0 \), and non-tolerant when \( \eta = 0 \).

Now we define the notions of non-concentrated distributions and non-concentrated properties.

**Definition 11 (Non-Concentrated Distribution):** A distribution \( D \) over the domain \( \Omega = [n] \) is said to be \((\alpha, \beta)\)-non-concentrated if for any set \( S \subseteq \Omega \) with size \( \beta n \), the probability mass on \( S \) is at least \( \alpha \), where \( \alpha \) and \( \beta \) are two parameters such that \( 0 < \alpha \leq \beta < \frac{1}{2} \).

**Definition 12 (Non-Concentrated Property):** Let \( 0 < \alpha \leq \beta < \frac{1}{2} \). A distribution property \( \mathcal{P} \) is defined to be \((\alpha, \beta)\)-non-concentrated, if all distributions in \( \mathcal{P} \) are \((\alpha, \beta)\)-non-concentrated.

Note that the uniform distribution is \((\alpha, \alpha)\)-non-concentrated for every \( \alpha \), and so is the property of being identical to the uniform distribution. Also, for any \( 0 < \alpha < \frac{1}{2} \) such that \( \alpha n \) is an integer, the uniform distribution is the only \((\alpha, \alpha)\)-non-concentrated one. Finally, observe that any arbitrary distribution is both \((0, \beta)\)-non-concentrated and \((\alpha, 1)\)-non-concentrated, for any \( \alpha, \beta \in (0, 1) \).

### A. Useful Concentration Bounds

Now we state some large deviation results that will be used throughout the paper.

**Lemma 13 (Multiplicative Chernoff Bound [38]):** Let \( X_1, \ldots, X_n \) be independent random variables such that \( X_i \in [0, 1] \). For \( X = \sum_{i=1}^{n} X_i \) and \( \mu = \mathbb{E}[X] \), the following holds for any \( 0 \leq \delta \leq 1 \).

\[
\mathbb{P}(|X - \mu| \geq \delta \mu) \leq 2 \exp \left( -\frac{\mu \delta^2}{3} \right)
\]

**Lemma 14 (Additive Chernoff Bound [38]):** Let \( X_1, \ldots, X_n \) be independent random variables such that \( X_i \in [0, 1] \). For \( X = \sum_{i=1}^{n} X_i \) and \( \mu_1 \leq \mathbb{E}[X] \leq \mu_h \), the following hold for any \( \delta > 0 \).

(i) \( \mathbb{P}(X \geq \mu_h + \delta) \leq \exp \left( -\frac{2\delta^2}{n} \right) \).

(ii) \( \mathbb{P}(X \leq \mu_h - \delta) \leq \exp \left( -\frac{2\delta^2}{n} \right) \).

### III. Overview of Our Results

In this section, we give an overview of our results as follows:

**A. Overview of Proof of Theorem 1 and Theorem 2**

We first show that for any label-invariant distribution property, the sample complexities of tolerant and non-tolerant testing are separated by at most a quadratic factor, ignoring poly-logarithmic terms. More specifically, in Theorem 1, we prove that for any label-invariant distribution property \( \mathcal{P} \) that has a non-tolerant tester with sample complexity \( \Lambda \), there exists a tolerant tester for \( \mathcal{P} \) that uses \( \tilde{O}(\Lambda^2) \) samples, ignoring poly-logarithmic factors. Since we can learn a distribution using \( \tilde{O}(n) \) samples, our proof is particularly useful when \( \Lambda = o(\sqrt{n}) \), where \( n \) is the size of the support of the distribution that is being tested.

To prove Theorem 1 (restated as Theorem 15), we provide an algorithm for tolerant testing of \( \mathcal{P} \) with sample complexity \( \tilde{O}(\Lambda^2) \), based on the existence of a non-tolerant tester of \( \mathcal{P} \) with sample complexity \( \tilde{O}(\Lambda) \). Given the existence of such a non-tolerant tester with sample complexity \( \tilde{O}(\Lambda) \), one crucial observation that we use here is that there cannot be two distributions \( D_1 \) and \( D_2 \) that are identical on the elements with mass \( \Omega \left( \frac{1}{\sqrt{n}} \right) \) (we call them high elements), where \( D_1 \) is in the property \( \mathcal{P} \) while \( D_2 \) is far from \( \mathcal{P} \). This is formally stated as Lemma 17.

Given that the two distributions \( D_1 \) and \( D_2 \) are identical on all elements with mass \( \Omega \left( \frac{1}{\sqrt{n}} \right) \), by the birthday paradox, we can say that \( \tilde{O}(\Lambda) \) samples are not enough to obtain any low elements, that is, elements with mass \( o \left( \frac{\Lambda}{\sqrt{n}} \right) \), that appear more than once. Since the property \( \mathcal{P} \) is label-invariant, we can apply uniformly random permutations over the low elements of both \( D_1 \) and \( D_2 \), making the samples obtained from both \( D_1 \) and \( D_2 \) appear as two uniformly random sequences. Thus, from the view of any tester that takes only \( \tilde{O}(\Lambda) \) samples, \( D_1 \) and \( D_2 \) will appear the same, which would contradict the existence of a non-tolerant tester that distinguishes \( D_1 \) from \( D_2 \) using \( \tilde{O}(\Lambda) \) samples. At this point, we would like to point out that the proof of Lemma 17 only assumes the existence of a non-tolerant tester, and is oblivious to its internal details. Later, in Lemma 18, we generalize this idea to show that when \( D_1 \) and \( D_2 \) are close with respect to the high elements, it cannot be the case that \( D_1 \) is in the property \( \mathcal{P} \), while \( D_2 \) is far from \( \mathcal{P} \). Although the proof follows a similar line to that of Lemma 17, more careful analysis is required to prove Lemma 18. Note that Lemma 18 is the main technical lemma required to prove Theorem 1.

Once we have Lemma 17 and Lemma 18, we can describe the algorithm of Theorem 1. Broadly speaking, we show that partial learning of the distribution is sufficient for constructing a tolerant tester for any label-invariant property, as opposed to the more familiar paradigm of testing by learning [39], [40]. Using Lemma 18, we show that estimating the masses of only the high elements is enough for us, along with the fact that the property \( \mathcal{P} \) that we are testing is label-invariant. Roughly, the algorithm has three steps. In the first step, we identify and measure the high elements of the unknown...
In the second step, we construct a new distribution \( \tilde{D} \) that adheres to the high mass elements obtained from the first step. Finally, in the third step, we check whether there exists any distribution \( D_1 \) in \( \mathcal{P} \) that is close to \( \tilde{D} \). If such a distribution exists, we accept, and otherwise we reject. In the first step, we need \( O(\Lambda^2) \) many samples to correctly estimate the masses of the high elements, which dominates the sample complexity of our tolerant tester.

It is important to note that the computational efficiency of the tolerant tester depends on how fast we can check whether the distribution \( \tilde{D} \) (constructed by the algorithm) is close to a known property \( \mathcal{P} \), where we have the complete description of \( \tilde{D} \). Later, in Theorem 2 (restated as Theorem 27), we show that when the property \( \mathcal{P} \) can be expressed as a feasible solution to a set of linear inequalities, there exists an algorithm that tolerantly tests for \( \mathcal{P} \) in time polynomial in the support size of the distribution and the number of linear inequalities required to represent it. The algorithm is similar to that of Theorem 1, whereas its polynomial running time follows by using the Ellipsoid method.

**B. Overview of Theorem 3 and Theorem 4**

In Theorem 3, we show that in order to non-tolerantly test any non-concentrated property (defined in Definition 12), \( \Omega(\sqrt{n}) \) samples are required, where \( n \) denotes the support size of the distribution. Before directly proceeding to prove the result, as a warm-up, we first show an analogous result for label-invariant non-concentrated properties in Theorem 29.

To prove the theorem, for any distribution \( D_{yes} \) in the label-invariant non-concentrated property \( \mathcal{P} \) that we are testing, we construct a new distribution \( D_{no} \) that is far from \( \mathcal{P} \), whose support is a subset of the support of \( D_{yes} \). The two distributions are identical over their high probability elements, and they are identical over their low probability elements, and the result follows. We note that the proof of Theorem 29 is a generalization of the lower bound proof for uniformity testing.

Though the proof of Theorem 3 (restated as Theorem 34) follows similarly to that of Theorem 29, delicate analysis is required to take care of the fact that the properties are no longer label-invariant. We also discuss briefly the reason why the technique used to prove Theorem 29 does not work to prove Theorem 3, in the beginning of Section V-B.

As a step further, in Theorem 4 (restated as Theorem 31), we show that \( \Omega(n^{1-\alpha(1)}) \) samples are necessary to tolerately test any non-concentrated label-invariant property. This property follows from an application of Valiant [14, Theorem 5]. The question of tolerant testing of general non-concentrated properties remains open.

**C. Overview of Theorem 5**

Finally, we consider the problem of learning a distribution \( D \), where \( D \) is concentrated over an unknown set \( S \subseteq \Omega \).

In Theorem 5 (restated as Theorem 38), we give an algorithm that achieves this with \( O(|S|) \) many samples, even when \( |S| \) is also unknown. Note that this problem is reminiscent of the folklore result of learning a distribution over any set \( S \) that takes \( O(|S|) \) samples. However, the folklore result holds only for the case where the set \( S \) is known.

Broadly, the algorithm iterates over possible values of \( |S| \). Starting from \( s = 1 \), we first take \( s \) many samples from the unknown distribution \( D \), and construct a new empirical distribution \( D_s \) based upon the samples obtained. Once we have the distribution \( D_s \), we apply the result of Valiant and Valiant [15] to test whether the unknown distribution \( D \) is close to the newly constructed distribution \( D_s \), by using number of samples that is slightly smaller than \( s \). If \( D_s \) is close to \( D \), we report the distribution \( D_s \) as the output and terminate the algorithm. Otherwise, we double the value of \( s \) and perform another iteration of the two steps as mentioned above. Finally, we show that when \( s \geq |S| \), where \( S \) is the unknown set on which \( D \) is concentrated, \( D_s \) will be close to \( D \) with high probability, and we will output a distribution satisfying the statement of Theorem 5. To the best of our knowledge, this is the first result of a tester of this kind that adapts to an unknown support size \( |S| \).

**IV. NON-TOLERANT VS. TOLERANT SAMPLE COMPLEXITIES OF LABEL-INARIANT PROPERTIES (PROOF OF THEOREM 1)**

We will prove that for any label-invariant property, the sample complexities of tolerant and non-tolerant testing are separated by at most a quadratic factor (ignoring some poly-logarithmic factors). Formally, the result is stated as follows:

**Theorem 15 (Theorem 1 Formalized):** Let \( \mathcal{P} \) be a label-invariant distribution property, for which there exists an \((0, \epsilon)\)-tester (non-tolerant tester) with sample complexity \( \Lambda(n, \epsilon) \), where \( \Lambda \in \mathbb{N} \) and \( 0 < \epsilon \leq 2 \). Then for any \( \gamma_1, \gamma_2 \) with \( \gamma_1 < \gamma_2 \) and \( 0 < \gamma_2 + \epsilon \leq 2 \), there exists a \((\gamma_1, \gamma_2 + \epsilon)\)-tester (tolerant tester) that has sample complexity \( O\left(\frac{1}{\sqrt{(\gamma_2 - \gamma_1)^2}} \cdot \min\left\{\Lambda^2 \log^2 \Lambda, n\right\}\right) \), where \( \Lambda = \Lambda(n, \epsilon) \), and \( n \) is the size of the support of the distribution.

Let us assume that \( D \) is the unknown distribution and \( \Lambda(n, \epsilon) \geq \Omega\left(\frac{1}{\sqrt{n}}\right) \).\(^9\) First note that if \( \Lambda = \Omega(\sqrt{n}) \), then we can construct a distribution \( \tilde{D} \) such that \(|D - \tilde{D}| < \frac{\gamma_2 - \gamma_1}{\sqrt{2}}\) \(\epsilon \) samples from \( D \). Hereafter we can report \( D \) to be \( \gamma_1 \)-close to the property if and only if \( \tilde{D} \) is \( \frac{\gamma_2 + \gamma_1 + \epsilon}{\sqrt{2}} \) close to the property. In what follows, we discuss an algorithm with sample complexity \( O\left(\Lambda^2\right) \) when \( \Lambda = o(\sqrt{n}) \). Also, we assume that \( n \) and \( \Lambda \) are larger than some suitable constant. Otherwise, the theorem becomes trivial.

The idea behind the proof is to classify the elements of \( \Omega \) with respect to their masses in \( D \) into \textit{high} and \textit{low}, as formally defined below in Definition 16. We argue that since \( \mathcal{P} \) is \((0, \epsilon)\)-testable using \( \Lambda(n, \epsilon) = O(q) \) samples, there cannot be two distributions \( D_1 \) and \( D_2 \) that are identical on all

---

\(^8\)There are also prior results where only \(|S|\) is known, such as in the work of Acharya et al. [41].

\(^9\)This is a reasonable assumption for any non-trivial property.
elements whose probability mass is at least $\frac{1}{q}$, for $q = \theta(\Lambda)$ (the set \text{High}_{1/q^2}$ defined below), where $D_1 \in \mathcal{P}$ but $D_2$ is $\varepsilon$-far from $\mathcal{P}$. We will formally show this in Lemma 17, where we will use the fact that $\mathcal{P}$ is label-invariant. Using Lemma 17, we prove Lemma 18, that (informally) says that if two distributions are close with respect to the high mass elements, then it is not possible that one distribution is close to $\mathcal{P}$ while the other one is far from it. In our algorithm, we intend to approximate the masses of the set \text{High}_{1/q^2}, and the term $\Lambda^2$ in the query complexity of our algorithm corresponds to that.

\textbf{Definition 16:} For a distribution $D$ over $\Omega$ and $0 < \kappa < 1$, we define

\[ \text{High}_\kappa(D) = \{ x \in \Omega \mid D(x) \geq \kappa \} \]

Now we define a quantity $q \in \mathbb{N}$ where $q = \Theta(\Lambda)$.\footnote{Note that $q$ and $\Lambda$ are of the same order of magnitude. We have introduced $q$ for writing proofs more rigorously.} Assume that $c^*$ is a large constant (independent of $\Lambda$) such that, if we take $\Lambda$ many samples from a distribution, then with probability at least $\frac{3}{4}$, we will not get any sample $x$ whose mass is at most $(\frac{c^*}{\Lambda})^2$ more than once. We define

\[ q := \frac{\Lambda}{c^*} \]  

(1)

We will complete the proof of Theorem 15 by using the following two lemmas which we will prove later.

\textbf{Lemma 17:} Let $\mathcal{P}$ be a label-invariant property that is $(0, \varepsilon)$-testable using $\Lambda(n, \varepsilon)$ samples and consider $q$ as defined in Equation 1. Let $D_1$ and $D_2$ be two distributions such that $\text{High}_{1/q^2}(D_1) = \text{High}_{1/q^2}(D_2)$, and for all $x \in \text{High}_{1/q^2}(D_1)$, the probability of $x$ is the same for both distributions, that is, $D_1(x) = D_2(x)$. Then it is not possible that $D_1$ satisfies $\mathcal{P}$ while $D_2$ is $\varepsilon$-far from satisfying $\mathcal{P}$.

\textbf{Lemma 18:} Let $\mathcal{P}$ be a label-invariant property that is $(0, \varepsilon)$-testable using $\Lambda(n, \varepsilon)$ samples, and consider $q$ as defined in Equation (1). Let $\overline{D}$ and $D$ be two distributions over $\Omega$, where $|\Omega| > 4q^2$, and let $H$ contain the top $q^2$ elements of $\overline{D}$. Also, assume that $D(\Omega \setminus H) - \overline{D}(\Omega \setminus H) \leq \gamma$. If

\[ \sum_{x \in H} |\overline{D}(x) - D(x)| \leq \alpha, \]  

(2)

then the following hold:

1) If $\overline{D}$ is $\beta$-close to $\mathcal{P}$, then there exists a distribution $D_1$ in $\mathcal{P}$ such that $\text{High}_{1/q^2}(D_1) \subseteq H$ and

\[ \sum_{x \in H} |D_1(x) - \overline{D}(x)| + |D_1(\Omega \setminus H) - \overline{D}(\Omega \setminus H)| \leq (\alpha + \beta + \gamma). \]  

(3)

2) If $\overline{D}$ is $(\varepsilon + 3\alpha + \beta + 2\gamma)$-far from $\mathcal{P}$ and $D_1$ is a distribution such that $\text{High}_{1/q^2}(D_1) \subseteq H$ and

\[ \sum_{x \in H} |D_1(x) - \overline{D}(x)| + |D_1(\Omega \setminus H) - \overline{D}(\Omega \setminus H)| \leq (\alpha + \beta + \gamma), \]  

(4)

then the distribution $D_1$ does not satisfy the property $\mathcal{P}$. Using the above two lemmas, we will prove Theorem 15 in Section IV-A. We present the proofs of Lemma 17 and Lemma 18 in Section IV-B.

\textbf{A. Proof of Theorem 15}

Let $D$ be the unknown distribution that we need to test, and assume that $\zeta = \gamma_1$, $\eta = \gamma_2 - \gamma_1$, and $q' = \frac{n}{64}$. We now provide a tolerant $(\gamma_1, \gamma_2 + \varepsilon)$-tester, that is, a $(\zeta, \zeta + \varepsilon)$-tester for the property $\mathcal{P}$, as follows:

1) Draw $W = \mathcal{O}\left(\frac{q^2}{\varepsilon^2} \log q\right)$ samples from the distribution $D$. Let $S \subseteq \Omega$ be the set of (distinct) samples obtained.

2) Draw additional $\mathcal{O}\left(\frac{W}{q^2} \log W\right)$ many samples $Z$ to estimate the value of $D(x)$ for all $x \in S$\footnote{Instead of two sets of random samples (where the first one is to generate the set $S$ and the other one is the multi-set $Z$), one can work with only one set of random samples. But in that case, the sample complexity becomes $\mathcal{O}(q^2 \log n)$, as opposed to $\mathcal{O}(q^2 \log q)$ that we are going to prove.}. Construct a set $H$ as the union of $S$ and arbitrary $q^2$ many elements from $\Omega \setminus (S \cup Z)$.

3) Draw $\mathcal{O}(\log \frac{q^2}{\zeta})$ many samples $x$ in $H$, $D(x) = \# x$ in the multi-set $Z$.

And for each $x \in \Omega \setminus H$,

\[ \tilde{D}(x) = \frac{1 - \sum_{x \in H} \tilde{D}(x)}{|\Omega| - |H|}. \]

(5)

5) If there exists a distribution $D_1$ in $\mathcal{P}$ that satisfies both the following conditions:

(A) $\sum_{x \in H} |D_1(x) - \tilde{D}(x)| + |D_1(\Omega \setminus H) - \tilde{D}(\Omega \setminus H)| \leq 2\eta + \zeta + \Omega(1 + q^2 - \zeta) \subseteq H$.

(B) $\text{High}_{1/q^2}(D_1) \subseteq H$.

then ACCEPT $D$.

6) If there does not exist any $D_1$ in $\mathcal{P}$ that satisfies both Conditions (A) and (B) above, then REJECT $D$.

\textbf{Remark 1:} Note that Step 5 as mentioned above is not completely constructive in a computational sense since we are applying that step while considering for all distributions in the property $\mathcal{P}$. In Section IV-C, we give a constructive variant of the tester where the property $\mathcal{P}$ can be expressed as a set of linear inequalities that runs in polynomial time. We also give an example of a natural property that can be expressed as a set of linear inequalities. In fact, we prove a slightly more general result, in the sense that, when the property $\mathcal{P}$ is a convex property (properties that can be expressed as the set of feasible solutions to a system of convex constraints), we design a tester that performs a single emptiness query to a suitably designed convex set (depending on $\mathcal{P}$). When the constraints are linear, we show that the emptiness query can be performed in polynomial time.

1) \textbf{Sample Complexity:} The sample complexity of the tester is $\mathcal{O}\left(\frac{q^2}{\varepsilon^2} \log^2 q\right) = \mathcal{O}\left(\frac{\Lambda^2 \log^2 \Lambda}{(\gamma_2 - \gamma_1)^2}\right)$, which follows from the above description.
2) Correctness of the Algorithm: The proof of correctness of our algorithm is divided into a sequence of lemmas.

Lemma 19: The set $H$ and the distribution $D$ satisfies the following three properties:

(i) With probability at least $1 - \frac{1}{q}$, $\text{High}_{\eta'/q^2}(D) \subseteq S \subseteq H$.

(ii) For any $x \in H$, if $D(x) \geq \frac{\eta}{1024}$, $(1 - \eta')D(x) \leq \bar{D}(x) \leq (1 + \eta')D(x)$ holds with probability at least $1 - \frac{1}{q}$. [Proof]

(iii) For any $x \in \Omega$ with $D(x) \leq \frac{\eta'}{1024}$, either $x \notin H$, or $\bar{D}(x) \leq (1 + \eta')\frac{\eta}{1024}$ holds with probability at least $1 - \frac{1}{q}$. [Proof]

Proof: Let us prove the three parts one by one:

- (i) Consider any $x \in \text{High}_{\eta'/q^2}(D)$, that is, $D(x) \geq \frac{\eta'}{q^2}$. Then the probability that $x \notin H$ is at most $1 - (1 - \frac{\eta}{q^2})|H| \leq 1 - \frac{1}{q}$. Applying the union bound over all the elements in $\text{High}_{\eta'/q^2}(D)$ (at most $\frac{q^2}{\eta} = O(q^3)$ many elements), the claim follows.

- (ii) Recall that $|Z| = O\left(\frac{n}{\eta^2} \log W\right)$. Applying Chernoff bound, we obtain that $(1 - \eta')D(x) \leq \bar{D}(x) \leq (1 + \eta')D(x)$ does not hold with probability at most $\frac{1}{q}$.

- (iii) Since $|Z| = O\left(\frac{n}{\eta^2} \log W\right)$, if $x$ is in $H$ (otherwise, we are already done), applying Chernoff bound (only on one side), the bound follows. [Proof]

We now bound the $\ell_1$-distance between $D$ and $\bar{D}$ with respect to $H$.

Lemma 20: $\sum_{x \in H} |D(x) - \bar{D}(x)| \leq 5\eta'(1 + \eta') \leq 10\eta'$ holds with probability at least $1 - \frac{3}{q}$.

Proof: Recall the definition of $\text{High}_{\eta'/10W}(D)$. Note that

$$\sum_{x \in H} |D(x) - \bar{D}(x)| = \sum_{x \in \text{High}_{\eta'/10W}(D)} |D(x) - \bar{D}(x)| + \sum_{x \in H \setminus \text{High}_{\eta'/10W}(D)} |D(x) - \bar{D}(x)|$$

Applying Lemma 19 (ii) for each $x \in \text{High}_{\eta'/10W}(D)$, and then using union bound over all such $x \in \text{High}_{\eta'/10W}(D)$, the first term is bounded by $\eta'$ with probability at least $1 - \frac{1}{q}$.

Now the second term, notice that for each $x \in H \setminus \text{High}_{\eta'/10W}(D)$, $D(x) \leq \frac{\eta}{1024}$. By Lemma 19 (iii), and using the union bound over all elements in $H \setminus \text{High}_{\eta'/10W}(D)$ (note that $|H| \leq 2W = O(q^3)$), with probability at least $1 - \frac{1}{q}$, $\bar{D}(x) \leq \eta'(1 + \eta')/10W$ for all $x \in H \setminus \text{High}_{\eta'/10W}(D)$. Since $|H| \leq 2W$, the second term is bounded by $4\eta'(1 + \eta')$ with probability at least $1 - \frac{1}{q}$. [Proof]

Now we prove a lemma that shows that for every distribution $D$, there is another distribution $\overline{D}$ that is “similar” to $D$, and for which $H$ contains the top $q^2$ elements of $\overline{D}$.

Lemma 21: There exists a distribution $\overline{D}$ such that $H$ contains the top $q^2$ elements of $\overline{D}$. Moreover, the following hold:

(i) $|D - \overline{D}|_1 \leq 2\eta'$, with probability at least $1 - \frac{1}{q}$.
exists a distribution $D$ that is $(\epsilon + \zeta + \eta - 2\eta')$-far from $P$, that is, $(\epsilon + 3\alpha + \beta + 2\gamma)$-far from $P$, where $\alpha = 12\eta'$, $\beta = \zeta + 2\eta'$. Here, we are using that $\eta = 64\eta'$ and $\gamma = 12\eta$. Also Lemma 21 guarantees that $H$ contains the top $q^2$ elements of $D$. Following Lemma 18, we know that there does not exist any such distribution $D_1$ in $P$ that satisfies both Condition (A) and Condition (B) of Step 5 of the algorithm. Thus the algorithm will REJECT the distribution $D$ in Step 6.

Note that the total failure probability of the algorithm is bounded by the probability that Lemma 21 does not hold, which is at most $\frac{12}{q}$.

**B. Proof of Lemma 17 and Lemma 18**

**Proof:** [Proof of Lemma 17] We will prove this by contradiction. Let us assume that there are two distributions $D_{yes}$ and $D_{no}$ such that

- $D_{yes} \in P$;
- $D_{no}$ is $\epsilon$-far from $P$;
- High$_1/\sqrt{q}(D_{yes}) = High_1/\sqrt{q}(D_{no}) = A$;
- For all $x \in A$, $D_{yes}(x) = D_{no}(x)$.

Now, we argue that any $(0, \epsilon)$-non-tolerant tester requires more than $\Lambda(n, \epsilon)$ samples from the unknown distribution $D$ to distinguish whether $D$ is in the property or $\epsilon$-far from it.

Let $D_Y$ be a distribution obtained from $D_{yes}$ by permuting the labels of $\Omega \setminus A$ using a uniformly random permutation. Specifically, consider a random permutation $\pi : \Omega \setminus A \to \Omega \setminus A$. The distribution $D_Y$ is as follows:

- $D_Y(x) = D_{yes}(x)$ for each $x \in A$ and
- $D_Y(\pi(x)) = D_{yes}(x)$ for each $x \in \Omega \setminus A$.

Similarly, consider the distribution $D_N$ obtained from $D_{no}$ by permuting the labels of $\Omega \setminus A$ using a uniformly random permutation. Note that $D_Y$ is in $P$, whereas $D_N$ is $\epsilon$-far from $P$, which follows from $P$ being label-invariant.

We will now prove that $D_Y$ and $D_N$ provide similar distributions over sample sequences. More formally, we will prove that any algorithm that takes at most $\Lambda(n, \epsilon)$ many samples, cannot distinguish $D_Y$ from $D_N$ with probability at least $\frac{1}{2}$. We argue that this claim holds even if the algorithm is provided with additional information about the input: Namely, for all $x \in A$, the algorithm is told the value of $D_Y(x)$ (which is the same as $D_N(x)$). When the algorithm is provided with this information, it can ignore all samples obtained from $A$.

By the definition of $A$, for all $x \in \Omega \setminus A$, both $D_Y(x)$ and $D_N(x)$ are at most $\frac{1}{q}$-close. Let $S_Y$ be a sequence of samples drawn according to $D_Y$. If $|S_Y| \leq A(n, \epsilon)$, then with probability at least $\frac{3}{4}$, the sequence $(\Omega \setminus A) \cap S_Y$ has no element that appears twice. In other words, the set $(\Omega \setminus A) \cap S_Y$ is a set of at most $A(n, \epsilon)$ distinct elements from $\Omega \setminus A$. Since the elements of $\Omega \setminus A$ were permuted using a uniformly random permutation, with probability at least $\frac{3}{4}$, the sequence $(\Omega \setminus A) \cap S_Y$ is a uniformly random sequence of distinct elements from $\Omega \setminus A$. Similarly, if $S_N$ is a sequence of samples drawn according to $D_N$, then with probability at least $\frac{3}{4}$, the sequence $(\Omega \setminus A) \cap S_N$ is a uniformly random sequence of distinct elements from $\Omega \setminus A$. Thus, the distributions over the received sample sequence obtained from $D_Y$ or $D_N$ are of distance $\frac{1}{q}$ of each other, which is strictly less than $\frac{1}{3}$.

Hence, if the algorithm obtains at most $\Lambda(n, \epsilon)$ many samples from the unknown distribution $D$, it cannot distinguish, with probability at least $\frac{3}{4}$, whether the samples are coming from $D_Y$ or $D_N$. □

For the proof of Lemma 18, we will need the following simple claim.

**Claim 22:** Let $\sigma : [n] \to [n]$ be a permutation and let $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ be two sets of $n$ positive real numbers. If $a_1 \geq a_2 \geq \cdots \geq a_n$ and $b_1 \geq b_2 \geq \cdots \geq b_n$, then the sum $\sum_{i \in [n]} |a_i - b_{\sigma(i)}|$ is minimized when $\sigma$ is the identity permutation.

**Proof:** First observe that if $a, b, c, d$ are four real numbers with $a \geq b$ and $c \geq d$, then the following holds:

$$|a - c| + |b - d| \leq |a - d| + |b - c|. \quad (5)$$

The above can be proved by checking all possible orderings of the numbers $a, b, c, d$.

Once we have the above observation, we can now proceed to prove the claim. Let us consider the set of permutations that minimize $\sum_{i \in [n]} |a_i - b_{\sigma(i)}|$. Let $\sigma$ be one such minimizing permutation that also minimizes the size for the following set $S$:

$$S = \{(i, j) : i < j \text{ and } \sigma(i) > \sigma(j)\}.$$  

Let $i$ be an index such that $\sigma(i) < \sigma(i+1)$ (such an index $i$ exists unless $\sigma$ is the identity permutation). Let $\sigma'$ be the permutation obtained from $\sigma$ by swapping $\sigma(i)$ and $\sigma(i+1)$. Then the sum $\sum_{i \in [n]} |a_i - b_{\sigma'(i)}|$ does not increase from $\sum_{i \in [n]} |a_i - b_{\sigma(i)}|$, because of Equation 5. However, the size of the set $S$ with respect to the permutation $\sigma'$ strictly decreases, and we have a contradiction. □

Now we present the proof of Lemma 18. **Proof:** [Proof of Lemma 18] We consider the two cases separately.

1. If $D$ is $\beta$-close to $P$, there exists a distribution $D_1$ in $P$ such that $\sum_x |D(x) - D_1(x)| \leq \beta$. Since $P$ is label-invariant, any permutation of $D_1$ is also in $P$. Without loss of generality, let us assume that the domain $\Omega$ is a subset of $\{1, \ldots, n\}$.

By Claim 22, the permutation $\sigma$ that minimizes $\sum_x |D(x) - D_1(\sigma(x))|$ is the one that orders the $i$-th largest element of $D_1$ with the $i$-th largest element of $D$, that is, if $x$ is the element with the $i$-th largest probability mass in $D_1$, then $\sigma(x)$ has the $i$-th largest probability mass in $D$. Consider the distribution $D_1^\sigma(x)$ that is defined by $D_1^\sigma(x) = D_1(\sigma(x))$. Clearly, $H$ contains the largest $q^2$ elements of $D_1^\sigma$, and hence also $High_1/\sqrt{q}(D_1^\sigma) \subseteq H$.

As $\sum_{x \in H} |D_1^\sigma(x) - D(x)| \leq \beta$, we have

$$\sum_{x \in H} |D_1^\sigma(x) - D(x)| + |D_1^\sigma(\Omega \setminus H) - D(\Omega \setminus H)| \leq |D_1^\sigma(\Omega \setminus H) - D(\Omega \setminus H)|,$$

by the triangle inequality, we obtain

$$\sum_{x \in H} |D_1^\sigma(x) - D(x)| + \sum_{x \in H} |D_1^\sigma(\Omega \setminus H) - D(\Omega \setminus H)| \leq \sum_{x \in H} |D_1^\sigma(x) - D(x)| + |D_1^\sigma(\Omega \setminus H) - D(\Omega \setminus H)|.$$
Observe that case also be a distribution such that $P_{\text{High}} \subseteq H$ and $\sum_{x \in H} |D_1(x) - \bar{D}(x)| + |D_1(\Omega \setminus H) - \bar{D}(\Omega \setminus H)| \leq \alpha + \beta + \gamma.

Then, as $\sum_{x \in H} |\bar{D}(x) - \bar{D}(x)| \leq \alpha$, by the triangle inequality, we have

$$\sum_{x \in H} |D_1(x) - \bar{D}(x)| + |D_1(\Omega \setminus H) - \bar{D}(\Omega \setminus H)| \leq 2\alpha + \beta + \gamma. \quad (6)$$

Consider the distribution $\hat{D}$ defined as follows:

- For all $x \in H$, $\hat{D}(x) = D_1(x)$.
- If $D_1(H) \geq \bar{D}(H)$, then for all $x \in \Omega \setminus H$,
  $$\hat{D}(x) = \bar{D}(x) \cdot \phi,$$
  where $\phi = \frac{1 - D_1(H)}{\bar{D}(H)}$. Notice that in this case $\phi \leq 1$.
- If $D_1(H) \leq \bar{D}(H)$, then pick the set $T \subset \Omega \setminus H$ with $|T| = 2q^2$ that minimizes $\bar{D}(T)$. Then for all $x \in T$,
  $$\hat{D}(x) = \bar{D}(x) + \frac{\bar{D}(H) - D_1(H)}{2q^2} \quad \text{and for all } x \in \Omega \setminus (T \cup H), \hat{D}(x) = \bar{D}(x)$$

Let us first prove that $\text{High}_{1/q^2}(\hat{D}) \subseteq H$. In the case where $D_1(H) \geq \bar{D}(H)$, for all $x \in \Omega \setminus H$, $\hat{D}(x) \leq \bar{D}(x)$. Since $\text{High}_{1/q^2}(\bar{D}) \subseteq H$, $\text{High}_{1/q^2}(\hat{D}) \subseteq H$. Now, in the case where $D_1(H) \leq \bar{D}(H)$, the only $x \in \Omega \setminus H$ for which $\hat{D}(x) > \bar{D}(x)$ are those in $T$. Since $|\Omega| > 4q^2$, the lowest $2q^2$ elements on $\bar{D}$ must each have mass less than $\frac{1}{4q^2}$. So even if we add $\frac{1}{2q^2}$ for any element $x \in T$, $\hat{D}(x) < 1/q^2$. Hence in this case also $\text{High}_{1/q^2}(\hat{D}) \subseteq H$ since $\text{High}_{1/q^2}(\bar{D}) \subseteq H$ and $\text{High}_{1/q^2}(D_1) \subseteq H$.

Now let us bound the $\ell_1$ distance between $\hat{D}$ and $\bar{D}$. Observe that

$$\sum_{x \in \Omega \setminus H} |\hat{D}(x) - \bar{D}(x)| = |\hat{D}(\Omega \setminus H) - \bar{D}(\Omega \setminus H)|.$$

This is because, in the case where $\hat{D}(H) \geq \bar{D}(H)$, we have $\hat{D}(x) = \phi \cdot \bar{D}(x) \leq \bar{D}(x)$ for all $x \in \Omega \setminus H$. On the other hand, in the case where $\hat{D}(H) \leq \bar{D}(H)$ then for all $x \in \Omega \setminus H$, $\hat{D}(x) \geq \bar{D}(x)$. Thus,

$$\sum_{x \in \Omega \setminus H} |\hat{D}(x) - \bar{D}(x)| = |\hat{D}(\Omega \setminus H) - \bar{D}(\Omega \setminus H)| \leq |\hat{D}(\Omega \setminus H) - \bar{D}(\Omega \setminus H)| + \gamma$$

Also note that, from the construction of $\hat{D}$, we have for all $x \in H$, $\hat{D}(x) = D_1(x)$ and thus $D(\Omega \setminus H) = D_1(\Omega \setminus H)$. Thus,

$$||\hat{D} - \bar{D}||_1 = \sum_{x \in H} |\hat{D}(x) - \bar{D}(x)| + \sum_{x \in \Omega \setminus H} |\hat{D}(x) - \bar{D}(x)| \leq \sum_{x \in H} |\hat{D}(x) - \bar{D}(x)| + |\hat{D}(\Omega \setminus H) - \bar{D}(\Omega \setminus H)| + \gamma = \sum_{x \in H} |D_1(x) - \bar{D}(x)| + |\hat{D}(\Omega \setminus H) - \bar{D}(\Omega \setminus H)| + \gamma \quad \text{(From the construction of $\hat{D}$)}$$

And so $\hat{D}$ is $\varepsilon + 2\alpha + \beta + 2\gamma$-close to $\bar{D}$, which is a contradiction.

C. Computationally Efficient Tolerant Testers

In this section we present a constructive variant of the tolerant tester studied in Section IV-A. Here, for any two vectors $a, b \in \mathbb{R}^N$, we say that $a \leq b$ if $a_i \leq b_i$ holds for every $i \in [N]$. Now let us recall the definitions of polyhedron and projection map.

**Definition 23 (Polyhedron):** Let $A$ be an $M \times N$ real matrix, $b \in \mathbb{R}^M$ be a real vector and $Ax \leq b$ be a system of linear inequalities. The solution set $\{x \in \mathbb{R}^N \mid Ax \leq b\}$ of the system of inequalities is called a polyhedron. The complexity of a polyhedron is defined as $MN$.

**Definition 24 (Projection Map):** Let $n$ be an integer. For all integers $N \geq n$, a projection map is denoted as $\pi_n : \mathbb{R}^N \rightarrow \mathbb{R}^n$ and is defined as the projection of the points in $\mathbb{R}^N$ on the first $n$ coordinates.

Before directly proceeding to our results, we first define two variants of distribution properties.

**Definition 25 (Linear Property):** Without loss of generality, let us assume that $\Omega = [n]$. A distribution property $\mathcal{P}$ is said to be a linear property if there exists a polyhedron $\mathcal{LP} = \{x \in \mathbb{R}^N \mid Ax \leq b\}$, where $A$ is a $M \times N$ real...
matrix and $b \in \mathbb{R}^M$ be a real vector, and $\pi_n(\mathcal{LP})$ is the set of distributions satisfying the property $\mathcal{P}$, that is, for every $z := (z_1, \ldots, z_n) \in \mathcal{LP}$, the distribution $D_z$, defined as

$$D_z(i) = z_i, \ \forall i \in [n]$$

satisfies the property $\mathcal{P}$, and conversely, for every distribution $D$ that satisfies $\mathcal{P}$, there exists some $z \in \mathcal{LP}$ such that $D = D_z$ as defined above. The complexity of $\mathcal{P}$ is defined as $M \times N$.

Similar to linear properties, we can also define properties that are feasible solutions to a system of convex constraints.

Definition 26 (Convex Property): A distribution property $\mathcal{P}$ is said to be a convex property if $\mathcal{P}$ is the set of all feasible solutions to a system of convex constraints over $D(i)$ for $i \in \Omega$, where $\Omega$ is the sample space of $D$. In other words, the set $\mathcal{P}$ forms a convex set.

Now we show that some well studied label-invariant distribution properties can be represented as linear or convex properties.

Remark 2 (An Example of a Linear Property: Approximate Uniformity Property): A distribution $D$ over $[n]$ is said to be uniform if $D(i) = \frac{1}{n}$ for all $i \in [n]$. Let the property $\mathcal{P}_{u, \varepsilon}$ denote the set of all distributions that are $\varepsilon$-close to uniform, where $\varepsilon \in (0, 1)$ is a parameter. Consider the following polyhedron $\mathcal{LP}_{u, \varepsilon}$ in $\mathbb{R}^{2n}$:

$$\sum_{i \in [n]} z_{n+i} \leq \varepsilon$$

$$z_i \geq 0 \ \forall i \in [2n]$$

$$-z_{n+i} \leq z_i - 1 \leq z_{n+i} \ \forall i \in [n]$$

Now, observe that $\pi_n(\mathcal{LP}_{u, \varepsilon})$ will give us the set of distributions that are $\varepsilon$-close to uniform, i.e., the set $\mathcal{P}_{u, \varepsilon}$ (this would serve as the linear transformation mentioned in Definition 25). Also, note that approximate uniformity property has complexity $O(n)$.

Now we present a property that can be expressed as a feasible solution to a system of convex inequalities, but that cannot be expressed as feasible solution to a system of linear inequalities.

Remark 3 (An example of a convex property: Entropy property): Let $D$ be a distribution supported on $[n]$. Given a parameter $k \in \mathbb{R}$, let $\mathcal{P}_{E, k}$ denote the set of all distributions with entropy at least $k$. $\mathcal{P}_{E, k}$ can be expressed as a convex inequality as follows:

$$\sum_{i \in [n]} D(i) \log \frac{1}{D(i)} \geq k.$$  

For a distribution property $\mathcal{P}$, let $CP \subset \mathbb{R}^n$ denote the geometric representation of the set of probability distributions over the set $[n]$ that satisfy $\mathcal{P}$ by considering each distribution over $[n]$ as a point in $\mathbb{R}^n$.

For all $\beta \in [0, 1]$, $k \leq n$ and $a \in \mathbb{R}^n$, we define the following convex set:

$$\Delta(k, q, a, \beta) := \{x \in \mathbb{R}^d : \left\lVert\sum_{i=1}^k |x_i - a_i| + \sum_{j=k}^n (x_j - a_j)\right\rVert_1 \leq \beta,$$

$$\forall i > k, x_i < \frac{1}{q^2}\}$$

Before proceeding to prove Theorem 2, we will first prove a more general result. We will show that if we have access to an emptiness oracle that takes a specific convex set (defined in the following statement) as input, and decides whether the convex set is empty or not, then we can design a tolerant tester for any label-invariant distribution property that takes $O(\Lambda^2)$ samples and performs a single emptiness query to the oracle. The result is formally stated and proved below.

Theorem 27: Let $\mathcal{P}$ be a label-invariant distribution property. If there is a $(0, \varepsilon)$-tester (non-tolerant tester) with sample complexity $\Lambda(n, \varepsilon)$, then for any $\gamma_1, \gamma_2$ with $\gamma_1 < \gamma_2$ and $0 < \gamma_1 < \gamma_2 + \varepsilon \leq 2$, there exists a $(\gamma_1, \gamma_2 + \varepsilon)$-tester (tolerant tester) that takes $s = O(\Lambda^2)$ many samples and makes a single emptiness query to the set $CP \cap \Delta(s, D, \beta)$, where $\beta$ is a known probability distribution and $\beta = \gamma_1 + \frac{\lambda - \gamma_2}{3}$.

Proof: Recall that in Step 5 of the tolerant tester presented in Section IV-A, the tester checks whether there is any distribution $D_1 \in \mathcal{P}$ that satisfies the following two conditions:

$$\sum_{x \in H} |D_1(x) - \bar{D}(x)| + |D_1(\Omega \setminus H) - \bar{D}(\Omega \setminus H)| \leq 2\eta'' + \zeta$$

and

$$High_{1/\eta'}(D_1) \subseteq H$$

where $\zeta = \gamma_1$, $\eta = \gamma_2 - \gamma_1$, $\eta' = \gamma_2 - \gamma_1$ and $\eta'' = \frac{n}{66}$. The set $H$ and the distribution $\bar{D}$ are defined in the tolerant tester presented in Section IV-A.

Without loss of generality, we can assume that $H = \{1, \ldots, |H|\}$. Therefore, in order to perform Step 5 of the tolerant tester, the following equations are needed to be satisfied:

$$D_1 \in CP$$

$$D_1 \in \Delta(|H|, \eta, \bar{D}, 26\eta'' + \zeta)$$

We now present the tolerant $(\gamma_1, \gamma_2 + \varepsilon)$-tester in its entirety, that is, a $(\zeta, \zeta + \varepsilon + \eta)$-tester for the property $\mathcal{P}$, where $\zeta = \gamma_1$, $\eta = \gamma_2 - \gamma_1$, and $\eta'' = \frac{n}{66}$.

1) Draw $W = \mathcal{O}\left(\frac{2n \log q}{\eta''}\right)$ many samples from the distribution $D$. Let $S \subseteq \Omega$ be the set of (distinct) samples obtained.

2) Draw additional $\mathcal{O}\left(\frac{W}{\eta'} \log W\right)$ many samples $Z$ to estimate the value of $D(x)$ for all $x \in S$.

3) Construct a set $H$ as the union of $S$ and $q^2$ many arbitrary elements from $\Omega \setminus (S \cup Z)$. 

---

13Note that $\pi_n(\mathcal{LP})$ will also be a polyhedron in $\mathbb{R}^n$, see, e.g., Corollary 2.5 in Chapter 2 from the book by Bertsimas and Tsitsiklis [42]. However, the number of linear inequalities defining the property, which affects the running time of the tester, can sometimes be greatly reduced by using a projection.
4) Define a distribution \( \tilde{D} \) such that, for \( x \in H \),
\[
\tilde{D}(x) = \frac{\# x \in Z}{|Z|}.
\]
And for each \( x \in \Omega \setminus H \),
\[
\tilde{D}(x) = \frac{1 - \sum_{z \in H} \tilde{D}(x)}{|\Omega| - |H|}.
\]
5) If there exists a distribution \( D_1 \in CP \cap \Delta \left( |H|, q, \tilde{D}, 26\eta' + \zeta \right) \), then ACCEPT \( D \).
6) If there does not exist any distribution \( D_1 \) that passes Step 5, then REJECT \( D \).

Observe that the sample complexity of the tester is \( O \left( \frac{q}{\eta'} \log^2 q \right) = O(\Lambda^2) \) in addition to a single emptiness query to the set \( P \in CP \cap \Delta \left( |H|, q, \tilde{D}, 26\eta' + \zeta \right) \) in Step 5. The correctness proof of the above tester follows from the correctness argument presented in Section IV-A. \( \square \)

**Emptiness checking when \( P \) is a linear property: Proof of Theorem 2:**

Now we proceed to analyze the time complexity of the \((\gamma_1, \gamma_2 + \varepsilon)\)-tester described in Theorem 27 when \( P \) is also a linear property. Recall that as \( P \) is a linear property, there exists a polyhedron \( LP = \{ x \in \mathbb{R}^N \mid Ax \leq b \} \), where \( A \) is a \( M \times N \) real matrix and \( b \in \mathbb{R}^M \) be a real vector, and \( \pi_n(LP) \) is the set of distributions satisfying the property \( P \). (See Definition 25)

Now in Observation 28, we show that checking emptiness of \( \pi_n(LP) \cap \Delta \left( |H|, q, \tilde{D}, 26\eta' + \zeta \right) \) is equivalent to testing the feasibility of a family of inequalities.

**Observation 28:** Without loss of generality, assume that \( H = \{ 1, \ldots, |H| \} \) and \( \Omega = \{ 1, \ldots, n \} \). Checking emptiness of \( \pi_n(LP) \cap \Delta \left( |H|, q, \tilde{D}, 26\eta' + \zeta \right) \) is equivalent to testing the feasibility of the following set of inequalities:

**Equation (9):**
\[
Az \leq b,
\]
\[
\sum_{i=1}^{|H|} z_{N+i} + z_{N+|H|+1} \leq 26\eta' + \zeta
\]
\[
z_{N+i} \geq 0 \quad \forall i \in [|H| + 1]
\]
\[
-z_{N+i} \leq z_i - \tilde{D}(i) \leq z_{N+i} \quad \forall i \in [|H|]
\]
\[
-z_{N+|H|+1} \leq \sum_{i=|H|+1}^n z_i - \sum_{i=|H|+1}^n \tilde{D}(i) \leq z_{N+|H|+1}
\]

Note that the inequality in Equation (10) can be expressed as the following set of linear inequalities using slack variables \( z_{N+i} \) for all \( i \in [|H| + 1] \):
\[
\sum_{i=1}^{|H|} z_{N+i} + z_{N+|H|+1} \leq 26\eta' + \zeta
\]
\[
z_{N+i} \geq 0 \quad \forall i \in [|H| + 1]
\]
\[
-z_{N+i} \leq z_i - \tilde{D}(i) \leq z_{N+i} \quad \forall i \in [|H|]
\]
\[
-z_{N+|H|+1} \leq \sum_{i=|H|+1}^n z_i - \sum_{i=|H|+1}^n \tilde{D}(i) \leq z_{N+|H|+1}
\]

Therefore checking the emptiness of \( \pi_n(LP) \cap \Delta \left( |H|, q, \tilde{D}, 26\eta' + \zeta \right) \) is equivalent to checking the feasibility of the following set of linear inequalities:

**Equation (10):**
\[
Az \leq b,
\]
\[
\sum_{i=1}^{|H|} z_{N+i} + z_{N+|H|+1} \leq 26\eta' + \zeta
\]
\[
z_{N+i} \geq 0 \quad \forall i \in [|H| + 1]
\]
\[
-z_{N+i} \leq z_i - \tilde{D}(i) \leq z_{N+i} \quad \forall i \in [|H|]
\]
\[
-z_{N+|H|+1} \leq \sum_{i=|H|+1}^n z_i - \sum_{i=|H|+1}^n \tilde{D}(i) \leq z_{N+|H|+1}
\]

The feasibility of the above set of linear inequalities can be solved in a polynomial time in the complexity of the polyhedron, that is, in a polynomial time in \( N \) and \( M \), using the Ellipsoid Method, where recall that \( A \) is a \( M \times N \) real matrix (see, e.g., [42], [43]). Thus, we have an efficient \((\gamma_1, \gamma_2 + \varepsilon)\)-tester for \( P \), that runs in time polynomial in the complexity of the label-invariant linear property \( P \). This concludes the proof of Theorem 2.

**V. SAMPLE COMPLEXITY OF TESTING NON–CONCENTRATED LABEL–INVARIANT PROPERTIES**

In this section we prove a lower bound of \( \Omega(\sqrt{n}) \) on the sample complexity of non-tolerant testing of any non-concentrated property. As a warm-up for our proof, we first show similar lower bound for non-tolerant testing of any non-concentrated label-invariant property. Although the construction of the hard distributions for the proof of non-concentrated label-invariant properties is the most natural one, it turns out that the same construction does not work for proving similar lower bound of general non-concentrated distribution properties. Along the way, we prove a near-linear lower bound on the sample complexity for tolerant testing of non-concentrated label-invariant properties.

**A. Sample Complexity of Testing Non–Concentrated Label–Invariant Properties**

Here we first prove a lower bound result analogous to Theorem 3, where the properties are non-concentrated and label-invariant. Then we proceed to prove a tolerant lower bound of \( \Omega(\sqrt{n}) \) samples for such properties.

**Theorem 29 (Analogous Result of Theorem 3 for non–concentrated label–invariant properties):** Let \( P \) be any \((\alpha, \beta)\)-non-concentrated label-invariant distribution property, where \( 0 < \alpha < \beta < \frac{1}{2} \). For \( \varepsilon \) with \( 0 < \varepsilon < \alpha \), any \((0, \varepsilon)\)-tester for property \( P \) requires \( \Omega(\sqrt{n}) \) many samples, where \( n \) is the size of the support of the distribution.

**Proof:** Let us first consider a distribution \( D_{res} \) that satisfies the property. Since \( P \) is an \((\alpha, \beta)\)-non-concentrated property, by Definition 12, \( D_{res} \) is an \((\alpha, \beta)\)-non-concentrated distribution. From \( D_{res} \), we generate a distribution \( D_{no} \) such that the support of \( D_{no} \) is a subset of that of \( D_{res} \), and \( D_{no} \) is \( \varepsilon \)-far from \( P \). Hence, if we apply a random permutation over the elements of \( \Omega \), we show that \( D_{res} \) and \( D_{no} \) are
indistinguishable, unless we query for $\Omega(\sqrt{n})$ many samples. Below we formally prove this idea.

We will partition the domain $\Omega$ into two parts, depending on the probability mass of $D_{\text{yes}}$ on the elements of $\Omega$. Given the distribution $D_{\text{yes}}$, let us first order the elements of $\Omega$ according to their probability masses. In this ordering, let $L$ be the smallest $2\beta n$ elements of $\Omega$. We denote $\Omega \setminus L$ by $H$. Before proceeding further, note that the following observation gives an upper bound on the probabilities of the elements in $L$.

**Observation 30:** For all $x \in L$, $D_{\text{yes}}(x) \leq \frac{1 - 2\alpha}{2\beta n} \frac{1}{n}$.

**Proof:** [Proof of Observation 30] By contradiction, assume that there exists $x \in L$ such that $D_{\text{yes}}(x) > \frac{1 - 2\alpha}{2\beta n} \frac{1}{n}$. This implies, for every $y \in H$, that $D_{\text{yes}}(y) > \frac{1 - 2\alpha}{2\beta n} \frac{1}{n}$. So,

$$1 = \sum_{x \in \Omega} D_{\text{yes}}(x)$$

$$= \sum_{x \in L} D_{\text{yes}}(x) + \sum_{y \in H} D_{\text{yes}}(y)$$

$$> D_{\text{yes}}(L) + |H| \left( \frac{1 - 2\alpha}{2\beta n} \frac{1}{n} \right)$$

As $|L| = 2\beta n$ and $D_{\text{yes}}$ is an $(\alpha, \beta)$-non-concentrated distribution, $D_{\text{yes}}(L) \geq 2\alpha$. Also, $|H| = (1 - 2\beta)n$. Plugging these into the above inequality, we get a contradiction. □

Note that Observation 30 implies that if $S$ is a multi-set of $o\left(\frac{1 - 2\alpha}{2\beta n}\right)$ samples from $D_{\text{yes}}$, then with probability $1 - o(1)$, no element in $S$ appears in $L$ more than once.

Now using the distribution $D_{\text{yes}}$ and the set $L$, let us define a distribution $D_{\text{no}}$ such that $D_{\text{no}}$ is $\varepsilon$-far from $P$. Note that $D_{\text{no}}$ is a distribution that comes from a distribution over a set of distributions, all of which are not $(\alpha, \beta)$-non-concentrated. The distribution $D_{\text{no}}$ is generated using the following random process:

- We partition $L$ randomly into two equal sets of size $\beta n$. Let the sets be $\{x_1, \ldots, x_{\beta n}\}$ and $\{y_1, \ldots, y_{\beta n}\}$.

- We pair the elements of $L$ randomly into $\beta n$ pairs. Let $(x_1, y_1), \ldots, (x_{\beta n}, y_{\beta n})$ be a random pairing of the elements in $L$, which is represented as $P_L$, that is, $P_L = \{(x_1, y_1), \ldots, (x_{\beta n}, y_{\beta n})\}$.

- The probability mass of $D_{\text{no}}$ at $z$ is defined as follows:
  - If $z \notin L$, then $D_{\text{no}}(z) = D_{\text{yes}}(z)$.
  - For every pair $(x_i, y_i) \in P_L$, $D_{\text{no}}(x_i) = D_{\text{yes}}(x_i) + D_{\text{yes}}(y_i)$, and $D_{\text{no}}(y_i) = 0$.

We start by observing that the distribution $D_{\text{no}}$ constructed above is supported on a set of at most $(1 - \beta)n$ elements. So, any distribution $D_{\text{no}}$ constructed using the above procedure is $\varepsilon$-far from satisfying the property $P$ for any $\varepsilon < \alpha$.

We will now prove that $D_{\text{yes}}$ and $D_{\text{no}}$ both have similar distributions over the sequences of samples. More formally, we will prove that any algorithm that takes $O(\sqrt{n})$ many samples, cannot distinguish between $D_{\text{yes}}$ from $D_{\text{no}}$ with probability at least $\frac{2}{3}$.

Since any $D_{\text{no}}$ produced using the above procedure has exactly the same probability mass on the elements in $H$ as $D_{\text{yes}}$, any tester that distinguishes between $D_{\text{yes}}$ and $D_{\text{no}}$ must rely on samples obtained from $L$. Recall that the algorithm is given a uniformly random permutation of the distribution. Since $\text{SUPP}(D_{\text{no}}) \subset \text{SUPP}(D_{\text{yes}})$ (particularly, $\text{SUPP}(D_{\text{no}}) \cap L \subset \text{SUPP}(D_{\text{yes}}) \cap L$), it is not possible to distinguish between $D_{\text{yes}}$ and $D_{\text{no}}$, unless an element of $L$ appears at least twice. Otherwise, as in the proof of Lemma 17, the elements drawn from $L$ are distributed identically to a uniformly random non-repeating sequence. But observe that $D_{\text{yes}}(i) = \mathcal{O}\left(\frac{1}{n}\right)$ and $D_{\text{no}}(i) = \mathcal{O}\left(\frac{1}{n}\right)$ when $i$ is in $L$. Thus any sequence of $O(\sqrt{n})$ samples will provide only a distance of $o(1)$ between the two distributions, completing the proof. □

Now, we prove our lower bound on the sample complexity of tolerant testers for non-concentrated label invariant properties, that is, Theorem 4.

**Theorem 31** (Theorem 4 formalized): Let $P$ be any $(\alpha, \beta)$-non-concentrated label-invariant distribution property, where $0 < \alpha \leq \beta < \frac{1}{2}$. For any constant $\varepsilon_1$ and $\varepsilon_2$ with $0 < \varepsilon_1 < \varepsilon_2 < \alpha$, any $(\varepsilon_1, \varepsilon_2)$-tester for $P$ requires $\Omega(n^{1-\alpha(1)})$ samples, where $n$ is the size of the support of the distribution.

To prove the above theorem, we recall some notions and a theorem from Valiant’s paper on a lower bound for the sample complexity of tolerant testing of symmetric properties [14].

These definitions refer to invariants of distributions, which are essentially a generalization of properties.

**Definition 32:** Let $\Pi : D_n \rightarrow \mathcal{R}$ denote a real-valued function over the set $D_n$ of all distributions over $[n]$.

1) $\Pi$ is said to be label-invariant if for any $D \in D_n$, the following holds: $\Pi(D) = \Pi(D_{\alpha})$ for any permutation $\sigma : [n] \rightarrow [n]$.

2) For any $\gamma, \delta$ with $\gamma \geq 0$ and $\delta \in [0, 2]$, $\Pi$ is said to be $(\gamma, \delta)$-weakly-continuous if for all distributions $p^+, p^-$ satisfying $||p^+ - p^-||_1 \leq \delta$, we have $||\Pi(p^+) - \Pi(p^-)|| \leq \gamma$.

For a property $\mathcal{P}$ of distributions, we define $\Pi_{\mathcal{P}} : D_n \rightarrow [0, 2]$ with respect to property $\mathcal{P}$ as follows:

$$\text{For } D \in D_n, \Pi_{\mathcal{P}}(D) := \text{the distance of } D \text{ from } \mathcal{P}.\text{ From the triangle inequality property of } \ell_1 \text{ distances, } \Pi_{\mathcal{P}} \text{ (which refers to the distance function from the property } \mathcal{P}) \text{ is } (\gamma, \gamma) \text{-weakly-continuous, for any } \gamma \in [0, 2].$$

**Theorem 33** (Valiant [14, Theorem 5.1]): Consider a function $\Pi : D_n \rightarrow \mathcal{R}$ that is label-invariant and $(\gamma, \gamma)$-weakly-continuous, where $\gamma \geq 0$ and $\delta \in [0, 2]$. Let there exist two distributions $p^+$ and $p^-$ in $D_n$ with $n$ being the size of their supports, such that $\Pi(p^+) > b$, $\Pi(p^-) < a$, and they are identical for any index occurring with probability at least $\frac{1}{2}$ in either distribution, where $a, b \in \mathbb{R}$. Then any tester that has sample access to an unknown distribution $D$ and distinguishes between $\Pi(D) > b - \gamma$ and $\Pi(D) < a + \gamma$, requires $\Omega(n^{1-\alpha(1)})$ many samples from $D$.¹⁴

Note that in Theorem 33, we have assumed that $p^+$ and $p^-$ are identical for any index that has probability mass at least $\frac{1}{2}$. We can actually replace this condition to $\mathcal{O}(\frac{1}{n})$ by adding $\mathcal{O}(\sqrt{n})$ many “dummy elements” to the support of $p^+$ and $p^-$. Now we are ready to prove Theorem 31.

**Proof:** [Proof of Theorem 31]

Consider $\Pi_{\mathcal{P}}$ as defined above. As $\mathcal{P}$ is a label-invariant property, the function $\Pi_{\mathcal{P}}$ is also label-invariant. We have $\text{SUPP}(D_{\text{no}}) \subset \text{SUPP}(D_{\text{yes}})$ (particularly, $\text{SUPP}(D_{\text{no}}) \cap L \subset \text{SUPP}(D_{\text{yes}}) \cap L$), it is not possible to distinguish between $D_{\text{yes}}$ and $D_{\text{no}}$, unless an element of $L$ appears at least twice. Otherwise, as in the proof of Lemma 17, the elements drawn from $L$ are distributed identically to a uniformly random non-repeating sequence. But observe that $D_{\text{yes}}(i) = \mathcal{O}\left(\frac{1}{n}\right)$ and $D_{\text{no}}(i) = \mathcal{O}\left(\frac{1}{n}\right)$ when $i$ is in $L$. Thus any sequence of $O(\sqrt{n})$ samples will provide only a distance of $o(1)$ between the two distributions, completing the proof. □

¹⁴$\alpha(\cdot)$ suppresses a term in $\delta$. 

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
already noted that $\Pi_P$ is $(\gamma, \gamma)$-weakly continuous as “distance from a property” satisfies the triangle inequality, for any $\gamma \in [0, 2]$. Now recall that the distributions $D_y$ and $D_o$ considered in the proof of Theorem 29. The probability mass of each element in the support of $D_y$ and $D_o$ is $O(1/\epsilon)$. Note that $D_y$ is in $P$ and $D_o$ is $\epsilon$-far from $P$, for any $\epsilon < \alpha$, and both of them have a support size of $\Theta(n)$. Here we take $\epsilon > \epsilon_2$. Now, we apply Theorem 33 with $a = 0$, some $b < \epsilon$ and $\gamma$ with $\gamma < \min\{\epsilon_1, \epsilon - \epsilon_2\}$. Observe that this completes the proof of Theorem 31.

\[ \Box \]

B. Sample Complexity of Non-Concentrated Properties
(Proof of Theorem 3)

Theorem 34 (Theorem 3 formalized): Let $P$ be any $(\alpha, \beta)$-non-concentrated distribution property for $0 < \alpha < \beta < \frac{1}{2}$. For any $\epsilon$ with $0 < \epsilon < \alpha$, any $(0, \epsilon)$-tester for $P$ requires $\Omega(\sqrt{n})$ many samples, where $n$ is the size of the support of the distribution.

1. Why Does the Proof of Theorem 29 Work Only for Label-Invariant Properties?: Note that the proof of Theorem 29 crucially uses the fact that the property $P$ is label-invariant. Recall that, while constructing $D_o$ from $D_y$ for each $i \in [\beta n]$, moving the masses of both $x_i$ and $y_i$ in $D_y$ to $x_i$ to produce $D_o$ is possible as the property $P$ is label-invariant. Due to this feature, we can apply a random permutation over $\Omega$, and still the permuted distribution will behave identically with respect to $P$. After applying the random permutation, the samples coming from $D_y$ and $D_o$ are indistinguishable as long as there are no collisions among the elements in $L$, which is the case when we take $o(\sqrt{n})$ samples. However, this technique does not work when the property is not label-invariant, as the value of the distribution with respect to $P$ may not be invariant under the random permutation over $\Omega$. This requires a new argument; although the proof is similar in spirit to the proof of Theorem 29, there are some crucial differences, and we present the proof next. In order to prove Theorem 34, instead of moving the masses of both $x_i$ and $y_i$ in $D_y$ to $x_i$ to produce $D_o$, we randomly move the sum to either $x_i$ or $y_i$, with probability proportional to the masses of $x_i$ and $y_i$.

C. Proof of Theorem 34
The proof of Theorem 34 starts off identically to the proof of Theorem 29, but there is a departure in the construction of $D_y$ and $D_o$.

Let us first consider $D_y$, $L$ and $P_L$ as discussed in the proof of Theorem 29, only here we cannot and will not pass $D_y$ through a random permutation. The difference starts from the description of the distribution $D_y$. In fact, $D_y$ will be randomly chosen according to a distribution over a set of distributions, all of which are not $(\alpha, \beta)$-non-concentrated. The distribution $D_y$ is generated using the following random process:

- We partition $L$ arbitrarily into two equal sets of size $\beta n$. Let the sets be $\{x_1, \ldots, x_{\beta n}\}$ and $\{y_1, \ldots, y_{\beta n}\}$.

We pair the elements of $L$ arbitrarily into $\beta n$ pairs. Let $(x_1, y_1), \ldots, (x_{\beta n}, y_{\beta n})$ be an arbitrary pairing of the elements in $L$. Let $P_L$ be the set of pairs. So $P_L = \{(x_1, y_1), \ldots, (x_{\beta n}, y_{\beta n})\}$. We refer to $x_i$ and $y_i$ as the elements corresponding to each other with respect to $P_L$, and denote $\pi(x_i) = y_i$ and $\pi(y_i) = x_i$.

- The probability mass of $D_o$ at $z$ is defined as follows:
  - If $z \notin L$, then $D_o(z) = D_o(z)$.
  - For every pair $(x_i, y_i) \in P_L$, use independent random coins and
    * With probability $\frac{D_y(x_i)}{D_y(x_i) + D_y(y_i)}$, set $D_o(x_i) = D_y(x_i) + D_y(y_i)$ and $D_o(y_i) = 0$.
    * With the remaining probability, that is, with probability $\frac{D_y(y_i)}{D_y(x_i) + D_y(y_i)}$, set $D_o(x_i) = 0$ and $D_o(y_i) = D_y(x_i) + D_y(y_i)$.

Observe that any $D_o$ constructed by the above procedure is supported over a set of at most $(1 - \beta)n$ elements. So any distribution $D_o$ constructed using the above procedure is $\epsilon$-far from satisfying the property $P$, for any $\epsilon < \alpha$. But since any $D_o$ produced using the above procedure has exactly the same probability mass on elements in $H$ as $D_y$, any tester that distinguishes between $D_y$ and $D_o$ must rely on samples obtained from $L$. However, we can prove that unless we receive two samples from the same pair in $L$ (which occurs with low probability), the sample sequence cannot distinguish $D_y$ from $D_o$.

Note that there is an upper bound of $O(\frac{1}{\epsilon})$ on the probability mass of any element in $L$. In fact, for any pair $(x_i, y_i) \in P_L$, the total probability mass of the pair is at most $O(\frac{1}{\epsilon})$.

Observation 35 (Follows Immediately From Observation 30): For all pairs $(x_i, y_i) \in P_L$, $D_o(x_i) + D_o(y_i) \leq \frac{1 - 2\beta}{1 - 2\alpha} \frac{1}{n}$. Also note that $D_o(x_i) = D_y(x_i) + D_y(y_i)$ and $D_o(y_i)$ with probability 1 over the construction of $D_o$.

From Observation 35, observe that if $S$ is a multi-set of $o\left(\frac{1 - 2\beta}{1 - 2\alpha} \frac{1}{n}\right)$ samples from $D_y$, then with probability $1 - o(1)$, no two elements in $S$ (identical or not) are from the same pair in $P_L$. The same holds for $D_o$ as well. Given that no two elements in $S$ are from the same pair in $P_L$, we will now prove that $D_y$ and $D_o$ have the same distributions over sample sequences. This implies that, for a sequence of $o\left(\frac{1 - 2\beta}{1 - 2\alpha} \frac{1}{n}\right)$ samples, $D_y$ and $D_o$ induce distributions over sample sequences that have $o(1)$ variation distance from each other.

Note that under the condition that at most one element is drawn from any pair $(x_i, y_i) \in P_L$, the probability that the sample is $x_i$ instead of $y_i$ is equal to $\frac{D_y(x_i)}{D_y(x_i) + D_y(y_i)}$, irrespective of whether the distribution is $D_y$ or $D_o$. So, we have the following lemma.

Lemma 36: Let $a_1, \ldots, a_q$ be a sequence of $q$ elements, where no element of $L$ appears twice, additionally containing no two elements from the same pair in $P_L$ (elements of $H$ can appear freely). Then

$$\Pr_{s_1, \ldots, s_q \sim D_y}\left[(s_1, \ldots, s_q) = (a_1, \ldots, a_q)\right] = \Pr_{s_1, \ldots, s_q \sim D_o}\left[(s_1, \ldots, s_q) = (a_1, \ldots, a_q)\right].$$

Proof: Let us begin by defining an event $E$ as follows:

$E := \text{no element of } L \text{ appears twice,}
\text{and no two elements from the same pair appear.}$

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Observe that we will be done by proving
\[
\Pr_{s_1, \ldots, s_q \sim D_{yes}} [s_i = a_i \text{ for each } i \in [q]] = \Pr_{s_1, \ldots, s_q \sim D_{no}} [s_i = a_i \text{ for each } i \in [q]],
\]
for each \( q \in [1, k] \) \( \Pr_{s_1, \ldots, s_q \sim D_{yes}} [s_i = a_i] = \Pr_{s_1, \ldots, s_q \sim D_{no}} [s_i = a_i] \text{ for each } i \in [q] \).

(12)

We will prove this by using induction over \( q \). Let us assume that we have generated samples \( s_1 = a_1, \ldots, s_k = a_k \) from the unknown distribution, where \( 1 \leq k < q \). Let \( X_k = \{s_1, \ldots, s_k\} \cap L \) be the samples we have seen until now from \( L \), and \( X_k' = \{\pi(x) : x \in X_k\} \). By the induction hypothesis, assume that Equation (12) holds for each \( q \) with \( q \leq k \). We will show that Equation (12) holds for \( q = k + 1 \).

To do so, let us now define two distributions \( D_{yes}^{k+1} \) and \( D_{no}^{k+1} \) as follows, and consider a claim (Claim 37) about them.

\[
D_{yes}^{k+1}(x) = \Pr_{s_1, \ldots, s_{k+1} \sim D_{yes}} [s_{k+1} = x | \mathcal{E} \text{ and } s_i = a_i \text{ for } i \leq k].
\]

Similarly,
\[
D_{no}^{k+1}(x) = \Pr_{s_1, \ldots, s_{k+1} \sim D_{no}} [s_{k+1} = x | \mathcal{E} \text{ and } s_i = a_i \text{ for } i \leq k].
\]

Claim 37: \( D_{yes}^{k+1}(x) = D_{no}^{k+1}(x) \) for every \( x \in \Omega \).

Proof: We prove the claim separately when \( x \in X_k \cup X_k' \), \( x \in L \setminus (X_k \cup X_k') \), and \( x \not\in L \).

(i) \( x \in X_k \cup X_k' \): As \( D_{yes}^{k+1}(x) = D_{no}^{k+1}(x) = 0 \) for every \( x \in X_k \cup X_k' \), we have the followings for each \( x \in L \setminus (X_k \cup X_k') \).

Assume that \( x = x_i \in L \setminus (X_k \cup X_k') \) for some \( i \in [\beta n] \) (using the notation defined for the partition of \( L \) into pairs while we have described the random process for generating \( D_{no} \)). The argument for the case where \( x = y_j \) for some \( j \in [\beta n] \) is analogous to this.

Under \( D_{yes} \), a direct calculation gives the probability for obtaining \( x = x_i \in L \setminus (X_k \cup X_k') \) as the \((k + 1)\)-th sample \( s_{k+1} \).

\[
D_{yes}^{k+1}(x) = D_{yes}(x | x \not\in X_k \cup X_k') = D_{yes}(x) = 1 - \sum_{y \in X_k \cup X_k'} D_{yes}(y).
\]

(12)

Now we are ready to prove Equation (12) for \( q = k + 1 \).

\[
\Pr_{s_1, \ldots, s_{k+1} \sim D_{yes}} [s_i = a_i \text{ for each } i \in [k+1]] = \Pr_{s_1, \ldots, s_{k+1} \sim D_{no}} [s_i = a_i \text{ for each } i \in [k]].
\]

(12)

Following the construction of \( D_{yes} \) and \( D_{no} \), we know that the two distributions differ only on the elements of \( L \). Moreover, following Observation 35, we know that if we take \( o\left(\sqrt{\frac{2n}{\epsilon^2}}\right) \) many samples, then with probability \( 1 - o(1) \), neither any element of \( L \) will appear more than once nor two elements of same pair in \( P_L \) will appear. Under these two conditions, Lemma 36 states that \( D_{yes} \) and \( D_{no} \) will appear to be the same. Thus we can say that any \((0, \epsilon)\)-tester that receives \( o\left(\sqrt{\frac{2n}{\epsilon^2}}\right) \) samples cannot distinguish between \( D_{yes} \) and \( D_{no} \), and obtain Theorem 34.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Remark 4: We would like to point out that the constructions and proofs of Theorem 34 would also work for proving the lower bound of non-tolerant testing of non-concentrated label-invariant properties (Theorem 29). However, we believe that the constructions of the hard distributions for Theorem 29 are more natural. As already noted, we first proved Theorem 29 as a warm up for the proof of Theorem 34.

VI. LEARNING A DISTRIBUTION (PROOF OF THEOREM 5)

In this section, we prove an upper bound related to the tolerant testing of more general properties. Following a folklore result, when provided with oracle access to an unknown distribution \( D \), we can always construct a distribution \( D' \), such that the \( \ell_1 \) distance between \( D' \) and \( D \) (the unknown distribution) is at most \( \varepsilon \), by using \( O(\frac{n}{\varepsilon^2}) \) samples from \( D \).\(^{15}\)

In this section, we provide a procedure that can be used for tolerant testing of properties, and in particular hints at how general tolerance gap bounds could be proved in the future. Our algorithm learns an unknown distribution approximately with high probability, adapting to the input, using as few samples as possible. Specifically, we prove that given a distribution \( D \), if there exists a subset \( S \subseteq [n] \) which holds most of the total probability mass of \( D \), then the distribution \( D \) can be learnt using \( O(|S|) \) samples, even if the algorithm is unaware of \( |S| \) in advance. Our result is formally stated as follows:

**Theorem 38 (Theorem 5 formalized):** Let \( D \) denote the unknown distribution over \( \Omega = [n] \), and assume that there exists a set \( S \subseteq [n] \) with \( D(S) \geq 1 - \frac{\eta}{2} \), where \( \eta \in [0,2] \) is known but \( S \) and \( |S| \) are unknown. Then there exists an algorithm that takes \( \delta \in (0,2] \) as input and constructs a distribution \( D' \) satisfying \( \|D - D'\|_1 \leq \eta + \delta \) with probability at least \( \frac{3}{4} \). Moreover, the algorithm uses, in expectation, \( O\left(\frac{|S|}{\delta^2}\right) \) many samples from \( D \).

Note that in the above theorem, the algorithm has no prior knowledge of \( |S| \). Before directly proving the above, we first show that if \( |S| \) is known, then \( O(|S|) \) many samples are enough to approximately learn the distribution \( D \). We would like to point out that a similar question has been studied under the local differential privacy model with communication constraints, by Acharya et al. [45] and by Chen et al. [46].

**Lemma 39 (Theorem 38 when \( |S| \) is known):** Let \( D \) be the unknown distribution over \( \Omega = [n] \) such that there exists a set \( S \subseteq [n] \) with \( |S| = s \), and \( \eta \in [0,2] \) such that \( D(S) \geq 1 - \frac{\eta}{2} \), where \( s \in [n] \) and \( \eta \in (0,1) \) are known. Then there exists an algorithm that takes \( \delta \in (0,2] \) as an input and constructs a distribution \( D' \) satisfying \( \|D - D'\|_1 \leq \eta + \delta \) with probability at least \( \frac{1}{4} \). Moreover, the algorithm uses \( O\left(\frac{s}{\delta^2}\right) \) many samples from \( D \).

We note that Lemma 39 can be obtained from the work of Acharya et al. [41, Theorem 2]. For completeness, we give a self-contained proof for this lemma below.

We later adapt the algorithm of Lemma 39 to give a proof to the scenario where \( |S| \) is unknown, using a guessing technique.\(^{16}\)

The idea is to guess \( |S| = s \) starting from \( s = 1 \), and then to query for \( O(s) \) many samples from the unknown distribution \( D \). From the samples obtained, we construct a distribution \( D_s \), and use Lemma 40 presented below to distinguish whether \( D_s \) and \( D \) are close or far. We argue that, for \( s \geq |S| \), \( D_s \) will be close to \( D \) with probability at least \( \frac{n}{10} \). We bound the total probability for the algorithm reporting a \( D' \) that is too far from \( D \) (for example when terminating before \( s \geq |S| \)), and also bound the probability of the algorithm not terminating in time when \( s \) becomes at least as large as \( |S| \).

**Lemma 40 (\([15]\)):** Let \( D_u \) and \( D_k \) denote unknown (input) and known (given in advance) distributions respectively over \( \Omega = [n] \), such that the support of \( D_u \) is a set of \( s \) elements of \( |n| \). Then there exists an algorithm \( \text{TOL-ALG}(D_u,D_k,\varepsilon_1,\varepsilon_2,\kappa) \) that takes the full description of \( D_k \), two proximity parameters \( \varepsilon_1,\varepsilon_2 \) with \( 0 \leq \varepsilon_1 < \varepsilon_2 \leq 2 \) and \( \kappa \in (0,1) \) as inputs, queries \( O\left(\frac{1}{((\varepsilon_2-\varepsilon_1)^2 \log s \log \kappa)}\right) \) many samples from \( D_u \), and distinguishes whether \( \|D_u - D_k\|_1 \leq \varepsilon_1 \) or \( \|D_u - D_k\|_1 \geq \varepsilon_2 \) with probability at least \( 1 - \kappa \).\(^{17}\)

Note that Theorem 38 talks about learning a distribution with \( O(|S|) \) samples, where there exists an unknown set \( S \) with \( D(S) \geq 1 - \eta/2 \). To prove Theorem 38, we use Lemma 40, that crucially uses less than \( s \) queries for tolerant identity testing (as opposed to learning).

The original bound following the paper of Valiant and Valiant [15] is \( O\left(\frac{n}{(\eta-\varepsilon_1)^2 \log n} \right) \), which holds for any general distributions \( D_u \) and \( D_k \) with constant success probability. When deploying Lemma 40, we “contract” the set \( \Omega \setminus \text{SUPP}(D_k) \) to a single element, which allows us to substitute \( s+1 \) for \( n \). Note that this does not change the distance between \( D_k \) and \( D_u \). Hence, \( O\left(\frac{1}{((\varepsilon_2-\varepsilon_1)^2 \log s)}\right) \) samples from \( D_u \) are enough for constant success probability. Following a recent work of Cannone et al. [31], the dependence on the proximity parameters can be slightly improved. However we are not using that result since the focus of this work is different.

We first prove Lemma 39, and then proceed to prove Theorem 38.

**Proof:** [Proof of Lemma 39] Let \( Z \) be a multi-set of \( O\left(\frac{s}{\delta^2}\right) \) samples taken from \( D \). The algorithm constructs a distribution \( D': [n] \rightarrow [0,1] \) such that

\[
D'(x) = \frac{\# \text{times } x \text{ appears in } Z}{|Z|}.
\]

Observe that \( \|D - D'\|_1 = 2 \max_{E \subseteq [n]} |D(E) - D'(E)|. \) So, we will be done by showing the following:

\[
\Pr\left(|D(E) - D'(E)| \leq \frac{\eta + \delta}{2}\right) \geq \frac{9}{10} \quad \text{for all } E \subseteq [n].
\]

Note that there are \( 2^n \) possibilities for \( E \). So, a direct application of the union bound would require a failure probability of at most \( O\left(\frac{1}{\delta^2}\right) \) for each \( E \) not satisfying \( |D(E) - D'(E)| \leq \frac{\eta + \delta}{2} \), that is, \( O\left(\frac{1}{\delta^2}\right) \) samples would be needed. Assuming that

\[\text{there is a writeup of this folklore result by Cannone [44].}\]

\[\text{Recall that the variation distance between two distribution is half than that of the } \ell_1 \text{ distance between them. So, we take } D(S) \geq 1 - \frac{\eta}{2} \text{ (with } \eta \in [0,2]) \text{ instead of } D(S) \geq 1 - \eta \text{ (with } \eta \in [0,1]).}\]
D is concentrated ($D(S) \geq 1 - \frac{s}{2}$), we argue below that it is enough to have a failure probability of $O(\frac{1}{2^2})$ for each $T$ not satisfying $|D(T) - D'(T)| \leq \frac{s}{4}$, but first we show that this is indeed the probability that we achieve.

**Observation 41:** Consider $T \subseteq [n]$. $|D(T) - D'(T)| \leq \frac{s}{4}$ holds with probability at least $1 - \frac{1}{100 \log^2 s}$.

**Proof:** Let $X_i$ denote the binary random variable that takes value 1 if and only if the $i$-th sample in $Z$ is an element of $T$, where $i \in [Z]$. So, $D'(T) = \frac{1}{|Z|} \sum_{i=1}^{|Z|} X_i$.

Observe that the expectation of $D'(T)$ is $\mathbb{E}[D'(T)] = D(T)$. Applying Chernoff bound (Lemma 14), we get the desired result.

By the above observation for every subset of $S$, applying the union bound over all possible subsets of $S$, we have $|D(T) - D'(T)| \leq \frac{s}{4}$ for every $T \subseteq S$ with probability at least $\frac{99}{100}$. Further applying the observation for $T = \Omega \setminus S$, we have $|D(\Omega \setminus S) - D'(\Omega \setminus S)| \leq \frac{s}{4}$ with probability at least $1 - \frac{1}{100 \log^2 s}$.

Let $E$ be the event that $|D(T) - D'(T)| \leq \frac{s}{4}$ for every $T \subseteq S$, and $|D(\Omega \setminus S) - D'(\Omega \setminus S)| \leq \frac{s}{4}$. Note that $\Pr(E) \geq \frac{99}{100}$. So, to prove Equation (13) and conclude the proof of Lemma 39, we show that $|D(E) - D'(E)| \leq \frac{s \eta}{4}$ holds, in the conditional probability space when $E$ occurs, for any $E \subseteq [n]$.

$$|D(E) - D'(E)| \leq |D(E \cap S) - D'(E \cap S)| + |D(E \cap (\Omega \setminus S)) - D'(E \cap (\Omega \setminus S))| \leq \frac{s}{4} + \max\{|D(\Omega \setminus S), D'(\Omega \setminus S)|\} \leq \frac{s}{4} + D(\Omega \setminus S) + \frac{s}{4} \leq \frac{\eta + \delta}{2}.$$ 

**Proof:** [Proof of Theorem 38] The algorithm is as follows:

1) Set $s = 1$.
2) Query for a multi-set $Z_s$ of $O\left(\frac{s}{\delta^2}\right)$ many samples from $D$.
3) Construct a distribution $D_s : [n] \rightarrow [0,1]$ such that $D_s(x) = \frac{\# \text{ times } x \text{ appears in } Z_s}{|Z_s|}$.
4) Call the algorithm $TOL-\text{ALG}(D_s, D, \eta + \delta/2, \eta + \delta, 1/100 \log^2 s)$ (corresponding to Lemma 40) to distinguish whether $||D - D_s||_1 \leq \eta \pm \frac{s}{2}$ or $||D - D_s||_1 \geq \eta \pm \delta$.

If we get $||D - D_s||_1 \leq \eta + \frac{s}{2}$ as the output of $TOL-\text{ALG}$, then we report $D'$ as the output and QUIT. Otherwise, we double the value of $s$. If $s \leq 2n$, go back to Step 2. Otherwise, report FAILURE.18

Let $S$ denote the event that the algorithm quits with the desired output. We first show that $\Pr(S) \geq \frac{s}{4}$. Then we analyze the expected sample complexity of the algorithm.

Observe that the algorithm quits after an iteration with guess $s$ such that $ALG-TOL$ reports $||D - D_s||_1 \leq \eta + \frac{s}{2}$. So, in that case, the probability that the algorithm exits with an output not satisfying $||D - D_s||_1 \leq \eta + \delta$ is at most $\frac{1}{100 \log^2 s}$.

When summing this up over all possible $s$ (all powers of $k$, even up to infinity), the probability that the algorithm does not produce the desired output, given that it quits, is at most $\sum_{k=1}^\infty \frac{1}{100 k^2} \leq \frac{1}{10}$. So, denoting $Q$ as the event that the algorithm quits without reporting FAILURE, $\Pr(S \mid Q) \geq \frac{s}{4}$.

For the lower bound on $\Pr(Q)$, consider the case where $s \geq |S|$. In this case, $||D_s - D||_1 \leq \frac{s}{2} + \frac{s}{2}$ with probability at least $\frac{9}{10}$, and $TOL-\text{ALG}$ quits by reporting $D_s$ as the output with probability at least $1 - \frac{1}{100 \log^2 s}$. So, for any guess $s \geq |S|$, the algorithm quits and reports the desired output with probability at least $\frac{s}{4}$. So, the probability that the algorithm quits without reporting failure is at least the probability that the algorithm quits with a desired output at some iteration with a guess $s \geq |S|$, which is at least $1 - \left(\frac{1}{3}\right)\log n - \log |S| + 1$. That is, $\Pr(Q) \geq \frac{s}{4}$.

Hence, the success probability of the algorithm can be lower-bounded as

$$\Pr(S) \geq \Pr(Q) \cdot \Pr(S \mid Q) \geq \frac{9}{10} \cdot \frac{4}{5} \geq \frac{2}{3}.$$ 

Now, we analyze the sample complexity of the algorithm. The algorithm queries for $O(s)$ samples when it runs the iteration whose guess is $s$. The algorithm goes to the iteration with guess $s > |S|$ if all prior iterations which guessed more than $|S|$ failed, which holds with probability at most $O\left(\left(\frac{1}{3}\right)\log s / |S|\right)$. Hence the expected sample complexity of the algorithm is at most

$$\sum_{k:s=2^k \leq |S|} O\left(\frac{s}{\delta^2}\right) + \sum_{k:s=2^k \geq |S|} O\left(\left(\frac{1}{3}\right)\log s / |S|\right) \cdot \frac{s}{\delta^2}$$ 

To explain the above equality, note that in the LHS of the above equation, each term of the second sum is bounded by $O\left(\left(\frac{1}{3}\right)\log s / |S|\right)$, hence the expected sample complexity is at most

$$O\left(\frac{|S|}{\delta^2}\right).$$ 

In Theorem 15 we proved that the gap between the sample complexities of tolerant and non-tolerant testing of label-invariant non-concentrated properties is essentially tight in the following sense:

- In Theorem 15 we proved that the gap between the sample complexities of tolerant and non-tolerant testing of any label-invariant property is at most quadratic;
- There exists a label-invariant distribution property which admits $\Theta(\sqrt{n})$ sample complexity for non-tolerant testing and almost linear sample complexity for non-tolerant testing.

Note that, in Theorem 15, we only tried to optimize the sample complexity of the tolerant tester in terms of the sample complexity of its corresponding non-tolerant tester. We did

18By Lemma 40, this step uses $O\left(\frac{s}{\delta^2}\right)$ samples.
not optimize the sample complexity in terms of proximity parameters. So, there are three natural open questions:

(i) Can we optimize all the parameters involved in Theorem 15 by improving the stated sample complexity and/or proving the relevant lower bound? In particular, can we optimize the sample complexity in terms of proximity parameters, as well as remove the logarithmic factors in the sample complexity of our tolerant tester?

(ii) For any \( q \in \mathbb{N} \) with \( 1 \leq q \leq \sqrt{n} \), can we show the existence of a label-invariant property whose non-tolerant sample complexity is \( \Theta(q) \) and tolerant sample complexity is \( \Theta(q^2) \)?

(iii) Can one prove that the gap between the sample complexities of tolerant and non-tolerant testing is at most quadratic for general distribution properties?

We believe that answering these questions would require substantial new insights and techniques.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers of IEEE TRANSACTIONS ON INFORMATION THEORY and RANDOM 2022 for their comments which improved the presentation of the article. Part of this work was done while Sayantan Sen was a Ph.D. student at the Indian Statistical Institute.

REFERENCES

[1] E. Fischer and A. Matsliah, “Testing graph isomorphism,” SIAM J. Comput., vol. 38, no. 1, pp. 207–225, Jan. 2008.

[2] H. Zhang, “A new approach for testing properties of distributions,” in Proc. IEEE 57th Annu. Symp. Foundations Comput. Sci. (FOCS), Oct. 2016, pp. 685–694.

[3] M. Aliakbarpour, I. Diakonikolas, D. Kane, and R. Rubinfeld, “Private testing of distributions via sample permutations,” in Proc. Neural Inf. Process. Syst. (NeurIPS), 2019, pp. 1012–1042.

[4] G. Valiant and P. Valiant, “The power of linear estimators,” in Proc. IEEE 52nd Annu. Symp. Found. Comput. Sci. (FOCS), Oct. 2011, pp. 403–412.

[5] G. Valiant and P. Valiant, “An automatic inequality prover and instance optimal identity testing,” SIAM J. Comput., vol. 46, no. 1, pp. 429–455, 2017.

[6] T. Batu and C. L. Canonne, “Generalized uniformity testing,” in Proc. IEEE 55th Annu. Symp. Found. Comput. Sci. (FOCS), Oct. 2017, pp. 880–889.

[7] T. Batu and C. L. Canonne, “Generalized uniformity testing,” in Proc. Neural Inf. Process. Syst. (NeurIPS), 2018, pp. 6204–6213.

[8] C. L. Canonne, I. Diakonikolas, T. Gouleakis, and R. Rubinfeld, “Testing shape restrictions of discrete distributions,” Theory Comput. Syst., vol. 62, pp. 4–62, Jan. 2018.

[9] E. Fischer, O. Lachisch, and V. Vasudev, “Improving and extending the testing of distributions for shape-restricted properties,” in Proc. Symp. Theor. Aspects Comput. Sci. (STACS), 2017, pp. 31:1–31:14.

[10] G. Valiant and P. Valiant, “Estimating the unseen: Improved estimators for entropy and other properties,” J. ACM, vol. 64, no. 6, pp. 37:1–37:41, 2017.

[11] T. King, A Guide To Chi-Squared Testing. New York, NY, USA: Taylor & Francis, 1997.

[12] E. Blais, C. L. Canonne, T. Eden, A. Levi, and D. Ron, “Tolerant junta testing and multi-testing scalability,” in Proc. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 3591–3599.

[13] C. Daskalakis, G. Kamath, and J. Wright, “Which distribution distances are sublinearly testable?” in Proc. Symp. Discrete Algorithms (SODA), 2018, pp. 2747–2764.

[14] C. L. Canonne, A. Jain, G. Kamath, and J. Li, “The price of tolerance in distribution testing,” in Proc. Conf. Learn. Theory (COLT), 2022, pp. 573–624.

[15] C. L. Canonne, “A survey on distribution testing: Your data is big. But is it blue?” Theory Comput., vol. 9, pp. 1–100, Aug. 2020.

[16] C. L. Canonne, “Topics and techniques in distribution testing: A biased but representative sample,” Found. Trends Commun. Inf. Theory, vol. 19, no. 6, pp. 1032–1198, 2022.

[17] M. Parnas, D. Ron, and R. Rubinfeld, “Tolerant property testing and distance approximation,” J. Comput. Syst. Sci., vol. 72, no. 6, pp. 1012–1042, Sep. 2006.

[18] E. Blais, C. L. Canonne, T. Eden, A. Levi, and D. Ron, “Tolerant junta testing and the connection to submodular optimization and function isomorphism,” ACM Trans. Comput. Theory, vol. 11, no. 4, pp. 24:1–24:33, Dec. 2019.

[19] O. Goldreich and D. Ron, “On sample-based testers,” ACM Trans. Comput. Theory, vol. 8, no. 2, pp. 1–54, May 2016.

[20] E. Fischer, O. Lachish, and V. Vasudev, “Trading query complexity for sample-based testing and multi-testing scalability,” in Proc. IEEE 56th Annu. Symp. Foundations Comput. Sci., Oct. 2015, pp. 1163–1182.

[21] D. P. Dubhashi and A. Panconesi, Concentration of Measure for the Analysis of Randomized Algorithms. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[22] I. Diakonikolas et al., “Testing for concise representations,” in Proc. 48th Annu. IEEE Symp. Found. Comput. Sci. (FOCS), Oct. 2007, pp. 549–558.

[23] R. A. Servedio, “Testing by implicit learning: A brief survey,” in Property Testing: Current Research and Surveys. Berlin, Germany: Springer, 2010, pp. 197–210.

[24] J. Acharya, I. Diakonikolas, J. Li, and L. Schmidt, “Sample-optimal density estimation in nearly-linear time,” in Proc. 28th Annu. ACM-SIAM Symp. Discrete Algorithms (SODA), Jan. 2017, pp. 1278–1289.
Sourav Chakraborty received the Ph.D. degree from the University of Chicago, under the supervision of Prof. László Babai. He is currently a Professor with the Advanced Computing and Microelectronics Unit, Indian Statistical Institute. Previously, he was a Faculty Member with Chennai Mathematical Institute, India. He was a Post-Doctoral Researcher with Centrum Wiskunde and Informatica (CWI), Amsterdam, The Netherlands, and Technion, Israel. His research interests include theoretical computer science, focusing on the classical and quantum complexity of Boolean functions, property testing, streaming algorithms, algorithmic game theory, and machine learning.

Arijit Ghosh received the dual B.Tech. and M.Tech. degree in computer science from Indian Institute of Technology Kharagpur and the Ph.D. degree in computer science from INRIA Sophia Antipolis, under the supervision of Prof. Jean-Daniel Boissonnat in 2012. He is currently an Associate Professor with the Advanced Computing and Microelectronics Unit, Indian Statistical Institute. His research interests include discrete and computational topology and geometry, randomized algorithms, and probabilistic methods.

Eldar Fischer received the Ph.D. degree from Tel-Aviv University in 1999, under the supervision of Prof. Noga Alon. He has been a member of the Faculty of Computer Science, Israel Institute of Technology (Technion), since 2001. His main CS interests lie in the analysis of algorithms, especially algorithms related to property testing and other models with restricted input access. His research interests include graph theory, the regularity lemma and its variants, and finite models in logic.

Gopinath Mishra received the Ph.D. degree from Indian Statistical Institute, Kolkata, India, in 2021, under the supervision of Prof. Arijit Bishnu and Arijit Ghosh. He has been a Post-Doctoral Fellow with the National University of Singapore since September 2023. Previously, he was a Post-Doctoral Fellow with the University of Warwick from August 2021 to August 2023. His research interests include theoretical computer science, with a focus on model-centric computation.

Sayantan Sen received the Ph.D. degree from Indian Statistical Institute in 2023, under the supervision of Prof. Sourav Chakraborty. He is currently a Research Fellow with the Centre for Quantum Technologies, National University of Singapore. Previously, he was a Research Fellow with the School of Computing, National University of Singapore, from April 2023 to April 2024. His research interests include randomized algorithms, with a focus on distribution and graph property testing in the classical and quantum models.