Exploring the Gap Between Tolerant and Non-Tolerant Distribution Testing

Sourav Chakraborty
Indian Statistical Institute, Kolkata, India

Eldar Fischer
Technion – Israel Institute of Technology, Haifa, Israel

Arijit Ghosh
Indian Statistical Institute, Kolkata, India

Gopinath Mishra
University of Warwick, Coventry, UK

Sayantan Sen
Indian Statistical Institute, Kolkata, India

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, not necessarily label-invariant properties, the situation is more complicated, and we show some partial results. We show that if a property requires the distributions to be non-concentrated, that is, the probability mass of the distribution is sufficiently spread out, then it cannot be non-tolerantly tested with $o(\sqrt{n})$ many samples, where n denotes the universe size. Clearly, this implies at most a quadratic gap, because a distribution can be learned (and hence tolerantly tested against any property) using $\mathcal{O}(n)$ many samples. Being non-concentrated is a strong requirement on properties, as we also prove a close to linear lower bound against their tolerant tests.

Apart from the case where the distribution is non-concentrated, we also show if an input distribution is very concentrated, in the sense that it is mostly supported on a subset of size s of the universe, then it can be learned using only $\mathcal{O}(s)$ many samples. The learning procedure adapts to the input, and works without knowing s in advance.

2012 ACM Subject Classification Theory of computation → Streaming, sublinear and near linear time algorithms

Keywords and phrases Distribution Testing, Tolerant Testing, Non-tolerant Testing, Sample Complexity

Digital Object Identifier 10.4230/LIPIcs.APPROX/RANDOM.2022.27

Category RANDOM

Related Version *Full Version*: https://arxiv.org/abs/2110.09972

Funding Gopinath Mishra's research is supported in part by the Centre for Discrete Mathematics and its Applications (DIMAP) and by EPSRC award EP/V01305X/1.

Acknowledgements The authors would like to thank the reviewers of RANDOM 2022 for their valuable suggestions which improved the presentation of the paper.

Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2022).
Editors: Amit Chakrabarti and Chaitanya Swamy; Article No. 27; pp. 27:1–27:23
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Let D be a distribution over a finite set Ω, and $\mathcal{P}$ be a property, that is, a set of distributions over Ω. Given access to independent random samples from Ω according to the distribution D, we are interested in the problem of distinguishing whether the distribution D is η-close to having the property $\mathcal{P}$, or is ε-far from having the property $\mathcal{P}$, where η and ε 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 $\min_{D' \in \mathcal{P}} ||D - D'||_1$, where $||D - D'||_1$ denotes the ℓ_1-*distance* between the distributions D and D' [1]. A distribution D is said to be η-close to $\mathcal{P}$, if the distance of D from $\mathcal{P}$ is at most η. Similarly, D is said to be ε-far if the distance of D from $\mathcal{P}$ is at least ε. 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 $\frac{2}{3}$) whether the distribution D is η-close to having the property $\mathcal{P}$, or is ε-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 [27, 29] and function isomorphism testing [6], learning theory [10, 23, 22], and differential privacy [5, 32, 41, 1]. 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 [9, 8, 35, 40, 37, 38]. Various other specific distribution properties have also been studied [7, 24]. Then, some works investigated general tests for the large class of all shape-restricted properties of distributions, which contains properties like monotonicity, log-concavity, modality etc. [15, 26]. This paper proves general results about the gap between tolerant and non-tolerant distribution testing that hold for large classes of properties.

1.1 Our results

We now informally present our results. The formal statements of the theorems are presented in the corresponding sections where they are proved, after the formal definitions are presented in Section 2. 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 Ω). 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 1.1** (Informal). *Any label-invariant distribution property that can be non-tolerantly tested using Λ samples, can also be tolerantly tested using $\widetilde{\mathcal{O}}(\min\{\Lambda^2, n\})$ samples, where n is the size of the support of the distribution* [2].

[1] Strictly speaking it is an infimum, but since all properties we consider are compact sets, it is equal to the minimum.

[2] $\widetilde{\mathcal{O}}(\cdot)$ hides a poly-logarithmic factor.

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 3. We also design a constructive variant of the tolerant tester of Theorem 1.1, when the property can be expressed as the feasible solution to a set of linear inequalities.

▶ **Theorem 1.2** (Informal). *Any label-invariant distribution property that can be non-tolerantly tested using Λ samples and can be expressed as a feasible solution to m linear inequalities, can also be tolerantly tested using $\widetilde{\mathcal{O}}(\min\{\Lambda^2, n\})$ samples and in time polynomial in m and n, where n is the size of the support of the distribution.*

We believe that this result can be generalized to the case where the property can be expressed as the feasible solution to a set of convex constraints, using more advanced techniques.

Note that if $\Lambda = \Omega(\sqrt{n})$, Theorem 1.1 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 ε-far from having support at most k. If $k = o(\sqrt{n})$, the query complexity for testing this problem is $\mathcal{O}(\frac{k}{\log k})$ [39].

It is a natural question to investigate the extent to which the above theorem can be generalized. Though we are not resolving this question completely, as a first step in the direction of extending the above theorem for properties that are not necessarily label-invariant, 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 βn with $\beta \in (0, \frac{1}{2})$), has probability mass of at least another threshold (say α 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 1.3** (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 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 $\mathcal{O}(n)$ many samples are required to learn any distribution approximately.

The proof of Theorem 1.3 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 5.

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 4.2. However, the question is left open for non-label-invariant properties.

▶ **Theorem 1.4** (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.*

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 $\mathcal{O}(n)$ 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 1.5** (Informal). *To learn a distribution approximately, $\mathcal{O}(|S|)$ 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.

Organization of the paper

Section 2 contains the definitions used throughout the paper. Section 3 contains the formal statement and proof of Theorem 1.1, where some of the lemmas, along with the proof of Theorem 1.2 (the constructive variant), are in the appendix. Theorems 1.3, 1.4 and 1.5 are formally stated and proved in Section 4, Section 5 and Section 6 respectively.

1.2 Related works

Several forms of distribution testing have been investigated for over a century in statistical theory [33, 19], while combinatorial properties of distributions have been explored over the last two decades in Algorithm Theory, Machine Learning and Information Theory [28, 34, 20]. In Algorithm Theory, the investigation into testing properties of distributions started with the work of Goldreich and Ron [30], even though it was not directly stated there in these terms. Batu, Fortnow, Rubinfeld, Smith, and White [9] launched the intensive study of property testing of distributions with the problem of equality testing [3]. Later, Batu, Fischer, Fortnow, Kumar, Rubinfeld and White [8] studied the problems of identity and independence testing of distributions [4]. Since then there has been a flurry of interesting works in this model. For example, Paninski [35] proved tight bounds on uniformity testing, Valiant and Valiant [37] resolved the tolerant sample complexity for a large class of label-invariant properties that includes uniformity testing, Acharya, Daskalakis, and Kamath [2] proved various optimal testing results under several distance measures, and Valiant and Valiant [38] studied the sample complexity of instance optimal identity testing. In [7], Batu and Cannone 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 $\widetilde{\Theta}(|S|^{2/3})$ 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, Kamath, and Wright [21] studied the problem of tolerant testing under various distance measures. Very recently, Canonne, Jain, Kamath, and Li [16] 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, Diakonikolas, Gouleakis, and Rubinfeld [15] studied the class of

---

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

[4] Given 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.

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, Lachish, and Vasudev [26]. See the survey of Cannone [14] 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, Ron and Rubinfeld [36], 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 [12] 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 [40, 37] [5]. 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 [31]. This model was investigated further in the work of Fischer, Lachish and Vasudev [25], where a general upper bound for non-tolerant sample-based testing of strongly testable properties was provided.

2 Notation and definitions

A probability distribution D over a universe $\Omega = [n]$ is a non-negative function $D : \Omega \to [0, 1]$ 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 Ω is denoted by $\textsc{Supp}(D)$. For any distribution D, by 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 the non-increasing order of their probability masses in D. When we write $\widetilde{\mathcal{O}}(\cdot)$, it suppresses a poly-logarithmic term in n and the inverse of the proximity parameter(s) [6]. Although there are several other distance measures, in this work, we mainly focus on the ℓ_1 distance. We subsume polynomial dependencies only on the proximity parameters in our results for clarity of presentation.

▶ **Definition 2.1** (Distribution property). *Let $\mathcal{D}$ denote the set of all distributions over Ω. A* distribution property *$\mathcal{P}$ is a topologically closed subset of $\mathcal{D}$* [7]*. 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 2.2** (Label-invariant property). *Let us consider a property $\mathcal{P}$. For a distribution D and a permutation $\sigma : \Omega \to \Omega$, consider the distribution D_σ 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_σ is also in $\mathcal{P}$ for every permutation σ, then the property $\mathcal{P}$ is said to be* label-invariant.

[5] To be precise, the exact lower bounds for non-tolerant uniformity testing is $\Omega(\sqrt{n})$, and for tolerant uniformity testing it is $\Omega(\frac{n}{\log n})$, where n is the support size of the distribution and the proximity parameter ε is constant.

[6] We will also use $\widetilde{\mathcal{O}}(\cdot)$ to suppress polynomials of the inverses of the differences of proximity parameters.

[7] We put this restriction to avoid formalism issues. In particular, the investigated distribution properties that we know of (such as monotonicity and being a k-histogram) are topologically closed.

▶ **Definition 2.3** (Distance between two distributions). *The distance between two distributions D_1 and D_2 over Ω is the standard ℓ_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 said to be η-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 ε-far from each other if $||D_1 - D_2||_1 \geq \varepsilon$.*

▶ **Definition 2.4** (Distance of a distribution from a property). *The distance of a distribution D from a property $\mathcal{P}$ is the minimum ℓ_1-distance between D and any distribution in $\mathcal{P}$. For $\eta \in [0,2]$, a distribution D is said to be η-close to $\mathcal{P}$ if the distance of D from $\mathcal{P}$ is at most η. Analogously, for $\varepsilon \in [0,2]$, a distribution D is said to be ε-far from $\mathcal{P}$ if the distance of D from $\mathcal{P}$ is at least ε.*

▶ **Definition 2.5** ((η, ε)-tester). *An* (η, ε)-tester *for a distribution property is a randomized algorithm that has sample access to the unknown distribution (upon query it can receive elements of Ω, each drawn according to the unknown distribution, independently of any previous query or the algorithm's private coins), and distinguishes whether the distribution is η-close to the property or ε-far from the property, with probability at least $\frac{2}{3}$, where η and ε 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 2.6** (Non-Concentrated distribution). *A distribution D over the domain $\Omega = [n]$ is said to be* (α, β)-non-concentrated *if for any set $S \subseteq \Omega$ with size βn, the probability mass on S is at least α, where α and β are two parameters such that $0 < \alpha \leq \beta < \frac{1}{2}$.*

▶ **Definition 2.7** (Non-Concentrated property). *Let $0 < \alpha \leq \beta < \frac{1}{2}$. A distribution property $\mathcal{P}$ is defined to be* (α, β)-non-concentrated, *if all distributions in $\mathcal{P}$ are (α, β)-non-concentrated.*

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

3 Non-tolerant vs. tolerant sample complexities of label-invariant properties (Proof of Theorem 1.1)

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 3.1** (Theorem 1.1 formalized). *Let $\mathcal{P}$ be a label-invariant distribution property. Also, let there exist an $(0, \varepsilon)$-tester (non-tolerant tester) for the property $\mathcal{P}$ with sample complexity $\Lambda(n, \varepsilon)$, where $\Lambda \in \mathbb{N}$ and $0 < \varepsilon \leq 2$. Then for any γ_1, γ_2 with $\gamma_1 < \gamma_2$ and $0 < \gamma_2 + \varepsilon < 2$, there exists a $(\gamma_1, \gamma_2 + \varepsilon)$-tester (tolerant tester) that has sample complexity $\mathcal{O}\left(\frac{1}{(\gamma_2 - \gamma_1)^2} \cdot \min\{\Lambda^2 \log^2 \Lambda, n\}\right)$, 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 $\Lambda(n,\epsilon) \geq \Omega(\frac{1}{\varepsilon})$ [8]. First note that if $\Lambda = \Omega(\sqrt{n})$, then we can construct a distribution $\widehat{D}$ such that $||D - \widehat{D}||_1 < \frac{\gamma_2-\gamma_1+\varepsilon}{2}$, by using $\mathcal{O}\left(\frac{n}{(\gamma_2-\gamma_1+\varepsilon)^2}\right)$ samples from D. Thereafter we can report D to be γ_1-close to the property if and only if $\widehat{D}$ is $\frac{\gamma_2+\gamma_1+\varepsilon}{2}$-close to the property. In what follows, we discuss an algorithm with sample complexity $\widetilde{\mathcal{O}}(\Lambda^2)$ when $\Lambda = o(\sqrt{n})$. Also, we assume that n and Λ are larger than some suitable constant. Otherwise, the theorem becomes trivial.

The idea behind the proof is to classify the elements of Ω with respect to their masses in D into *high* and *low*, as formally defined below in Definition 3.2. We argue that since $\mathcal{P}$ is $(0,\varepsilon)$-testable using $\Lambda(n,\varepsilon) = \mathcal{O}(q)$ samples, there cannot be two distributions D_1 and D_2 that are identical on all elements whose probability mass is at least $\frac{1}{q^2}$, for $q = \theta(\Lambda)$ (the set High_{1/q^2} defined below), where $D_1 \in \mathcal{P}$ but D_2 is ε-far from $\mathcal{P}$. We will formally show this in Lemma 3.3, where we will use the fact that $\mathcal{P}$ is label-invariant. Using Lemma 3.3, we prove Lemma 3.4, 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 maintain the masses of the set High_{1/q^2}, and the term Λ^2 in the query complexity of our algorithm corresponds to that.

▶ **Definition 3.2.** *For a distribution D over Ω and $0 < \kappa < 1$, we define*

$$\mathsf{High}_\kappa(D) = \{x \in \Omega \ \mid \ D(x) \geq \kappa\}$$

Now we define a quantity $q \in \mathbb{N}$ where $q = \Theta(\Lambda)$ [9]. Assume that c^* is a suitable large constant (independent of Λ) such that, if we take Λ 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

$$q := \frac{\Lambda}{c^*}. \tag{1}$$

We will complete the proof of Theorem 3.1 by using the following two lemmas which we will prove later.

▶ **Lemma 3.3.** *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 $\mathsf{High}_{1/q^2}(D_1) = \mathsf{High}_{1/q^2}(D_2)$, and for all $x \in \mathsf{High}_{1/q^2}(D_1)$, 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 ε-far from satisfying $\mathcal{P}$.*

▶ **Lemma 3.4.** *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 $\widetilde{D}$ be two distributions over Ω ($|\Omega| > 4q^2$) and let H contain the top q^2 elements of $\overline{D}$. Also, assume that $\left|\widetilde{D}(\Omega \setminus H) - \overline{D}(\Omega \setminus H)\right| \leq \gamma$. If*

$$\sum_{x \in H} \left|\overline{D}(x) - \widetilde{D}(x)\right| \leq \alpha, \tag{2}$$

then the following hold:

[8] This is a reasonable assumption for any non-trivial property.
[9] Note that q and Λ essentially denotes the same quantity. We have introduced q for writing proofs more rigorously.

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

$$\sum_{x \in H} \left| D_1(x) - \widetilde{D}(x) \right| + \left| D_1(\Omega \setminus H) - \widetilde{D}(\Omega \setminus H) \right| \leq (\alpha + \beta + \gamma). \tag{3}$$

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

$$\sum_{x \in H} \left| D_1(x) - \widetilde{D}(x) \right| + \left| D_1(\Omega \setminus H) - \widetilde{D}(\Omega \setminus H) \right| \leq (\alpha + \beta + \gamma), \tag{4}$$

then the distribution D_1 does not satisfy the property $\mathcal{P}$.

Using the above two lemmas, we will prove Theorem 3.1 in Section 3.1. We present the proofs of Lemma 3.3 and Lemma 3.4 in Appendix A.

3.1 Proof of Theorem 3.1

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

1. Draw $W = \mathcal{O}\left(\frac{q^2}{\eta'} \log q\right)$ many samples from the distribution D. Let $S \subseteq \Omega$ be the set of (distinct) samples obtained.
2. Draw additional $\mathcal{O}\left(\frac{W}{\eta'^2} \log W\right)$ many samples Z to estimate the value of $D(x)$ for all $x \in S$ [10].
3. Construct a set H as the union of S and arbitrary q^2 many elements from $\Omega \setminus (S \cup Z)$.
4. Define a distribution $\widetilde{D}$ such that, for $x \in H$,

$$\widetilde{D}(x) = \frac{\#\ x \text{ in the multi-set } Z}{|Z|}.$$

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

$$\widetilde{D}(x) = \frac{1 - \sum\limits_{x \in H} \widetilde{D}(x)}{|\Omega| - |H|}.$$

5. If there exists a distribution D_1 in $\mathcal{P}$ that satisfies both the following conditions:
   (A) $\sum\limits_{x \in H} \left| D_1(x) - \widetilde{D}(x) \right| + |D_1(\Omega \setminus H) - \widetilde{D}(\Omega \setminus H)| \leq 26\eta' + \zeta$.
   (B) $\mathsf{High}_{1/q^2}(D_1) \subseteq H$.
   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.

Note that Step 5 as mentioned above is not completely constructive in a computational sense. In Appendix B, we give a constructive variant of the tester where the property $\mathcal{P}$ can be expressed as a set of linear inequalities. We also give an example of a natural property that can be expressed as a set of linear inequalities.

[10] 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.

Sample Complexity

The sample complexity of the tester is $\mathcal{O}(\frac{q^2}{\eta^2}\log^2 q) = \mathcal{O}(\frac{\Lambda^2\log^2\Lambda}{(\gamma_2-\gamma_1)^2})$, which follows from the above description.

Correctness of the algorithm

The proof of correctness of our algorithm is divided into a sequence of lemmas.

▶ **Lemma 3.5.** *The set H and the distribution $\widetilde{D}$ satisfies the following three properties:*

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

(ii) *For any $x \in H$, if $D(x) \geq \frac{\eta'}{10W}$, $(1-\eta')D(x) \leq \widetilde{D}(x) \leq (1+\eta')D(x)$ holds with probability at least $1-\frac{1}{q^4}$.*

(iii) *For any $x \in \Omega$ with $D(x) \leq \frac{\eta'}{10W}$, either $x \notin H$, or $\widetilde{D}(x) \leq (1+\eta')\frac{\eta'}{10W}$ holds with probability at least $1-\frac{1}{q^4}$.*

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

- **(i)** Consider any $x \in \mathsf{High}_{\eta'/q^2}(D)$, that is, $D(x) \geq \frac{\eta'}{q^2}$. Then the probability that $x \notin H$ is at most $(1-\frac{\eta'}{q^2})^{|H|} \leq \frac{1}{q^4}$. Applying the union bound over all the elements in $\mathsf{High}_{\eta'/q^2}(D)$ (at most $\frac{q^2}{\eta'} = \mathcal{O}(q^3)$ [11] many elements), the claim follows.
- **(ii)** Since $|Z| = \mathcal{O}(\frac{W}{\eta'^2}\log W)$, applying Chernoff bound, we see that $(1-\eta')D(x) \leq \widetilde{D}(x) \leq (1+\eta')D(x)$ does not hold with probability at most $\frac{1}{q^4}$.
- **(iii)** Since $|Z| = \mathcal{O}(\frac{W}{\eta'^2}\log W)$, if x is in H (otherwise, we are already done), applying Chernoff bound (only on one side), the bound follows. ◀

We now bound the ℓ_1-distance between D and $\widetilde{D}$ with respect to H.

▶ **Lemma 3.6.** $\sum_{x\in H}\left|D(x)-\widetilde{D}(x)\right| \leq 5\eta'(1+\eta') \leq 10\eta'$ *holds with probability at least* $1-\frac{3}{q}$.

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

$$\sum_{x\in H}\left|D(x)-\widetilde{D}(x)\right| = \sum_{x\in \mathsf{High}_{\eta'/10W}(D)}\left|D(x)-\widetilde{D}(x)\right| + \sum_{x\in H\setminus \mathsf{High}_{\eta'/10W}(D)}\left|D(x)-\widetilde{D}(x)\right|$$

Applying Lemma 3.5 (*ii*) for each $x \in \mathsf{High}_{\eta'/10W}(D)$, and then using union bound over all such $x \in \mathsf{High}_{\eta'/10W}(D)$, the first term is bounded by η' with probability at least $1-\frac{1}{q}$.

Now the second term, notice that for each $x \in H \setminus \mathsf{High}_{\eta'/10W}(D)$, $D(x) \leq \frac{\eta'}{10W}$. By Lemma 3.5 (iii), and using the union bound over all elements in $H \setminus \mathsf{High}_{\eta'/10W}(D)$ (note that $|H| \leq 2W = \mathcal{O}(q^3)$), with probability at least $1-\frac{2}{q}$, $\widetilde{D}(x) \leq \eta'(1+\eta')/10W$ for all $x \in H \setminus \mathsf{High}_{\eta'/10W}(D)$. Since $|H| \leq 2W$, the second term is bounded by $4\eta'(1+\eta')$ with probability at least $1-\frac{2}{q}$. ◀

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

▶ **Lemma 3.7.** *There exists a distribution $\overline{D}$ such that H contains top q^2 elements of $\overline{D}$. Moreover, the following hold:*

(i) *$||D-\overline{D}||_1 \leq 2\eta'$, with probability at least $1-\frac{2}{q}$.*

[11] This follows from the assumption that $\Lambda(n,\epsilon)$ is at least $\Omega(1/\epsilon)$.

(ii) $\sum_{x \in H} \left|\overline{D}(x) - \widetilde{D}(x)\right| \leq 12\eta'$, *with probability at least* $1 - \frac{5}{q}$.

(iii) $|\overline{D}(\Omega \setminus H) - \widetilde{D}(\Omega \setminus H)| \leq 12\eta'$, *with probability at least* $1 - \frac{5}{q}$.

Proof. Let T be the set of q^2 largest elements of D. If $T \subseteq S$, H (as $S \subset H$) contains the largest q^2 elements of D. In that case, setting $\overline{D}$ to be D gives us the above results.

Now, let us consider the case where $T \not\subseteq S$. By Lemma 3.5 (part (i)), with probability at least $1 - \frac{2}{q}$, $\mathsf{High}_{\eta'/q^2}(D) \subseteq S$. Thus for any $x \in H \setminus S$, $D(x) < \frac{\eta'}{q^2}$. Consider the set $U = T \setminus H$. Notice that since $|H \setminus S| = q^2$ and $|T| = q^2$, $|U| \leq |H \setminus (T \cup S)|$. Let $U = \{y_1, \ldots, y_{|U|}\} \subset \Omega \setminus H$, and let $z_1 \ldots, z_{|U|}$ be some $|U|$ elements of $H \setminus (T \cup S)$. Note, by definition of T and U, the set $\{z_1 \ldots, z_{|U|}\}$ and the set $\{y_1 \ldots, y_{|U|}\}$ are disjoint.

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

- For elements in $\{z_1 \ldots, z_{|U|}\}$, we define $\overline{D}(z_i) = D(y_i)$.
- For elements in $\{y_1 \ldots, y_{|U|}\}$, we define $\overline{D}(y_i) = D(z_i)$.
- For all other x, we define $\overline{D}(x) = D(x)$.

Note that since all the elements in the sets $\{z_1 \ldots, z_{|U|}\}$ and $\{y_1 \ldots, y_{|U|}\}$ were from $\Omega \setminus S$, from Lemma 3.5 (part (i)), with probability at least $1 - \frac{2}{q}$, $D(y_i) \leq \frac{\eta'}{q^2}$ and $D(z_i) \leq \frac{\eta'}{q^2}$, for all $i \in \Omega \setminus S$. Moreover, as $|U| \leq q^2$, we have condition (i) as well. Furthermore, H contains the largest q^2 elements of $\overline{D}$ due to its construction.

Using the triangle inequality (relative to H) along with Lemma 3.6 and the above expression, we can say that, with probability at least $1 - \frac{5}{q}$, (ii) follows.

Let us now prove (iii). Since $\overline{D}$ and $\widetilde{D}$ are distributions, $\sum_{x \in H} \overline{D}(x) + \sum_{x \in \Omega \setminus H} \overline{D}(x) = \sum_{x \in H} \widetilde{D}(x) + \sum_{x \in \Omega \setminus H} \widetilde{D}(x)$. Thus,

$$\left|\overline{D}(\Omega \setminus H) - \widetilde{D}(\Omega \setminus H)\right| = \left|\sum_{x \in H} \widetilde{D}(x) - \sum_{x \in H} \overline{D}(x)\right| \leq \sum_{x \in H} \left|\widetilde{D}(x) - \overline{D}(x)\right| \leq 12\eta'$$

The last inequality follows from (ii). ◀

Now we finally establish the correctness of the algorithm.

Proof of correctness of the algorithm. For completeness, consider the case where D is ζ-close to $\mathcal{P}$. By Lemma 3.7 (i) and the triangle inequality, we know that there exists a distribution $\overline{D}$ that is $(\zeta + 2\eta')$-close to $\mathcal{P}$ and H contain the largest q^2 elements of $\overline{D}$. Since $\sum_{x \in H} \left|\overline{D}(x) - \widetilde{D}(x)\right| \leq 12\eta'$ and $\left|\overline{D}(\Omega \setminus H) - \widetilde{D}(\Omega \setminus H)\right| \leq 12\eta'$ hold from Lemma 3.7 (ii) and (iii), following Lemma 3.4 for $\alpha = 12\eta'$, $\beta = \zeta + 2\eta'$ and $\gamma = 12\eta'$, we can say that there exists a distribution D_1 in $\mathcal{P}$ satisfying Equation (3) (which is same as satisfying **Condition (A)** and **Condition (B)** in Step 5 of the algorithm). Hence, our algorithm accepts D in Step 5.

For soundness, consider a distribution D that is $(\varepsilon + \zeta + \eta)$-far from $\mathcal{P}$. Then following Lemma 3.7 (i), we know that there exists a distribution $\overline{D}$ that is $(\varepsilon + \zeta + \eta - 2\eta')$-far from $\mathcal{P}$, that is, $(\varepsilon + 3\alpha + \beta + 2\gamma)$-far from $\mathcal{P}$, where $\alpha = 12\eta'$, $\beta = \zeta + 2\eta'$. Here, we are using that $\eta = 64\eta'$ and $\gamma = 12\eta'$. Also Lemma 3.7 guarantees that H contains the top q^2 elements of $\overline{D}$. Following Lemma 3.4, we know that there does not exist any such distribution D_1 in $\mathcal{P}$ 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 3.7 does not hold, which is at most $\frac{12}{q}$. ◀

4 Sample complexity of testing non-concentrated label-invariant properties

In this section we first prove a lower bound of $\Omega(\sqrt{n})$ on the sample complexity of non-tolerant testing of any non-concentrated label-invariant property. Then we proceed to prove a tolerant lower bound of $\Omega(n^{1-o(1)})$ samples for such properties in Section 4.2.

4.1 Non-tolerant lower bound (Proof of Theorem 1.3 for label-invariant properties)

Here we first prove a lower bound result analogous to Theorem 1.3, where the properties are non-concentrated and label-invariant. In Section 5, we discuss why the proof of Theorem 4.1 does not directly work for Theorem 1.3, and then prove Theorem 1.3 using a different argument.

▶ **Theorem 4.1** (Analogous result of Theorem 1.3 for non-concentrated label-invariant properties)**.** *Let $\mathcal{P}$ be any (α, β)-non-concentrated label-invariant distribution property, where $0 < \alpha \leq \beta < \frac{1}{2}$. For ε with $0 < \varepsilon < \alpha$, any $(0, \varepsilon)$-tester for property $\mathcal{P}$ 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_{yes} that satisfies the property. Since $\mathcal{P}$ is an (α, β)-non-concentrated property, by Definition 2.7, D_{yes} is an (α, β)-non-concentrated distribution. From D_{yes}, we generate a distribution D_{no} such that the support of D_{no} is a subset of that of D_{yes}, and D_{no} is ε-far from $\mathcal{P}$. Hence, if we apply a random permutation over the elements of Ω, we show that D_{yes} 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 Ω into two parts, depending on the probability mass of D_{yes} on the elements of Ω. Given the distribution D_{yes}, let us first order the elements of Ω according to their probability masses. In this ordering, let L be the smallest $2\beta n$ elements of Ω. 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 4.2.** *For all $x \in L$, $D_{yes}(x) \leq \frac{1-2\alpha}{1-2\beta}\frac{1}{n}$.*

Proof of Observation 4.2. By contradiction, assume that there exists $x \in L$ such that $D_{yes}(x) > \frac{1-2\alpha}{1-2\beta}\frac{1}{n}$. This implies, for every $y \in H$, that $D_{yes}(y) > \frac{1-2\alpha}{1-2\beta}\frac{1}{n}$. So,

$$1 = \sum_{x\in\Omega} D_{yes}(x) = \sum_{x\in L} D_{yes}(x) + \sum_{y\in H} D_{yes}(y) > D_{yes}(L) + |H|\,\frac{1-2\alpha}{1-2\beta}\frac{1}{n}.$$

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

Note that Observation 4.2 implies that if S is a multi-set of $o\left(\sqrt{\frac{1-2\beta}{1-2\alpha}n}\right)$ samples from D_{yes}, then with probability $1 - o(1)$, no element from L appears in S more than once. Now using the distribution D_{yes} and the set L, let us define a distribution D_{no} such that D_{no} is ε-far from $\mathcal{P}$. Note that D_{no} is a distribution that comes from a distribution over a set of distributions, all of which are not (α, β)-non-concentrated. The distribution D_{no} is generated using the following random process:

- We partition L randomly into two equal sets of size βn. Let the sets be $\{x_1, \ldots, x_{\beta n}\}$ and $\{y_1, \ldots, y_{\beta n}\}$. We first pair the elements of L randomly into βn 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_{no} at z is defined as follows:
  - If $z \notin L$, then $D_{no}(z) = D_{yes}(z)$.
  - For every pair $(x_i, y_i) \in P_L$, $D_{no}(x_i) = D_{yes}(x_i) + D_{yes}(y_i)$, and $D_{no}(y_i) = 0$.

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

We will now prove that D_{yes} and D_{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_{yes} from D_{no} with probability at least $\frac{2}{3}$.

Since any D_{no} produced using the above procedure has exactly the same probability mass on the elements in H as D_{yes}, any tester that distinguishes between D_{yes} and D_{no} must rely on samples obtained from L. Recall that the algorithm is given a uniformly random permutation of the distribution. Since $\textsc{Supp}(D_{no}) \subset \textsc{Supp}(D_{yes})$ (particularly, $\textsc{Supp}(D_{no}) \cap L \subset \textsc{Supp}(D_{yes}) \cap L$), it is not possible to distinguish between D_{yes} and D_{no}, unless an element of L appears at least twice. Otherwise, as in the proof of Lemma 3.3, the elements drawn from L are distributed identically to a uniformly random non-repeating sequence. But observe that $D_{yes}(i) = \mathcal{O}(\frac{1}{n})$ and $D_{no}(i) = \mathcal{O}(\frac{1}{n})$ 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. ◀

4.2 Tolerant lower bound (Proof of Theorem 1.4)

▶ **Theorem 4.3** (Theorem 1.4 formalized). *Let $\mathcal{P}$ be any (α, β)-non-concentrated label-invariant distribution property, where $0 < \alpha \leq \beta < \frac{1}{2}$. For any constant ε_1 and ε_2 with $0 < \varepsilon_1 < \varepsilon_2 < \alpha$, any $(\varepsilon_1, \varepsilon_2)$-tester for $\mathcal{P}$ requires $\Omega(n^{1-o(1)})$ 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 [40]. These definitions refer to invariants of distributions, which are essentially a generalization of properties.

▶ **Definition 4.4.** *Let $\Pi : \mathcal{D}_n \to \mathbb{R}$ denote a real-valued function over the set $\mathcal{D}_n$ of all distributions over $[n]$.*

1. Π *is said to be* label-invariant *if for any $D \in \mathcal{D}_n$ the following holds:* $\Pi(D) = \Pi(D_\sigma)$ *for any permutation $\sigma : [n] \to [n]$.*
2. *For any γ, δ with $\gamma \geq 0$ and $\delta \in [0, 2]$,* Π *is said to be* (γ, δ)-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 $\Pi_{\mathcal{P}} : \mathcal{D}_n \to [0, 2]$ with respect to property $\mathcal{P}$ as follows:

For $D \in \mathcal{D}_n, \Pi_{\mathcal{P}}(D) :=$ the distance of D from $\mathcal{P}$.

From the triangle inequality property of ℓ_1 distances, $\Pi_{\mathcal{P}}$ (which refers to the distance function from the property $\mathcal{P}$) is (γ, γ)-weakly continuous, for any $\gamma \in [0, 2]$.

▶ **Theorem 4.5** (Low Frequency Blindness [40]). *Consider a function $\Pi : \mathcal{D}_n \to \mathcal{R}$ that is label-invariant and (γ, δ)-weakly-continuous, where $\gamma \geq 0$ and $\delta \in [0, 2]$. Let there exist two distributions p^+ and p^- in $\mathcal{D}_n$ 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 $\frac{1}{n}$ 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 $\Omega(n^{1-o_\delta(1)})$ many samples from D* [12].

Note that in Theorem 4.5, we have assumed that p^+ and p^- are identical for any index that has probability mass at least $\frac{1}{n}$. We can actually replace this condition to $\mathcal{O}(\frac{1}{n})$ by adding $\mathcal{O}(n)$ many "dummy elements" to the support of p^+ and p^-. Now we are ready to prove Theorem 4.3.

Proof of Theorem 4.3. 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 already noted that $\Pi_\mathcal{P}$ is (γ, γ)-weakly continuous as "*distance from a property*" satisfies the triangle inequality, for any $\gamma \in [0, 2]$. Now recall that the distributions D_{yes} and D_{no} considered in the proof of Theorem 4.1. The probability mass of each element in the support of D_{yes} and D_{no} is $\mathcal{O}(\frac{1}{n})$. Note that D_{yes} is in $\mathcal{P}$ and D_{no} is ε-far from $\mathcal{P}$, for any $\varepsilon < \alpha$, and both of them have a support size of $\Theta(n)$. Here we take $\varepsilon > \varepsilon_2$. Now, we apply Theorem 4.5 with $a = 0$, some $b < \varepsilon$ and γ with $\gamma < \min\{\varepsilon_1, \varepsilon - \varepsilon_2\}$. Observe that this completes the proof of Theorem 4.3. ◀

5 Sample complexity of non-concentrated properties (Proof of Theorem 1.3)

▶ **Theorem 5.1** (Theorem 1.3 formalized). *Let $\mathcal{P}$ be any (α, β)-non-concentrated distribution property for $0 < \alpha < \beta < \frac{1}{2}$. For any ε with $0 < \varepsilon < \alpha$, any $(0, \varepsilon)$-tester for $\mathcal{P}$ requires $\Omega(\sqrt{n})$ many samples, where n is the size of the support of the distribution.*

Why does the proof of Theorem 4.1 work only for label-invariant properties?

Note that the proof of Theorem 4.1 crucially uses the fact that the property $\mathcal{P}$ is label-invariant. Recall that, while constructing D_{no} from D_{yes}, for each $i \in [\beta n]$, moving the masses of both x_i and y_i in D_{yes} to x_i to produce D_{no} is possible as the property $\mathcal{P}$ is label-invariant. Because of this feature, we can apply a random permutation over Ω, and still the permuted distribution will behave identically with respect to $\mathcal{P}$. After applying the random permutation, the samples coming from D_{yes} and D_{no} 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 $\mathcal{P}$ may not be invariant under the random permutation over Ω. This requires a new argument; although the proof is similar in spirit to the proof of Theorem 4.1, there are some crucial differences, and we present the proof next. In order to prove Theorem 5.1, instead of moving the masses of both x_i and y_i in D_{yes} to x_i to produce D_{no}, we randomly move the sum to either x_i or y_i proportionally to the masses of x_i and y_i.

[12] $o_\delta(\cdot)$ suppresses a term in δ.

5.1 Proof of Theorem 5.1

The proof of Theorem 5.1 starts off identically to the proof of Theorem 4.1, but there is a departure in the construction of D_{yes} and D_{no}. Due to shortage of space, we only give the constructions of D_{yes} and D_{no} here, along with a brief sketch of the proof. See the full version of the paper [17] for the complete proof.

Let us first consider D_{yes}, L and P_L as discussed in the proof of Theorem 4.1, only here we cannot and will not pass D_{yes} through a random permutation. The difference starts from the description of the distribution D_{no}. In fact, D_{no} will be randomly chosen according to a distribution over a set of distributions, all of which are not (α, β)-non-concentrated. The distribution D_{no} is generated using the following random process:

- We partition L arbitrarily into two equal sets of size βn. Let the sets be $\{x_1, \dots, x_{\beta n}\}$ and $\{y_1, \dots, y_{\beta n}\}$. We first pair the elements of L arbitrarily into βn pairs. Let $(x_1, y_1), \dots, (x_{\beta n}, y_{\beta n})$ be an arbitrary pairing of the elements in L. Let P_L be the set of pairs. So $P_L = \{(x_1, y_1), \dots, (x_{\beta n}, y_{\beta n})\}$. We refer to x_i and y_i as the corresponding elements of 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_{no} at z is defined as follows:
  - If $z \notin L$, then $D_{no}(z) = D_{yes}(z)$.
  - For every pair $(x_i, y_i) \in P_L$, use independent random coins and
    - With probability $\frac{D_{yes}(x_i)}{D_{yes}(x_i)+D_{yes}(y_i)}$, set $D_{no}(x_i) = D_{yes}(x_i)+D_{yes}(y_i)$ and $D_{no}(y_i) = 0$.
    - With the remaining probability, that is, with probability $\frac{D_{yes}(y_i)}{D_{yes}(x_i)+D_{yes}(y_i)}$, set $D_{no}(x_i) = 0$ and $D_{no}(y_i) = D_{yes}(x_i) + D_{yes}(y_i)$.

Observe that any D_{no} constructed by the above procedure is supported on a set of at most $(1-\beta)n$ elements. So, any distribution D_{no} constructed using the above procedure is ε-far from satisfying the property $\mathcal{P}$, for any $\varepsilon < \alpha$. But since any D_{no} produced using the above procedure has exactly the same probability mass on elements in H as D_{yes}, any tester that distinguishes between D_{yes} and D_{no} 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_{yes} from D_{no}.

6 Learning a distribution (Proof of Theorem 1.5)

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 ℓ_1 distance between D' and D (the unknown distribution) is at most ε, by using $\mathcal{O}(\frac{n}{\varepsilon^2})$ samples from D [13]. 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 $\mathcal{O}(|S|)$ samples (even if the algorithm is unaware of $|S|$ in advance). Our result is formally stated as follows:

[13] There is a writeup of this folklore result by Cannone [13].

▶ **Theorem 6.1** (Theorem 1.5 formalized). *Let D denote the unknown distribution over $\Omega = [n]$, and assume that there exists a set $S \subseteq [n]$ with $D(S) \geq 1 - \frac{\eta}{2}$*[14]*, 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 $\frac{2}{3}$. Moreover, the algorithm uses, in expectation, $\mathcal{O}\left(\frac{|S|}{\delta^2}\right)$ 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 $\mathcal{O}(|S|)$ many samples are enough to approximately learn the distribution D. We would like to point out that similar question has been studied under the local differential privacy model with communication constraints by Acharya, Kairouz, Liu and Sun [4] and by Chen, Kairouz and Özgür [18].

▶ **Lemma 6.2** (Theorem 6.1 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 $\frac{9}{10}$. Moreover, the algorithm uses $\mathcal{O}\left(\frac{s}{\delta^2}\right)$ many samples from D.*

We note that Lemma 6.2 can be obtained from the work of Acharya, Diakonikolas, Li and Schmidt [3] (Theorem 2). For completeness, we give a self-contained proof of this lemma in the full version of the paper [17].

We later adapt the algorithm of Lemma 6.2 to give a proof to the scenario where $|S|$ is unknown, using a guessing technique. The idea is to guess $|S| = s$ starting from $s = 1$, and then to query for $\mathcal{O}(s)$ many samples from the unknown distribution D. From the samples obtained, we construct a distribution D_s, and use Lemma 6.3 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 $\frac{9}{10}$. 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 6.3** ([37]). *Let D_u and D_k denote the unknown and known distributions over $\Omega = [n]$ such that the support of D_u is a set of s elements of $[n]$. Then there exists an algorithm* 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 $\mathcal{O}\left(\frac{1}{(\varepsilon_2 - \varepsilon_1)^2} \frac{s}{\log s} \log \frac{1}{\kappa}\right)$ 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 $1 - \kappa$* [15].

Note that Theorem 6.1 talks about learning a distribution with $\mathcal{O}(s)$ samples, where there exists an unknown set S with s elements and $D(S) \geq 1 - \eta/2$. To prove Theorem 6.1, we use Lemma 6.3 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 [37] is $\mathcal{O}\left(\frac{1}{(\varepsilon_2 - \varepsilon_1)^2} \frac{n}{\log n}\right)$, which holds for any general distributions D_u and D_k with constant success probability. When deploying Lemma 6.3, we "contract" the set $\Omega \setminus \textsc{Supp}(D_k)$ 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

[14] Recall that the variation distance between two distribution is half than that of ℓ_1 distance between them. So, we take $D(S) \geq 1 - \frac{\eta}{2}$ (with $\eta \in [0, 2)$) instead of $D(S) \geq 1 - \eta$ (with $\eta \in [0, 1)$) .

[15] The multiplicative factor $\log \frac{1}{\kappa}$ is for amplifying the success probability from $\frac{2}{3}$ to $1 - \kappa$.

D_u. Hence, $\mathcal{O}\left(\frac{1}{(\varepsilon_2-\varepsilon_1)^2}\frac{s}{\log s}\right)$ samples from D_u are enough for constant success probability. Following a recent work of Cannone, Jain, Kamath and Li [16], 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.

Proof of Theorem 6.1. The algorithm is as follows:

1. Set $s = 1$.
2. Query for a multi-set Z_s of $\mathcal{O}\left(\frac{s}{\delta^2}\right)$ many samples from D.
3. Construct a distribution $D_s : [n] \to [0,1]$ such that
$$D_s(x) = \frac{\#\text{ times } x \text{ appears in } Z_s}{|Z_s|}$$
4. Call the algorithm Tol-Alg $\left(D_s, D, \eta + \frac{\delta}{2}, \eta + \delta, \frac{1}{100\log^2 s}\right)$ (corresponding to Lemma 6.3) to distinguish whether $||D - D_s||_1 \leq \eta + \frac{\delta}{2}$ or $||D - D_s||_1 \geq \eta + \delta$. If we get $||D - D_s||_1 \leq \eta + \frac{\delta}{2}$ as the output of Tol-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.

Let $\mathcal{S}$ denote the event that the algorithm quits with the desired output. We first show that $\Pr(S) \geq \frac{2}{3}$. 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 $||D - D_s||_1 \leq \eta + \frac{\delta}{2}$. 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 $\sum_{k=1}^{\infty} \frac{1}{100k^2} \leq \frac{1}{10}$. So, denoting $\mathcal{Q}$ as the event that the algorithm quits without reporting Failure, $\Pr(\mathcal{S} \mid \mathcal{Q}) \geq \frac{9}{10}$.

For the lower bound on $\Pr(\mathcal{Q})$, consider the case where $s \geq |S|$. In this case, $||D_s - D||_1 \leq \eta + \frac{\delta}{2}$ with probability at least $\frac{9}{10}$, and Tol-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 $\frac{4}{5}$. 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 $1 - (\frac{1}{5})^{(\log n - \log|S|+1)}$. That is, $\Pr(\mathcal{Q}) \geq \frac{4}{5}$.

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

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

Now, we analyze the sample complexity of the algorithm. The algorithm queries for $\mathcal{O}(s)$ 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 $\mathcal{O}\left((\frac{1}{5})^{\lfloor \log s/|S| \rfloor}\right)$. Hence the expected sample complexity of the algorithm is at most

$$\sum_{k: s=2^k<|S|} \mathcal{O}(s) + \sum_{k:s=2^k \geq |S|} \mathcal{O}\left(\left(\frac{1}{5}\right)^{\lfloor \log(s/|S|) \rfloor} \cdot s\right) = \mathcal{O}(|S|).$$

To explain the above equality, note that in the LHS of the above equation, each term of the second sum is bounded by $\mathcal{O}((\frac{1}{5})^{(k-\log|S|)} \cdot 2^{(k-\log|S|)} \cdot |S|)$. Thus, substituting $k - \log(|S|)$ by r, we see that the second part of the LHS is upper bounded by $\sum_{r\geq 0} \mathcal{O}\left((\frac{2}{5})^r \cdot |S|\right)$ which is clearly $\mathcal{O}(|S|)$. Thus we have the above bound. ◀

References

1 Jayadev Acharya, Clément L. Canonne, Cody Freitag, Ziteng Sun, and Himanshu Tyagi. Inference under information constraints III: local privacy constraints. *IEEE J. Sel. Areas Inf. Theory*, pages 253–267, 2021. doi:10.1109/JSAIT.2021.3053569.

2 Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. In *NIPS*, pages 3591–3599, 2015. URL: https://proceedings.neurips.cc/paper/2015/hash/1f36c15d6a3d18d52e8d493bc8187cb9-Abstract.html.

3 Jayadev Acharya, Ilias Diakonikolas, Jerry Li, and Ludwig Schmidt. Sample-optimal density estimation in nearly-linear time. In *SODA*, pages 1278–1289, 2017. doi:10.1137/1.9781611974782.83.

4 Jayadev Acharya, Peter Kairouz, Yuhan Liu, and Ziteng Sun. Estimating sparse discrete distributions under privacy and communication constraints. In *ALT*, pages 79–98, 2021. URL: http://proceedings.mlr.press/v132/acharya21b.html.

5 Maryam Aliakbarpour, Ilias Diakonikolas, Daniel Kane, and Ronitt Rubinfeld. Private testing of distributions via sample permutations. In *NeurIPS*, pages 10877–10888, 2019. URL: https://proceedings.neurips.cc/paper/2019/hash/8e036cc193d0af59aa9b22821248292b-Abstract.html.

6 Noga Alon, Eric Blais, Sourav Chakraborty, David García-Soriano, and Arie Matsliah. Nearly tight bounds for testing function isomorphism. *SIAM J. Comput.*, 42(2):459–493, 2013. doi:10.1137/110832677.

7 Tugkan Batu and Clément L. Canonne. Generalized uniformity testing. In *FOCS*, pages 880–889, 2017. doi:10.1109/FOCS.2017.86.

8 Tugkan Batu, Lance Fortnow, Eldar Fischer, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. Testing random variables for independence and identity. In *FOCS*, pages 442–451, 2001. doi:10.1109/SFCS.2001.959920.

9 Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing that distributions are close. In *FOCS*, pages 259–269, 2000. doi:10.1109/SFCS.2000.892113.

10 Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Mach. Learn.*, pages 151–175, 2010. doi:10.1007/s10994-009-5152-4.

11 Dimitris Bertsimas and John N Tsitsiklis. *Introduction to linear optimization*, volume 6. Athena Scientific Belmont, MA, 1997.

12 Eric Blais, Clément L. Canonne, Talya Eden, Amit Levi, and Dana Ron. Tolerant junta testing and the connection to submodular optimization and function isomorphism. *ACM Trans. Comput. Theory*, pages 24:1–24:33, 2019. doi:10.1145/3337789.

13 Clément L Canonne. A short note on learning discrete distributions. *https://github.com/ccanonne/probabilitydistributiontoolbox/blob/master/learning.pdf*, 2020.

14 Clément L. Canonne. *A Survey on Distribution Testing: Your Data is Big. But is it Blue?* Theory of Computing Library, 2020. doi:10.4086/toc.gs.2020.009.

15 Clément L. Canonne, Ilias Diakonikolas, Themis Gouleakis, and Ronitt Rubinfeld. Testing shape restrictions of discrete distributions. *Theory Comput. Syst.*, pages 4–62, 2018. doi:10.1007/s00224-017-9785-6.

16 Clément L. Canonne, Ayush Jain, Gautam Kamath, and Jerry Li. The price of tolerance in distribution testing. In *COLT*, pages 573–624, 2022. URL: https://proceedings.mlr.press/v178/canonne22a.html.

17 Sourav Chakraborty, Eldar Fischer, Arijit Ghosh, Gopinath Mishra, and Sayantan Sen. Exploring the gap between tolerant and non-tolerant distribution testing. *CoRR*, 2021. arXiv:2110.09972.

18 Wei-Ning Chen, Peter Kairouz, and Ayfer Özgür. Breaking the communication-privacy-accuracy trilemma. In *NeurIPS*, 2020. URL: https://proceedings.neurips.cc/paper/2020/hash/222afbe0d68c61de60374b96f1d86715-Abstract.html.

19 Gregory W Corder and Dale I Foreman. *Nonparametric statistics: A step-by-step approach.* John Wiley & Sons, 2014.

20 Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* Wiley, 2001. doi:10.1002/0471200611.

21 Constantinos Daskalakis, Gautam Kamath, and John Wright. Which distribution distances are sublinearly testable? In *SODA*, pages 2747–2764, 2018. doi:10.1137/1.9781611975031.175.

22 Ilias Diakonikolas and Daniel M. Kane. A new approach for testing properties of discrete distributions. In *FOCS*, pages 685–694, 2016. doi:10.1109/FOCS.2016.78.

23 Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *FOCS*, pages 73–84, 2017. doi:10.1109/FOCS.2017.16.

24 Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Sharp bounds for generalized uniformity testing. In *NeurIPS*, pages 6204–6213, 2018. URL: https://proceedings.neurips.cc/paper/2018/hash/fc325d4b598aaede18b53dca4ecfcb9c-Abstract.html.

25 Eldar Fischer, Oded Lachish, and Yadu Vasudev. Trading query complexity for sample-based testing and multi-testing scalability. In *FOCS*, pages 1163–1182, 2015. doi:10.1109/FOCS.2015.75.

26 Eldar Fischer, Oded Lachish, and Yadu Vasudev. Improving and extending the testing of distributions for shape-restricted properties. In *STACS*, pages 31:1–31:14, 2017. doi:10.4230/LIPIcs.STACS.2017.31.

27 Eldar Fischer and Arie Matsliah. Testing graph isomorphism. *SIAM J. Comput.*, pages 207–225, 2008. doi:10.1137/070680795.

28 Oded Goldreich. *Introduction to Property Testing.* Cambridge University Press, 2017. doi:10.1017/9781108135252.

29 Oded Goldreich. Testing isomorphism in the bounded-degree graph model. *Electron. Colloquium Comput. Complex.*, page 102, 2019. URL: https://eccc.weizmann.ac.il/report/2019/102.

30 Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. *Electron. Colloquium Comput. Complex.*, 7(20), 2000. URL: https://eccc.weizmann.ac.il/eccc-reports/2000/TR00-020/index.html.

31 Oded Goldreich and Dana Ron. On sample-based testers. *ACM Trans. Comput. Theory*, pages 7:1–7:54, 2016. doi:10.1145/2898355.

32 Sivakanth Gopi, Gautam Kamath, Janardhan Kulkarni, Aleksandar Nikolov, Zhiwei Steven Wu, and Huanyu Zhang. Locally private hypothesis selection. In *COLT*, pages 1785–1816, 2020. URL: http://proceedings.mlr.press/v125/gopi20a.html.

33 Terry King. *A guide to chi-squared testing.* Taylor & Francis, 1997.

34 David J. C. MacKay. *Information theory, inference, and learning algorithms.* Cambridge University Press, 2003.

35 Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Trans. Inf. Theory*, pages 4750–4755, 2008. doi:10.1109/TIT.2008.928987.

36 Michal Parnas, Dana Ron, and Ronitt Rubinfeld. Tolerant property testing and distance approximation. *J. Comput. Syst. Sci.*, pages 1012–1042, 2006. doi:10.1016/j.jcss.2006.03.002.

37 Gregory Valiant and Paul Valiant. The power of linear estimators. In *FOCS*, pages 403–412, 2011. doi:10.1109/FOCS.2011.81.

38 Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM J. Comput.*, pages 429–455, 2017. doi:10.1137/151002526.

39 Gregory Valiant and Paul Valiant. Estimating the unseen: Improved estimators for entropy and other properties. *J. ACM*, pages 37:1–37:41, 2017. doi:10.1145/3125643.

40 Paul Valiant. Testing symmetric properties of distributions. *SIAM J. Comput.*, pages 1927–1968, 2011. doi:10.1137/080734066.

41 Huanyu Zhang. Statistical inference in the differential privacy model. *CoRR*, abs/2108.05000, 2021. arXiv:2108.05000.

A Proof of Lemma 3.3 and Lemma 3.4

Proof of Lemma 3.3. We will prove this by contradiction. Let us assume that there are two distributions D_{yes} and D_{no} such that (i) $D_{yes} \in \mathcal{P}$, (ii) D_{no} is ε-far from $\mathcal{P}$, (iii) $\mathsf{High}_{1/q^2}(D_{yes}) = \mathsf{High}_{1/q^2}(D_{no}) = A$, and (iv) for all $x \in A$, $D_{yes}(x) = D_{no}(x)$. Now we argue that any $(0, \varepsilon)$-non-tolerant tester requires more than $\Lambda(n, \varepsilon)$ samples from the unknown distribution D to distinguish whether D is in the property or ε-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: (i) $D_Y(x) = D_{yes}(x)$ for each $x \in A$ and (ii) $D_Y(\pi(x)) = D_{yes}(x)$ 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 $\mathcal{P}$, whereas D_N is ε-far from $\mathcal{P}$, which follows from $\mathcal{P}$ 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 $\Lambda(n, \varepsilon)$ many samples, cannot distinguish D_Y from D_N with probability at least $2/3$. We argue that this claim holds even if the algorithm is provided with additional information about the input: Namely, for all $x \in A$, it 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 $1/q^2$. Let S_Y be a sequence of samples drawn according to D_Y. If $|S_Y| \leq \Lambda(n, \varepsilon)$, then with probability at least $3/4$, 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 $\Lambda(n, \varepsilon)$ distinct elements from $\Omega \setminus A$. Since the elements of $\Omega \setminus A$ were permuted using a uniformly random permutation, with probability at least $3/4$, 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 $3/4$, 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 $1/4$ of each other, which is strictly less than $1/3$.

Hence, if the algorithm obtains at most $\Lambda(n, \varepsilon)$ many samples from the unknown distribution D, it cannot distinguish, with probability at least $2/3$, whether the samples are coming from D_Y or D_N. ◀

To prove Lemma 3.4, we will need the following simple claim, whose proof we omit here.

▷ Claim A.1. 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 $b_1 \geq b_2 \geq \cdots \geq b_n$ and $\sum_{i \in [n]} a_i = \sum_{i \in [n]} b_i = 1$, then the sum $\sum_{i \in [n]} \left| a_i - b_{\sigma(i)} \right|$ is minimized when σ is the identity permutation.

Now we present the proof of Lemma 3.4.

Proof of Lemma 3.4. We consider the two cases separately. **(1)** If $\overline{D}$ is β-close to $\mathcal{P}$, there exists a distribution D_1 in $\mathcal{P}$ such that $\sum_x \left| \overline{D}(x) - D_1(x) \right| \leq \beta$. Since $\mathcal{P}$ is label-invariant, any permutation of D_1 is also in $\mathcal{P}$. Without loss of generality, let us assume that the domain Ω is a subset of $\{1, \ldots, n\}$.

By Claim A.1, the permutation σ that minimizes $\sum_x \left| \overline{D}(x) - D_1(\sigma(x)) \right| \leq \beta$ is the one that orders the i-th largest element of D_1 with the i-th largest element of $\overline{D}$, 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 $\overline{D}$. Consider the distribution D_1^σ that is defined by $D_1^\sigma(x) = D_1(\sigma(x))$.

Clearly, H contains the largest q^2 elements of $\overline{D}$, and hence also $\mathsf{High}_{1/q^2}(D_1^\sigma) \subseteq H$. As $\sum_{x\in\Omega} |D_1^\sigma(x) - \overline{D}(x)| \leq \beta$, $\sum_{x\in H} |\overline{D}(x) - \widetilde{D}(x)| \leq \alpha$ and $|\overline{D}(\Omega \setminus H) - \widetilde{D}(\Omega \setminus H)| \leq \gamma$, by the triangle inequality, we obtain

$$\begin{aligned}
& \textstyle\sum_{x\in H} |D_1^\sigma(x) - \widetilde{D}(x)| + && |D_1^\sigma(\Omega \setminus H) - \widetilde{D}(\Omega \setminus H)| \\
\leq\ & \textstyle\sum_{x\in H} |D_1^\sigma(x) - \overline{D}(x)| + && \sum_{x\in H} |\overline{D}(x) - \widetilde{D}(x)| \\
& && + |D_1^\sigma(\Omega \setminus H) - \overline{D}(\Omega \setminus H)| + |\overline{D}(\Omega \setminus H) - \widetilde{D}(\Omega \setminus H)| \\
\leq\ & \textstyle\sum_{x\in H} |D_1^\sigma(x) - \overline{D}(x)| + && \sum_{x\in H} |\overline{D}(x) - \widetilde{D}(x)| \\
& && + \sum_{x\in\Omega\setminus H} |D_1^\sigma(x) - \overline{D}(x)| + |\overline{D}(\Omega \setminus H) - \widetilde{D}(\Omega \setminus H)| \\
=\ & \textstyle\sum_{x\in\Omega} |D_1^\sigma(x) - \overline{D}(x)| + && \sum_{x\in H} |\overline{D}(x) - \widetilde{D}(x)| + |\overline{D}(\Omega \setminus H) - \widetilde{D}(\Omega \setminus H)| \\
\leq\ & \qquad \alpha + \beta + \gamma
\end{aligned}$$

(2) We will prove this case by contradiction. Let $D_1 \in \mathcal{P}$ be a distribution such that $\mathsf{High}_{1/q^2}(D_1) \subseteq H$ and $\sum_{x\in H} |D_1(x) - \widetilde{D}(x)| + |D_1(\Omega \setminus H) - \widetilde{D}(\Omega \setminus H)| \leq \alpha + \beta + \gamma$. Then, as $\sum_{x\in H} |\overline{D}(x) - \widetilde{D}(x)| \leq \alpha$, by the triangle inequality, we have

$$\sum_{x\in H} |D_1(x) - \overline{D}(x)| + |D_1(\Omega \setminus H) - \widetilde{D}(\Omega \setminus H)| \leq 2\alpha + \beta + \gamma. \tag{5}$$

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

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

Let us first prove that $\mathsf{High}_{1/q^2}(\widehat{D}) \subseteq H$. In the case where $D_1(H) \geq \overline{D}(H)$, for all $x \in \Omega \setminus H$, $\widehat{D}(x) \leq \overline{D}(x)$. Since $\mathsf{High}_{1/q^2}(\overline{D}) \subseteq H$, $\mathsf{High}_{1/q^2}(\widehat{D}) \subseteq H$. Now, in the case where $D_1(H) \leq \overline{D}(H)$, the only $x \in \Omega \setminus H$ for which $\widehat{D}(x) > \overline{D}(x)$ are those in T. Since $|\Omega| > 4q^2$, the lowest $2q^2$ elements on $\overline{D}$ must each have mass less than $\frac{1}{2q^2}$. So even if we add $\frac{1}{2q^2}$ for any element $x \in T$, $\widehat{D}(x) < 1/q^2$. Hence in this case also $\mathsf{High}_{1/q^2}(\widehat{D}) \subseteq H$ since $\mathsf{High}_{1/q^2}(\overline{D}) \subseteq H$ and $\mathsf{High}_{1/q^2}(D_1) \subseteq H$. Now let us bound the ℓ_1 distance between $\widehat{D}$ and $\overline{D}$. Observe that $\sum_{x\in\Omega\setminus H} |\widehat{D}(x) - \overline{D}(x)| = |\widehat{D}(\Omega \setminus H) - \overline{D}(\Omega \setminus H)|$. This is because, in the case where $\widehat{D}(H) \geq \overline{D}(H)$, we have $\widehat{D}(x) = \phi \cdot \overline{D}(x) \leq \overline{D}(x)$ for all $x \in \Omega \setminus H$. On the other hand, in the case where $\widehat{D}(H) \leq \overline{D}(H)$ then for all $x \in \Omega \setminus H$, $\widehat{D}(x) \geq \overline{D}(x)$. Thus,

$$\begin{aligned}
\sum_{x\in\Omega\setminus H} |\widehat{D}(x) - \overline{D}(x)| &= |\widehat{D}(\Omega \setminus H) - \overline{D}(\Omega \setminus H)| \\
&\leq |\widehat{D}(\Omega \setminus H) - \widetilde{D}(\Omega \setminus H)| + |\overline{D}(\Omega \setminus H) - \widetilde{D}(\Omega \setminus H)| \\
&\leq |\widehat{D}(\Omega \setminus H) - \widetilde{D}(\Omega \setminus H)| + \gamma
\end{aligned}$$

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

$$
\begin{aligned}
||\widehat{D}-\overline{D}||_1 &= \sum_{x\in H}|\widehat{D}(x)-\overline{D}(x)| + \sum_{x\in\Omega\setminus H}|\widehat{D}(x)-\overline{D}(x)| \\
&\leq \sum_{x\in H}|\widehat{D}(x)-\overline{D}(x)| + |\widehat{D}(\Omega\setminus H)-\widetilde{D}(\Omega\setminus H)| + \gamma \\
&= (\sum_{x\in H}|D_1(x)-\overline{D}(x)| + |D_1(\Omega\setminus H)-\widetilde{D}(\Omega\setminus H)|) + \gamma \\
&\qquad\qquad \text{(From the construction of } \widehat{D}) \\
&\leq 2\alpha+\beta+2\gamma \quad \text{(By Equation (5))}
\end{aligned}
$$

Moreover, as $\mathsf{High}_{1/q^2}(D_1) \subseteq H$ and by the construction of $\widehat{D}$, we have $\mathsf{High}_{1/q^2}(D_1) = \mathsf{High}_{1/q^2}(\widehat{D})$ and for all $x \in \mathsf{High}_{1/q^2}(D_1)$, $D_1(x) = \widehat{D}(x)$. Since we assumed that D_1 is in $\mathcal{P}$, using Lemma 3.3, $\widehat{D}$ is ε-close to $\mathcal{P}$. And since $||\widehat{D}-\overline{D}||_1 \leq 2\alpha+\beta+2\gamma$, we conclude that $\overline{D}$ is $(\varepsilon+2\alpha+\beta+2\gamma)$-close to $\mathcal{P}$, which is a contradiction. ◀

B Computationally efficient tolerant testers

In this section we present a constructive variant of the tolerant tester studied in Section 3.1. Let us first recall the definitions of *polyhedron* and *projection map*.

▶ **Definition B.1** (Polyhedron). *Let A be a $M \times N$ 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 B.2** (Projection map). *Let n be an integer. For all integers $N \leq n$, a* projection map *is denoted as $\pi_n : \mathbb{R}^N \to \mathbb{R}^n$ and is defined as the projection of the points in $\mathbb{R}^N$ on the first n coordinates.*

Before proceeding to our results, we first define linear property.

▶ **Definition B.3** (Linear property). *Without loss of generality, let us assume $\Omega = [n]$. 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 $\pi_n(\mathcal{LP})$ [16] is the set of distributions satisfying the property $\mathcal{P}$, that is, for every $z := (z_1,\ldots,z_n,\ldots,z_N) \in \mathcal{LP}$, the distribution D_z, defined as $D_z(i) = z_i, \quad \forall i \in [n]$ satisfies the property $\mathcal{P}$. 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 $M \times \max\{N, n\}$.*

Now we give an example of a linear property.

▶ Remark B.4 (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 ε-close to the uniform distribution, where $\varepsilon \in (0,1)$ is a parameter. Consider the following polyhedron $\mathcal{LP}_{u,\varepsilon}$ in $\mathbb{R}^{2n}$:

$$\left(\sum_{i\in[n]} z_{n+i} \leq \varepsilon\right) \quad \bigwedge \quad \left(z_i \geq 0 \ \forall i \in [2n]\right) \quad \bigwedge \quad \left(-z_{n+i} \leq z_i - \tfrac{1}{n} \leq z_{n+i} \quad \forall i \in [n]\right)$$

[16] Note 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 [11]. However, the number of linear inequalities defining the property, which affects the running time of the tester, can sometimes be greatly reduced by using this projection.

Now observe that $\pi_n\left(\mathcal{LP}_{u,\varepsilon}\right)$ will give us the set of distributions that are ε-close to uniform, i.e., the set $\mathcal{P}_{u,\varepsilon}$ (this would serve as the linear transformation mentioned in Definition B.3). Also note that approximate uniformity property has complexity $\mathcal{O}\left(n^2\right)$.

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

$$\Delta\left(k,q,a,\beta\right) := \left\{ x \in \mathbb{R}^d \ : \ \sum_{i=1}^{k} |x_i - a_i| + \left|\sum_{j>k} x_j - \sum_{j>k} a_j\right| \leq \beta \ \wedge \ \forall_{i>k} \ x_i < \frac{1}{q^2} \right\}$$

▶ **Theorem B.5.** *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 γ_1, γ_2 with $\gamma_1 < \gamma_2$ and $0 < \gamma_1 < \gamma_2 + \varepsilon < 2$, there exists a $(\gamma_1, \gamma_2+\varepsilon)$-tester (tolerant tester) that takes $s = \widetilde{\mathcal{O}}(\Lambda^2)$ many samples and makes a single emptiness query to the set $\mathcal{CP} \cap \Delta(\widetilde{\mathcal{O}}(s), \Lambda, \widetilde{D}, \beta)$, where $\widetilde{D}$ is a known probability distribution and $\beta = \gamma_1 + \frac{\gamma_2-\gamma_1}{3}$.*

Proof. Recall that in Step 5 of the tolerant tester presented in Section 3.1, the tester checks whether there is any distribution $D_1 \in \mathcal{P}$ that satisfies the following two conditions:

(i) $\sum_{x\in H} \left|D_1(x) - \widetilde{D}(x)\right| + \left|D_1(\Omega \setminus H) - \widetilde{D}(\Omega\setminus H)\right| \leq 26\eta' + \zeta$

(ii) $\mathsf{High}_{1/q^2}(D_1) \subseteq H$

where $\zeta = \gamma_1$, $\eta = \gamma_2 - \gamma_1$, $\eta = \gamma_2 - \gamma_1$ and $\eta' = \frac{\eta}{64}$. The set H and the distribution $\widetilde{D}$ are defined in the tolerant tester presented in Section 3.1.

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:

$$D_1 \in \mathcal{CP} \text{ and } D_1 \in \Delta\left(|H|, q, \widetilde{D}, 26\eta' + \zeta\right)$$

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 $\eta' = \frac{\eta}{64}$.

1. Draw $W = \mathcal{O}\left(\frac{q^2}{\eta'}\log q\right)$ many samples from the distribution D. Let $S \subseteq \Omega$ be the set of (distinct) samples obtained.
2. Draw additional $\mathcal{O}\left(\frac{W}{\eta'^2}\log W\right)$ 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 arbitrary q^2 many elements from $\Omega \setminus (S \cup Z)$.
4. Define a distribution $\widetilde{D}$ such that, for $x \in H$, $\widetilde{D}(x) = \frac{\#\ x \text{ in the multi-set } Z}{|Z|}$, and for each $x \in \Omega\setminus H$, $\widetilde{D}(x) = \frac{1-\sum\limits_{x\in H}\widetilde{D}(x)}{|\Omega|-|H|}$.
5. If there exists a distribution $D_1 \in \mathcal{CP} \cap \Delta\left(|H|, q, \widetilde{D}, 26\eta' + \zeta\right)$, 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 $\mathcal{O}\left(q^2\log^2 q/\eta^2\right) = \widetilde{\mathcal{O}}(\Lambda^2)$ in addition to a single emptiness query to the set $\mathcal{P} \in \mathcal{CP} \cap \Delta\left(|H|, q, \widetilde{D}, 26\eta' + \zeta\right)$ in Step 5. The correctness proof of the above tester follows from the correctness argument presented in Section 3.1. ◀

B.1 Emptiness checking when $\mathcal{P}$ is a linear property: Proof of Theorem 1.2

Now we proceed to analyze the time complexity of the $(\gamma_1, \gamma_2 + \varepsilon)$-tester described in Theorem B.5 when $\mathcal{P}$ is also a linear property. Recall that as $\mathcal{P}$ is a linear property, there exists a polyhedron $\mathcal{LP} = \left\{x \in \mathbb{R}^N \;\mid\; Ax \leq b\right\}$, where A is a $M \times N$ real matrix and $b \in \mathbb{R}^M$ be a real vector, and $\pi_n\left(\mathcal{LP}\right)$ is the set of distributions satisfying the property $\mathcal{P}$ (See, Definition B.3). Now in Observation B.6, we show that checking emptiness of $\pi_n(\mathcal{LP}) \cap \Delta\left(|H|, q, \widetilde{D}, 26\eta' + \zeta\right)$ is equivalent to testing the feasibility of a family of inequalities.

▶ **Observation B.6.** *Without loss of generality, assume that $H = \{1, \ldots, |H|\}$ and $\Omega = \{1, \ldots, n\}$. Checking emptiness of $\pi_n(\mathcal{LP}) \cap \Delta(|H|, q, \widetilde{D}, 26\eta' + \zeta)$ is equivalent to testing the feasibility of the following set of inequalities:*

$$\left(Az \leq b\right) \quad \bigwedge \quad \left(z_i < \tfrac{1}{q^2} \quad \forall i \in [n] \setminus \{1, \ldots, |H|\}\right)$$

$$\sum_{i=1}^{|H|} \left|z_i - \widetilde{D}(i)\right| + \left|\sum_{i=|H|+1}^{n} z_i - \sum_{i=|H|+1}^{n} \widetilde{D}(i)\right| \leq 26\eta' + \zeta \tag{6}$$

Note that the inequality in Equation (6) can be expressed as the following set of linear inequalities using slack variables z_{N+i} for all $i \in [|H|+1]$:

$$\left(\sum_{i=1}^{|H|} z_{N+i} \;+\; z_{N+|H|+1} \leq 26\eta' + \zeta\right) \quad \bigwedge \quad \left(z_{N+i} \geq 0 \;\; \forall i \in [|H|+1]\right)$$

$$-\, z_{N+i} \leq z_i - \widetilde{D}(i) \leq z_{N+i} \qquad \forall i \in [|H|]$$

$$-\, z_{N+|H|+1} \leq \sum_{i=|H|+1}^{n} z_i - \sum_{i=|H|+1}^{n} \widetilde{D}(i) \leq z_{N+|H|+1}$$

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

$$\left(Az \leq b\right) \bigwedge \left(\sum_{i=1}^{|H|} z_{N+i} + z_{N+|H|+1} \leq 26\eta' + \zeta\right) \bigwedge \left(z_i < \tfrac{1}{q^2} \;\forall i \in [n] \setminus \{1, \ldots, |H|\}\right)$$

$$\left(z_{N+i} \geq 0 \;\; \forall i \in [|H|+1]\right) \quad \bigwedge \quad \left(-\, z_{N+i} \leq z_i - \widetilde{D}(i) \leq z_{N+i} \;\; \forall i \in [|H|]\right)$$

$$-\, z_{N+|H|+1} \leq \sum_{i=|H|+1}^{n} z_i - \sum_{i=|H|+1}^{n} \widetilde{D}(i) \leq z_{N+|H|+1}$$

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., [11]). Thus, we have an efficient $(\gamma_1, \gamma_2 + \varepsilon)$-tester for $\mathcal{P}$, that runs in time polynomial in the complexity of the linear property $\mathcal{P}$ which is also label-invariant. This concludes the proof of Theorem 1.2.