TESTING SIMILAR MEANS

REUT LEVI [*], DANA RON [†], AND RONITT RUBINFELD [‡]

Abstract. We consider the problem of testing a basic property of collections of distributions: having similar means. Namely, the algorithm should accept collections of distributions in which all distributions have means that do not differ by more than some given parameter, and should reject collections that are relatively far from having this property. By 'far' we mean that it is necessary to modify the distributions in a relatively significant manner (according to some predetermined distance measure) so as to obtain the property. We study this problem in two models. In the first model (the *query model*) the algorithm may ask for samples from any distribution of its choice, and in the second model (the *sampling model*) the distributions from which it gets samples are selected randomly. We provide upper and lower bounds in both models. In particular, in the query model, the complexity of the problem is polynomial in $1/\epsilon$ (where ϵ is the given distance parameter), while in the sampling model, the complexity grows roughly as $m^{1-\text{poly}(\epsilon)}$, where m is the number of distributions.

1. Introduction. We consider testing a basic property of collections of distributions: having similar means. Namely, given a collection $\mathcal{D} = (D_1, \ldots, D_m)$ of distributions over $\{0, \ldots, n\}$, and parameters γ and ϵ, we would like to determine whether the means of all distributions reside in an interval of size γn (in which case they have the property of γ-*similar means*), or whether the collection is ϵ-far from having this property. By "ϵ-far" we mean that for every collection $\mathcal{D}^* = (D_1^*, \ldots, D_m^*)$ that has the property, $\frac{1}{m} \sum_{i=1}^{m} d(D_i, D_i^*) > \epsilon$, where $d(\cdot, \cdot)$ is some predetermined distance measure between distributions.

The problem of determining whether a collection of distributions consists of distributions that have similar means arises contexts such as the following: Suppose one is given a collection of coins and would like to determine whether they have the same (or very similar) bias. Alternatively, suppose one would like to compare mean behavior of multiple groups in a scientific experiment. As described above we consider a relaxed version of this problem. One relaxation is in the desired property of the collection. Namely, while one desired property of the collection might be that all the distributions have the same mean, i.e., $\gamma = 0$, it may also be that it is only required that the difference between the maximum mean and minimum mean in the collection is small, i.e., that γ is small. Since the statistical distance between the distribution of samples of a collection that has the property and the distribution of samples of a collection that does not have the property can be infinitesimally small, one can not hope to distinguish between every pair of such instances. Therefore the second relaxation is in weakening the guarantee of the test. More specifically, the test is guaranteed (with high constant probability) to accept collections that have the property, the YES instances, and reject collections that are far from having the property (that is, are far from any collection having the property), the NO instances. If a collections neither has the property nor is far from having the property, the algorithm may either reject or accept. In other words, if the algorithm rejects, then we have high confidence that the collection does not have the property, and if it accepts, then we have high confidence that it is close to having the property. According to

our definition, distance between collections of distributions (and hence to having the property) is defined as the average pairwise distance between corresponding distributions in the collection. This can be viewed as allowing an additional relaxation (as compared to taking the maximum over all pairwise distances), as it implies that a collection is close to having the property if almost every distribution in the collection is close to a distribution whose mean is in the allowed range (rather than requiring this from every distribution).

As we discuss in some more detail in Subsection 1.2, related questions have been studied in the Statistics literature, resulting in particular in the commonly used family of procedures ANOVA (Analysis of Variance), used for deciding whether a collection of normal distributions all have the same mean. As stated above, we consider distributions over a discrete domain but other than that we do not make any assumptions regarding the distributions. Our formulation of the problem falls within the framework of property testing [17, 8, 4], so that in particular it allows for a small fraction of "outlier" distributions.

1.1. Our Contributions. We consider two models, proposed in previous work [13], that describe possible access patterns to multiple distributions $D_1, \ldots, D_m$ over the same domain $\{0, \ldots, n\}$. In the *query model* the algorithm is allowed to specify $i \in \{1, \ldots, m\}$ and receives j that is distributed according to D_i. We refer to each such request for a sample from D_i as a *query*. In the *(uniform) sampling model*, the algorithm receives pairs of the form (i, j) where i is selected uniformly in $\{1, \ldots, m\}$ and j is distributed according to D_i.

The ℓ_1 distance between two probability distributions, $d(D_1, D_2) = \sum_{j=0}^{n} |D_1(j) - D_2(j)|$, is perhaps the most standard measure of distance between distributions, as it measures the maximum difference between the probability of *any event* (i.e., set $S \subseteq \{0, \ldots, n\}$) occurring according to one distribution as compared to the other distribution. In other words, if the distance is small, then the distributions are essentially indistinguishable in terms of their behavior. Hence, we take it as our default distance measure when testing properties of distributions. However, for specific properties one may consider other distance measures that are appropriate. In this study, since the property is related to the means of the distributions and thus the numerical values of the domain elements are meaningful (as opposed to symmetric properties of distributions), we also consider the Earth Mover's Distance (EMD).[1] We prove our upper and lower bounds for the case where the underlying distance measure is the ℓ_1 distance and show by a simple observation that all our results hold for the case which the underlying distance measure is EMD. Hence, unless stated explicitly otherwise, in all that follows the underlying distance measure is the ℓ_1 distance.

Another distance measure between two distributions, that one might consider in the context of testing similar means, is the absolute difference of their means. However, in many natural settings the domain size is bounded, for example, the distributions of grades of students in school, the distribution of the number of rainy days in a year, and so on. Thus, we focus our attention on discrete distributions over a bounded domain for which the above-mentioned distance measure is less applicable, as a pair of distributions can have very close means and nonetheless the mean of one distribution can not get closer to the mean of the other without changing the probability distribution function significantly (for more discussion see Subsection 3.1).

RESULTS IN THE QUERY MODEL. We give an algorithm whose query complexity is $\tilde{O}(1/\epsilon^2)$. which is almost tight as there is a simple lower bound of $\Omega(1/\epsilon^2)$ (even for the $\{0, 1\}$ case).

---

[1] Informally, if the distributions are interpreted as two different ways of piling up a certain amount of earth over the region D, the EMD is the minimum cost of turning one pile into the other, where the cost is assumed to be amount of earth moved times the distance by which it is moved. A formal definition appears in Section 5.

Consider first a basic algorithm that works by obtaining very good estimates of the means of a sufficient number of randomly selected distributions. If the collection is ϵ-far from having γ-similar means, then (with high probability) after performing $\tilde{\Theta}(1/\epsilon^3)$ queries, the algorithm will obtain two distributions whose estimated means are sufficiently far from each other. Thus, this algorithm essentially uses estimates of means as estimates of the distance to having a certain mean.

We design and analyze an improved (almost optimal) algorithm that, roughly speaking, tries to directly estimate the distance to having a certain mean. The more direct estimate is done by estimating means as well, albeit these are means of "mutations" of the original distribution in which the weight of the distribution is either shifted higher or lower. By obtaining such estimates we can apply a "bucketing" technique that allows us to save a factor of $\tilde{\Theta}(1/\epsilon)$ in the query complexity.

RESULTS IN THE SAMPLING MODEL. While in the query model the complexity of the problem of testing similar means has no dependence on the number of distributions, m, this is no longer the case in the sampling model. We prove that the number of samples required is lower bounded by $(1-\gamma)m^{1-\tilde{O}((\epsilon/\gamma)^{1/2})}$. Thus, for any fixed γ (bounded away from 0 and 1), the sample complexity approaches a linear dependence on m as ϵ is decreased. On the positive side, we can show the following. First, by simulating the algorithm designed for the query model, we get an algorithm in the sampling model whose sample complexity is $\tilde{O}(1/\epsilon^2)m^{1-\tilde{\Omega}(\epsilon^2)}$. If we restrict our attention to the case where the domain is $\{0,1\}$, then we can get a better dependence on ϵ in the exponent (at a cost of a higher dependence in the factor that depends only on ϵ). We also observe that (for the $\{0,1\}$ case), if $\gamma < \epsilon/c$ for some sufficiently large constant c, then we can use an algorithm from [13] whose sample complexity is $\mathrm{poly}(1/\epsilon)\sqrt{m}$ (we note that it is not possible to go below $\sqrt{m}$ even for $\gamma = 0$).

In order to prove the abovementioned lower bound we construct a pair of collections of distributions, one that has the property of γ-similar means, and one that is ϵ-far from having this property. We prove that when taking $(1-\gamma)m^{1-\tilde{O}((\epsilon/\gamma)^{1/2})}$ samples, these two collections are indistinguishable. The heart of the proof is the construction of two random variables that on one hand have the same first t moments (for $t = \tilde{O}((\gamma/\epsilon)^{1/2})$) and on the other hand differ in the maximal distance between pairs of elements in the support. These random variables can then be transformed into collections of distributions that cannot be distinguished (with the abovementioned number of samples) but differ in the distance between the maximal and minimal means in the collection. We use tools from [20] and [19] in order to prove that the collections of distributions cannot be distinguished. The construction of the random variables is based on Chebyshev polynomials [6], whose roots, and their derivatives at the roots, have useful properties that we exploit.

EXTENSION TO k INTERVALS. We extend our results for testing whether the means of the distributions in the collection reside in at most k intervals of size at most γn, a property which we refer to as (k, γ)-*similar means*. In particular we provide an algorithm that works in the query model which performs $\tilde{O}(k^2/\epsilon^2)$ queries and an algorithm that works in the sampling model and takes $\tilde{O}(k^2/\epsilon^2)m^{1-\tilde{\Omega}(\epsilon^2/k^2)}$ samples.

1.2. Related Work. The work that is most closely related to the work presented in this paper appears in [13]. The testing models used here were introduced in [13], where the main focus was on the property of equivalence of a collection of distributions. Namely, the goal is to distinguish between the case that all distributions in the collection are the same (or possibly very similar), and the case in which the collection is far from having this property.

The sample complexity (in terms of the dependence on n and m) of testing equivalence grows as $\widetilde{O}(n^{2/3}m^{1/3})$, as long as $n \geq m$, and when $m \geq n$ it grows as $\widetilde{O}(n^{1/2}m^{1/2})$. When the domain of the distributions is $\{0,1\}$, then the problem of testing similar means for $\gamma = 0$ is the same as testing whether the distributions are equivalent. Therefore, an algorithm with sample complexity $\text{poly}(1/\epsilon)\sqrt{m}$ that is given in [13] for testing equivalence in the sampling model, carries over directly to our problem, when $\gamma = 0$ and the domain is $\{0,1\}$). In fact, a *tolerant* version of this algorithm [13] implies the same complexity for $\gamma \leq \epsilon/c$ for a sufficiently large constant c. However, these results do not have any implications for larger γ, and the problems are very different when the domain is larger.

Testing and approximating properties of single and pairs of distributions has been studied quite extensively in the past (see [9, 15, 4, 3, 1, 2, 11, 5, 16, 20, 19]).

Statistical techniques for determining whether sets of populations have the same mean are in wide use. Paired difference tests, and in particular the Student's and Welch's t-tests, are commonly used to study whether the means of two normally distributed populations are equal [14, 18, 21]. The family of procedures ANOVA (Analysis of Variance), applies when there are more than two normally distributed populations (see [12, Chapter 12]), where the difficulty is that the pairwise comparison of all the populations increases the chance of incorrectly failing collections of populations that do in fact all have the same mean, i.e., rejecting the null hypothesis. In all of these procedures, the assumptions made are quite strong, e.g., assuming the normality of the distributions and assuming that all distributions have the same variance. In the settings of ANOVA, the user determines the significance level of the test, α, i.e., the probability to reject the null hypothesis when it is true. The guarantee is the following: When the null hypothesis is false, the probability that the test rejects is at least α. While the probability to accept when the null hypothesis is false decreases as the number of samples and the significance level of the test increases, there is no explicit guarantee on the probability (except that it is smaller than $1 - \alpha$). Furthermore, regardless of the significance level of the test, the user is required to provide samples from every distribution in the collection.

1.3. Extensions and Further Research. A natural question that arises is what is the *exact* complexity of testing γ-similar means in the sampling model. One sub-question is the dependence on ϵ in the exponent of m. Perhaps a more intriguing question is the dependence on γ. Namely, our lower bound (which also holds for the domain $\{0,1\}$) are meaningful only when γ is (sufficiently) larger than ϵ. As noted previously, when $\gamma \leq \epsilon/c$ (for some constant c), the testing problem can be solved for the domain $\{0,1\}$ using a number of samples that grows as $\sqrt{m}$. The question is whether for the domain $\{0,1\}$, and possibly more generally, we can give an algorithm whose complexity is a function of ϵ/γ rather than ϵ, so that, in particular, the complexity decreases as γ decreases. To gain some intuition on why for the domain $\{0,1\}$ the complexity decreases as γ decreases consider for every collection of distributions the corresponding vector of frequencies of means in the collection. When γ is sufficiently smaller than ϵ, we show that the ℓ_2 norm (or the second moment) of the vector of frequencies of means to be sufficiently different between the NO instances and the YES instances, implying that it is easy to distinguish between a YES instance and a NO instance. However, when γ is sufficiently greater than ϵ, we construct a pair of collections, a YES instance and a NO instance, where both collections have corresponding vectors of frequencies with identical first, second, and higher moments, implying that the complexity of the testing problem becomes higher.

We also note that our results (in particular in the query model) easily extend to a generalization of the similar-means problem, where we ask whether the distributions can be *clustered* into at most k clusters, such that within each clusters all distributions have means that differ

by at most γn.

2. Preliminaries. Let $\mathcal{D} = (D_1, \ldots, D_m)$ be a collection of m distributions, where $D_i : \{0, \ldots, n\} \to [0, 1]$ and $\sum_{j=1}^{n} D_i(j) = 1$ for every $1 \le i \le m$. For a vector $\mathbf{v} = (v_1, \ldots, v_n) \in \mathbb{R}^n$, let $\|\mathbf{v}\|_1 = \sum_{i=1}^{n} |v_i|$ denote the ℓ_1 norm of the vector v.

Following [13], for a property $\mathcal{P}$ of collections of distributions and $0 \le \epsilon \le 1$, we say that $\mathcal{D}$ is ϵ-*far from* (*having*) $\mathcal{P}$ if $\frac{1}{m} \sum_{i=1}^{m} \|D_i - D_i^*\|_1 > \epsilon$ for every collection $\mathcal{D}^* = (D_1^*, \ldots, D_m^*)$ that has the property $\mathcal{P}$ (note that $\|D_i - D_i^*\|_1$ is twice the the statistical distance between the two distributions).

Given a distance parameter ϵ, a testing algorithm for a property $\mathcal{P}$ should distinguish between the case that $\mathcal{D}$ has the property $\mathcal{P}$ and the case that it is ϵ-far from $\mathcal{P}$. We consider two models within which this task is performed.

1. **The Query Model.** In this model the testing algorithm may indicate an index $1 \le i \le m$ of its choice and it gets a sample j distributed according to D_i. We refer to each such request as a *query*.
2. **The Sampling Model.** In this model the algorithm cannot select (query) distributions of its choice. Rather, it may obtain a pair (i, j) where i is selected uniformly (we refer to this as the *uniform sampling model*) and j is distributed according to D_i.

For a distribution D over the domain $\{0, \ldots, n\}$, let $\mu(D) \overset{\text{def}}{=} \frac{1}{n} \sum_{j=1}^{n} i \cdot D(j)$. For a collection $\mathcal{D} = (D_1, \ldots, D_m)$ of distributions over $\{0, \ldots, n\}$ let $\mu_j = \mu(D_j)$. For a parameter $0 \le \gamma \le 1$, we say that $\mathcal{D}$ has γ-*similar means* if $|\mu_j - \mu_{j'}| \le \gamma n$ for every $j \ne j'$.

3. Results for the Query Model. In this section we provide an algorithm for testing γ-similar means in the query model, and give an almost matching simple lower bound.

For a distribution D over $\{0, \ldots, n\}$, we shall use the notation $\mu(D) \overset{\text{def}}{=} \sum_{i=1}^{n} i \cdot D(i)$ for the mean of D. for a value $0 \le z \le n$ let $d_1(D, z) \overset{\text{def}}{=} \min_{D':\mu(D')=z} \{\|D - D'\|_1\}$ denote the minimum ℓ_1 distance between D and a distribution that has mean z.

3.1. A Basic Algorithm. Consider first a basic algorithm that works by randomly selecting $\Theta(1/\epsilon)$ distributions, and estimating each of their means to within $O(\epsilon n)$ additive error. This can be done by querying each selected distribution so as to obtain $\tilde{O}(1/\epsilon^2)$ samples from each. The resulting query complexity is $\tilde{O}(1/\epsilon^3)$. The correctness of this algorithm is based on Lemma 1 (stated below), which gives an upper bound on the "cost", in terms of the ℓ_1-distance, for modifying the mean of a distribution by ϵn. Note that in general this cost is not necessarily linear in ϵ. For example, consider the case in which ϵn is an integer and D has all its weight on $n(1 - \epsilon)$, so that $\mu(D) = n(1 - \epsilon)$. Suppose we want to *increase* D's mean by ϵn. The only distribution whose mean is n is the distribution whose weight is all on n, and the ℓ_1 distance between D and this distribution is 1. On the other hand, if we wanted to *decrease* the mean of D by ϵn, then this can easily be done with a cost linear in ϵ, by moving $\epsilon/(1 - \epsilon)$ weight from $n(1 - \epsilon)$ to 0.

LEMMA 1. *Let D be a distribution over $\{0, \ldots, n\}$, let $\mu = \mu(D)$, and let $\alpha \le 1/16$. If $\mu \ge n/2$, then for every $\mu' \in [\mu - \alpha n, \mu]$ there exists D' such that $\mu(D') = \mu'$ and $\|D - D'\|_1 \le 4\alpha$. If $\mu \le n/2$, then for every $\mu' \in [\mu, \mu + \alpha n]$ there exists D' such that $\mu(D') = \mu'$ and $\|D - D'\|_1 \le 4\alpha$.*

Proof. Let $X \sim D$. In the case that $\mu \ge n/2$ we have that $\Pr[X \ge n/4] \ge 1/4 \ge 4\alpha$. Therefore, we can decrease the mean of D by at most αn by moving at most 4α weight from its support on points in the interval $[n/4, n]$, to 0. In the case that $\mu \le n/2$, we have that

$\Pr[X \leq 3n/4] > 1/3 \geq 4\alpha$. Therefore, we can increase the mean of D by at most αn by moving at most 4α weight from its support on the interval $[0, 3n/4]$, to n. □

Lemma 2 can be shown to follow from Lemma 1.

LEMMA 2. *Let $\mathcal{D}$ be a collection of distributions. If $\mathcal{D}$ is ϵ-far from having γ-similar means, then there exists an interval $[x, y] \subseteq [n]$ where $y - x = \gamma n$ such that $\sum_{i:\mu(D_i)>y} d_1(D_i, y) > (\epsilon/4)m$ and $\sum_{i:\mu(D_i)<x} d_1(D_i, x) > (\epsilon/4)m$.*

Proof. Let $\mathcal{D}$ be a collection which is ϵ-far from having γ-similar means. By definition, for every interval $[x, y]$ such that $y - x \leq \gamma n$ we have that $\sum_{i:\mu(D_i)<x} d_1(D_i, x) + \sum_{i:\mu(D_i)>y} d_1(D_i, y) > \epsilon m$. We claim that for every interval $[x, y]$ such that $y - x < \gamma n + (\epsilon/8)n$, we have that

$$\sum_{i:\mu(D_i)<x} d_1(D_i, x) + \sum_{i:\mu(D_i)>y} d_1(D_i, y) > (\epsilon/2)m . \tag{3.1}$$

This is true because otherwise, we could modify the distributions so that their means belong to an interval $[x', y'] \subseteq [x, y]$ of size γn at a total cost upper bounded by ϵm. In particular, we would first increase the means of all those distributions D_i such that $\mu(D_i) < x$ to x, and decrease the means of all D_i such that $\mu(D_i) > y$ to y, at a total cost of $(\epsilon/2)m$. We then increase/decrease the means further by at most $(\epsilon/8)n$ each using Lemma 1.

Let x be the minimum value of the form $\ell \cdot (\epsilon/8)n$, where ℓ is an integer, such that $\sum_{i:\mu(D_i)<x} d_1(D_i, x) > (\epsilon/4)m$, and let $y = x + \gamma n$. By the definition of x, if we consider $x' = x - (\epsilon/16)n$, then $\sum_{i:\mu(D_i)<x'} d_1(D_i, x') \leq (\epsilon/4)m$, thus by Equation 3.1, $\sum_{i:\mu(D_i)>y} d_1(D_i, y) > (\epsilon/4)m$ as desired. □The correctness of the basic algorithm follows

from Lemma 2: If $\mathcal{D}$ is ϵ-far from having γ-similar means, then there exist an interval $[a, b]$ of size γn such that at least $\Omega(\epsilon m)$ distributions in the collection have mean that is at most $a - \epsilon/c$ and at least $\Omega(\epsilon m)$ distributions in the collection have mean that is at least $b + \epsilon/c$, for some absolute constant c. Therefore, by selecting $\Theta(1/\epsilon)$ distributions and estimating the means of each to within $O(\epsilon n)$, with high constant probability the algorithm finds evidence for a pair of distributions with means outside both sides of the interval defined in Lemma 2. On the other hand, if $\mathcal{D}$ has γ-similar means, then the probability of such an event is small.

3.2. An Improved Algorithm. We can modify the basic algorithm so as to obtain a lower complexity (which we later show is almost optimal). One ingredient of the modification (similar to that applied for example in [10]) is roughly the following. Consider the following two (extreme) cases where the collection is ϵ-far from having γ-similar means. In the first case, there is an interval $[x, y]$ of size $\gamma n + 2\epsilon n$, such that half of the distributions have mean x and half of the distributions have mean y. If we select just a constant number of distributions, and for each we estimate its mean to within $\epsilon n/2$, then we shall have sufficient evidence for rejection. In the second case, all but $2\epsilon m$ of the distributions have a mean that resides in an interval of size γn, say, $[0, \gamma n]$ and the remaining $2\epsilon m$ distributions have a mean of n. In this case we need to sample $\Theta(1/\epsilon)$ distributions so as to "hit" one of the high-mean distributions, but then it suffices to take a constant size sample so as to detect that it has a high mean.

If the distributions were over $\{0, 1\}$, then by generalizing the above discussion we can get a certain trade-off between the number of selected distributions and the required quality of the estimate of their means. When dealing with general domains, estimating the means might not suffice. As noted previously, a distribution might have a mean that is very close to a certain value, while the distribution is very far, in terms of the ℓ_1 distance, from any

distribution that has this mean. Therefore, rather than estimating means as a "proxy" for estimating the ℓ_1 distance to having a certain mean, we estimate the latter directly.

SOME NOTATION. To make the above notion of estimation more precise, we introduce some notation. Consider any value $0 \leq \beta \leq 1$ and distribution D. If $d_1(D, n) \geq \beta$, (where $d_1(\cdot, \cdot)$ is as defined at the beginning of this section), then define $\mu_\beta^>(D)$ to be the value $\mu > \mu(D)$ for which $d_1(D, \mu) = \beta$. If $d_1(D, n) < \beta$, then $\mu_\beta^>(D) \stackrel{\text{def}}{=} n$. That is, $\mu_\beta^>(D)$ is the largest mean that can be obtained if we are allowed to modify D so that the resulting distribution is at distance (at most) β from D. Similarly, if $d_1(D, 0) \geq \beta$, then we define $\mu_\beta^<(D)$ to be the value $\mu < \mu(D)$ for which $d_1(D, \mu) = \beta$, and if $d_1(D, 0) < \beta$, then $\mu_\beta^<(D) \stackrel{\text{def}}{=} 0$. Observe that if the domain is $\{0, 1\}$, then $\mu_\beta^>(D) = \min\{\mu(D) + \beta, 1\}$ and $\mu_\beta^<(D) = \max\{\mu(D) - \beta, 0\}$ (while if the domain is larger, then $\mu_\beta^>(D) - \mu(D)$ and $\mu(D) - \mu_\beta^<(D)$ might be much smaller than βn even when $d_1(D, n)$ and $d_1(D, 0)$ are large).

We first describe an algorithm that given sampling access to a distribution D and a parameter β, outputs a pair of estimates such that with high probability one is between $\mu(D)$ and $\mu_\beta^>(D)$ and the other is between $\mu_\beta^<(D)$ and $\mu(D)$. In particular, if the domain is $\{0, 1\}$, then one estimate is between $\mu(D)$ and $\mu(D) + \beta$ and the other is between $\mu(D) - \beta$ and $\mu(D)$. For $n > 1$ the estimates may be required to be much closer to $\mu(D)$. The number of samples that the algorithm takes is quadratic in $1/\beta$. Note that if the domain is $\{0, 1\}$ (or any constant size domain), then the algorithm can simply estimate the mean of D to within β. However, for a general domain, the algorithm is different. We later show how to apply this algorithm so as to obtain a testing algorithm with query complexity $\tilde{O}(1/\epsilon^2)$. The testing algorithm is based on showing that if $\mathcal{D}$ is ϵ-far from having γ-similar means then we can efficiently obtain two distributions D_1 and D_2 for which the following holds. There exist an interval $[x, y]$ of size greater than γn such that $\mu(D_1) < x$, $\mu(D_2) > y$ and for certain (relatively large) values β_1, β_2 we have that $d_1(D_1, x) > \beta_1$ and $d_1(D_2, y) > \beta_2$. This implies that $x > \mu_{\beta_1}^>(D_1)$ and $y < \mu_{\beta_2}^<(D_2)$. Therefore, if we apply the algorithm to D_1 with β_1, and to D_2 with β_2, we shall obtain (with high probability) points x_1 and y_2, where $\mu(D_1) \leq x_1 \leq \mu_{\beta_1}^>(D) \leq x$ and $y \leq y_2 \leq \mu_{\beta_2}^<(D_2) \leq \mu(D_2)$. These points can serve as evidence that $\mathcal{D}$ does not have γ-similar means.

THE ALGORITHM. The high-level idea behind the algorithm is that we consider certain "mutations" of the distribution D. One mutation has a higher mean than D but not higher than $\mu_\beta^>(D)$, and the other has a lower mean than D but not lower than $\mu_\beta^<(D)$. Moreover, the mean of the first is bounded away from both $\mu(D)$ and $\mu_\beta^>(D)$, and the mean of the second is bounded away from $\mu(D)$ and $\mu_\beta^<(D)$. By estimating the means of these distributions sufficiently well, we get the desired outcome of the algorithm. Since $\mu_\beta^>(D) - \mu(D)$ might be much smaller than βn (and similarly $\mu(D) - \mu_\beta^<(D)$) it is not immediately clear what we gain as compared to simply estimating $\mu(D)$. The gain is due to the fact that the support of each of the two distributions is contained in an interval that is strictly smaller than $[0, n]$, and it suffices to estimate each of their means to within β/c *times the size of the corresponding interval* (for some constant c) so as to ensure that the estimate is in the range $[\mu(D), \mu_\beta^>(D)]$ (similarly, the range $[\mu_\beta^<(D), \mu(D)]$). More details follow, where we focus here on the first mutation.

Consider a distribution D. For any $0 \leq a \leq a' \leq n$,

$$\sum_{i > a} i \cdot D(i) + \sum_{i \leq a} a' \cdot D(i) \geq \mu(D) . \tag{3.2}$$

On the other hand, by the definition of $\mu_\beta^>(D)$, if a is such that $\Pr_D[i \leq a] \leq \beta$, then (setting $a' = n$),

$$\sum_{i>a} i \cdot D(i) + \sum_{i \leq a} n \cdot D(i) \leq \mu_\beta^>(D) . \tag{3.3}$$

Let a indeed be a value that satisfies $\Pr_D[i \leq a] \leq \beta$. If we could obtain the *exact* evaluation of the expression in Equation (3.3), which corresponds to the mean of a distribution that is a certain modification (mutation) of D, then we would have a value between $\mu(D)$ and $\mu_\beta^>(D)$. However, if we can obtain only an (additive) estimate of this value, then it might not satisfy the requirements of the algorithm. Hence we further modify Equation (3.3) as follows. Let $a' = a + (n-a)/2 = (a+n)/2$ and let

$$\mu^a(D) \stackrel{\text{def}}{=} \sum_{i>a} i \cdot D(i) + a' \cdot \Pr_D[i \leq a] . \tag{3.4}$$

Then on one hand

$$\mu^a(D) \geq \mu(D) + ((n-a)/2) \cdot \Pr_D[i \leq a] \tag{3.5}$$

and on the other hand

$$\mu^a(D) \leq \mu_\beta^>(D) - ((n-a)/2) \cdot \Pr_D[i \leq a] . \tag{3.6}$$

If $\Pr_D[i \leq a] \geq \beta/c$ for some constant c, then by estimating $\mu^a(D)$ to within an additive error of $(n-a)\beta/4c$, we get a value between $\mu(D)$ and $\mu_\beta^>(D)$. Since $\mu^a(D)$ is the mean of a distribution (we describe this distribution formally in the proof of Lemma 3) whose support is in the interval $[a, n]$, this can be done by taking a sample of size $\Theta(1/\beta^2)$. A technical issue that arises is that it is possible that no such value a exists because $\Pr_D[i = a]$ is relatively large. But then we can slightly modify the definition of $\mu^a(D)$ and still obtain the desired estimate. A similar argument can give us $\mu_\beta^<(D) \leq y \leq \mu(D)$ (with high probability).

Algorithm 1 GetBounds(D, β, δ)

1. Take a sample of size $s_1 = \Theta(\log(1/\delta)/\beta)$ from D and let $i_1 \leq \cdots \leq i_{s_1}$ be the selected points (ordered from small to large).
2. Set $a = i_{(\beta/4)s_1}$, $b = i_{(1-\beta/4)s_1}$, $a' = (a+n)/2$ and $b' = b/2$.
3. Take a sample of size $s_2 = \Theta(\log(1/\delta)/\beta^2)$ from D. Let $\hat{\alpha}(a)$ be the fraction of sampled points $i = a$ and let $\hat{\alpha}(b)$ be the fraction of sampled points $i = b$.
4. If $\hat{\alpha}(a) \leq \beta/4$, then let $a'' = a'$, else let $a'' = \frac{\beta}{4\hat{\alpha}(a)} \cdot a' + (1 - \frac{\beta}{4\hat{\alpha}(a)}) \cdot a$. Similarly, if $\hat{\alpha}(b) \leq \beta/4$ then let $b'' = b'$, else let $b'' = \frac{\beta}{4\hat{\alpha}(b)} \cdot b' + (1 - \frac{\beta}{4\hat{\alpha}(b)}) \cdot b$.
5. Take a sample of size $s_3 = \Theta(\log(1/\delta)/\beta^2)$ from D, and denote the sampled points by $i_1, \ldots, i_{s_3}$. Let $u = \frac{1}{s_3} \left(\sum_{i_j > a} i_j + \sum_{i_j < a} a' + \sum_{i_j = a} a'' \right)$ and $\ell = \frac{1}{s_3} \left(\sum_{i_j < b} i_j + \sum_{i_j > b} b' + \sum_{i_j = b} b'' \right)$.
6. Return (u, ℓ).

LEMMA 3. *The algorithm* GetBounds(D, β, δ) *returns u and ℓ such that with probability at least $1 - \delta$ (over its internal coin flips), $\mu(D) \leq u \leq \mu_\beta^>(D)$ and $\mu_\beta^<(D) \leq \ell \leq \mu(D)$.*

Proof. We prove the claim for u, and an analogous (symmetric) analysis holds for ℓ. Let

a, a', a'' be as determined in the algorithm GetBounds, and let

$$\widetilde{D}(i) \stackrel{\text{def}}{=} \begin{cases} D(i) & \text{if } i > a, i \neq a', i \neq a'' \\ D(i) + \Pr_D[i < a] & \text{if } i = a' \\ D(i) + D(a) & \text{if } i = a'' \\ 0 & \text{o.w.} \end{cases} \tag{3.7}$$

By the definition of the distribution $\widetilde{D}$ we have that

$$\mu(\widetilde{D}) = \sum_{i>a} i \cdot D(i) + a' \cdot \Pr_D[i < a] + a'' \cdot \Pr_D[i = a] . \tag{3.8}$$

Observe that in Step 5, the algorithm takes s_3 independent samples from $\widetilde{D}$ and that $\mathrm{E}[u] = \mu(\widetilde{D})$. We note that the support of $\widetilde{D}$ may not be only on points in $\{0, 1, \ldots, n\}$ (as is the case with D), but this is immaterial to the argument.

By a multiplicative Chernoff bound, with probability at least $1 - \delta/4$ (for a sufficiently large constant in the Θ notation for s_1) we have that $\Pr_D[i < a] \leq \beta/3$ and $\Pr_D[i \leq a] \geq \beta/8$. To verify this, for any $0 < \alpha < 1$, let $e_\alpha(D) \in \{0, \ldots, n\}$ be such that $\Pr_D[i < e_\alpha(D)] < \alpha$ and $\Pr_D[i \leq e_\alpha(D)] \geq \alpha$. Given that $s_1 = c \log(1/\delta)/\beta$, the probability that a sample of size s_1 contains less than $(\beta/4)s_1$ points $i \leq e_{\beta/3}(D)$ or at least $(\beta/4)s_1$ points $i < e_{\beta/8}(D)$ is at most $\delta/4$ (for a sufficiently large constant c). But if neither of these events occur, then, given the choice of a in Step 2, we have that $\Pr_D[i < a] \leq \beta/3$ and $\Pr_D[i \leq a] \geq \beta/8$, as claimed.

Next, by an additive Chernoff bound, with probability at least $1 - \delta/4$ (for a sufficiently large constant in the Θ notation for s_2) we have that $\Pr_D[i = a] - \beta/4 \leq \hat{\alpha}(a) \leq \Pr_D[i = a] + \beta/4$. From this point on assume that the above inequalities indeed hold. If $\hat{\alpha}(a) \leq \beta/4$ (so that $\Pr_D[i \leq a] \leq \beta/3 + \beta/4 + \beta/4 < \beta$), then (as explained in the discussion preceding the algorithm), on one hand,

$$\mu(\widetilde{D}) \geq \mu(D) + \frac{n-a}{2} \cdot \Pr_D[i \leq a] \geq \mu(D) + (n-a) \cdot (\beta/16) , \tag{3.9}$$

and on the other hand

$$\mu(\widetilde{D}) < \mu^>(D) - \frac{n-a}{2} \cdot \Pr_D[i \leq a] \leq \mu^>(D) - (n-a) \cdot (\beta/16) . \tag{3.10}$$

If $\hat{\alpha}(a) > \beta/4$ (so that $\Pr_D[i < a] + \min\{1, \beta/4\hat{\alpha}(a)\} \cdot \Pr_D[i = a] \leq \beta/3 + \beta/2 < \beta$), then

$$\mu(\widetilde{D}) \geq \mu(D) + \frac{n-a}{2} \cdot \left(\Pr_D[i < a] + \frac{\beta}{4\hat{\alpha}(a)} \cdot \Pr_D[i = a] \right) \geq \mu(D) + \frac{n-a}{2} \cdot \frac{\beta}{32} \tag{3.11}$$

and similarly $\mu(\widetilde{D}) < \mu_\beta^>(D) - (n-a) \cdot (\beta/32)$. By the definition of u and an additive Chernoff bound, with probability at least $1 - \delta/4$ (for a sufficiently large constant in the Θ notation for s_3), we have that $|u - \mu(\widetilde{D})| \leq (n-a) \cdot (\beta/32)$ implying that $\mu(D) \leq u \leq \mu_\beta^>(D)$. $\square$

THEOREM 1. *Algorithm 2 tests γ-similar means in the query model. The algorithm's query complexity is $O(\log^2(1/\epsilon)/\epsilon^2)$.*

Proof. Let E_g denote the event that all pairs (u_j^q, ℓ_j^q) returned by the algorithm Get-Bounds are as specified in Lemma 3. Since each call to GetBounds in iteration q is done with

---

Algorithm 2 testing γ-similar means

---

1. For $q = 1$ to r, where $r = \lceil \log(8/\epsilon) \rceil$ do:
   - Select $t(q) = \Theta\left(2^q \log(1/\epsilon)\right)$ distributions from the collection, and denote them by $D_1^q, \ldots, D_{t(q)}^q$.
   - For each D_j^q selected let $(u_j^q, \ell_j^q) = \text{GetBounds}\left(D_j^q, (\epsilon/8)2^{q-1}, \frac{1}{(6rt(q))}\right)$
2. Let $\hat{x} = \max_{q,j}\{u_j^q\}$ and $\hat{y} = \min_{q,j}\{\ell_j^q\}$. If $\hat{y} - \hat{x} > \gamma n$, then REJECT, otherwise, ACCEPT.

$\delta = 1/(6rt(q))$, by Lemma 3 the probability that E_g holds is at least $5/6$. If $\mathcal{D}$ has γ-similar means, then, conditioned on E_g, the algorithm accepts.

We now turn to the case that $\mathcal{D}$ is ϵ-far from having γ-similar means. Let $[x, y]$ be an interval as described in Lemma 2. We partition the distributions D_i such that $\mu(D_i) < x$ into *buckets* B_q^L, for $1 \leq q \leq r$, where

$$B_q^L = \{i : (\epsilon/8)2^{q-1} < d_1(D_i, x) \leq (\epsilon/8)2^q\}, \tag{3.12}$$

and similarly we partition the distributions D_i such that $\mu(D_i) > y$ into buckets B_q^R, where

$$B_q^R = \{i : (\epsilon/16)2^{q-1} < d_1(D_i, y) \leq (\epsilon/8)2^q\}. \tag{3.13}$$

Since

$$\sum_{i:\mu(D_i)<x} d_1(D_i, x) > (\epsilon/4)m \tag{3.14}$$

and

$$\sum_{i:d_1(D_i,x)\leq\epsilon/8} d_1(D_i, x) \leq (\epsilon/8)m \tag{3.15}$$

we have that

$$\sum_{i:\mu(D_i)<x,d_1(D_i,x)>\epsilon/8} d_1(D_i, x) \geq (\epsilon/8)m, \tag{3.16}$$

and so there exists an index q^L such that

$$\sum_{i\in B_{q^L}^L} d_1(D_i, x) \geq (\epsilon/8)m/r \tag{3.17}$$

(where r is the number of buckets). By the definition of the buckets, $r = O(\log(1/\epsilon))$, and for the bucket $B_{q^L}^L$ we have

$$|B_{q^L}^L| > \frac{(\epsilon/8)m/r}{(\epsilon/8)2^{q^L}} = \Omega\left(\frac{m}{\log(1/\epsilon)2^{q^L}}\right). \tag{3.18}$$

Similarly there exists an index q^R such that $|B_{q^R}^R| = \Omega\left(\frac{m}{\log(1/\epsilon)2^{q^R}}\right)$. But in such a case, since the algorithm selects $t(q) = \Theta\left(2^q \log(1/\epsilon)\right)$ distributions in each iteration q, with high constant probability, the algorithm will select a distribution D_i such that $i \in B_{q^L}^L$ in iteration

10

q^L, and a distribution D_j such that $j \in B_{q^R}^R$ in iteration q^R, and conditioned on the event E_g, will reject, as required.

Let $s(q)$ denote the number of queries performed in iteration q by the algorithm GetBounds for each distribution it is called on. The query complexity of the algorithm is

$$\sum_{q=1}^{r} t(q) \cdot s(q) = O\left(\sum_{q=1}^{r} 2^q \log(1/\epsilon) \cdot \frac{\log(1/\epsilon)}{2^{2q}\epsilon^2}\right) = O(\log^2(1/\epsilon)/\epsilon^2) \qquad (3.19)$$

and the theorem follows. $\square$

3.3. A Lower Bound. We next establish a lower bound (almost matching our upper bound) for testing γ-similar means in the query model by reducing the testing problem to the problem of distinguishing two coins.

FACT 4. *Distinguishing an unbiased coin from a coin with bias ϵ with constant success probability requires $\Omega(1/\epsilon^2)$ samples.*

COROLLARY 2. *Testing γ-similar means in the query model requires $\Omega(1/\epsilon^2)$ samples.*

Proof. Let $D_1(0) = D_1(n) = \frac{1}{2}$, $D_2(0) = \frac{1}{2} - \epsilon$, $D_2(n) = \frac{1}{2} + \epsilon$ and $D_3(0) = \frac{1}{2} + \gamma$, $D_3(n) = \frac{1}{2} - \gamma$. By their definition, the collection (pair) (D_1, D_3) has γ-similar means, while the pair (D_2, D_3) is at least $\epsilon/2$-far from having γ-similar means. Distinguishing between the case that we are given an unbiased coin and the case that we are given a coin with bias ϵ reduces to testing whether the collection of two distributions, (D, D_3) has γ-similar means (when $D = D_1$, which is simulated when the coin is unbiased) or is $\epsilon/2$-far from γ-similar means (when $D = D_2$, which is simulated when the coin is biased). $\square$

For the sake of completeness we include the proof of Fact 4 next.

Proof of Fact 4. We shall use the *KL-divergence* between distributions. Namely, for two distributions p_1 and p_2 over a domain X, $D_{\mathrm{KL}}(p_1 \| p_2) \stackrel{\mathrm{def}}{=} \sum_{x \in X} p_1(x) \cdot \ln \frac{p_1(x)}{p_2(x)}$.

$$D_{\mathrm{KL}}\left(Bin\left(n, \frac{1}{2} + \epsilon\right) \| Bin\left(n, \frac{1}{2}\right)\right)$$
$$= \sum_{0 \le k \le n} \binom{n}{k}\left(\frac{1}{2} + \epsilon\right)^k \left(\frac{1}{2} - \epsilon\right)^{n-k} \cdot \ln\left(1 + 2\epsilon\right)^k \left(1 - 2\epsilon\right)^{n-k}$$
$$= \frac{n}{2}\left((1 + 2\epsilon)\ln(1 + 2\epsilon) + (1 - 2\epsilon)\ln(1 - 2\epsilon)\right)$$
$$\le \frac{n}{2} \cdot 4\epsilon \ln(1 + 2\epsilon)$$
$$\le \frac{n}{2} \cdot 4\epsilon \cdot 2\epsilon$$
$$= 4n\epsilon^2 \, .$$

The ℓ_1 distance is related to the KL-divergence by $\|p_1 - p_2\|_1 \le 2\sqrt{2D_{\mathrm{KL}}(p_1 \| p_2)}$ and thus the fact is established. $\square$

4. Results for the Sampling Model. As opposed to the query model, where the algorithms had no dependence on the number of distributions, m, we show that in the sampling model there is a strong dependence on m. We start by giving a lower bound for the sample complexity of this problem, and continue with several upper bounds.

4.1. A Lower Bound. In this section we prove the following theorem.

THEOREM 3. *For every $n \geq 1$, testing γ-similar means in the uniform sampling model requires $(1 - \gamma) \cdot m^{1-\tilde{O}((\epsilon/\gamma)^{1/2})}$ samples. In particular, when γ is a constant we get a lower bound of $m^{1-\tilde{O}(\epsilon^{1/2})}$. We also note that we may assume without loss of generality that $1 - \gamma = \Omega(\epsilon)$, or else the algorithm can accept automatically.*

In order to prove Theorem 3 we construct a pair of collections of distributions, one that has the property of γ-similar means, the YES instance, and one that is ϵ-far from having this property, the NO instance. We prove that when taking $m^{1-\tilde{O}((\epsilon/\gamma)^{1/2})}$ samples, these pair of collections are indistinguishable and thus prove a lower bound on the sample complexity of the problem. The main part of this proof is the construction of two random variables that on one hand have the same first t moments (where t will be defined later) and on the other hand differ in the maximal distance between pairs of elements in the support, which we call the *diameter* of the random variable. These random variables can then be transformed into collections of distributions that cannot be distinguished (with the abovementioned number of samples) but differ in the distance between the maximal and minimal means in the collection. The random variable that is at the core of the construction of the YES instance, i.e. a collection of distributions where γ is the maximal distance between means of pairs of distributions in the collection, has diameter of γ. While the random variable which corresponds to the NO instance has a diameter which is greater than γ. The next lemma is central to the proof of Theorem 3. In the lemma and what follows we shall use the notation $[k] \stackrel{\text{def}}{=} \{1, \ldots, k\}$.

LEMMA 5. *Given sequences $\{d_i\}_{i=1}^{t}$ and $\{\alpha_i\}_{i=1}^{t}$ that satisfy $0 \leq |d_i|, \alpha_i \leq 1$ for every $i \in [t]$ and $\sum_{i=1}^{t} \alpha_i = 1$, we define a random variable $X = X(\{d_i\}, \{\alpha_i\})$ over $[0, 1]$ as follows: $\Pr[X = d_i] = \alpha_i$. For every even integer t, there exist sequences $\{d_i^+\}_{i=1}^{t}$, $\{\alpha_i^+\}_{i=1}^{t}$, $\{d_i^-\}_{i=1}^{t+1}$ and $\{\alpha_i^-\}_{i=1}^{t+1}$ that obey the aforementioned constraints and for which the following holds:*

1. *For the random variables $X^+ = X(\{d_i^+\}, \{\alpha_i^+\})$ and $X^- = X(\{d_i^-\}, \{\alpha_i^-\})$ we have*

$$\mathrm{E}\left[\left(X^+\right)^i\right] = \mathrm{E}\left[\left(X^-\right)^i\right] \quad \forall i \in [t] . \tag{4.1}$$

2. *The sequences are symmetric around zero. Namely, $d_{t/2+1}^- = 0$, and for every $1 \leq i \leq t/2$, we have that $d_i^+ = -d_{t+1-i}^+$ and $\alpha_i^+ = \alpha_{t+1-i}^+$ as well as $d_i^- = -d_{t+2-i}^-$ and $\alpha_i^- = \alpha_{t+2-i}^-$.*

3. *If we denote by $d_{\max}^+$ ($d_{\min}^+$) the maximal (minimal) non-negative element in the support of X^+ (so that $d_{\max}^+ = d_1^+$ and $d_{\min}^+ = d_{t/2}^+$) and by $\alpha_{\max}^+$ ($\alpha_{\min}^+$) the corresponding probability, and let $d_{\max}^-$, $d_{\min}^-$, $\alpha_{\max}^-$, $\alpha_{\min}^-$ be defined analogously, then*

$$\alpha_{\max}^-(d_{\max}^- - d_{\max}^+) = \tilde{\Theta}\left(\frac{1}{t^2}\right) , \tag{4.2}$$

and

$$d_{\max}^+ - d_{\min}^+ = \Theta(1) . \tag{4.3}$$

We prove Lemma 5 in Subsection 4.1.1 and first show how Theorem 3 follows from it. Let $\{\alpha_i^+\}_{i=1}^{t}$, $\{d_i^+\}_{i=1}^{t}$, $\{\alpha_i^-\}_{i=1}^{t+1}$ and $\{d_i^-\}_{i=1}^{t+1}$ be as defined in Lemma 5. Let $\{\tilde{\alpha}^+\}_{i=1}^{t}$ and $\{\tilde{\alpha}^-\}_{i=1}^{t+1}$ satisfy:

1. $\tilde{\alpha}_i^+ m$ and $\tilde{\alpha}_i^- m$ are integers.

12

2. $\sum_{i=1}^{t} \tilde{\alpha}_i^+ = \sum_{i=1}^{t+1} \tilde{\alpha}_i^- = 1$.
3. $\tilde{\alpha}_i^+ = \tilde{\alpha}_{t-i+1}^+$ and $\tilde{\alpha}_i^- = \tilde{\alpha}_{t-i+1}^-$ (for $1 \le i \le t/2$).
4. $|\tilde{\alpha}_i^+ - \alpha_i^+| \le 1/m$ and $|\tilde{\alpha}_i^- - \alpha_i^-| \le 1/m$.

For a parameter δ, we define the collection of distributions $\mathcal{D}_t^+$ (the YES instance) as follows. For every $1 \le i \le t/2$ there are $\tilde{\alpha}_i^+ m$ distributions $D \in \mathcal{D}_t^+$ of the following form:

$$D(j) \stackrel{\text{def}}{=} \begin{cases} \frac{1}{2} \cdot \left(1 + d_i^+ \delta\right) & \text{if } j = 0 \\ \frac{1}{2} \cdot \left(1 - d_i^+ \delta\right) & \text{if } j = n \\ 0 & \text{o.w.} \end{cases} \tag{4.4}$$

and another $\tilde{\alpha}_i^+ m$ of the distributions $D \in \mathcal{D}_t^+$ are of the following form:

$$D(j) \stackrel{\text{def}}{=} \begin{cases} \frac{1}{2} \cdot \left(1 - d_i^+ \delta\right) & \text{if } j = 0 \\ \frac{1}{2} \cdot \left(1 + d_i^+ \delta\right) & \text{if } j = n \\ 0 & \text{o.w.} \end{cases} \tag{4.5}$$

The collection $\mathcal{D}_t^-$ is defined analogously based on $\{\tilde{\alpha}_i^-\}_{i=1}^{t+1}$ and $\{d_i^-\}_{i=1}^{t+1}$, where for $i = t/2 + 1$ there are $\tilde{\alpha}_{t/2+1}^- m$ distributions $D \in \mathcal{D}_t^-$ such that $D(0) = D(n) = 1/2$ (recall that $d_{t/2+1}^- = 0$).

LEMMA 6. *For every even integer $t \le m^{1/2}$, in order to distinguish between $\mathcal{D}_t^+$ and $\mathcal{D}_t^-$ in the uniform sampling model (with success probability at least $2/3$), it is necessary to take $\Omega\left(m^{1-1/t}(1 - d_{\max}^+ \delta)\right)$ samples.*

The proof of Lemma 6 is given in Subsection 4.1.2.

Proof of Theorem 3. Define γ such that $\mathcal{D}_t^+$ has the property of γ-similar means, i.e.

$$\gamma = \frac{1}{2} \cdot (1 + d_{\max}^+ \delta) - \frac{1}{2} \cdot (1 - d_{\max}^+ \delta) = d_{\max}^+ \delta. \tag{4.6}$$

To change $\mathcal{D}_t^-$ into a γ-similar means instance, we have to either change the means of $\alpha_{\max}^-$ fraction of the distributions from $\frac{1}{2} \cdot (1 + d_{\max}^-) \cdot n$ to $\frac{1}{2} \cdot (1 + d_{\max}^+) \cdot n$ or change the means of $\alpha_{\max}^-$ distributions from $\frac{1}{2} \cdot (1 - d_{\max}^-) \cdot n$ to $\frac{1}{2} \cdot (1 - d_{\max}^+) \cdot n$. Letting $\epsilon = \alpha_{\max}^+ \cdot (d_{\max}^+ - d_{\max}^-) \delta$, we get that $\mathcal{D}^-$ is at least ϵ-far from γ-similar means. By Lemma 5 we have that

$$\frac{\epsilon}{\gamma} = \frac{\alpha_{\max}^+ \cdot (d_{\max}^+ - d_{\max}^-)}{d_{\max}^+} = \tilde{\Theta}\left(\frac{1}{t^2}\right). \tag{4.7}$$

We note that for every $\epsilon/\gamma \le 1/\log^2 m$ we get that $m^{1-\tilde{O}((\epsilon/\gamma)^{1/2})} = \Omega(m)$. Hence we can assume without loss of generality that $\epsilon/\gamma = \tilde{\Omega}(m^{-1/2})$ and thus by setting $1/t = \tilde{\Theta}(\epsilon^{1/2}/\gamma^{1/2})$, the theorem follows from Lemma 6. $\square$

4.1.1. Proof of Lemma 5. The random variables described in Lemma 5 are constructed via a polynomial f: the support of X^+ (respectively, X^-) is the set of roots of f with a negative (respectively, positive) derivative. If f has an odd number of roots then the sign of the derivative at the largest root is the same as the sign at the smallest root. If it is positive, then the support of X^- resides in an interval which contains the support of X^+. To prove a lower bound, X^- needs to be far from similar means (more precisely, the collection of coins that corresponds to X^-) and indistinguishable from X^+. To make X^- far from having similar means, f should maximize the size of X^-'s interval (compared to X^+'s interval) and the weight on the extreme roots.

As suggested by Lemma 7 (stated below) X^- and X^+ have matching moments if the probability to take the value x_i, where x_i is a root of f, is $1/|f'(x_i)|$, up to normalization. In this case, a small derivative on the extreme roots would result with X^- which is far from having similar means. When the roots of f is taken to be the value of the Sine function at equal distances, the derivative at the extreme roots, that is at -1 and 1, is small. As we see next, these roots are the roots of the Chebyshev polynomials.

The proof of Lemma 5 requires some preliminaries concerning Chebyshev polynomials, which we provide next. Let T_ℓ be the ℓ-th Chebyshev polynomial of the first kind, which is defined by the recurrence relation:

$$T_0(x) = 1 \tag{4.8}$$

$$T_1(x) = x \tag{4.9}$$

$$T_{\ell+1}(x) = 2xT_\ell(x) - T_{\ell-1}(x) . \tag{4.10}$$

Let U_ℓ be the ℓ-th Chebyshev polynomial of the second kind, which is defined by the recurrence relation:

$$U_0(x) = 1 \tag{4.11}$$

$$U_1(x) = 2x \tag{4.12}$$

$$U_{\ell+1}(x) = 2xU_\ell(x) - U_{\ell-1}(x) . \tag{4.13}$$

Then we have that

$$\frac{dT_\ell(x)}{dx} = \ell \cdot U_{\ell-1} , \tag{4.14}$$

and that

$$U_{\ell-1}(\cos(x)) = \frac{\sin(\ell x)}{\sin x} . \tag{4.15}$$

T_ℓ has ℓ different simple roots:

$$x_i = \cos\left(\frac{\pi}{2} \cdot \frac{2i-1}{\ell}\right) \quad i = 1, \ldots, \ell \tag{4.16}$$

and the following equalities hold:

$$T_\ell(1) = 1, \quad \text{and} \quad T_\ell(-1) = (-1)^\ell . \tag{4.17}$$

We shall also use the next lemma concerning properties of (derivatives of) polynomials.

LEMMA 7 ([19]). *Let $f(x)$ be a polynomial of degree ℓ whose roots $\{x_i\}$ are real and distinct. Letting f' denote the derivative of f, for every $j \le \ell-2$ we have that $\sum_{i=1}^\ell \frac{x_i^j}{f'(x_i)} = 0$.*

We are now ready to prove Lemma 5.

Proof of Lemma 5. Consider the following polynomial:

$$f(x) \stackrel{\text{def}}{=} (x-1)(x+1) \cdot T_\ell(x) , \tag{4.18}$$

where $T_\ell(\cdot)$ is the ℓ-th Chebyshev polynomial of the first kind and $\ell = 2t-1$. The polynomial $f(\cdot)$ has $\ell+2$ roots, which, by decreasing order, are: $1, \cos\left(\frac{\pi}{2} \cdot \frac{1}{\ell}\right), \cos\left(\frac{\pi}{2} \cdot \frac{3}{\ell}\right), \ldots, -1$. The derivative of $f(\cdot)$ is $f'(x) = 2x \cdot T_\ell(x) + (x^2 - 1) \cdot T'_\ell(x)$ and thus

$$\frac{1}{|f'(1)|} = \frac{1}{2T_\ell(1)} = \frac{1}{2} , \tag{4.19}$$

and

$$\frac{1}{|f'(-1)|} = \frac{1}{|2T_\ell(-1)|} = \frac{1}{2} . \tag{4.20}$$

While for x_i which is a root of T_ℓ we have:

$$\frac{1}{|f'(x_i)|} = \frac{1}{|(x_i^2 - 1) \cdot T'_\ell(x_i)|} . \tag{4.21}$$

By Equations (4.14) and (4.15):

$$\frac{1}{|T'_\ell(x_i)|} = \left| \frac{\sin\left(\frac{\pi}{2} \cdot \frac{2i-1}{\ell}\right)}{\ell \cdot \sin\left(\frac{\pi}{2} \cdot (2i-1)\right)} \right| = \frac{1}{\ell} \cdot \left| \sin\left(\frac{\pi}{2} \cdot \frac{2i-1}{\ell}\right) \right| \tag{4.22}$$

where we used the fact that $\left|\sin\left(\frac{\pi}{2} \cdot (2i-1)\right)\right| = 1$. Therefore by Equation (4.16) and the identity $1 - \cos^2 x = \sin^2 x$ we obtain:

$$\frac{1}{|f'(x_i)|} = \frac{1}{\ell} \cdot \frac{1}{\left|\sin\left(\frac{\pi}{2} \cdot \frac{2i-1}{\ell}\right)\right|} . \tag{4.23}$$

Since $g(x) = \sin x / x$ is monotone decreasing for $0 < x \leq \pi/2$, from the fact that $g(\pi/2) = 2/\pi$ we get that $\sin x > (2/\pi)x$ for $0 < x \leq \pi/2$. Thus for $i \leq \ell/2$,

$$\frac{1}{|f'(x_i)|} \leq \frac{1}{\ell} \cdot \frac{\pi}{2} \cdot \frac{1}{\left(\frac{\pi}{2} \cdot \frac{2i-1}{\ell}\right)} = \frac{1}{2i-1} . \tag{4.24}$$

Therefore, for $\{x_i\}$, the roots of $T_\ell(\cdot)$,

$$\sum_{i=1}^{\ell} \frac{1}{|f'(x_i)|} \leq 2 \sum_{x_i \geq 0} \frac{1}{|f'(x_i)|} = O(\log \ell) . \tag{4.25}$$

We take $\{d_i^-\}_{i=1}^{t+1}$ to be those roots x_j of $f(\cdot)$ for which $f'(x_j) > 0$ and set

$$\alpha_i^- = \frac{1}{|f'(d_i^-)|} \cdot \beta^- , \tag{4.26}$$

where $\beta^- = 1/\left(\sum_{i=1}^{t+1} \frac{1}{|f'(d_i^-)|}\right)$ is a normalization factor. Similarly we take $\{d_i^+\}_{i=1}^{t}$ to be the roots with the negative derivative. Then $d_{\max}^- = 1$ and by Equation (4.25), $\alpha_{\max}^- = \Omega(1/\log \ell)$. On the other hand, $d_{\max}^+ = \cos\left(\frac{\pi}{2} \cdot \frac{1}{\ell}\right)$. Due to the identity $1 - \cos x = \sin x \cdot \tan(x/2)$, we get that:

$$\lim_{x \to 0} \frac{1 - \cos x}{x^2} = \lim_{x \to 0} \frac{\sin x}{x} \cdot \frac{\sin(x/2)}{x} \cdot \frac{1}{\cos(x/2)} = \frac{1}{2} , \tag{4.27}$$

15

and so $d_{\max}^- - d_{\max}^+ = \Theta(1/\ell^2)$. Since ℓ is odd and the sign of the derivative alternates between roots we get that $d_{\min}^- = \cos\left(\frac{\pi}{2}\right) = 0$ while $d_{\min}^+ = \cos\left(\frac{\pi}{2} \cdot \frac{\ell-2}{\ell}\right) = \sin\left(\frac{\pi}{2} \cdot \frac{2}{\ell}\right)$. Thus $d_{\min}^- - d_{\min}^+ = \Theta(1/\ell)$. By Equations (4.23) and (4.25) we get $\alpha_{\min}^+ = \tilde{\Theta}(1/\ell)$. Therefore, Equations (4.2) and (4.3) hold. Equation (4.1) follows from Lemma 7. Since the roots of the Chebyshev polynomials are symmetric around zero, we get that the roots of $f(\cdot)$ are also symmetric. For an odd ℓ we get that zero is one of the roots and thus each one of the sequences $\{d_i^+\}_{i=1}^t$, $\{d_i^-\}_{i=1}^{t+1}$ is symmetric around zero, as desired. $\quad\square$

4.1.2. Proof of Lemma 6. We first recall the definition of the Poisson distribution, and introduce one more lemma (from [20]). For a positive real number λ, the Poisson distribution $\text{poi}(\lambda)$ takes the value $x \in \mathbb{N}$ (where $\mathbb{N} = \{0, 1, 2, \ldots\}$) with probability $\text{poi}(x; \lambda) = e^{-\lambda}\lambda^x/x!$. For the collection $\mathcal{D} = (D_1, D_2)$ define

$$\vec{\lambda}^{\mathcal{D},k}(a,b) \stackrel{\text{def}}{=} \sum_i \text{poi}(a; k \cdot D_1(i))\text{poi}(b; k \cdot D_2(i)) , \tag{4.28}$$

which is the expected number of elements i for which we get a samples of the form $(1, i)$ and b samples of the form $(2, i)$ if we take $\text{poi}(k)$ samples from $\mathcal{D}$ in the uniform sampling model (see [13]).

LEMMA 8 ([20]). *Given a positive integer k and two distribution pairs $p_1^+, p_2^+, p_1^-, p_2^-$ all of whose frequencies are at most $\frac{1}{1000k}$, let $\vec{\lambda}^+(a,b) = \sum_i \text{poi}(a; k \cdot p_1^+(i))\text{poi}(b; k \cdot p_2^+(i))$ and $\vec{\lambda}^-(a,b) = \sum_i \text{poi}(a; k \cdot p_1^-(i))\text{poi}(b; k \cdot p_2^-(i))$ for $a + b > 0$. If it is the case that*

$$\sum_{a+b>0} \frac{\left|\vec{\lambda}^+(a,b) - \vec{\lambda}^-(a,b)\right|}{\sqrt{1 + \max\{\vec{\lambda}^+(a,b), \vec{\lambda}^-(a,b)\}}} < \frac{1}{36} , \tag{4.29}$$

then it is impossible to test any symmetric property that is true for (p_1^+, p_2^+) and false for (p_1^-, p_2^-) in k samples.

Proof of Lemma 6. Since for every $D \in \mathcal{D}_t^+$, the size of the support is 2, and due to the symmetry between n and 0 in the construction, we can perceive samples (i, j) taken from $\mathcal{D}_t^+$ as samples taken from the collection (p_0^+, p_n^+) of distributions over the domain $[m]$, where p_0^+ and p_n^+ are defined as follows

$$p_0^+(i) \stackrel{\text{def}}{=} 2D_i^+(0)/m \tag{4.30}$$

$$p_n^+(i) \stackrel{\text{def}}{=} 2D_i^+(n)/m . \tag{4.31}$$

That is, instead of perceiving i as an index of a distribution in the collection and j as an element of the domain $\{0, \ldots, n\}$ we perceive i as an element in the domain $[m]$ and j as an index of a distribution in the collection (p_0^+, p_n^+) denoted $\mathcal{P}^+$. The collection (p_0^-, p_n^-), denoted $\mathcal{P}^-$, is defined similarly. In this settings we can apply Lemma 8 to give a lower bound on the number of samples required to test γ-similar means. We note that even though the γ-similar means property is not a symmetric property we can still use Lemma 8 because there is a symmetry between n and 0 in the construction. We next turn to calculating $\vec{\lambda}^{\mathcal{P}^+,k}(a,b)$ defined in Equation (4.28). We shall first assume that $\{\alpha_i^+ m\}$ (and $\{\alpha_i^- m\}$) are integers (in

other words, $\tilde{\alpha}_i^+ = \alpha_i^+$ and $\tilde{\alpha}_i^- = \alpha_i^-$ for all i), and later deal with the issue of rounding.

$$\vec{\lambda}^{\mathcal{P}^+,k}(a,b) = \sum_{i=1}^{m} \mathrm{poi}(a; k \cdot p_0^+(i)) \cdot \mathrm{poi}(b; k \cdot p_n^+(i)) \tag{4.32}$$

$$= \sum_{i=1}^{t} \alpha_i^+ m \cdot \left(\mathrm{poi}(a; k \cdot (1 + d_i^+\delta)/m) \cdot \mathrm{poi}(b; k \cdot (1 - d_i^+\delta)/m) \right) \tag{4.33}$$

$$+ \sum_{i=1}^{t} \alpha_i^+ m \cdot \left(\mathrm{poi}(a; k \cdot (1 - d_i^+\delta)/m) \cdot \mathrm{poi}(b; k \cdot (1 + d_i^+\delta)/m) \right) \tag{4.34}$$

$$= \frac{e^{-k/m}}{a!b!} \left(\frac{k}{m} \right)^{a+b} \sum_{i=1}^{t} \alpha_i^+ m \cdot \left(\frac{(1 + d_i^+\delta)^a}{(1 - d_i^+\delta)^{-b}} + \frac{(1 - d_i^+\delta)^a}{(1 + d_i^+\delta)^{-b}} \right) \tag{4.35}$$

Similarly,

$$\vec{\lambda}^{\mathcal{P}^-,k}(a,b) = \frac{e^{-k/m}}{a!b!} \left(\frac{k}{m} \right)^{a+b} \sum_{i=1}^{t} \alpha_i^- m \cdot \left(\frac{(1 + d_i^-\delta)^a}{(1 - d_i^-\delta)^{-b}} + \frac{(1 - d_i^-\delta)^a}{(1 + d_i^-\delta)^{-b}} \right) \tag{4.36}$$

Therefore, there exist coefficients β_j such that,

$$\vec{\lambda}^{\mathcal{P}^+,k}(a,b) = \frac{e^{-k/m}}{a!b!} \left(\frac{k}{m} \right)^{a+b} \cdot \sum_{i=1}^{t} \alpha_i^+ m \cdot \sum_{j=0}^{a+b} \beta_j \cdot (d_i^+\delta)^j \tag{4.37}$$

$$= \frac{e^{-k/m}}{a!b!} \left(\frac{k}{m} \right)^{a+b} \cdot \sum_{j=0}^{a+b} m\beta_j \delta^j \cdot \sum_{i=1}^{t} \alpha_i^+ (d_i^+)^j , \tag{4.38}$$

and

$$\vec{\lambda}^{\mathcal{P}^-,k}(a,b) = \frac{e^{-k/m}}{a!b!} \left(\frac{k}{m} \right)^{a+b} \cdot \sum_{j=0}^{a+b} m\beta_j \delta^j \cdot \sum_{i=1}^{t+1} \alpha_i^- (d_i^-)^j \tag{4.39}$$

By Equation (4.1) we have that $\sum_{i=1}^{t} \alpha_i^+ (d_i^+)^j = \sum_{i=1}^{t+1} \alpha_i^- (d_i^-)^j$ for every $0 \leq j \leq t$. By Equations (4.38) and (4.39) we get that for every $a + b \leq t$,

$$\vec{\lambda}^{\mathcal{P}^+,k}(a,b) = \vec{\lambda}^{\mathcal{P}^-,k}(a,b), \tag{4.40}$$

Set $k = \frac{1-\delta}{4} \cdot m^{1-1/t}$ and let $\lambda(a,b) = \max\left\{ \vec{\lambda}^{\mathcal{P}^-,k}(a,b), \vec{\lambda}^{\mathcal{P}^+,k}(a,b) \right\}$. Then

$$\sum_{a+b>t} \lambda(a,b) \leq \sum_{a+b>t} \left(\frac{k}{m} \right)^{a+b} \cdot m = \sum_{a+b>t} \left(\frac{1-\delta}{4m^{1/t}} \right)^{a+b} \cdot m \tag{4.41}$$

$$\leq \sum_{x>t} 2^x \left(\frac{1}{4m^{1/t}} \right)^x \cdot m \ll 1 . \tag{4.42}$$

In the above construction we assumed that $\{\alpha_i^+ m\}$ (and $\{\alpha_i^- m\}$) are integers. If this is not the case, then by Equations (4.35) and (4.36), calculating the sum in Equation (4.29)

where we replace $\{\alpha_i^+\}_{i=1}^t$ and $\{\alpha_i^-\}_{i=1}^t$ by $\{\tilde{\alpha}_i^+\}_{i=1}^t$ and $\{\tilde{\alpha}_i^-\}_{i=1}^t$ respectively, can be off by at most

$$\sum_{a+b>0} \frac{\frac{e^{-k/m}}{a!b!}\left(\frac{k}{m}\right)^{a+b}\sum_{i=1}^t m(|\alpha_i^+ - \tilde{\alpha}_i^+| + |\alpha_i^- - \tilde{\alpha}_i^-|)\cdot 2^{a+b}}{(\vec{\lambda}^{\mathcal{P}^-,k}(a,b))^{1/2}}$$

$$\leq \sum_{a+b>0}\left(\frac{e^{-k/m}}{a!b!}\left(\frac{k}{m}\right)^{a+b}\right)^{1/2}\cdot\frac{\sum_{i=1}^t m(|\alpha_i^+ - \tilde{\alpha}_i^+| + |\alpha_i^- - \tilde{\alpha}_i^-|)\cdot 2^{a+b}}{\left(\sum_{i=1}^t m\tilde{\alpha}_i^+\cdot(1-d_{\max}^+\delta)^{a+b}\right)^{1/2}} \quad (4.43)$$

$$\leq \frac{t}{m^{1/2}}\sum_{a+b>0}\left(\frac{e^{-k/m}}{a!b!}\left(\frac{k}{m}\cdot\frac{4}{1-\delta}\right)^{a+b}\right)^{1/2} \quad (4.44)$$

$$\leq \frac{t}{m^{1/2}}\sum_{a+b>0}\left(\frac{1}{m^{1/t}}\right)^{(a+b)/2} \ll 1 , \quad (4.45)$$

where the last inequality holds for every $t \leq m^{1/2}$. By Lemma 8, $\mathcal{D}_t^+$ and $\mathcal{D}_t^-$ are indistinguishable when $k \leq cm^{1-1/t}(1-d_{\max}^+\delta)$ (for some constant $c \leq 1$). $\quad\square$

4.2. An Upper Bound. The lower bound stated in Theorem 3 does not leave much room for an algorithm in the sampling model with sample complexity that is sublinear in m. In particular note that for a constant γ, if $\epsilon = o(1/\log^2 m \log\log m)$, then we get a linear dependence on m. However, for ϵ that is not too small, we may still ask whether we can get an upper bound that is sublinear in m. We start by observing that given samples as provided in the sampling model it is possible to simulate any algorithm that works in the query model. This observation immediately provides a test for γ-similar means in the sampling model that has $m^{1-\tilde{\Omega}(\epsilon^2)}$ sample complexity (conditioned on $\epsilon > c\log\log m/\log m$ for some sufficiently large constant c.) The following is a well known fact, and we provide a proof for completeness.

LEMMA 9. *We say that we have a t-way collision on element i if we sampled i exactly t times. Let u_m be the uniform distribution over $[m]$ and let t be a non-negative integer such that $t \leq \frac{\log m}{\log\log m}$, if we take $s = \Theta(tm^{1-1/t})$ samples from u_m we will have t-way collisions on $\Omega(1)$ elements with high constant probability.*

The proof of Lemma 9 applies the Efron-Stein inequality, which we state next:

THEOREM 4 (Efron-Stein inequality [7]). *Let χ be some set, and let $g : \chi^s \to \mathbb{R}$ be a measurable function of s variables. Define the random variable $Z = g(X_1,\ldots,X_s)$ where $X_1,\ldots,X_s$ are arbitrary independent random variables taking values in χ. Let $X_1',\ldots,X_s'$ form an independent copy of $X_1,\ldots,X_s$ and write $Z_i' = g(X_1,\ldots,X_i',\ldots,X_s)$. Then*

$$\mathrm{V}\left[Z\right] \leq \frac{1}{2}\sum_{i=1}^s \mathrm{E}\left[(Z-Z_i')^2\right] . \quad (4.46)$$

Proof of Lemma 9. Define the indicator variable $I_i \in \{0,1\}$ to take the value 1 if and only if we have a t-way collision on i and the random variable $Z = \sum_{i=1}^m I_i$, namely, Z is the

18

number of t-collisions. Then,

$$\mathrm{E}\,[Z] = \sum_{i=1}^{m} \mathrm{E}\,[I_i] \tag{4.47}$$

$$= \sum_{i=1}^{m} \Pr\,[Bin(s, 1/m) = t] \tag{4.48}$$

$$= m \cdot \Pr\,[Bin(s, 1/m) = t] \tag{4.49}$$

$$= m \cdot \binom{s}{t} \left(\frac{1}{m}\right)^t \left(1 - \frac{1}{m}\right)^{s-t} \tag{4.50}$$

By the inequality $\binom{s}{t} \geq \left(\frac{s}{t}\right)^t$ and the fact that $\left(1 - \frac{1}{m}\right)^{s-t} \geq \frac{1}{4}$ we obtain

$$\Pr\,[Bin(s, 1/m) = t] \geq \frac{1}{4} \cdot \left(\frac{s}{t \cdot m}\right)^t . \tag{4.51}$$

Therefore for $s = 64tm^{1-1/t}$,

$$\mathrm{E}\,[Z] \geq m \cdot \frac{1}{4} \cdot \left(\frac{64}{m^{1/t}}\right)^t = \Omega(1) \tag{4.52}$$

Applying the Efron-Stein inequality (Theorem 4) we get that:

$$\mathrm{V}\,[Z] \leq \sum_{i=1}^{s} \mathrm{E}\left[(Z - Z_i')^2\right] \tag{4.53}$$

$$= \sum_{i=1}^{s} \sum_{a,b \in [m]} \frac{\mathrm{E}\left[(Z - Z_i')^2 \,|X_i = a, X_i' = b\right]}{m^2} \tag{4.54}$$

Conditioned on $X_i = a$ and $X_i' = b$ we have that $(Z - Z_i')^2 = 1$ if the number of occurrences of a in $X_1, \ldots, X_s$ is t and the number of occurrences of b in $X_1, \ldots, X_i', \ldots, X_s$ is not t, or alternatively, if the number of occurrences of a in $X_1, \ldots, X_s$ is not t and the number of occurrences of b in $X_1, \ldots, X_i', \ldots, X_s$ is t. Otherwise, $(Z - Z_i')^2 = 0$. So we have that for $a \neq b$,

$$\mathrm{E}\left[(Z - Z_i')^2 \,|X_i = a, X_i' = b\right] \leq 2\Pr\,[Bin(s - 1, 1/m) = t - 1]$$

and for $a = b$ clearly

$$\mathrm{E}\left[(Z - Z_i')^2 \,|X_i = a, X_i' = b\right] = 0 .$$

Therefore, from Equation (4.54) we get that $\mathrm{V}\,[Z] \leq 2s \cdot \Pr\,[Bin(s - 1, 1/m) = t - 1]$. By Chebyshev's inequality,

$$\Pr\,[|Z - \mathrm{E}[Z]| \geq \mathrm{E}[Z]/2] \leq \frac{8s \cdot \Pr\,[Bin(s - 1, 1/m) = t - 1]}{(m \cdot \Pr\,[Bin(s, 1/m) = t])^2} \tag{4.55}$$

$$= \frac{8t}{m \cdot \Pr\,[Bin(s, 1/m) = t]} \tag{4.56}$$

$$\leq \frac{32t}{64^t} \ll 1 \tag{4.57}$$

Recall that, by Equation (4.52), $E[Z]$ is bounded from below by a positive constant, and hence the lemma follows. $\square$

COROLLARY 5. *The sample complexity of testing γ-similar means in the uniform sampling model is upper bounded by $m^{1-\tilde{\Omega}(\epsilon^2)} \cdot \tilde{O}(1/\epsilon^2)$.*

Proof. Recall that Algorithm 2 which works in the query model performs $r = \log(4/\epsilon)$ iterations and in the q-th iteration, selects $\ell(q) = \Theta\left(2^q \log(1/\epsilon)\right)$ distributions. From each selected distribution it then takes $s(q) = \Theta\left(\frac{\log(1/\epsilon)}{2^{2q}\epsilon^2}\right)$ samples. By applying Lemma 9 for $t = s(q)$ we get that by taking $s(q)m^{1-1/s(q)}$ samples we get $s(q)$-way collisions on $\Omega(1)$ distributions. If for each iteration q, we repeat this process $\ell(q)\log(\ell(q)^2)$ times then by union bound over r iterations, with high constant probability we would have enough samples to simulate Algorithm 2. Thus the sample complexity we get is bounded by

$$\sum_{q=1}^{r} \ell(q)\log(\ell(q)^2) \cdot s(q)m^{1-1/s(q)} = m^{1-\tilde{\Omega}(\epsilon^2)} \cdot \tilde{O}(1/\epsilon^2)\,, \tag{4.58}$$

as desired. $\square$

4.3. An Improved Upper Bound for the Domain $\{0,1\}$ and Large m. When the domain of the distributions is $\{0,1\}$ and m is sufficiently larger than $1/\epsilon$, we can get an improved upper bound by applying an algorithm that does not work by reducing to the query model. This algorithm does not require that the sample contain $\tilde{O}(1/\epsilon^2)$-way collisions but rather can make do with $\tilde{O}(1/\epsilon)$-way collisions (though it asks for many more of them). It uses such collisions in which either all samples are 1 or all samples are 0. These relatively extreme events give sufficient indications as to whether the collection has γ-similar means or is ϵ-far from having the property.

Algorithm 3 testing γ-similar means

1. Take $s = r \ln r \cdot m^{1-1/\ell}$ samples where $\ell = \Theta(\ln(1/\epsilon)/\epsilon)$ and $r = \Theta((16/\epsilon)^{\ell+1})$.
2. For each distribution D_i, divide the samples obtained from D_i into blocks of ℓ samples and ignore blocks with less than ℓ samples.
3. Let t be the number of blocks, let t_0 be the number of blocks where all the samples are 0 and let t_1 be the number of blocks where all the samples are 1.
4. If there exist $x, y \in \{0, \epsilon/16, 2\epsilon/16, \ldots, 1\}$ such that $y - x \leq \gamma + \epsilon/8$ and $\left(\frac{2t_1}{3t}\right)^{1/\ell} < y$ and $\left(\frac{2t_0}{3t}\right)^{1/\ell} < 1 - x$, then ACCEPT, otherwise REJECT.

THEOREM 6. *Algorithm 3 tests γ-similar means in the sampling model. The algorithm's sample complexity is $r \ln r \cdot m^{1-1/\ell}$ where $\ell = c_1 \ln(1/\epsilon)/\epsilon$ and $r = c_2(16/\epsilon)^{\ell+1}$ where c_1 and c_2 are absolute constants.*

Proof. Denote by $p_1 \leq \ldots \leq p_m$ the means in the collection $\mathcal{D}$ in increasing order and let $q_i = 1 - p_{m-i+1}$. Thus $q_1 \leq \ldots \leq q_m$. By the definitions of t_1 and t_0,

$$E\left[\frac{t_1}{t}\right] = \frac{1}{m} \cdot \sum_{i=1}^{m} (D_i(1))^\ell \leq (p_m)^\ell\,, \tag{4.59}$$

and similarly

$$E\left[\frac{t_0}{t}\right] = \frac{1}{m} \cdot \sum_{i=1}^{m} (D_i(0))^\ell \leq (q_m)^\ell\,. \tag{4.60}$$

By Chernoff's bound, for every $b \in \{0,1\}$ if $t \geq c \left(\mathrm{E} \left[\frac{t_b}{t} \right] \right)^{-1}$, for some sufficiently large constant c, then with high constant probability it holds that

$$\left| \frac{t_b}{t} - \mathrm{E} \left[\frac{t_b}{t} \right] \right| \leq \frac{\mathrm{E} \left[\frac{t_b}{t} \right]}{2} . \tag{4.61}$$

By Lemma 9, with high constant probability $t = \Omega((16/\epsilon)^{\ell+1})$, so henceforth we assume that this is the case.

Suppose that $\mathcal{D}$ has γ-similar means, so that $p_m - p_1 = p_m - (1 - q_m) \leq \gamma$. Let $x, y \in \{0, \epsilon/16, 2\epsilon/16, \ldots, 1\}$ be such that $0 < y - p_m \leq \epsilon/16$ and $0 < p_1 - x \leq \epsilon/16$. By Equation (4.59) and Equation (4.60) we get

$$\mathrm{E} \left[\frac{t_1}{t} \right] \leq (p_m)^\ell \leq (y)^\ell \quad \text{and} \quad \mathrm{E} \left[\frac{t_0}{t} \right] \leq (1 - p_1)^\ell \leq (1 - x)^\ell . \tag{4.62}$$

Since $y, 1 - x > \epsilon/16$ and $t = \Omega \left((16/\epsilon)^\ell \right)$, from Equation (4.61) we get that

$$\frac{t_1}{t} \leq (3/2)(y)^\ell \quad \text{and} \quad \frac{t_0}{t} \leq (3/2)(1 - x)^\ell \tag{4.63}$$

Since $y - x \leq \gamma + \epsilon/8$ the algorithm accepts.

Suppose that $\mathcal{D}$ is ϵ-far from γ-similar means, and let $x, y \in [0,1]$ be such that $y - x \leq \gamma + \epsilon/8$. Let $y' = \min\{y + \epsilon/16, 1\}$ and let $x' = \max\{0, x - \epsilon/16\}$, thus $y' - x' \leq \gamma + \epsilon/4$. Since $\mathcal{D}$ is ϵ-far from γ-similar means, at least $(\epsilon/2)m$ of the elements in $\{p_i\}$ are outside the interval $[x', y']$. Thus either at least $(\epsilon/4)m$ of the elements in $\{p_i\}$ are bigger than y' or at least $(\epsilon/4)m$ of the elements in $\{p_i\}$ are smaller than x'. Since the blocks are distributed uniformly over the distributions we get that in the former case,

$$\mathrm{E} \left[\frac{t_1}{t} \right] \geq (\epsilon/4) \cdot (y')^\ell = (\epsilon/4) \cdot (y + \epsilon/16)^\ell > 3y^\ell . \tag{4.64}$$

The latter case implies that at least $(\epsilon/4)m$ of the elements in $\{q_i\}$ are bigger than $1 - x'$, thus

$$\mathrm{E} \left[\frac{t_0}{t} \right] \geq (\epsilon/4) \cdot (1 - x')^\ell = (\epsilon/4) \cdot (1 - x + \epsilon/16)^\ell > 3(1 - x)^\ell . \tag{4.65}$$

Since $y' \geq \epsilon/16$ and $t = \Omega \left((16/\epsilon)^{\ell+1} \right)$, from Equation (4.61) and Equation (4.64) we get that

$$\frac{t_1}{t} > (3/2)y^\ell . \tag{4.66}$$

Similarly, since $1 - x' \geq \epsilon/16$, we get from Equation (4.61) and Equation (4.65) we get that

$$\frac{t_0}{t} > (3/2)(1 - x)^\ell , \tag{4.67}$$

thus the algorithm rejects. $\square$

4.4. An Improved (Tight) Upper bound for the $\{0,1\}$ case when $\gamma \leq \epsilon/c$. If $\gamma \leq \epsilon/c$ for a sufficiently large constant c (a special case of interest is $\gamma = 0$), then we can significantly improve the bound we obtained in the previous subsection (for the domain $\{0,1\}$). The problem of testing γ-similar means under these conditions reduces to a certain tolerant version

of testing uniformity of a distribution over domain $[m]$, which was studied in [13] For this problem there is an algorithm that uses $\tilde{O}(\sqrt{m} \cdot \text{poly}(1/\epsilon))$ samples (see [13, Theorem 12], based on [22]).

We note that in general, for every γ and ϵ such that $\gamma + 2\epsilon < 1$ (and in particular, even for $\gamma = 0$ and $\epsilon < 1/2$), any algorithm must take a sample of size $\Omega(\sqrt{m})$. This is true because there exists a constant $c < 1$ such that less than $c\sqrt{m}$ samples are insufficient to distinguish (with constant success probability) between the following two collections of distributions. In the first, $\mathcal{D}^+ = \{D_1^+, \ldots, D_m^+\}$, all distributions are the same: $D_j^+(0) = D_j^+(n) = 1/2$, while in the second, $\mathcal{D}^- = \{D_1^-, \ldots, D_m^-\}$, we have $D_j^-(0) = 1, D_j^-(n) = 0$ for half of the distributions and $D_j^-(0) = 0, D_j^-(n) = 1$ for the other half (the choice of which distribution belongs to which half is done uniformly at random).

5. Earth Mover's Distance. Let D and D' be probability distributions over $\{0, \ldots, n\}$. The Earth Mover's Distance between D and D' with respect to the ℓ_1-distance is

$$\text{EMD}(D, D') \stackrel{\text{def}}{=} \min_F \left\{ \sum_{i=0}^{n} \sum_{j=0}^{n} f_{i,j} |i - j| \right\} , \tag{5.1}$$

where the flow, $F = (f_{i,j})$, is subject to the following constraints:

$$f_{i,j} \geq 0, 1 \leq i \leq m, 1 \leq j \leq n \tag{5.2}$$

$$\sum_{j=0}^{n} f_{i,j} = D(i), 1 \leq i \leq m \tag{5.3}$$

$$\sum_{i=0}^{n} f_{i,j} = D'(j), 1 \leq j \leq m \tag{5.4}$$

$$\sum_{i=0}^{n} \sum_{j=0}^{n} f_{i,j} = 1 . \tag{5.5}$$

We define the normalized Earth Mover's Distance to be $\mathcal{E}(D, D') \stackrel{\text{def}}{=} \frac{1}{n}\text{EMD}(D, D')$, where we normalize by the the maximum distance, n. Given a distribution D and a value $0 \leq z \leq n$ the function $d_{\mathcal{E}}$ returns the "work", normalized by n, required to move earth from D in order to obtain a distribution with mean equals z. Formally $d_{\mathcal{E}}(D, z) \stackrel{\text{def}}{=} \min_{D' : \mu(D') = z} \{\mathcal{E}(D, D')\}$. We claim that $d_{\mathcal{E}}(D, z) = |\mu(D) - z|/n$. To see why this true, assume without loss of generality that $z > \mu(D)$. Let D_z denote a distribution such that $\mu(D_z) = z$ and $\mathcal{E}(D, D_z) = d_{\mathcal{E}}(D, z)$. Let $F^* = (f_{i,j}^*)$ denote an optimal flow between D and D_z. Clearly, for every $i > j$, $f_{i,j}^* = 0$, therefore by Equations (5.3) and (5.4) we get that $\sum_{i=0}^{n} \sum_{j=0}^{n} f_{i,j}^* |i - j| = z - \mu(D)$. This implies that $EMD(D, D_z) = z - \mu(D)$ and the claim follows. The fact that the minimum EMD distance between a distribution D, and a distribution with mean $\mu(D) + \delta$, where the minimum is taken over all distributions over $\{0, \ldots, n\}$ with mean $\mu(D) + \delta$, is exactly δ without regard to any property of D other than its mean, makes EMD a very natural distance measure for the property of γ-similar means.

In particular, it is so natural that all the results in this paper continue to hold when we change the underlying distance measure form ℓ_1 distance to the normalized EMD. The completeness of all the algorithms is immediate for any distance measure. The soundness of the algorithms follows from the fact that $\mathcal{E}(D, D') \leq \|D - D'\|_1$. Moreover, Algorithm 2 can be simplified as follows. As stated in Lemma 3, the algorithm GetBounds returns x

and y such that $\mu(D) \le x \le \mu_\beta^>(D)$ and $\mu_\beta^<(D) \le y \le \mu(D)$. Due to the fact that $d_\mathcal{E}(D, z) = |\mu(D) - z|/n$, when we define $\mu_\beta^>$ and $\mu_\beta^<$ with respect to $d_\mathcal{E}(\cdot, \cdot)$ instead of the $d_1(\cdot, \cdot)$, $\mu_\beta^>(D)$ is simply $\mu(D) + \beta n$ and $\mu_\beta^<(D)$ is $\mu(D) - \beta n$. Therefore the algorithm GetBounds for EMD goes as follows:

- Take $\Theta(\log(1/\delta)/\beta^2)$ samples from D
- Return $(\hat\mu - \beta n/2, \hat\mu + \beta n/2)$, where $\hat\mu$ is the average value of the samples taken from D

By Chernoff's bounds, with high probability it holds that $|\hat\mu - \mu(D)| < \beta n/2$ and thus the return values of GetBounds for EMD satisfy the requirements stated in Lemma 3. The rest of the proof of correctness of Algorithm 2 holds for normalized EMD.

Both lower bounds, stated in Theorem 3 and in Corollary 2, are proved via a construction of NO instance collections of distributions with support $\{0, n\}$. Since for every D with support $\{0, n\}$ and every $0 \le z \le n$ we have that $d_1(D, z) = d_\mathcal{E}(D, z)$ the above NO instance collections are also NO instances for the normalized EMD. Therefore, the lower bounds are valid for the normalized EMD.

6. Clusters of Similar Means. A collection $\mathcal{D} = \{D_1, \dots, D_m\}$ of distributions over $\{0, \dots, n\}$ has the property of (k, γ)-similar means if the means of all distributions reside in k intervals, each of size γn. In this section we provide an algorithm for testing (k, γ)-similar means in the query model (Algorithm 4) and using Lemma 9 we extend the algorithm to the sampling model. The algorithm and the proofs are a straightforward generalization of testing γ-similar means.

The idea behind the algorithm in the query model is that if the collection is far from having (k, γ)-similar means, then by taking enough samples we are guaranteed to find, with high constant probability, a subset of distributions of $\mathcal{D}$, $D_{i_1}, \dots, D_{i_{k+1}}$, such that $\mu(D_{i_{s+1}}) - \mu(D_{i_s}) > \gamma n$ for every integer $1 \le s \le k$ and thus have evidence that $\mathcal{D}$ does not have (k, γ)-similar means. To prove that $\tilde{O}(k^2/\epsilon^2)$ queries suffice in order to find evidence when the collection is ϵ-far from (k, γ)-similar we extend Lemma 2 and prove that there exist k disjoint intervals each of size γn such that between any adjacent intervals there are enough distributions that are far in the ℓ_1-norm from having mean within one of the intervals (see Lemma 10). Similarly to γ-similar means testing algorithm, here too we use the algorithm GetBounds and a bucketing technique to save a factor of $\tilde{\Omega}(k/\epsilon)$ samples.

Algorithm 4 testing (k, γ)-similar means

1. For $q = 1$ to r, where $r = \lceil \log(8k/\epsilon) \rceil$ do:
   - Select $t(q) = \Theta\left(2^q \log(k/\epsilon) \cdot \log k\right)$ distributions from the collection, and denote them by $D_1^q, \dots, D_{t(q)}^q$.
   - For each D_j^q selected let $(u_j^q, \ell_j^q) = \text{GetBounds}\left(D_j^q, (\epsilon/(8k))2^{q-1}, \frac{1}{(6rt(q))}\right)$
2. If there exist $(q_1, j_1), \dots, (q_{k+1}, j_{k+1})$ such that $\ell_{j_{s+1}}^{q_{s+1}} - u_{j_s}^{q_s} > \gamma n$ for every integer $1 \le s \le k$ then REJECT, otherwise ACCEPT.

LEMMA 10. *Let $\mathcal{D}$ be a collection of distributions. If $\mathcal{D}$ is ϵ-far from having (k, γ)-similar means, then there exist k intervals of $[n]$, denoted $[x_1, y_1], \dots, [x_k, y_k]$ where $0 = y_0 < x_1 < y_1 < \dots x_k < y_k < x_{k+1} = \infty$ such that $y_\ell - x_\ell = \gamma n$ and $\sum_{i: y_{j-1} < \mu(D_i) < x_j} \min(d_1(D_i, x_j), d_1(D_i, y_{j-1})) > (\epsilon/(4k))m$ for every $\ell \in [k]$ and $j \in [k+1]$.*

Proof. Let $\mathcal{D}$ be a collection that is ϵ-far from having (k, γ)-similar means. Let x_1 be the minimum value of the form $\ell \cdot (\epsilon/(8k))n$, where ℓ is an integer, such that

$\sum_{i:\mu(D_i)<x_1} d_1(D_i, x_1) > (\epsilon/(4k))m$, and let $y_1 = x_1 + \gamma n$. By the definition of x_1, if we consider $x_1' = x_1 - (\epsilon/(16k))n$, then $\sum_{i:\mu(D_i)<x_1'} d_1(D_i, x_1') \leq (\epsilon/(4k))m$. Thus, by Lemma 1,

$$\sum_{i:\mu(D_i)>y_1} d_1(D_i, y_1) > 1 - (\epsilon/(4k))m - 4 \cdot (\epsilon/(16k)) = (1 - \epsilon/(2k))m \qquad (6.1)$$

(see elaboration in the proof of Lemma 2). Now, given y_j, for each integer $j \geq 1$, iteratively define x_{j+1} to be the minimum value of the form $\ell \cdot (\epsilon/(8k))n$, where ℓ is an integer, such that $\sum_{i:y_j<\mu(D_i)<x_{j+1}} \min(d_1(D_i, x_{j+1}), d_1(D_i, y_j)) > (\epsilon/(4k))m$, and define $y_{j+1} = x_{j+1} + \gamma n$. By the definition of x_{j+1}, if we consider any $y_j < x_{j+1}' < x_{j+1}$, in particular $x_{j+1}' = \max(x_{j+1} - (\epsilon/(16k))n, (y_j + x_{j+1})/2)$, then $\sum_{i:y_j<\mu(D_i)<x_{j+1}'} \min(d_1(D_i, x_{j+1}'), d_1(D_i, y_j)) \leq (\epsilon/(4k))m$. Since $x_{j+1} - x_{j+1}' \leq (\epsilon/(16k))n$, it is implied by Lemma 1 that $\sum_{i:\mu(D_i)>y_{j+1}} d_1(D_i, y_{j+1}) > (1 - \epsilon \cdot (j + 1)/(2k))m$. Therefore, the k intervals are well-defined. $\square$

THEOREM 7. *Algorithm 4 tests (k, γ)-similar means in the query model. The algorithm's query complexity is $\tilde{O}(k^2/\epsilon^2)$.*

Proof. Let E_g denote the event that all pairs (u_j^q, ℓ_j^q) returned by the algorithm Get-Bounds are as specified in Lemma 3. Since each call to GetBounds in iteration q is done with $\delta = 1/(6rt(q))$, by Lemma 3 the probability that E_g holds is at least $5/6$. Observe that if there exists a subset of distribution, $D_{i_1}, \ldots, D_{i_{k+1}}$, of $\mathcal{D}$ such that $\mu(D_{i_{s+1}}) - \mu(D_{i_s}) > \gamma n$ for every integer $1 \leq s \leq k$ then $\mathcal{D}$ does not have (k, γ)-similar means. Therefore, if $\mathcal{D}$ has (k, γ)-similar means, then, conditioned on E_g, the algorithm accepts.

We now turn to the case that $\mathcal{D}$ is ϵ-far from having (k, γ)-similar means. Let $[x_1, y_1], \ldots, [x_k, y_k]$ be intervals as described in Lemma 10. For every integer $1 \leq j \leq k+1$, we partition the distributions D_i such that $y_{j-1} < \mu(D_i) < x_j$ into *buckets* B_q^j, for $1 \leq q \leq r$,

$$B_q^j = \{i : (\epsilon/(8k))2^{q-1} < \min(d_1(D_i, x_j), d_1(D_i, y_{j-1})) \leq (\epsilon/(8k))2^q\} . \qquad (6.2)$$

Define $A_1 = \{i : y_{j-1} < \mu(D_i) < x_j\}$. By Lemma 10

$$\sum_{i \in A_1} \min(d_1(D_i, x_j), d_1(D_i, y_{j-1})) > (\epsilon/(4k))m . \qquad (6.3)$$

Define $A_2 = \{i : \min(d_1(D_i, x_j), d_1(D_i, y_{j-1})) \leq \epsilon/(8k)\}$. Clearly,

$$\sum_{i \in A_2} \min(d_1(D_i, x_j), d_1(D_i, y_{j-1})) \leq (\epsilon/(8k))m . \qquad (6.4)$$

Define $A_3 = \{i : y_{j-1} < \mu(D_i) < x_j, \min(d_1(D_i, x_j), d_1(D_i, y_{j-1})) > \epsilon/(8k)\}$. It follows that

$$\sum_{i \in A_3} \min(d_1(D_i, x_j), d_1(D_i, y_{j-1})) \geq (\epsilon/(8k))m , \qquad (6.5)$$

and so there exists an index q_j such that

$$\sum_{i \in B_{q_j}^j} \min(d_1(D_i, x_j), d_1(D_i, y_{j-1})) \geq (\epsilon/(8k))m/r \qquad (6.6)$$

(where r is the number of buckets). By the definition of the buckets, $r = O(\log(k/\epsilon))$, and for the bucket $B^j_{q_j}$ we have

$$|B^j_{q_j}| > \frac{(\epsilon/(8k))m/r}{(\epsilon/(8k))2^{q_j}} = \Omega\left(\frac{m}{\log(k/\epsilon)2^{q_j}}\right) \ . \tag{6.7}$$

But in such a case, since the algorithm selects $t(q) = \Theta\left(2^q \log(k/\epsilon) \cdot \log k\right)$ distributions in each iteration q, by union bound, with high constant probability, for every j, the algorithm will select a distribution D_i such that $i \in B^j_{q_j}$ in iteration q_j, and conditioned on the event E_g, will reject, as required.

Let $s(q)$ denote the number of queries performed in iteration q by the algorithm GetBounds for each distribution it is called on. The query complexity of the algorithm is

$$\sum_{q=1}^{r} t(q) \cdot s(q) = O\left(\sum_{q=1}^{r} 2^q \log(k/\epsilon) \cdot \log k \cdot \frac{k^2 \log(k/\epsilon)}{2^{2q}\epsilon^2}\right) = \tilde{O}(k^2/\epsilon^2) \tag{6.8}$$

and the theorem follows. $\square$

THEOREM 8. *The sample complexity of testing (k, γ)-similar means in the uniform sampling model is upper bounded by $m^{1-\tilde{\Omega}(\epsilon^2/k^2)} \cdot \tilde{O}(k^2/\epsilon^2)$.*

Proof. Algorithm 4 which works in the query model performs $r = \lceil \log(8/\epsilon) \rceil$ iterations and in the q-th iteration, selects $t(q) = \Theta\left(2^q \log(k/\epsilon) \cdot \log k\right)$ distributions. From each selected distribution it then takes $s(q) = O\left(\frac{k^2 \log(k/\epsilon)}{2^{2q}\epsilon^2}\right)$ samples. By Lemma 9 we get that by taking $s(q)m^{1-1/s(q)}$ samples we get $s(q)$-way collisions on $\Omega(1)$ distributions. If for each iteration q, we repeat this process $t(q)\log(t(q)^2)$ times then by applying the union bound over r iterations, with high constant probability we would have enough samples to simulate Algorithm 2. Thus the sample complexity we get is bounded by

$$\sum_{q=1}^{r} t(q)\log(t(q)^2) \cdot s(q)m^{1-1/s(q)} = m^{1-\tilde{\Omega}(\epsilon^2/k^2)} \cdot \tilde{O}(k^2/\epsilon^2) \ , \tag{6.9}$$

as desired. $\square$

REFERENCES

[1] N. ALON, A. ANDONI, T. KAUFMAN, K. MATULEF, R. RUBINFELD, AND N. XIE, *Testing k-wise and almost k-wise independence*, in Proceedings of STOC, 2007, pp. 496–505.

[2] T. BATU, S. DASGUPTA, R. KUMAR, AND R. RUBINFELD, *The complexity of approximating the entropy.*, SIAM J. Comput., 35 (2005), pp. 132–150.

[3] T. BATU, L. FORTNOW, E. FISCHER, R. KUMAR, R. RUBINFELD, AND P. WHITE, *Testing random variables for independence and identity*, in Proceedings of FOCS, 2001, pp. 442–451.

[4] T. BATU, L. FORTNOW, R. RUBINFELD, W.D. SMITH, AND P. WHITE, *Testing closeness of discrete distributions*, J. ACM, 60 (2013), p. 4.

[5] T. BATU, R. KUMAR, AND R. RUBINFELD, *Sublinear algorithms for testing monotone and unimodal distributions*, in Proceedings of STOC, 2004, pp. 381–390.

[6] P. L. CHEBYSHEV, *Théorie des mécanismes connus sous le nom de parallélogrammes*, Mémoires des Savants étrangers présentés â l'Académie de Saint-Pétersbourg, 7 (1854), pp. 539–586.

[7] B. EFORN AND C. STEIN, *The jacknife estimate of variance*, Annals of Statistics, 9 (1981), pp. 586–596.

[8] O. GOLDREICH, S. GOLDWASSER, AND D. RON, *Property testing and its connection to learning and approximation*, J. ACM, 45 (1998), pp. 653–750.

[9] O. GOLDREICH AND D. RON, *On testing expansion in bounded-degree graphs*, Electronic Colloqium on Computational Complexity, 7 (2000).

[10] ——, *Property testing in bounded degree graphs*, Algorithmica, (2002), pp. 302–343.

[11] S. Guha, A. McGregor, and S. Venkatasubramanian, *Sub-linear estimation of entropy and information distances*, ACM Trans. Algorithms, 5 (2009).

[12] R.J. Larsen and M.L. Marx, *An introduction to mathematical statistics and its applications Volume 1*, Pearson Prentice Hall, 2006.

[13] Reut Levi, Dana Ron, and Ronitt Rubinfeld, *Testing properties of collections of distributions*, Theory Comput., 9 (2013), pp. 295–347.

[14] W. Mendenhall, R.J. Beaver, and B.M. Beaver, *Introduction to probability and statistics*, Brooks/Cole, Cengage Learning, 2009.

[15] L. Paninski, *Testing for uniformity given very sparsely-sampled discrete data*, IEEE Transactions on Information Theory, 54 (2008), pp. 4750–4755.

[16] S. Raskhodnikova, D. Ron, A. Shpilka, and A. Smith, *Strong lower bonds for approximating distributions support size and the distinct elements problem*, SIAM J. Comput., 39 (2009), pp. 813–842.

[17] R. Rubinfeld and M. Sudan, *Robust characterization of polynomials with applications to program testing*, SIAM J. Comput., 25 (1996), pp. 252–271.

[18] Student, *The probable error of a mean*, Biometrika, 6 (1908), pp. 1–25.

[19] G. Valiant and P. Valiant, *Estimating the unseen: an $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs*, in Proceedings of STOC, 2011, pp. 685–694.

[20] P. Valiant, *Testing symmetric properties of distributions*, SIAM J. Comput., 40 (2011), pp. 1927–1968.

[21] B. L. Welch, *The generalization of 'student's' problem when several different population variances are involved*, Biometrika, 34 (1947), pp. 28–35.

[22] P. White, *Testing random variables for independence and identity*. Unpublished manuscript.