Electronic Colloquium on Computational Complexity, Report No. 55 (2012)

Testing Similar Means

Reut Levi * Dana Ron † Ronitt Rubinfeld ‡

May 4, 2012

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 (measured according to the ℓ_1 distance averaged over the distributions) 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 $m^{1-\mathrm{poly}(\epsilon)}$, where m is the number of distributions.

*School of Computer Science, Tel Aviv University. Tel Aviv 69978, Israel. E-mail: reuti.levi@gmail.com. Research supported by the Israel Science Foundation grant nos. 1147/09 and 246/08

†School of Electrical Engineering, Tel Aviv University. Tel Aviv 69978, Israel. E-mail: danar@eng.tau.ac.il. Research supported by the Israel Science Foundation grant number 246/08.

‡CSAIL, MIT, Cambridge MA 02139 and the Blavatnik School of Computer Science, Tel Aviv University. E-mail: ronitt@csail.mit.edu. Research supported by NSF grants CCF-1065125 and CCF-0728645, Marie Curie Reintegration grant PIRG03-GA-2008-231077 and the Israel Science Foundation grant nos. 1147/09 and 1675/09.

ISSN 1433-8092

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 in many contexts: 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 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 [19, 9, 5], 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 [15], 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.

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 emulating 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 [14] 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. The construction of the random variables is based on Chebyshev polynomials [7], whose roots, and their derivatives at the roots, have useful properties that we exploit.

1.2 Related Work

The work that is most closely related to the work presented in this paper appears in [15]. The testing models used here were introduced in [15], 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. 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 sampling complexity $\mathrm{poly}(1/\epsilon)\sqrt{m}$ that is given in [15] 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 [14] 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 exten-

sively in the past (see [10, 17, 4, 3, 1, 2, 12, 6, 18, 22, 21]).

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 mean of two normally distributed populations are equal [16, 20, 24]. The family of procedures ANOVA (Analysis of Variance), applies when there are more than two normally distributed populations (see [13, 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. In all of these procedures, the problem solved is more stringent than in our property testing setting, but the assumptions made in all settings are quite strong, e.g., assuming the normality of the distributions and assuming that all distributions have the same variance, and thus the sample complexity bounds are incomparable to those in our setting.

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).

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 $1 \leq i \leq m$. 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 [15], for a property $\mathcal{P}$ of collections of distributions and $0 \leq \epsilon \leq 1$, 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 $1 \leq i \leq m$ 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 $\mu(D) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{j=1}^{n} i \cdot D(j)$. 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 $0 \leq \gamma \leq 1$, we say that $\mathcal{D}$ has γ-*similar means* if $|\mu_j - \mu_{j'}| \leq \gamma n$ for every $j \neq j'$.

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 $\mu(D) \stackrel{\text{def}}{=} \sum_{i=1}^{n} i \cdot D(i)$ for the mean of D. for a value $0 \leq z \leq n$ let $d_1(D,z) \stackrel{\text{def}}{=} \min_{D':\mu(D')=z}\{\|D-D'\|_1\}$ 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 $\epsilon \leq 1/16$. If $\mu \geq n/2$, then for every $\mu' \in [\mu - \epsilon n, \mu]$ there exists D' such that $\mu(D') = \mu'$ and $\|D - D'\|_1 \leq 4\epsilon$. If $\mu \leq n/2$, then for every $\mu' \in [\mu, \mu + \epsilon n]$ there exists D' such that $\mu(D') = \mu'$ and $\|D - D'\|_1 \leq 4\epsilon$.*

Proof: Let $X \sim D$. In the case that $\mu \geq n/2$ we have that $\Pr[X \geq n/4] \geq 1/4 \geq 4\epsilon$. Therefore, we can decrease the mean of D by at most ϵn by moving at most 4ϵ weight from its support on points in the interval $[n/4, n]$, to 0. In the case that $\mu \leq n/2$, we have that $\Pr[X \leq 3n/4] > 1/3 \geq 4\epsilon$. Therefore, we can increase the mean of D by at most ϵn by moving at most 4ϵ 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 $y - x \geq \gamma n + \epsilon n/8$ such that $\sum_{i:\mu(D_i)>y} d_1(D_i,y) > (\epsilon/4)m$ and $\sum_{i:\mu(D_i)<x} d_1(D_i,x) > (\epsilon/4)m$. In particular, there are more than $(\epsilon/4)m$ distributions whose mean is at most x and more than $(\epsilon/4)m$ distributions whose mean is at least y.*

Proof: Let $\mathcal{D}$ be a collection which is ϵ-far from having γ-similar means. By defintion, 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 $y - x > \gamma + \epsilon/8$, we have that

$$\sum_{i:\mu(D_i)<x} d_1(D_i, x) + \sum_{i:\mu(D_i)>y} d_1(D_i, y) > (\epsilon/2)m \ . \tag{1}$$

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 maximum value such that $\sum_{i:\mu(D_i)<x'} d_1(D_i, x') \leq (\epsilon/4)m$, and let $y = x + \gamma n + (\epsilon/2)n$. By the claim we have just established, $\sum_{i:\mu(D_i)>y} d_1(D_i, y) > (\epsilon/4)m$. By the definition of x', if we consider any $x > x'$, in particular $x = x' + (3\epsilon/8)n$, then $\sum_{i:\mu(D_i)<x} d_1(D_i, x) > (\epsilon/4)m$. ■

The correctness of the basic algorithm follows from Lemma 2: If $\mathcal{D}$ is ϵ-far from having γ-similar means, then 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). The intuition underlying the modification (similar to that applied for example in [11]) 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.

To make the above notion of estimation more precise, we introduce some notations. For $0 \leq \beta \leq 1$ and D such that $d_1(D, n) \geq \beta$ (where $d_1(\cdot, \cdot)$ is as defined at the beginning of this section), let $\mu^{>}_{\beta}(D)$ equal $\mu > \mu(D)$ such that $d_1(D, \mu) = \beta$ and for D such that $d_1(D, 0) \geq \beta$, let $\mu^{<}_{\beta}(D)$ equal $\mu < \mu(D)$ such that $d_1(D, \mu) = \beta$. If $d_1(D, n) < \beta$, then $\mu^{>}_{\beta}(D) \stackrel{\text{def}}{=} n$ 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).

We first describe a procedure 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 $\mu(D)$. The number of samples that the procedure takes is quadratic in $1/\beta$. Note that if the domain is $\{0,1\}$ (or any constant size domain), then the procedure can simply estimate the mean of D to within β. However, for a general domain, the procedure is different. We later show how to apply this procedure so as to obtain a testing algorithm with query complexity $\tilde{O}(1/\epsilon^2)$.

The idea behind the procedure is the following. Consider a distribution D. For any value $0 \leq a \leq n$ we have that

$$\sum_{i>a} i \cdot D(i) + \sum_{i \leq a} a \cdot D(i) \geq \mu(D) \; . \tag{2}$$

On the other hand, by the definition of $\mu^{>}_{\beta}(D)$, if a is such that $\Pr_D[i \leq a] \leq \beta$, then

$$\sum_{i>a} i \cdot D(i) + \sum_{i \leq a} n \cdot D(i) \leq \mu^{>}_{\beta}(D) \; . \tag{3}$$

Let a indeed be a value that satisfies $\Pr_D[i \leq a] \leq \beta$, let $a' = a + (n-a)/2 = (a+n)/2$ and let

$$\mu^a(D) \stackrel{\text{def}}{=} \sum_{i>a} i \cdot D(i) + a' \cdot \Pr_D[i \leq a] \; . \tag{4}$$

Then on one hand

$$\mu^a(D) \geq \mu(D) + ((n-a)/2) \cdot \Pr_D[i \leq a] \tag{5}$$

and on the other hand

$$\mu^a(D) \leq \mu^{>}_{\beta}(D) - ((n-a)/2) \cdot \Pr_D[i \leq a] \; . \tag{6}$$

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 x 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).

Lemma 3 *The procedure* GetBounds(D,β,δ) *returns x and y such that with probability at least $1-\delta$ (over its internal coin flips), $\mu(D) \leq x \leq \mu^{>}_{\beta}(D)$ and $\mu^{<}_{\beta}(D) \leq y \leq \mu(D)$.*

Proof: We prove the claim for x, and an analogous (symmetric) analysis holds for y. Let a'' be as defined in Step 4 of Procedure GetBounds, and let

$$D^{a,a''}(i) \stackrel{\text{def}}{=} \begin{cases} D(i) & \text{if } i > a, i \neq a', i \neq a'' \\ D(i) + \Pr_D[i < a] & \text{if } i = a' \\ D(i) + D(a) & \text{if } i = a'' \\ 0 & \text{o.w.} \end{cases} \tag{7}$$

By the definition of the distribution $D^{a,a''}$ we have that

$$\mu(D^{a,a''}) = \sum_{i>a} i \cdot D(i) + a' \cdot \Pr_D[i < a] + a'' \cdot \Pr_D[i = a] \; . \tag{8}$$

Procedure 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 $x = \frac{1}{s_3}\left(\sum_{i_j > a} i_j + \sum_{i_j < a} a' + \sum_{i_j = a} a''\right)$ and $y = \frac{1}{s_3}\left(\sum_{i_j < b} i_j + \sum_{i_j > b} b' + \sum_{i_j = b} b''\right)$.
6. Return (x, y).

Observe that in Step 5, the procedure takes s_3 independent samples from $D^{a,a''}$ and that $\mathrm{E}[x] = \mu(D^{a,a''})$.

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 $\Pr_D[i \leq a] \geq \beta/8$. 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,

$$\mu(D^{a,a''}) \geq \mu(D) + \frac{n-a}{2} \cdot \Pr_D[i \leq a] \geq \mu(D) + (n-a) \cdot (\beta/16)\ , \tag{9}$$

and on the other hand

$$\mu(D^{a,a''}) < \mu^{>}(D) - \frac{n-a}{2} \cdot \Pr_D[i \leq a] \leq \mu^{>}(D) - (n-a) \cdot (\beta/16)\ . \tag{10}$$

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

$$\mu(D^{a,a''}) \geq \mu(D) + \frac{n-a}{2} \cdot \left(\Pr_D[i < a] + \frac{\beta}{4\hat{\alpha}(a)} \cdot \Pr_D[i = a]\right) \geq \mu(D) + \frac{n-a}{2} \cdot \frac{\beta}{32} \tag{11}$$

and similarly $\mu(D^{a,a''}) < \mu^{>}_{\beta}(D) - (n-a) \cdot (\beta/32)$. By the definition of x 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 $|x - \mu(D^{a,a''})| \leq (n-a) \cdot (\beta/32)$ implying that $\mu(D) \leq x \leq \mu^{>}_{\beta}(D)$. ■

Theorem 1 *Algorithm 1 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 (x_j^q, y_j^q) returned by the procedure GetBounds are as specified in Lemma 3. Since each call to GetBounds in iteration q is done with $\delta = 1/(6r\ell(q))$, by Lemma 3 the

Algorithm 1: testing γ-similar means

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

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

$$B_q^L = \{i : (\epsilon/8)2^{q-1} < d_1(D_i, x) \leq (\epsilon/8)2^q\} \ , \tag{12}$$

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

$$B_q^R = \{i : (\epsilon/8)2^{q-1} < d_1(D_i, y) \leq (\epsilon/8)2^q\} \ . \tag{13}$$

Since $\sum_{i:\mu(D_i)<x} d_1(D_i, x) > (\epsilon/4)m$ and $\sum_{i:\mu(D_i)<x, d_1(D_i,x)\leq\epsilon/8} d_1(D_i, x) \leq (\epsilon/8)m$ we have that there exists an index q^L such that

$$|B_{q^L}^L| > \Big((\epsilon/4)m\Big) / \Big((\epsilon/8)2^{q^L}\log(8/\epsilon)\Big) = \Omega\left(2^{-q^L} m / \log(1/\epsilon)\right) \ , \tag{14}$$

and similarly there exists an index q^R such that $|B_{q^R}^R| = \Omega\left(2^{-q^R} m / \log(1/\epsilon)\right)$. But in such a case, with high constant probability, the algorithm will select a distribution D_i such that $i \in B_{q^L}^L$ in iteration 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 procedure GetBounds for each distribution it is called on. The query complexity of the algorithm is

$$\sum_{q=1}^{r} \ell(q) \cdot s(q) = O\left(\sum_{q=1}^{r} 2^q \log(1/\epsilon) \cdot \frac{\log(1/\epsilon)}{2^{2q}\epsilon^2}\right) = O(\log^2(1/\epsilon)/\epsilon^2) \tag{15}$$

and the theorem follows. ■

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 emulated when the coin is unbiased) or is $\epsilon/2$-far from γ-similar means (when $D = D_2$, which is emulated when the coin is biased). ■

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, $D_{\mathrm{KL}}(p_1 \| p_2) \stackrel{\mathrm{def}}{=} \sum_{x \in X} p_1(x) \cdot \ln \frac{p_1(x)}{p_2(x)}$.

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

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

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 sampling 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 $n \ge 1$, 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 $X^+ = X\left(\{d_i^+\}, \{\alpha_i^+\}\right)$ and $X^- = X\left(\{d_i^-\}, \{\alpha_i^-\}\right)$ we have*

$$\mathrm{E}\left[\left(X^+\right)^i\right] = \mathrm{E}\left[\left(X^-\right)^i\right] \quad \forall i \in [t] \; . \tag{16}$$

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*

$$\alpha_{\max}^-(d_{\max}^- - d_{\max}^+) = \tilde{\Theta}\left(\frac{1}{t^2}\right) \; , \tag{17}$$

and

$$d_{\max}^+ - d_{\min}^+ = \Theta(1) \; . \tag{18}$$

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.
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 $1 \leq i \leq t/2$).
4. $|\tilde{\alpha}_i^+ - \alpha_i^+| \leq 1/m$ and $|\tilde{\alpha}_i^- - \alpha_i^-| \leq 1/m$.

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

$$D(j) \stackrel{\text{def}}{=} \begin{cases} \frac{1}{2} \cdot \left(1 + d_i^+ \delta\right) & \text{if } j = 0 \\ \frac{1}{2} \cdot \left(1 - d_i^+ \delta\right) & \text{if } j = n \\ 0 & \text{o.w.} \end{cases} \tag{19}$$

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

$$D(j) \stackrel{\text{def}}{=} \begin{cases} \frac{1}{2} \cdot \left(1 - d_i^+ \delta\right) & \text{if } j = 0 \\ \frac{1}{2} \cdot \left(1 + d_i^+ \delta\right) & \text{if } j = n \\ 0 & \text{o.w.} \end{cases} \tag{20}$$

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 $t \leq m^{1/2}$, 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.

$$\gamma \;=\; \frac{1}{2} \cdot (1 + d_{\max}^+ \delta) - \frac{1}{2} \cdot (1 - d_{\max}^+ \delta) \;=\; d_{\max}^+ \delta \; . \tag{21}$$

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

$$\frac{\epsilon}{\gamma} = \frac{\alpha_{\max}^+ \cdot (d_{\max}^+ - d_{\max}^-)}{d_{\max}^+} = \tilde{\Theta}\left(\frac{1}{t^2}\right) \; . \tag{22}$$

We note that for every $\epsilon/\gamma \leq 1/\log^2 m$ 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. ■

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:

$$\begin{aligned} T_0(x) &= 1 && (23)\\ T_1(x) &= x && (24)\\ T_{\ell+1}(x) &= 2xT_\ell(x) - T_{\ell-1}(x)\ . && (25) \end{aligned}$$

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

$$\begin{aligned} U_0(x) &= 1 && (26)\\ U_1(x) &= 2x && (27)\\ U_{\ell+1}(x) &= 2xU_\ell(x) - U_{\ell-1}(x)\ . && (28) \end{aligned}$$

Then we have that

$$\frac{dT_\ell(x)}{dx} = \ell \cdot U_{\ell-1}\ , \tag{29}$$

and that

$$U_{\ell-1}(\cos(x)) = \frac{\sin(\ell x)}{\sin x}\ . \tag{30}$$

T_ℓ has ℓ different simple roots:

$$x_i = \cos\left(\frac{\pi}{2}\cdot\frac{2i-1}{\ell}\right) \quad i = 1,\ldots,\ell \tag{31}$$

and the following equalities hold:

$$T_\ell(1) = 1, \quad \text{and} \quad T_\ell(-1) = (-1)^\ell\ . \tag{32}$$

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

Lemma 7 ([21]) *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 $j \leq \ell - 2$ 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:

$$f(x) \stackrel{\text{def}}{=} (x-1)(x+1)\cdot T_\ell(x)\ , \tag{33}$$

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

$$\frac{1}{|f'(1)|} = \frac{1}{2T_\ell(1)} = \frac{1}{2}\ , \tag{34}$$

and

$$\frac{1}{|f'(-1)|} = \frac{1}{|2T_\ell(-1)|} = \frac{1}{2} \ . \tag{35}$$

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

$$\frac{1}{|f'(x_i)|} \ = \ \frac{1}{\left|(x_i^2-1)\cdot T'_\ell(x_i)\right|} \ . \tag{36}$$

By Equations (29) and (30):

$$\frac{1}{\left|T'_\ell(x_i)\right|} \ = \ \left|\frac{\sin\left(\frac{\pi}{2}\cdot\frac{2i-1}{\ell}\right)}{\ell\cdot\sin\left(\frac{\pi}{2}\cdot(2i-1)\right)}\right| = \frac{1}{\ell}\cdot\left|\sin\left(\frac{\pi}{2}\cdot\frac{2i-1}{\ell}\right)\right| \tag{37}$$

where we used the fact that $\left|\sin\left(\frac{\pi}{2}\cdot(2i-1)\right)\right| = 1$. Therefore by Equation (31) and the identity $1-\cos^2 x = \sin^2 x$ we obtain:

$$\frac{1}{|f'(x_i)|} \ = \ \frac{1}{\ell}\cdot\frac{1}{\left|\sin\left(\frac{\pi}{2}\cdot\frac{2i-1}{\ell}\right)\right|} \ . \tag{38}$$

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

$$\frac{1}{|f'(x_i)|} \ \le \ \frac{1}{\ell}\cdot\frac{\pi}{2}\cdot\frac{1}{\left(\frac{\pi}{2}\cdot\frac{2i-1}{\ell}\right)} \ = \ \frac{1}{2i-1} \ . \tag{39}$$

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

$$\sum_{i=1}^{\ell}\frac{1}{|f'(x_i)|} \le 2\sum_{x_i\ge 0}\frac{1}{|f'(x_i)|} = O(\log\ell) \ . \tag{40}$$

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

$$\alpha_i^- = \frac{1}{\left|f'(d_i^-)\right|}\cdot\beta^- \ , \tag{41}$$

where $\beta^- = 1/\left(\sum_{i=1}^{t+1}\frac{1}{\left|f'(d_i^-)\right|}\right)$ 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 (40), $\alpha^-_{\max} = \Omega(1/\log\ell)$. 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:

$$\lim_{x\to 0}\frac{1-\cos x}{x^2} = \lim_{x\to 0}\frac{\sin x}{x}\cdot\frac{\sin(x/2)}{x}\cdot\frac{1}{\cos(x/2)} = \frac{1}{2} \ , \tag{42}$$

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 $d^-_{\min} - d^+_{\min} = \Theta\left(1/\ell\right)$. By Equations (38) and (40) we get $\alpha^+_{\min} = \tilde{\Theta}(1/\ell)$. Therefore, Equations (17) and (18) hold. Equation (16) 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. ■

4.1.2 Proof of Lemma 6

We first recall the definition of the Poisson distribution, and introduce one more lemma (from [23]). For a positive real number λ, the Poisson distribution $\mathrm{poi}(\lambda)$ takes the value $x \in \mathbb{N}$ (where $\mathbb{N} = \{0, 1, 2, \ldots\}$) with probability $\mathrm{poi}(x; \lambda) = e^{-\lambda}\lambda^x/x!$. For the collection $\mathcal{D} = (D_1, D_2)$ define

$$\vec{\lambda}^{\mathcal{D},k}(a, b) \stackrel{\text{def}}{=} \sum_i \mathrm{poi}(a; k \cdot D_1(i))\mathrm{poi}(b; k \cdot D_2(i)) \,, \tag{43}$$

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 $\mathrm{poi}(k)$ samples from $\mathcal{D}$ in the uniform sampling model (see [14]).

Lemma 8 ([23]) *Given a positive integer k and two distribution pairs $p_1^+, p_2^+, p_1^-, p_2^-$ all of whose frequencies are at most $\frac{1}{2000k}$, let $\vec{\lambda}^+(a, b) = \sum_i \mathrm{poi}(a; k \cdot p_1^+(i))\mathrm{poi}(b; k \cdot p_2^+(i))$ and $\vec{\lambda}^-(a, b) = \sum_i \mathrm{poi}(a; k \cdot p_1^-(i))\mathrm{poi}(b; k \cdot p_2^-(i))$ for $a + b > 0$. If it is the case that*

$$\sum_{a+b>0} \frac{\left|\vec{\lambda}^+(a, b) - \vec{\lambda}^-(a, b)\right|}{\sqrt{1 + \max\{\vec{\lambda}^+(a, b), \vec{\lambda}^-(a, b)\}}} < \frac{1}{50} \,, \tag{44}$$

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

$$p_0^+(i) \stackrel{\text{def}}{=} 2D_i^+(0)/m \tag{45}$$

$$p_n^+(i) \stackrel{\text{def}}{=} 2D_i^+(n)/m \,. \tag{46}$$

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 (43). 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.

$$\begin{aligned}
\vec{\lambda}^{\mathcal{P}^+,k}(a, b) &= \sum_{i=1}^{m} \mathrm{poi}(a; k \cdot p_0^+(i)) \cdot \mathrm{poi}(b; k \cdot p_n^+(i)) && (47)\\
&= \sum_{i=1}^{t} \alpha_i^+ m \cdot \big(\mathrm{poi}(a; k \cdot (1 + d_i^+\delta)/m) \cdot \mathrm{poi}(b; k \cdot (1 - d_i^+\delta)/m)\big) && (48)\\
&\quad + \sum_{i=1}^{t} \alpha_i^+ m \cdot \big(\mathrm{poi}(a; k \cdot (1 - d_i^+\delta)/m) \cdot \mathrm{poi}(b; k \cdot (1 + d_i^+\delta)/m)\big) && (49)\\
&= \frac{e^{-k/m}}{a!b!}\left(\frac{k}{m}\right)^{a+b} \sum_{i=1}^{t} \alpha_i^+ m \cdot \Big(\left(1 + d_i^+\delta\right)^a \left(1 - d_i^+\delta\right)^b + \left(1 - d_i^+\delta\right)^a \left(1 + d_i^+\delta\right)^b\Big) && (50)
\end{aligned}$$

Similarly,

$$\vec{\lambda}^{\mathcal{P}^-,k}(a,b) \;=\; \frac{e^{-k/m}}{a!b!}\left(\frac{k}{m}\right)^{a+b}\sum_{i=1}^{t}\alpha_i^- m\cdot\Big(\big(1+d_i^-\delta\big)^a\big(1-d_i^-\delta\big)^b+\big(1-d_i^-\delta\big)^a\big(1+d_i^-\delta\big)^b\Big)\;. \quad (51)$$

Therefore, there exist coefficients β_j such that,

$$\begin{aligned}\vec{\lambda}^{\mathcal{P}^+,k}(a,b) &= \frac{e^{-k/m}}{a!b!}\left(\frac{k}{m}\right)^{a+b}\cdot\sum_{i=1}^{t}\alpha_i^+ m\cdot\sum_{j=0}^{a+b}\beta_j\cdot(d_i^+\delta)^j &(52)\\ &= \frac{e^{-k/m}}{a!b!}\left(\frac{k}{m}\right)^{a+b}\cdot\sum_{j=0}^{a+b}m\beta_j\delta^j\cdot\sum_{i=1}^{t}\alpha_i^+(d_i^+)^j\;, &(53)\end{aligned}$$

and

$$\vec{\lambda}^{\mathcal{P}^-,k}(a,b) \;=\; \frac{e^{-k/m}}{a!b!}\left(\frac{k}{m}\right)^{a+b}\cdot\sum_{j=0}^{a+b}m\beta_j\delta^j\cdot\sum_{i=1}^{t+1}\alpha_i^-(d_i^-)^j \quad (54)$$

By Equation (16) we have that $\sum_{i=1}^{t}\alpha_i^+(d_i^+)^j=\sum_{i=1}^{t+1}\alpha_i^-(d_i^-)^j$ for every $0\le j\le t$. By Equations (53) and (54) we get that for every $a+b\le t$,

$$\vec{\lambda}^{\mathcal{P}^+,k}(a,b)=\vec{\lambda}^{\mathcal{P}^-,k}(a,b), \quad (55)$$

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

$$\begin{aligned}\sum_{a+b>t}\lambda(a,b)\le\sum_{a+b>t}\left(\frac{k}{m}\right)^{a+b}\cdot m &= \sum_{a+b>t}\left(\frac{1-\delta}{4m^{1/t}}\right)^{a+b}\cdot m &(56)\\ &\le\sum_{x>t}2^x\left(\frac{1}{4m^{1/t}}\right)^x\cdot m\ll 1\;. &(57)\end{aligned}$$

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 (50) and (51), calculating the sum in Equation (44) 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

$$\begin{aligned}&\sum_{a+b>0}\frac{\frac{e^{-k/m}}{a!b!}\left(\frac{k}{m}\right)^{a+b}\sum_{i=1}^{t}m(|\alpha_i^+-\tilde{\alpha}_i^+|+|\alpha_i^--\tilde{\alpha}_i^-|)\cdot 2^{a+b}}{(\vec{\lambda}^{\mathcal{P}^-,k}(a,b))^{1/2}}\\ &\le\sum_{a+b>0}\left(\frac{e^{-k/m}}{a!b!}\left(\frac{k}{m}\right)^{a+b}\right)^{1/2}\cdot\frac{\sum_{i=1}^{t}m(|\alpha_i^+-\tilde{\alpha}_i^+|+|\alpha_i^--\tilde{\alpha}_i^-|)\cdot 2^{a+b}}{\left(\sum_{i=1}^{t}m\tilde{\alpha}_i^+\cdot(1-d_{\max}^+\delta)^{a+b}\right)^{1/2}} &(58)\\ &\le\frac{t}{m^{1/2}}\sum_{a+b>0}\left(\frac{e^{-k/m}}{a!b!}\left(\frac{k}{m}\cdot\frac{4}{1-\delta}\right)^{a+b}\right)^{1/2} &(59)\\ &\le\frac{t}{m^{1/2}}\sum_{a+b>0}\left(\frac{1}{m^{1/t}}\right)^{(a+b)/2}\ll 1\;, &(60)\end{aligned}$$

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

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 emulate 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 [8]) *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*

$$\mathrm{V}[Z] \leq \frac{1}{2}\sum_{i=1}^{s} \mathrm{E}\left[\left(Z - Z'_i\right)^2\right] . \tag{61}$$

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 number of t-collisions. Then,

$$\begin{aligned}
\mathrm{E}[Z] &= \sum_{i=1}^{m} \mathrm{E}[I_i] && (62)\\
&= \sum_{i=1}^{m} \Pr[Bin(s,1/m) = t] && (63)\\
&= m \cdot \Pr[Bin(s,1/m) = t] && (64)\\
&= m \cdot \binom{s}{t}\left(\frac{1}{m}\right)^t \left(1-\frac{1}{m}\right)^{s-t} && (65)
\end{aligned}$$

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

$$\Pr[Bin(s,1/m) = t] \geq \frac{1}{4} \cdot \left(\frac{s}{t \cdot m}\right)^t . \tag{66}$$

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

$$\mathrm{E}[Z] \geq m \cdot \frac{1}{4} \cdot \left(\frac{64}{m^{1/t}}\right)^t = \Omega(1) \tag{67}$$

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

$$\begin{aligned} \mathrm{V}[Z] &\leq \sum_{i=1}^{s} \mathrm{E}\left[\left(Z - Z_i'\right)^2\right] && (68) \\ &= \sum_{i=1}^{s} \sum_{a,b\in[m]} \frac{\mathrm{E}\left[\left(Z - Z_i'\right)^2 | X_i = a, X_i' = b\right]}{m^2} && (69) \end{aligned}$$

Conditioned on $X_i = a$ and $X_i' = b$ we have that $\left(Z - Z_i'\right)^2 = 1$ 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, $\left(Z - Z_i'\right)^2 = 0$. So we have that for $a \neq b$,

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

and for $a = b$ clearly

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

Therefore, form Equation (69) we get that $\mathrm{V}[Z] \leq 2s \cdot \mathrm{Pr}\left[Bin(s-1, 1/m) = t-1\right]$. By Chebyshev's inequality,

$$\begin{aligned} \mathrm{Pr}\left[|Z - \mathrm{E}[Z]| \geq \mathrm{E}[Z]/2\right] &\leq \frac{8s \cdot \mathrm{Pr}\left[Bin(s-1, 1/m) = t-1\right]}{(m \cdot \mathrm{Pr}\left[Bin(s, 1/m) = t\right])^2} && (70) \\ &= \frac{8t}{m \cdot \mathrm{Pr}\left[Bin(s, 1/m) = t\right]} && (71) \\ &\leq \frac{32t}{64^t} \ll 1 && (72) \end{aligned}$$

Thus by Equation (66) we have that $\sigma(Z) = o(\mathrm{E}[Z])$. Therefore the lemma follows from ■

Corollary 5 *The sampling 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 1 which works in the query model performs $r = \log(4/\epsilon)$ iterations and in each iteration, q, 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 emulate Algorithm 1. Thus the sample complexity we get is bounded by

$$\sum_{q=1}^{r} \ell(q)\log(\ell(q)^2) \cdot s(q)m^{1-1/s(q)} = m^{1-\tilde{\Omega}(\epsilon^2)} \cdot \tilde{O}(1/\epsilon^2) \ , \quad (73)$$

as desired. ■

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 2: 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 2 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,

$$\mathrm{E}\left[\frac{t_1}{t}\right] = \frac{1}{m} \cdot \sum_{i=1}^{m} (D_i(1))^{\ell} \leq (p_m)^{\ell} \,, \tag{74}$$

and similarly

$$\mathrm{E}\left[\frac{t_0}{t}\right] = \frac{1}{m} \cdot \sum_{i=1}^{m} (D_i(0))^{\ell} \leq (q_m)^{\ell} \,. \tag{75}$$

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

$$\left|\frac{t_b}{t} - \mathrm{E}\left[\frac{t_b}{t}\right]\right| \leq \frac{\mathrm{E}\left[\frac{t_b}{t}\right]}{2} \,. \tag{76}$$

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 (74) and Equation (75) we get

$$\mathrm{E}\left[\frac{t_1}{t}\right] \leq (p_m)^\ell \leq (y)^\ell \quad \text{and} \quad \mathrm{E}\left[\frac{t_0}{t}\right] \leq (1-p_1)^\ell \leq (1-x)^\ell \,. \tag{77}$$

Since $y, 1 - x > \epsilon/16$ and $t = \Omega\left((16/\epsilon)^\ell\right)$, from Equation (76) we get that

$$\frac{t_1}{t} \leq (3/2)(y)^\ell \quad \text{and} \quad \frac{t_0}{t} \leq (3/2)(1-x)^\ell \tag{78}$$

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,

$$\mathrm{E}\left[\frac{t_1}{t}\right] \geq (\epsilon/4) \cdot (y')^\ell = (\epsilon/4) \cdot (y + \epsilon/16)^\ell > 3y^\ell \,. \tag{79}$$

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

$$\mathrm{E}\left[\frac{t_0}{t}\right] \geq (\epsilon/4) \cdot (1-x')^\ell = (\epsilon/4) \cdot (1 - x + \epsilon/16)^\ell > 3(1-x)^\ell \,. \tag{80}$$

Since $y' \geq \epsilon/16$ and $t = \Omega\left((16/\epsilon)^{\ell+1}\right)$, from Equation (76) and Equation (79) we get that

$$\frac{t_1}{t} > (3/2)y^\ell \,. \tag{81}$$

Similarly, since $1 - x' \geq \epsilon/16$, we get from Equation (76) and Equation (80) we get that

$$\frac{t_0}{t} > (3/2)(1-x)^\ell \,, \tag{82}$$

thus the algorithm rejects. ■

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 significatly 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 [14] For this problem there is an algorithm that uses $\tilde{O}(\sqrt{m} \cdot \mathrm{poly}(1/\epsilon))$ samples (see [14, Theorem 12], based on [25]).

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

$$\mathrm{EMD}(D, D') \stackrel{\text{def}}{=} \min_{F} \left\{ \sum_{i=0}^{n} \sum_{j=0}^{n} f_{i,j}|i-j| \right\} , \tag{83}$$

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

$$f_{i,j} \geq 0, \quad 1 \leq i \leq m, \quad 1 \leq j \leq n \tag{84}$$

$$\sum_{j=0}^{n} f_{i,j} = D(i), \quad 1 \leq i \leq m \tag{85}$$

$$\sum_{i=0}^{n} f_{i,j} = D'(j), \quad 1 \leq j \leq m \tag{86}$$

$$\sum_{i=0}^{n} \sum_{j=0}^{n} f_{i,j} = 1 \, . \tag{87}$$

We define the normalized Earth Mover's Distance to be $\mathcal{E}(D, D') \stackrel{\text{def}}{=} \frac{1}{n}\mathrm{EMD}(D, D')$, 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 (85) and (86) 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 1 can be simplified as follows. As stated in Lemma 3, the procedure GetBounds returns x and y such that $\mu(D) \leq x \leq \mu^{>}_{\beta}(D)$ and $\mu^{<}_{\beta}(D) \leq y \leq \mu(D)$. 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 procedure 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 1 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 $0 \leq z \leq n$ 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.

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*, pages 496–505, 2007.

[2] T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld. The complexity of approximating the entropy. *SIAM Journal on Computing*, 35(1):132–150, 2005.

[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*, pages 442–451, 2001.

[4] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing closeness of discrete distributions. *CoRR*, abs/1009.5397, 2010. This is a long version of [5].

[5] T. Batu, L. Fortnow, R. Rubinfeld, W.D. Smith, and P. White. Testing that distributions are close. In *Proceedings of FOCS*, pages 259–269, Los Alamitos, CA, USA, 2000. IEEE Computer Society.

[6] T. Batu, R. Kumar, and R. Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *Proceedings of STOC*, pages 381–390, 2004.

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

[8] B. Eforn and C. Stein. The jacknife estimate of variance. *Annals of Statistics*, 9:586–596, 1981.

[9] O. Goldreich, S. Goldwasser, and D. Ron. Property testing and its connection to learning and approximation. *Journal of the ACM*, 45(4):653–750, 1998.

[10] O. Goldreich and D. Ron. On testing expansion in bounded-degree graphs. *Electronic Colloqium on Computational Complexity*, 7(20), 2000.

[11] O. Goldreich and D. Ron. Property testing in bounded degree graphs. *Algorithmica*, pages 302–343, 2002.

[12] S. Guha, A. McGregor, and S. Venkatasubramanian. Sub-linear estimation of entropy and information distances. *ACM Transactions on Algorithms*, 5, 2009.

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

[14] R. Levi, D. Ron, and R. Rubinfeld. Testing properties of collections of distributions. Technical Report TR10-157, Electronic Colloquium on Computational Complexity (ECCC), 2010.

[15] R. Levi, D. Ron, and R. Rubinfeld. Testing properties of collections of distributions. In *Proceedings of ICS*, pages 179–194, 2011. See also ECCC TR10-157.

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

[17] L. Paninski. Testing for uniformity given very sparsely-sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008.

[18] S. Raskhodnikova, D. Ron, A. Shpilka, and A. Smith. Strong lower bonds for approximating distributions support size and the distinct elements problem. *SIAM Journal on Computing*, 39(3):813–842, 2009.

[19] R. Rubinfeld and M. Sudan. Robust characterization of polynomials with applications to program testing. *SIAM Journal on Computing*, 25(2):252–271, 1996.

[20] Student. The probable error of a mean. *Biometrika*, 6:1–25, 1908.

[21] 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*, pages 685–694, 2011.

[22] P. Valiant. Testing symmetric properties of distributions. In *Proceedings of STOC*, pages 383–392, 2008.

[23] P. Valiant. *Testing symmetric properties of distributions*. PhD thesis, CSAIL, MIT, 2008.

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

[25] P. White. Testing random variables for independence and identity. Unpublished manuscript.

ECCC ISSN 1433-8092

http://eccc.hpi-web.de