Settling the Complexity of Testing Grainedness of Distributions, and Application to Uniformity Testing in the Huge Object Model

Clément L. Canonne
School of Computer Science, University of Sydney, Australia

Sayantan Sen
Centre for Quantum Technologies, National University of Singapore, Singapore

Joy Qiping Yang
School of Computer Science, University of Sydney, Australia

Abstract

In this work, we study the problem of testing *m-grainedness* of probability distributions over an n-element universe $\mathcal{U}$, or, equivalently, of whether a probability distribution is induced by a multiset $S \subseteq \mathcal{U}$ of size $|S| = m$. Recently, Goldreich and Ron (Computational Complexity, 2023) proved that $\Omega(n^c)$ samples are necessary for testing this property, for any $c < 1$ and $m = \Theta(n)$. They also conjectured that $\Omega(\frac{m}{\log m})$ samples are necessary for testing this property when $m = \Theta(n)$. In this work, we positively settle this conjecture.

Using a known connection to the Distribution over Huge objects (DoHo) model introduced by Goldreich and Ron (TheoretiCS, 2023), we leverage our results to provide improved bounds for uniformity testing in the DoHo model.

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

Keywords and phrases Distribution testing, Uniformity testing, Huge Object Model, Lower bounds

Digital Object Identifier 10.4230/LIPIcs.ITCS.2025.26

Related Version *Full Version*: `https://eccc.weizmann.ac.il/report/2024/196` [12]

Funding *Clément L. Canonne*: Supported by an ARC DECRA (DE230101329).
Sayantan Sen: Supported by the National Research Foundation, Singapore and A*STAR under its Quantum Engineering Programme NRF2021-QEP2-02-P05.
Joy Qiping Yang: Supported by a JD Technology Research Scholarship in Artificial intelligence.

Acknowledgements We would like to thank the anonymous reviewers of ITCS 2025 for their suggestions which improved the presentation of the paper. SS would like to thank Clément Canonne and the Theory CS group at the University of Sydney for the warm hospitality during his academic visit, where this work was initiated.

1 Introduction

The field of distribution testing [7, 6] is concerned with providing statistically accurate information about large datasets or their underlying probability distributions, given very scarce data (sample size). Drawing insights from property testing [20, 27], distribution testing lies at the intersection of theoretical computer science, statistics, and learning theory; and has received significant attention over the past two decades, with many algorithms, insights, and new theoretical access models to the data being proposed and analyzed. We refer the reader to recent surveys [8, 4, 9] and textbook [19, Chapter 11] for more on distribution testing and property testing.

In the most standard access model, the algorithm accesses the underlying unknown probability distribution D (usually assumed to be over a known discrete domain of size n) by obtaining independent, identically distributed (i.i.d.) samples from it. The goal of the algorithm is then, given a fixed property $\Pi \subseteq \Delta_n$ (a subset of the n-dimensional probability simplex Δ_n) and a distance parameter $\varepsilon \in (0, 1]$, as follows:

- If $D \in \Pi$, the algorithm must output accept with probability at least $2/3$;
- If $\mathsf{d_{TV}}(D, Q) > \varepsilon$ for all $Q \in \Pi$, the algorithm must output reject with probability at least $2/3$;

where $\mathsf{d_{TV}}$ denotes the *total variation distance* between distributions, and the probability is taken over the choice of the samples and the internal randomness of the algorithm. This is defined as the *ε-testing* of the property Π. The minimum number of samples $s = s(n, \varepsilon)$ necessary to achieve this task in the worst case (over all possible distributions D) is the *sample complexity* of testing Π. That is, the testing task requires the algorithm, given as few samples as possible, to distinguish with high probability between distributions which have the property, and those which are "far" from having it.

Many other access models have been introduced, providing additional types of queries to the algorithm, or changing the distance metric, or both (see [8] for an overview): among them is the Distribution over Huge Objects (DoHO) model, recently introduced by Goldreich and Ron [23] to capture settings where full access to the data sampled is itself costly or impractical, due to the size of these objects. In the simplest version of this model, the distribution D is defined over the n-dimensional hypercube $\{0, 1\}^n$, where n is assumed to be very large; given a sample $x \sim D$, the algorithm must then choose which bits of x to observe, paying a unit cost for every such query made. The distance metric to quantify "farness" between distributions also differs from that of the standard formulation, and instead is taken to be the Earthmover distance (EMD) between probability distributions, with underlying metric chosen to be the (relative) Hamming distance between n-bit strings. (For the formal definition of this model, and the relation to the standard sampling model, see Section 1.3.)

Many different properties of distributions have been studied, some of them quite extensively: among those, *uniformity* ($\Pi = \{U_\Omega\}$, consisting of the single uniform distribution over the whole domain Ω), *generalized uniformity* [5] ($\Pi_U = \{U_S : S \subseteq \Omega\}$, consisting of all distributions uniform over *some* subset of the domain), and *parameterized uniformity* ($\Pi_m = \{U_S : S \subseteq \Omega, |S| = m\}$, consisting of all distributions uniform over *some* size-m subset of the domain) [23], and m-grainedness (denoted Π°_m, which we will define shortly) [18, 22] are the most relevant to this work.

In this paper, we will focus on two distribution testing tasks, intimately related:

- Testing m-grainedness of distributions in the standard sampling model (property Π°_m); and
- Testing parameterized uniformity in the DoHO model.

Grainedness of distributions

A probability distribution $D \in \Delta_n$ over a discrete set Ω of n elements is said to be *m-grained*, for a given parameter m, if all its probabilities are integer multiples of $1/m$. That is,

$$mD(x) \in \mathbb{N}^{\geq 0}, \qquad x \in \Omega\,.$$

Such distributions naturally arise due to quantization (e.g., binning of continuous or discrete distributions), or when sampling from datasets: that is, if $S \subseteq \Omega$ is a multiset of size m, uniformly sampling from S with replacement yields an m-grained distribution D_S over Ω.

Throughout this work, we will assume $\varepsilon \in (0,1)$ is a constant. Thus, $m < n/\varepsilon = O(n)$. Previous works fixed $m = \Theta(n)$ but these results can also be extended for $m = O(n)$, and therefore our results are stated in terms of m, instead of n.

Recent work of Goldreich and Ron [22] showed that $\Omega(m^c)$ samples are necessary for testing this property, for any fixed $c < 1$. A sample complexity upper bound of $O(\frac{m}{\log m})$ also follows from previous work of Valiant and Valiant [29], which led [22] to conjecture a lower bound of $\Omega(\frac{m}{\log m})$ samples. Our main contribution is to resolve this conjecture, showing that, indeed, $\Theta(\frac{m}{\log m})$ samples are necessary and sufficient for m-grainedness testing:

▶ **Theorem 1.1.** *Let n be a sufficiently large integer, $m = O(n)$, and let $\varepsilon \in (0,1)$ be a sufficiently small constant. Then, ε-testing m-grainedness of distributions over $[n]$ requires $\Omega(\frac{m}{\log m})$ samples.*

We note that the restriction $m = O(n)$ is necessary, as if $m = n/\alpha$ for some $\alpha < 1$, *every* distribution is α-close to being m-grained. Recall that we assume ε is a small constant. Moreover, the problem of testing m-grainedness becomes trivial when $m \geq n/\varepsilon$ as in that case every distribution is ε-close to being m-grained. Along the way, we also present as a warmup a new proof of the lower bound of $\Omega(n^c)$ samples for m-grainedness testing:

▶ **Theorem 1.2.** *Let n be a sufficiently large integer, $m = O(n)$, and let $\varepsilon \in (0,1)$ be a sufficiently small constant. Then, for any fixed constant $c \in (0,1)$, ε-testing m-grainedness of distributions over $[n]$ requires $\Omega(m^c)$ samples.*

While this is a strictly weaker result than our main lower bound, our proof strategy differs significantly from that of [22], and may be of independent interest.

Parameterized uniformity testing in the DoHO model

In a separate work [23], Goldreich and Ron studied the problem of parameterized uniformity testing in the Huge Object model. Among other results, they established a connection between m-grainedness testing in the *standard* model and testing Π_m in the DoHO model:

▶ **Theorem 1.3** ([23, Theorem 2.13])**.** *Assume that, for constant $\varepsilon \in (0,1)$, ε-testing m-grainedness in the standard model has sample complexity $\Omega(\frac{m}{\log m})$. Then, for every $1 \leq m \leq n$ and constant $\varepsilon' \in (0, \frac{\varepsilon}{2})$, ε'-testing Π_m for distribution over $\{0,1\}^n$ in the DoHO model has* query *complexity $\Omega(\frac{m}{\log m})$.*

Our lower bound for m-grainedness immediately implies that this result holds unconditionally. In the same paper, the authors provided an algorithm for testing Π_m using $\widetilde{O}(m)$ samples and queries (for constant $\varepsilon = \Omega(1)$), which together with the above – now unconditional – lower bound settles the complexity of testing Π_m in the DoHO model, up to logarithmic factors, *for $m \leq n$*. One may be tempted to assume the same query complexity lower bound holds in all parameter regimes: perhaps surprisingly, our next result is a new and simple algorithm for testing Π_m in the DoHO model which takes $O(m^{2/3})$ samples and performs $O(m^{2/3}n)$ queries:

▶ **Theorem 1.4.** *There is an algorithm which, given an integer $m \in \mathbb{N}$ and constant $\varepsilon \in (0,1)$, ε-tests the property Π_m of distributions over $\{0,1\}^n$ in the DoHO model, taking $\mathcal{O}(m^{2/3})$ samples and performing $\mathcal{O}(m^{2/3} \cdot n)$ queries. Moreover, any ε-tester for Π_m in the DoHO model must take $\Omega(m^{2/3})$ samples.*

Notably, this improves the upper bound of Goldreich and Ron whenever $n^3 \leq m \leq 2^n$, and rules out any $\tilde{\Omega}(m)$ query complexity lower bound for the whole range of m.

1.1 Overview of our techniques

Lower bound for m-grainedness testing

Our lower bound approach follows Le Cam's two-point method: we will design two distributions of distributions D_{yes} and D_{no} such that the distributions in D_{yes} are m-grained and the distributions in D_{no} are "far" from being m-grained in variation distance. The name of the game is then to prove that it is information-theoretically impossible to distinguish between D_{yes} and D_{no}.

To achieve this, we will rely on the *moment-matching* technique, particularly suited to properties like m-grained (which are symmetric, i.e., invariant to relabelling of the domain elements). Broadly, if two probability distributions D_{yes} and D_{no} have sufficiently similar moments $\|D_{\mathsf{yes}}\|_t^t \approx \|D_{\mathsf{no}}\|_t^t$, for say $1 \leq t \leq K$, then it is hard to distinguish (a permutation of) D_{yes} from (a permutation of) D_{no} by using $o(n^{1-1/K}/K)$ samples. That is, the more moments one can match, the more samples one needs to tell two distributions apart. Note that setting $K = \Theta(\log n)$ would then lead to an $\Omega(\frac{n}{\log n})$ lower bound (and that, as remarked later, in our case we can assume without loss of generality, for the sake of the lower bound argument, that $m = \Theta(n)$.) This approach has been used in the literature extensively [26, 29, 28, 30, 31, 32, 10], and has proven to be very successful in obtaining tight lower bounds for a range of symmetric properties. We will rely on the formulation of the technique described by [32], which maps the problem of matching moments of two full probability distributions (an n-dimensional object) to that of matching moments of two single-dimensional random variables U and V: these two random variables will then be used to generate the probability distributions: intuitively (and as described below), "n independent draws from U will give $D_{\mathsf{yes}}(1), \ldots, D_{\mathsf{yes}}(n)$" (and similarly for D_{no} and V).

Thus, a crucial ingredient in our lower bound is the construction of a pair of discrete random variables U and V whose first logarithmically many moments are identical. Using n independent copies of U (resp. V), we will then define a random measure over $[n]$, which corresponds to a "yes"-instance D_{yes} (resp., a "no"-instance D_{no}). Thus, another requirement on our construction of U and V is that the first one should yield an m-grained distribution, while the second should correspond to a distribution far from being m-grained. We will formalize this in Lemma 2.2.

To build these two random variables U and V, we will, as previously done in the literature, rely on the properties of the Chebyshev polynomial T_d of degree $d = O(\log n)$, defining U and V to be supported on disjoint subsets of the roots of a suitable polynomial $P(x)$, where the probability mass assigned to a given root r is proportional to $1/|P'(r)|$. The idea is that some of the roots of P, those corresponding to U, will be multiples of $1/m$ (leading to m-grained probabilities) while the others, corresponding to V, will be odd multiples of $1/(2m)$ (leading to "far-from-m-grained" probabilities): for instance, we would take

$$P(x) = x\left(x - \frac{1}{2m}\right)\left(1 - \frac{1}{m}\right)T_d(c \cdot x)$$

for some scaling constant $c > 0$. The remaining roots will be that of the Chebyshev polynomial T_d, which are there to ensure that we can match sufficiently many moments of U and V.

However, this approach, which underlies most of previous work, cannot be directly used here: indeed, while the resulting two random variables U, V *could* be made to have many matching moments, and as such be hard to distinguish, doing so will put a lot of probability mass on the roots of the Chebyshev polynomial T_d both for U and V; and these roots are not multiples of $1/m$. This would have the undesirable effect of making *both* our distributions far from being m-grained. Trying to fix this the obvious requires to ensure very little mass is

put by U (and so V) on the roots of T_d, which in turns limits the number of moments that can be matched. (Note that fixing this by choosing not to use the Chebyshev polynomial T_d at all but instead choosing $P(x)$ of the form $P(x) = x\left(x - \frac{1}{2m}\right) \cdot \left(x - \frac{1}{m}\right) \cdots \left(x - \frac{L}{m}\right)$, for some large constant integer L, does get part of the way there and provides a non-trivial result, leading to our (weaker) $\Omega(m^c)$ lower bound for every $c < 1$).

To remedy this, we take a different route, using an idea we believe to be of independent interest and applicable to other lower bounds: instead of using Chebyshev polynomials directly, we will be using a *modified* version of Chebyshev polynomials, $\tilde{T}_d$, defined from T_d by first rounding its roots to multiples of $1/m$. By carefully analyzing the resulting polynomial, we can show that it behaves, for our lower bound construction purposes, similarly to the Chebyshev polynomial, and so we can use it to define both U and V. This allows us to match more moments, leading to our final $\Omega(m/\log m)$ lower bound.

Upper bound for parameterized uniformity testing

In contrast, the idea underlying our upper bound for testing Π_m in the DoHO model is relatively straightforward: namely, by making n queries per sample, we can simulate any testing algorithm in the *standard* sampling model, as we now have observed the full samples. This, along with the relation between TV distance and EMD distance, allows us to lift any s-sample tester for Π_m in the standard sampling model to an s-sample, $s \cdot n$-query tester for Π_m in the DoHO model.

The only issue with this plan is that *there is no known (nontrivial) tester for* Π_m *in the standard sampling model.* There is, however, a known testing algorithm for the related property of generalized uniformity, Π_U. Our key contribution here is then to use this known tester A to derive a tester B for Π_m in the standard sampling model. Notably, this is not as immediate as it seems, and our algorithm needs to use A in a whitebox way, and combine this with an additional test to estimate the ℓ_2 norm of the unknown distribution D. The subtlety here (and the reason for which we cannot use *any* A in a blackbox fashion, but need to use a specific algorithm due to [5]) is that being close to *some* uniform distribution over *some* subset does not immediately allow to conclude about being close to *some* uniform distribution over some m*-size* subset – even if we are guaranteed the ℓ_2 norm of D is close to $1/m$.

1.2 Related work

The field of distribution testing has its roots in theoretical computer science with the work of Goldreich and Ron [21], who designed a uniformity testing algorithm as a tool to check whether a graph is an expander; and formally defined and introduced in [7]. [6] studied the problem of identity testing, which generalizes the problem of uniformity testing. Over the last two decades, this field has seen significant growth, and a host of new tools and techniques have emerged, see [25, 1, 18, 16, 14, 15, 11] to name a few. See the surveys [8, 4, 9] and the book of [19, Chapter 11] for more details.

The study of grained distributions was done implicitly in the work of [26], and later [18] studied it in more detail.

As mentioned earlier, the Huge Object model was introduced by [23]. Since then, there have been several works focusing on this model. [13] defined the notion of *index-invariant properties*, a set of properties that are invariant under the permutation of the indices (these properties are different than label-invariant properties). They showed that index-invariant properties whose VC dimension of the support set is bounded can be learned using a constant

number of queries, depending only on the VC dimension. They also gave tight separation between adaptive and non-adaptive testers for index and non-index-invariant properties. Later, [2] studied various different notions of adaptivity, and showed several separation results. Very recently, [3] studied the problem of support size testing in this model.

1.3 Preliminaries

For a positive integer n, let $[n]$ denote the set $\{1, \ldots, n\}$. We will use the standard asymptotic notation $\mathcal{O}(\cdot), \Omega(\cdot), \Theta(\cdot)$, and, in some cases, the (somewhat less) standard notation $\widetilde{O}(\cdot)$, which omits poly-logarithmic dependencies in the parameters for readability.

We will use several notions of distance between two distributions.

▶ **Definition 1.5.** *Let D_1 and D_2 be two probability distributions over a domain Ω. The ℓ_1* distance *between D_1 and D_2 is defined as*

$$\|D_1 - D_2\|_1 = \sum_{x \in \Omega} |D_1(x) - D_2(x)|.$$

The total variation distance *between D_1 and D_2 is defined as:*

$$\mathsf{d}_{\mathsf{TV}}(D_1, D_2) = \frac{1}{2} \cdot \|D_1 - D_2\|_1 = \sup_{S \subseteq \Omega} (D_1(S) - D_2(S)).$$

▶ **Definition 1.6** (EMD with respect to Hamming distance)**.** *Let D_1, D_2 be two distributions defined over the n-dimensional Hamming cube $\{0,1\}^n$ and d_H be the (relative) Hamming distance over $\{0,1\}^n$. Then the* Earth Mover distance *(EMD) between D_1 and D_2 is defined as follows:*

$$\mathsf{d}_{\mathsf{EM}}(D_1, D_2) := \inf_{T \in \mathcal{T}(D_1, D_2)} \mathbb{E}_{(x,y) \sim T} \left[d_H(x, y) \right]$$

where $\mathcal{T}(D_1, D_2)$ denotes the set of all couplings between μ and τ.

Since we are working with the (relative) Hamming distance, $d_H(x, y) \leq 1$. Then directly from the definitions of $\mathsf{d}_{\mathsf{EM}}(D_1, D_2)$ and $d_H(x, y)$, we get the following simple yet useful observation connecting total variation and EMD distances between two distributions.

▶ **Observation 1.7** (Relation between EMD and TV distances)**.** *Let D_1 and D_2 be two distributions over the n-dimensional Hamming cube $\{0,1\}^n$. Then,*

$$\mathsf{d}_{\mathsf{EM}}(D_1, D_2) \leq \mathsf{d}_{\mathsf{TV}}(D_1, D_2).$$

▶ **Definition 1.8** (Huge Object Model)**.** *Consider a discrete distribution D supported over the n-dimensional Hamming cube $\{0,1\}^n$. D is accessed by obtaining iid samples, where each sample obtained is an n-bit Boolean string. In addition to the sampling oracle, there is also a query oracle associated with D, where the tester can query any index $i \in [n]$ for any samples (say j-th sample s_j) obtained, and the query oracle will return the i-th bit of s_j. The goal is then to minimize both the total sample and query complexities required to decide a property.*

Note that this Huge Object model is particularly suited for studying high-dimensional distributions, where the dimension of the underlying domain is so large that even reading a single sample in its entirety might not be feasible.

Finally, we state here some technical result which will be used in our lower bounds proofs:

▶ **Lemma 1.9** ([31, Lemma 4])**.** *Let W_1, W_2 be discrete random variables taking values in $[0, \Lambda]$. If $\mathbb{E}[W_1^t] = \mathbb{E}[W_2^t]$ for $1 \leq t \leq T$, then*

$$\mathsf{d}_{\mathsf{TV}}\Big(\mathbb{E}_{W_1}[\mathsf{Poi}(W_1)], \mathbb{E}_{W_2}[\mathsf{Poi}(W_2)]\Big) \leq \left(\frac{e\Lambda}{2T}\right)^T.$$

▶ **Fact 1.10** ([10, Fact 7], [24]). *Let p be a degree-d polynomial with distinct roots $r_1, \ldots, r_d$. Then, for every $0 \le k \le d-2$, we have that $\sum_{i=1}^{d} \frac{r_i^k}{p'(r_i)} = 0$.*

We will be using the following two well-known trigonometric inequalities.

▶ **Fact 1.11.**

(i) *For $x \in [0, \pi]$, the following holds:*

$$\sin x \ge \frac{2}{\pi} \min(x, \pi - x) \tag{1}$$

(ii) *For $x \in [-\pi/2, \pi/2]$, the following holds:*

$$|\frac{4}{\pi} \cdot x| \ge |\sin x| \ge |\frac{2}{\pi} \cdot x|, \tag{2}$$

Organization

The rest of the paper is organized as follows. In Section 2, we present a new proof of the lower bound of $\Omega(m^c)$ samples. Then in Section 3, we present the $\Omega(m/\log m)$ lower bound for testing m-grainedness. In Section 4, we present our result of uniformity testing in the Huge Object Model. We conclude in Section 5 with some open questions. Due to the shortage of space, the formal proofs of some claims are deferred to the full version of the paper [12].

2 An $\Omega(m^c)$ lower bound for testing m-grainedness

In this section, we prove Theorem 1.2, thereby presenting a new proof of the lower bound of [22].

▶ **Theorem 1.2.** *Let n be a sufficiently large integer, $m = O(n)$, and let $\varepsilon \in (0,1)$ be a sufficiently small constant. Then, for any fixed constant $c \in (0,1)$, ε-testing m-grainedness of distributions over $[n]$ requires $\Omega(m^c)$ samples.*

Note that it suffices to consider the case $m = \Theta(n)$, as if $m \ll n$ one can then embed the hard instances in a larger domain, lifting the lower bound. As a result, we hereafter assume without loss of generality that $m = \Theta(n)$. We start by stating a relatively standard theorem which will be crucial in the proof of our sample complexity lower bound, and whose proof is deferred to the full version [12].

▶ **Theorem 2.1.** *Fix positive integers m, n, s, where $n > 4.3 \times 10^8$ and $m \le C_0(n-1)$ for some absolute constant $C_0 \in \mathbb{R}$. Suppose there exist random variables U and V, where U is supported on $\{0, \frac{1}{m}, \ldots, \frac{m}{m}\}$; $\Pr\left[V \in \{\frac{1}{2m}, \frac{3}{2m}, \ldots, \frac{2m-1}{2m}\}\right] \ge \frac{2}{3}$; and they satisfy the following three conditions,*

$$\max(U, V) \le \frac{20 \log^2(n-1)}{(n-1)}, \tag{3}$$

$$\mathbb{E}[U] = \mathbb{E}[V] \le \frac{1}{2(n-1)}. \tag{4}$$

$$\mathsf{d}_{\mathsf{TV}}\left(\mathbb{E}[\mathrm{Poi}(sU)], \mathbb{E}[\mathrm{Poi}(sV)]\right) \le \frac{1}{20(n-1)}, \tag{5}$$

Then any tester taking less than $s/2$ number of samples from unknown distribution p cannot distinguish between the cases that p is m-grained and p is at least $\frac{1}{8C_0}$-far from any m-grained distributions in TV distance with probability $3/5$.

In order to prove Theorem 1.2, we will be using the following lemma.

▶ **Lemma 2.2.** *There exist constants L, K, such that for any n and $m = L(n-1)$, there exist two discrete random variables U and V such that the following Equations* (6), (7), (8) *and* (9) *hold.*

$$\mathsf{Support}(U) \subseteq \frac{\mathbb{Z}}{m} \text{ and } \mathsf{Support}(V) \subseteq \frac{2\mathbb{Z}+1}{2m} \tag{6}$$

$$U, V \leq \frac{L}{m} \tag{7}$$

$$\mathbb{E}[U] = \mathbb{E}[V] \leq \frac{1}{2(n-1)}. \tag{8}$$

$$\mathbb{E}[U^t] = \mathbb{E}[V^t] \quad \text{for} \quad t \in [K]. \tag{9}$$

In order to prove the above lemma, we will use the following polynomial to construct our U and V (fix any positive integer L):

$$P(x) = x\left(x - \frac{1}{2m}\right)\left(x - \frac{2}{2m}\right)\left(x - \frac{3}{2m}\right)\cdots\left(x - \frac{2L}{2m}\right) \tag{10}$$

In the context of Lemma 2.2, the random variable U will be supported on the roots $0, \frac{1}{m}, \frac{2}{m}, \ldots, \frac{L}{m}$. Similarly, the random variable V will be supported on the roots $\frac{1}{2m}, \frac{3}{2m}, \ldots, \ldots, \frac{2L-1}{2m}$.

Proof of Lemma 2.2. We will first describe the construction of random variables U and V, associated with the polynomial P. Namely, let $\Pr[U = r] \propto \frac{1}{|P'(r)|}$ for r in the support of U (every derivative is positive when r is in the support of U, i.e., $r \in \{0, \frac{1}{m}, \frac{2}{m}, \ldots, \frac{L}{m}\}$). Similarly, we have $\Pr[V = r] \propto \frac{1}{|P'(r)|}$ for r in the support of V (likewise, all derivative will be negative). The normalization term is simply

$$Z_P := \sum_{i=0}^{L} \frac{1}{|P'(\frac{i}{m})|} = \sum_{i=1}^{L} \frac{1}{|P'(\frac{2i-1}{2m})|}.$$

And we can show that

$$Z_P = \sum_{k=0}^{L} \frac{\binom{2L}{2k}(2m)^{2L}}{(2L)!} = \frac{(2m)^{2L}}{(2L)!}\sum_{k=0}^{L}\binom{2L}{2k} = \frac{(2m)^{2L}}{(2L)!}\cdot 2^{2L-1}.$$

Applying Fact 1.10 with $k = 0$, we can notice that the two summations are equal (U has all positive roots and V all negative).

Proof of Equation (7). The largest value of U and V is the largest root of the polynomial P: $\max\{U, V\} \leq L/m$.

Proof of Equation (8). Let us compute the derivative of the polynomial P in Equation (10). For any $k \in [2L] \cup \{0\}$, we have the following:

$$\begin{aligned} P'\left(\frac{k}{2m}\right) &= \prod_{\ell=0}^{k-1}\frac{k-\ell}{2m}\prod_{\ell=k+1}^{2L}\frac{\ell-k}{2m}(-1)^{2\ell-k} \\ &= \frac{(-1)^k}{(2m)^{2L}}k!(2L-k)! \\ &= (-1)^k\frac{(2L)!}{\binom{2L}{k}(2m)^{2L}} \end{aligned}$$

Now let us compute $\mathbb{E}[U]$ as follows:

$$\begin{aligned}\mathbb{E}[U] &= \Pr[U=0]\cdot 0 + \sum_{\ell:\ \ell \text{ even},\ \ell\in[2m]} \Pr\left[U=\frac{\ell}{m}\right]\cdot\frac{\ell}{m}\\ &= \frac{1}{Z_P}\sum_{k=0}^{L}\frac{2k}{2m}\frac{\binom{2L}{2k}(2m)^{2L}}{(2L)!}\\ &= \frac{1}{m}\frac{\sum_{k=0}^{L}k\binom{2L}{2k}}{2^{2L-1}}\\ &= \frac{L}{2m}\end{aligned}$$

We finally upper bound the expectation:

$$\frac{1}{2(n-1)} \geq \frac{L}{2m} = \mathbb{E}[U] \Rightarrow m \geq L(n-1). \tag{11}$$

Proof of Equation (9). Their first $K = 2L-1$ moments are matched by recalling Fact 1.10 (applied for $k = t$) and how U and V are constructed. ◀

Proof of Theorem 1.2. Using Lemma 1.9 on random variables constructed in Lemma 2.2, by setting $m = 2L(n-1) = \Theta(n)$, we know $\Lambda = s\cdot\frac{L}{m}$, we get that

$$\mathsf{d}_{\mathsf{TV}}(\mathbb{E}[\mathrm{Poi}(sU),\mathrm{Poi}(sV)]) \leq \left(\frac{es\cdot\frac{L}{m}}{2K}\right)^{K} \leq \frac{1}{2(n-1)},$$

which gives,

$$s \leq 2(n-1)\left(\frac{2K}{e}\right)\cdot\left(\frac{1}{2(n-1)}\right)^{1/K}.$$

Further simplification yields,

$$s \leq \left(\frac{4L-2}{e}\right)(2(n-1))^{1-\frac{1}{2L-1}}.$$

Applying Theorem 2.1, gives the lower bound $s = \Omega\left(n^{1-\frac{1}{2L-1}}\right)$. For any constant $c < 1$, one can choose L large enough such that $c \leq 1-\frac{1}{2L-1}$, and thus concludes our proof. ◀

3 $\Omega(\frac{m}{\log m})$ lower bound for m-grainedness (Proof of Theorem 1.1)

In this section, we will be proving the following theorem:

▶ **Theorem 1.1.** *Let n be a sufficiently large integer, $m = O(n)$, and let $\varepsilon \in (0,1)$ be a sufficiently small constant. Then, ε-testing m-grainedness of distributions over $[n]$ requires $\Omega(\frac{m}{\log m})$ samples.*

Similar to the proof of Theorem 1.2, we will make use of Theorem 2.1 to prove our $\Omega(\frac{m}{\log m})$ lower bound. Namely, we will construct two random variables U and V such that they can be used to construct approximate probability vector for m-grained and mostly $2m$-grained and their first $\log(m)$ moments match. In particular, we will prove the following lemma:

▶ **Lemma 3.1.** *For $m = 70(n-1)$ and $n > 20$, there exist discrete random variables U and V such that the following hold:*

$$\Pr[V = \frac{1}{2m}] \geq 2/3. \tag{12}$$

$$\mathsf{Support}(U) \subseteq \frac{\mathbb{Z}}{m} \tag{13}$$

$$0 \leq U, V \leq \frac{20\log^2(n-1)}{n-1}. \tag{14}$$

$$\mathbb{E}[U] = \mathbb{E}[V] \leq \frac{1}{20(n-1)}. \tag{15}$$

$$\mathbb{E}[U^t] = \mathbb{E}[V^t] \text{ for } \ t \in [3 \cdot \log(n-1)] \tag{16}$$

3.1 Construction of U and V

The main focus of this subsection is to establish Lemma 3.1.

Fact 1.10 allows us to define two random variables U and V with matching moments up to the degree of some polynomial – similar to the analysis in Section 2, by simply defining U and V through its roots and partitioning them by the signs of the derivatives at those roots. However, the challenge here is, we need to match even higher degree (in Section 2, we can match them up to a constant degree C_1, here we need to match them up to $\Theta(\log m) = \Theta(\log n)$ degree), all while keeping every other conditions unchanged: U is m-grained and V on the most part is $(2m)$-grained; maximum value of the support size does not blow up; and more importantly, the expectation could be bounded by $O(1/n)$. Notably, if we try to match higher degree: say instead of constant, we make $L = O(\log n)$ in Equation (11), with $m = \Theta(n)$, using the construction in Section 2. It will fail miserably – the expectation $\mathbb{E}[U]$ will exceed $O(1/n)$, and so when we take n copies of U or V, it would not be a distribution in an approximate sense.

To work around this and match degree (by implication, moments) as high as possible, we take the (shifted) Chebyshev polynomial of degree d and we will later use properties of this polynomial to construct our prior variables U and V. Recall that the shifted Chebyshev polynomial of the first kind is defined as follows:

$$p_{T_d}(x) = T_d(1 - \Delta x) = 2^{d-1}\Delta^d \cdot \prod_{i=1}^{d}(x - t_i), \tag{17}$$

where $T_d(\theta) = \cos(d \cdot \arccos\theta)$, for $\theta \in [-1, 1]$; t_i denotes the roots of the degree-d (shifted) Chebyshev polynomial for every $i \in [d]$; $\Delta \in \mathbb{R}$, a parameter to be chosen in the analysis later.

However, as mentioned in the introduction, instead of using the shifted Chebyshev polynomial directly, we will be using a variant of it for our proof. For this purpose, we will define the polynomial $\tilde{p}_{T_d}$, a slightly modified Chebyshev polynomial of degree d, by "rounding up to the nearest multiple of $1/m$" its d roots:

$$\tilde{p}_{T_d}(x) = \tilde{T}_d(1 - \Delta \cdot x) = 2^{d-1}\Delta^d \cdot \prod_{i=1}^{d}(x - \tilde{t}_i) \tag{18}$$

where $\frac{1}{m} + t_i \geq \tilde{t}_i = \frac{1}{m}\lceil mt_i \rceil \geq t_i$.

Putting these ideas together, we will focus on two polynomials based on the Chebyshev polynomial and modified Chebyshev polynomial respectively (obtained by appending Chebyshev polynomial to $x(x-\frac{1}{2m})(x-\frac{1}{m})$):

$$p(x) = x\left(x-\frac{1}{2m}\right)\left(x-\frac{1}{m}\right)\cdot p_{T_d}(x) \tag{19}$$

$$\tilde{p}(x) = x\left(x-\frac{1}{2m}\right)\left(x-\frac{1}{m}\right)\cdot \tilde{p}_{T_d}(x) \tag{20}$$

Through polynomial $\tilde{p}$ in Equation (20), we can construct two random variables U and V identified by their probability mass function. Namely, we define U as follows:

$$\Pr[U=r] \propto \frac{1}{|\tilde{p}'(r)|} \quad \text{if } \tilde{p}'(r)>0 \text{ and } r \text{ is a root of } \tilde{p} \tag{21}$$

Similarly, we define the random variable V as follows:

$$\Pr[V=r] \propto \frac{1}{|\tilde{p}'(r)|} \quad \text{if } \tilde{p}'(r)<0 \text{ and } r \text{ is a root of } \tilde{p} \tag{22}$$

For both U and V, we assign probability 0 to the roots not specified. Thus, by setting $k=0$ in Fact 1.10 and noting that the negative and positive terms sum to 0, the normalization factor are equal for both U and V, and can be expressed as follows:

▶ **Observation 3.2.**

$$Z_{\tilde{p}} = Z(\Delta,m,d) = \frac{1}{\tilde{p}'(0)} + \frac{1}{\tilde{p}'(\frac{1}{m})} + \sum_{j=1}^{d}\frac{\mathbb{1}_{\{\tilde{p}'(\tilde{t}_j)>0\}}}{\tilde{p}'(\tilde{t}_j)} = \frac{1}{|\tilde{p}'(\frac{1}{2m})|} + \sum_{j=1}^{d}\frac{\mathbb{1}_{\{\tilde{p}'(\tilde{t}_j)<0\}}}{|\tilde{p}'(\tilde{t}_j)|}. \tag{23}$$

The reasons we are taking the shifted Chebyshev polynomial and "round the roots up" are two-fold: first, this kind of construction (before rounding) was used to prove lower bounds with $\Omega(n/\log n)$ complexity before, by constructing a polynomial of degree $\Theta(\log n)$ with certain constraints [10]. Second, we need to make sure that the distributions derived from U are m-grained with very high probability. Yet, using the shifted Chebyshev directly, some of the roots of U (the positive roots) resulting from T_d will not be m-grained with non-trivial probability. Intuitively, we want to leverage the known properties of the shifted Chebyshev p_{T_d} and argue that rounding up and creating $\tilde{p}_{T_d}$ will only "mildly" change those properties.

Since we are deciding the support of U, V based on the evaluation of derivative at each root of $\tilde{p}_{T_d}$ and argue that it is somewhat close to p_{T_d}, we need to first make sure that the signs of derivative at the main three roots remain the same as p_{T_d}'s (the rest of the roots' derivative's signs, the ones induced by $\tilde{p}_{T_d}$, will not affect the argument). Note that the sign of the polynomial $\tilde{p}_{T_d}$ when evaluated at $0, 1/2m, 1/m$ is not changed (same as p_{T_d}), as $(0-t_j)$ and $(0-\tilde{t}_j)$ have the same sign; we also have $1/2m, 1/m \le t_j \le \tilde{t}_j$, via a constraint on setting Δ and therefore we can conclude that

$$p_{T_d}(0)\cdot\tilde{p}_{T_d}(0)>0; \quad p_{T_d}(1/2m)\cdot\tilde{p}_{T_d}(1/2m)>0; \quad p_{T_d}(1/m)\cdot\tilde{p}_{T_d}(1/m)>0.$$

Next, we compute their corresponding derivatives of the two polynomials evaluated at the roots t_ℓ and $\tilde{t}_\ell$.

$$p'(t_\ell) = t_\ell\left(t_\ell-\frac{1}{2m}\right)\left(t_\ell-\frac{1}{m}\right)2^{d-1}\Delta^d\prod_{j\neq\ell}(t_j-t_\ell) \tag{24}$$

Table 1 Properties of the polynomial p and its distinct $d+3$ roots. $|p'(r_1)|, |p'(r_2)|, |p'(r_3)|$'s upper and lower bounds are proven in Claim A.7 of the full version [12].

roots	value of $p'(r_\ell)$	bound on $\lvert p'(r_\ell)\rvert$
$r_1 = 0$	$\frac{1}{2m^2} \cdot T_d(1)$	$\Theta\left(\frac{1}{m^2}\right)$
$r_2 = \frac{1}{2m}$	$-\frac{1}{4m^2} \cdot T_d\left(1 - \frac{\Delta}{2m}\right)$	$\Theta\left(\frac{1}{m^2}\right)$
$r_3 = \frac{1}{m}$	$\frac{1}{2m^2} \cdot T_d\left(1 - \frac{\Delta}{m}\right)$	$\Theta\left(\frac{1}{m^2}\right)$
$r_{\ell+3} = t_\ell = \frac{2}{\Delta}\sin^2\left(\frac{\pi}{2}\left(\frac{2\ell-1}{2d}\right)\right)$	$p'(t_\ell) = t_\ell \cdot \left(t_\ell - \frac{1}{2m}\right) \cdot \left(t_\ell - \frac{1}{m}\right) \cdot T'_d(t_\ell)$	$\Theta\left(\frac{\ell^5}{\Delta^2 d^4}\right)$

$$\tilde{p}'(\tilde{t}_\ell) = \tilde{t}_\ell \left(\tilde{t}_\ell - \frac{1}{2m}\right)\left(\tilde{t}_\ell - \frac{1}{m}\right) 2^{d-1}\Delta^d \prod_{j\neq\ell}(\tilde{t}_j - \tilde{t}_\ell) \tag{25}$$

Here, note that by construction, all the roots of $\tilde{p}_{T_d}$ will be m-grained (see Equation (18)). Moreover, 0 and $\frac{1}{m}$ are m-grained roots of $\widetilde{p}(x)$ (see Equation (20)). On the other hand, the root $\frac{1}{2m}$ of the polynomial $\widetilde{p}(x)$ is not m-grained by definition.

In the context of Lemma 3.1, the random variable U will be supported on the roots 0, $\frac{1}{m}$, and a subset of the roots $\widetilde{t}_\ell$ of the modified Chebyshev polynomial $\widetilde{p}_{T_d}$ from Equation (18), for $\ell \in [d]$. On the other hand, the random variable V will be supported over $\frac{1}{2m}$ and a disjoint (from support of U constructed by modified Chebeshev roots of $\tilde{p}_{T_d}$) subset of the modified Chebyshev polynomial roots. U and V's support form a partition for all roots of $\tilde{p}_{T_d}$.

To prove Lemma 3.1, we need a few claims, that we state below and whose proof can be found in the full version [12].

▷ Claim 3.3 (Relation between roots at $\tilde{p}'$ and p').

$$\frac{|\tilde{p}'(\tilde{t}_\ell)|}{|p'(t_\ell)|} \geq \exp\left(-\frac{12\Delta}{m/d^2}\right) \text{ for every } \ell \in [d]. \tag{26}$$

$$1 \leq \frac{|\tilde{p}'(0)|}{|p'(0)|}, \quad 1 \leq \frac{|\tilde{p}'(1/m)|}{|p'(1/m)|} \leq 6, \quad 1 \leq \frac{|\widetilde{p}'(1/2m)|}{|p'(1/2m)|} \leq 6. \tag{27}$$

The proofs of Equation (26) and Equation (27) can be found in Appendix A.1 of the full version (Claims A.2 to A.5). From now onwards, we will assume that Claim 3.3 holds.

For the lower bound, we will be choosing $\Delta = \frac{m}{20d^2}$ and $d = \sqrt{10}\log(n-1)$.

▶ **Observation 3.4.** *Let t_j and $\tilde{t}_j$ be the roots of the Cheybyshev and modified Chebyshev polynomials, for any $j \in [d]$. When $\Delta \leq \frac{m}{2d^2}$, we have that*

$$\tilde{t}_j \geq t_j \geq \frac{1}{m}.$$

Proof. Using Equation (2), we can lower bound t_j for any j,

$$t_j = \frac{1}{\Delta}\left(1 - \cos\left(\frac{2j-1}{2d}\pi\right)\right) = \frac{2}{\Delta}\sin^2\left(\frac{\pi}{2}\left(\frac{2j-1}{2d}\right)\right) \geq \frac{2}{\Delta}\left(\frac{2j-1}{2d}\right)^2 \geq \frac{1}{m}, \tag{28}$$

where the last inequality follows from $2j-1 \geq 1$ (which holds for all $j \geq 1$), and our assumption on Δ. Moreover, we also have:

$$\tilde{t}_j = \frac{1}{m}\lceil mt_j \rceil \geq t_j \geq \frac{1}{m}$$

where the last inequality follows from Equation (28). This completes the proof. ◀

We need the following two claims (Claim 3.5 and Claim 3.6) whose proofs are deferred to Appendix A.2 of the full version [12], to show that $\Pr[V = \frac{1}{2m}] \geq \Omega(1)$ in Claim 3.7 as required by Equation (12).

▷ Claim 3.5. Suppose $1 \leq \frac{|\widetilde{p'}(\frac{1}{2m})|}{|p'(\frac{1}{2m})|} \leq 6$ holds and $\Delta \leq \frac{2m}{9d^2}$. Then we have,

$$\Pr\left[V = \frac{1}{2m}\right] \propto \frac{1}{|\tilde{p}'(\frac{1}{2m})|} \geq \frac{2}{3}m^2.$$

▷ Claim 3.6. Suppose that $\Delta \leq \frac{2m}{9d^2}$. For any $j \in [d]$, let $\tilde{t}_j$ denote the j-th root of the modified Chebyshev polynomial $\widetilde{p}$. Then we have:

$$\Pr\left[V \neq \frac{1}{2m}\right] \propto \sum_{j=1}^{d} \frac{\mathbb{1}_{\{\tilde{p}'(\tilde{t}_j)<0\}}}{|\tilde{p}'(\tilde{t}_j)|} \leq \exp\left(\frac{12\Delta}{m/d^2}\right) \cdot \frac{66}{\pi} \cdot \Delta^2 d^4.$$

▷ Claim 3.7. If $\Delta \leq \frac{m}{20d^2}$, then

$$\Pr\left[V = \frac{1}{2m}\right] = \frac{1}{|\tilde{p}'(\frac{1}{2m})| \cdot Z} \geq \frac{5}{6},$$

where Z is the normalizing constant defined in Equation (23).

Proof. Following the definition of the random variable V (Equation (22)) and the normalization term $Z_{\tilde{p}}$ (Equation (23)), we can say that:

$$\Pr\left[V = \frac{1}{2m}\right] = \frac{1}{|\tilde{p}'(\frac{1}{2m})| \cdot Z} = \frac{1}{1 + \left(\sum_{j=1}^{d} \frac{\mathbb{1}_{\{\tilde{p}'(\tilde{t}_j)<0\}}}{|\tilde{p}'(\tilde{t}_j)|}\right) \cdot |\tilde{p}'(\frac{1}{2m})|}.$$

Combining Claim 3.5 and Claim 3.6, we see that

$$\begin{aligned}
\left(\sum_{j=1}^{d} \frac{\mathbb{1}_{\{\tilde{p}'(\tilde{t}_j)<0\}}}{|\tilde{p}'(\tilde{t}_j)|}\right) \cdot |\tilde{p}'(\frac{1}{2m})| &\leq \frac{3\exp\left(\frac{12\Delta}{m/d^2}\right)}{2m^2} \cdot \frac{66}{\pi} \cdot \Delta^2 d^4 \\
&= \frac{99}{\pi}\exp\left(\frac{12\Delta}{m/d^2}\right) \cdot \frac{\Delta^2 d^4}{m^2} \\
&\leq \frac{99}{\pi}\exp\left(\frac{12\frac{m}{20d^2}}{m/d^2}\right) \cdot \frac{\left(\frac{m}{20d^2}\right)^2 d^4}{m^2} \quad [\because \Delta \leq \frac{m}{20d^2}] \\
&\leq \frac{99}{\pi}\exp(\frac{12}{20}) \cdot \frac{1}{400} \leq \frac{1}{5}.
\end{aligned}$$

Thus, we can say that

$$\Pr\left[V = \frac{1}{2m}\right] = \frac{1}{1 + \left(\sum_{j=1}^{d} \frac{\mathbb{1}_{\{\tilde{p}'(\tilde{t}_j)<0\}}}{|\tilde{p}'(\tilde{t}_j)|}\right) \cdot |\tilde{p}'(\frac{1}{2m})|} \geq \frac{5}{6}.$$

This completes the proof. ◁

Now we are ready to prove Lemma 3.1.

Proof of Lemma 3.1. We now prove that the random variables U and V defined in Section 3.1 satisfies Equation (12)-Equation (16). We begin by setting $\Delta = \frac{m}{20d^2}$.

Proof of Equation (12). Using Claim 3.7, we know that

$$\Pr\left[V = \frac{1}{2m}\right] \geq 5/6.$$

Proof of Equation (14). The largest value of U and V are the largest root of the modified Chebyshev polynomial. Now let us first upper bound the largest value of the roots of the Chebyshev polynomial. The largest value of the roots of Chebyshev polynomial is:

$$\max_{\ell\in[d]} t_\ell = \max_{\ell\in[d]} \frac{1}{\Delta}\left(1 - \cos\left(\frac{2\ell-1}{2d}\pi\right)\right) = \frac{2}{\Delta}\sin^2\left(\frac{\pi}{2}\left(\frac{2d-1}{2d}\right)\right) \leq \frac{2}{\Delta},$$

where the last inequality is obtained via Equation (2). So, the largest value of the roots of the modified Chebyshev polynomial is:

$$\max_{\ell\in[d]} \tilde{t}_\ell = \frac{1}{m}\lceil mt_j\rceil \leq \frac{2}{\Delta} + \frac{1}{m} = \frac{40d^2+1}{m},$$

The last equality follows from the fact that $\Delta = \frac{m}{20d^2}$. Plugging the values of $m = 70(n-1)$ and $d = \sqrt{10}\log(n-1)$, we have the result.

Proof of Equation (15). Recall that the random variable U is supported on the roots 0, $1/m$ and a subset of the roots of the modified Chebyshev polynomial (Equation (18)). Let us start by computing an upper bound on $\mathbb{E}[U]$.

$$\begin{aligned}
\mathbb{E}[U] &\leq \Pr[U=0]\cdot 0 + \Pr[U=\frac{1}{m}]\cdot\frac{1}{m} + \sum_{j=1}^{d}\Pr[U=\tilde{t}_j]\cdot\tilde{t}_j \\
&= \Pr[U=\frac{1}{m}]\cdot\frac{1}{m} + \frac{1}{Z_{\tilde{p}}}\left(\sum_{j=1}^{d}\tilde{t}_j\cdot\frac{1}{|\tilde{p}'(\tilde{t}_j)|}\right) \quad [Z_{\tilde{p}} \text{ is the normalization term, see Equation (23)}] \\
&\leq \frac{1}{m} + \frac{1}{Z_{\tilde{p}}}\left(\sum_{j=1}^{d}\left(t_j+\frac{1}{m}\right)\cdot\frac{\exp\left(\frac{12\Delta}{m/d^2}\right)}{|p'(t_j)|}\right) \qquad \left[\because \tilde{t}_j = \frac{1}{m}\lceil mt_j\rceil \text{ and Equation (26)}\right] \\
&\leq \frac{1}{m} + \frac{1}{Z_{\tilde{p}}}\left(\sum_{j=1}^{d}\left(1+\frac{1}{2}\right)t_j\cdot\frac{\exp\left(\frac{12\Delta}{m/d^2}\right)}{\frac{\Delta}{4}\cdot t_j^3\cdot\frac{d}{\sin\frac{2j-1}{2d}}}\right) \\
&\qquad [\because \text{using the proof of Claim 3.6 (Appendix A.2, full version [12])}] \\
&= \frac{1}{m} + \frac{1}{Z_{\tilde{p}}}\left(\sum_{j=1}^{d}\frac{6\cdot\exp\left(\frac{12\Delta}{m/d^2}\right)}{\Delta\cdot\left(\frac{4}{\Delta^2}\sin^4\left(\frac{\pi}{2}\cdot\frac{2j-1}{2d}\right)\right)\cdot\frac{d}{\sin\frac{2j-1}{2d}}}\right) \qquad [\because t_j = \tfrac{2}{\Delta}\sin^2\left(\tfrac{\pi}{2}\left(\tfrac{2j-1}{2d}\right)\right)] \\
&\leq \frac{1}{m} + \frac{1}{Z_{\tilde{p}}}\left(\sum_{j=1}^{d}\frac{6\cdot\exp\left(\frac{12\Delta}{m/d^2}\right)}{\left(\frac{4}{\Delta}\cdot\left(\frac{2j-1}{2d}\right)^4\right)\cdot\frac{d}{\frac{4}{\pi}\cdot\frac{(2j-1)}{2d}}}\right) \qquad [\because \text{Via Equation (2)}] \\
&= \frac{1}{m} + \frac{\exp\left(\frac{12\Delta}{m/d^2}\right)\cdot\Delta d^2}{Z_{\tilde{p}}}\cdot\left(\sum_{j=1}^{d}\frac{6}{\pi(j-1/2)^3}\right) \\
&\leq \frac{1}{m} + 17\cdot\frac{\exp\left(\frac{12\Delta}{m/d^2}\right)\cdot\Delta d^2}{Z_{\tilde{p}}} \qquad \left[\because \sum_{i=1}^{\infty}\frac{6}{\pi\cdot(i-1/2)^3} \leq 17\right]
\end{aligned}$$

$$= \frac{1}{m} + \frac{17\times 3}{2}\cdot\frac{\exp\left(\frac{12\Delta}{m/d^2}\right)\cdot\Delta d^2}{m^2} \qquad \left[\because Z_{\tilde{p}} \geq \frac{1}{|\tilde{p}'(\frac{1}{2m})|} \geq \frac{2}{3}m^2 \text{ via Equation (23) \& Claim 3.5}\right]$$

$$\leq \frac{7}{2m} = \frac{1}{20(n-1)} \qquad \left[\because \Delta \leq \frac{m}{20d^2} \;\&\; m = 70(n-1)\right]$$

Since Equation (16) holds, as we will prove below, we know that $\mathbb{E}[U] = \mathbb{E}[V]$. This implies that $\mathbb{E}[V] \leq \frac{1}{20(n-1)}$ holds as well.

Proof of Equation (16). We have $\mathbb{E}[U^t] = \mathbb{E}[V^t]$ for $1 \leq t \leq d+1$ from Fact 1.10 and Observation 3.2. ◀

3.2 Putting it together: proof of the $\Omega(m/\log m)$ lower bound

We are now ready to prove the lower bound of m-grainedness testing (Theorem 1.1): we will combine the results from Lemma 3.1, Lemma 1.9 and Theorem 2.1 to show our main $\Omega(m/\log m)$ lower bound in Theorem 1.1.

Proof of Theorem 1.1. Let us fix $n > 20, m = 70(n-1)$, and $s \geq 1$ be a parameter (we will later on set it to $s = \Theta(n/\log n)$). From Lemma 3.1, we know that there exist two discrete random variables U and V with the matching moments properties. Now we will use Lemma 1.9 for sU, sV. Following Lemma 3.1 and Lemma 1.9, we have $\Lambda \leq s \cdot \frac{20\log^2(n-1)}{n-1}$ and $T = 3\cdot\log(n-1)$

$$\mathsf{d}_{\mathsf{TV}}\left(\mathbb{E}[\mathrm{Poi}(sU), \mathbb{E}[\mathrm{Poi}(sV)]]\right) \leq \left(\frac{e\cdot s\cdot\frac{20\log^2(n-1)}{n-1}}{3\cdot\log(n-1)}\right)^{3\cdot\log(n-1)}.$$

And we want to satisfy Equation (5), as all others have been matched to use Theorem 2.1.

$$\left(\frac{e\cdot s\cdot\frac{20\log^2(n-1)}{n-1}}{3\cdot\log(n-1)}\right)^{3\cdot\log(n-1)} \leq \frac{1}{20(n-1)} \Leftrightarrow \frac{e\cdot s\cdot 20\log(n-1)}{3(n-1)} \leq \left(\frac{1}{20(n-1)}\right)^{\frac{1}{3\cdot\log(n-1)}}$$

For $n \geq 21$, we have

$$\frac{e\cdot s\cdot 20\log(n-1)}{3(n-1)} \leq \frac{1}{4} \leq \left(\frac{1}{20(n-1)}\right)^{\frac{1}{3\cdot\log(n-1)}},$$

and thus for $s = \frac{3(n-1)}{80e\log(n-1)}$, we can invoke, and get a sample complexity lower bound in Theorem 1.1, noting that $m = 70(n-1) = \Theta(n)$. Thus

$$\frac{s}{2} = \frac{3(n-1)}{160e\log(n-1)} = \Omega\left(\frac{m}{\log m}\right).$$

This concludes our proof. ◀

4 Uniformity Testing in the DoHo Model

In this section, we prove the following result on parameterized uniformity testing in the DoHO model, improving on the known $\tilde{O}(m)$-query upper bound in the regime $m \geq n^3$.

▶ **Theorem 4.1.** *For any m and constant $\varepsilon > 0$, the property Π_m of distributions over $\{0,1\}^n$ defined as*

$$\Pi_m = \{U_S : S \subseteq \{0,1\}^n, |S| = m\}$$

can be ε-tested in the DoHO model using $s = \mathcal{O}(m^{2/3}/\varepsilon^6)$ samples and $s\cdot n$ queries. Moreover, $\Omega(m^{2/3})$ samples are necessary.

Proof. The algorithm itself is quite simple, and follows from combining existing algorithms for *generalized uniformity* testing in the standard sampling model [5, 17] with an additional "check" of the ℓ_2 norm of the distribution: the tester is described in Algorithm 1. The query complexity follows trivially from the stated sample complexity, as n queries suffice to read the full sample.

Specifically, the algorithm is as follows:

Algorithm 1 Algorithm to test Π_m.

1: Set $\delta \leftarrow 1/6$, and $\alpha = \Theta(\varepsilon)$.
2: Run the adaptive ℓ_2 norm estimator of [5, Lemma 3.1] to obtain an $(1 \pm \varepsilon/10)$ estimate $\hat{\rho}$ of $\|D\|_2^2$ (with error probability δ) using $O(\sqrt{m}/\varepsilon^2)$ samples. If the algorithm does not terminate with this number of samples, or if the estimate is not in $(1 \pm \varepsilon/10)/m$, output reject.
3: Run the generalized uniformity tester of [5] on s (fully queried) samples from D with (TV) distance parameter α and error probability δ. If it rejects, output reject.
4: Output accept.

We now argue correctness. By a union bound, both subroutines behave as they should with overall probability 2/3: we hereafter assume their output is correct.

- *Completeness:* Assume $D \in \Pi_m$. Then $\|D\|_2^2 = \frac{1}{m}$, so $\hat{\rho} \in [\frac{1-\varepsilon/10}{m}, \frac{1+\varepsilon/10}{m}]$ and the first step does not reject. Similarly, $D \in \Pi_U = \bigcup_{k=1}^{\infty} \Pi_k$, so the generalized uniformity tester (which is by definition a tester for Π_U) accepts. Overall, Algorithm 1 returns accept.
- *Soundness:* By contrapositive, suppose Algorithm 1 returns accept. Since the second subroutine did not reject, it must be the case that D is $\frac{\varepsilon}{2}$-close (in total variation distance) to a distribution U_T, uniform on some subset T of a given size k. Moreover, the algorithm of [5] provides the extra guarantee that $(1 - O(\varepsilon))\|D\|_2^2 \leq \frac{1}{k} \leq (1 + O(\varepsilon))\|D\|_2^2$ (see [5, Lemma 3.4]).[1]

  Now, let $U_S \in \Pi_m$ be a closest distribution to D, over all subsets S of size m:

$$\mathsf{d}_{\mathsf{TV}}(D, U_S) \leq \mathsf{d}_{\mathsf{TV}}(D, U_T) + \mathsf{d}_{\mathsf{TV}}(U_T, U_S) \leq \frac{\varepsilon}{2} + 1 - \frac{\min(m,k)}{\max(m,k)}$$

  where the second inequality follows the minimum TV distance between two uniform distributions on supports k and m, which occurs when the supports of the distributions overlap as much as possible, and is equal to $1 - \frac{\min(m,k)}{\max(m,k)}$. In particular, we have $1 - \frac{\min(m,k)}{\max(m,k)} \leq \frac{\varepsilon}{2}$ if

$$\frac{1}{1 - \frac{\varepsilon}{2}} \cdot m \leq k \leq (1 - \tfrac{\varepsilon}{2}) \cdot m$$

  Now, to see why this holds, observe that by our first check, $\|D\|_2^2$ is within a $(1 \pm \frac{\varepsilon}{10})$ factor of $\frac{1}{m}$, and, by the additional guarantee of the [5] tester (adjusting the constant in the setting of α), within a $(1 \pm \frac{\varepsilon}{10})$ factor of $\frac{1}{k}$. This implies that m and k are within a $1 \pm \frac{\varepsilon}{4}$ factor of each other, and thus that $1 - \frac{\min(m,k)}{\max(m,k)} \leq \frac{\varepsilon}{2}$. Overall, this establishes that $\mathsf{d}_{\mathsf{TV}}(D, U_S) \leq \varepsilon$.

Finally, the claimed lower bound follows directly from the lower bound on generalized uniformity testing of [17], which holds even when the target size of the support is given. ◀

[1] This is why we chose to use this specific algorithm, instead of that of [17], which has better sample complexity (with respect to ε) but does not provide this extra guarantee.

5 Conclusion and open problems

In this work, we studied the problem of testing m-grainedness of distributions. We established a new lower bound of $\Omega(m/\log m)$ samples, improving on the previous lower bound of $\Omega(m^c)$ due to [22], thereby settling a conjecture by [22]. Along the way, we also obtained an alternative, simpler proof of the $\Omega(m^c)$ lower bound. By leveraging a reduction between the testing models due to [23], our result implies an optimal lower bound for uniformity testing in the DoHO (Distributions over Huge Objects) model, settling another conjecture of [23]. Finally, we provided a simple tester for uniformity testing in this DoHO model, with improved sample complexity for a large range of the parameters.

Our work leaves open several new avenues of research in this context; we list two of them below:

- **(i)** Our lower bound for m-grainedness testing establishes the optimal dependence on the parameter m, for constant proximity parameter $\varepsilon = \Omega(1)$. It would be interesting to fully characterize the complexity of the question, including the dependence on ε.
- **(ii)** From [23], it is known that testing uniformity over an m-element subset of $\{0,1\}^n$ requires $\Omega(m/\log m)$ queries, for $m \leq n$; and that $\tilde{O}(m)$ queries are sufficient for all m. Our own tester (Theorem 4.1) shows an upper bound of $O(m^{2/3}n)$ queries, better in the regime $m \geq n^3$. This leaves understanding the landscape of parameterized uniformity testing, especially in the intermediate regime $n \ll m \ll n^3$, as an open and intriguing question.

References

1 Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. *Advances in Neural Information Processing Systems (NeurIPS)*, 28, 2015.

2 Tomer Adar and Eldar Fischer. Refining the adaptivity notion in the huge object model. In *International Conference on Randomization and Computation (RANDOM)*, pages 45:1–45:16, 2024. doi:10.4230/LIPICS.APPROX/RANDOM.2024.45.

3 Tomer Adar, Eldar Fischer, and Amit Levi. Support testing in the huge object model. In *International Conference on Randomization and Computation (RANDOM)*, pages 46:1–46:16, 2024. doi:10.4230/LIPICS.APPROX/RANDOM.2024.46.

4 Sivaraman Balakrishnan and Larry Wasserman. Hypothesis testing for high-dimensional multinomials: A selective review, 2018.

5 Tugkan Batu and Clément L Canonne. Generalized uniformity testing. In *Symposium on Foundations of Computer Science (FOCS)*, pages 880–889, 2017. doi:10.1109/FOCS.2017.86.

6 Tugkan Batu, Eldar Fischer, Lance Fortnow, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. Testing random variables for independence and identity. In *Symposium on Foundations of Computer Science (FOCS)*, pages 442–451, 2001. doi:10.1109/SFCS.2001.959920.

7 Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D Smith, and Patrick White. Testing that distributions are close. In *Symposium on Foundations of Computer Science (FOCS)*, pages 259–269, 2000. doi:10.1109/SFCS.2000.892113.

8 Clément L Canonne. A survey on distribution testing: Your data is big. but is it blue? *Theory of Computing*, pages 1–100, 2020.

9 Clément L Canonne. Topics and techniques in distribution testing: A biased but representative sample. *Foundations and Trends® in Communications and Information Theory*, pages 1032–1198, 2022. doi:10.1561/0100000114.

10 Clément L Canonne, Ilias Diakonikolas, Daniel Kane, and Sihan Liu. Nearly-tight bounds for testing histogram distributions. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:31599–31611, 2022.

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

12 Clément L. Canonne, Sayantan Sen, and Joy Qiping Yang. Settling the complexity of testing grainedness of distributions, and application to uniformity testing in the huge object model. *ECCC preprint*, 2024. URL: https://eccc.weizmann.ac.il/report/2024/196.

13 Sourav Chakraborty, Eldar Fischer, Arijit Ghosh, Gopinath Mishra, and Sayantan Sen. Testing of index-invariant properties in the huge object model. In *Conference on Learning Theory (COLT)*, pages 3065–3136, 2023. URL: https://proceedings.mlr.press/v195/chakraborty23a.html.

14 Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. Sample-optimal identity testing with high probability. In *International Colloquium on Automata, Languages, and Programming (ICALP)*, 2018.

15 Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. Collision-based testers are optimal for uniformity and closeness. *Chic. J. Theor. Comput. Sci*, 25:1–21, 2019. URL: http://cjtcs.cs.uchicago.edu/articles/2019/1/contents.html.

16 Ilias Diakonikolas and Daniel M Kane. A new approach for testing properties of discrete distributions. In *Symposium on Foundations of Computer Science (FOCS)*, pages 685–694, 2016. doi:10.1109/FOCS.2016.78.

17 Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Sharp bounds for generalized uniformity testing. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

18 Oded Goldreich. The uniform distribution is complete with respect to testing identity to a fixed distribution. In *Electronic Colloquium on Computational Complexity (ECCC)*, page 1, 2016.

19 Oded Goldreich. *Introduction to property testing*. Cambridge University Press, 2017. doi:10.1017/9781108135252.

20 Oded Goldreich, Shafi Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. *Journal of the ACM (JACM)*, pages 653–750, 1998. doi:10.1145/285055.285060.

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

22 Oded Goldreich and Dana Ron. A lower bound on the complexity of testing grained distributions. *Computational Complexity (CC)*, page 11, 2023. doi:10.1007/S00037-023-00245-W.

23 Oded Goldreich and Dana Ron. Testing distributions of huge objects. In *TheoretiCS*, page 78, 2023.

24 metamorphy. Proving a formula for n-th degree polynomial with n distinct real roots, 2021. URL: https://math.stackexchange.com/questions/4074098/proving-a-formula-for-sum-j-1n-fracx-jkfx-j-for-f-an-n-th-degr.

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

26 Sofya Raskhodnikova, Dana Ron, Amir Shpilka, and Adam Smith. Strong lower bounds for approximating distribution support size and the distinct elements problem. *SIAM Journal on Computing (SICOMP)*, pages 813–842, 2009. doi:10.1137/070701649.

27 Ronitt Rubinfeld and Madhu Sudan. Robust characterizations of polynomials with applications to program testing. *SIAM Journal on Computing (SICOMP)*, pages 252–271, 1996. doi:10.1137/S0097539793255151.

28 Gregory Valiant and Paul Valiant. Estimating the unseen: improved estimators for entropy and other properties. *Journal of the ACM (JACM)*, pages 1–41, 2017. doi:10.1145/3125643.

29 Paul Valiant. Testing symmetric properties of distributions. *SIAM Journal on Computing (SICOMP)*, pages 1927–1968, 2011. doi:10.1137/080734066.

30 Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, pages 3702–3720, 2016. doi:10.1109/TIT.2016.2548468.

31 Yihong Wu and Pengkun Yang. Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *The Annals of Statistics*, pages 857–883, 2019.

32 Yihong Wu, Pengkun Yang, et al. Polynomial methods in statistical inference: theory and practice. *Foundations and Trends® in Communications and Information Theory*, pages 402–586, 2020.