

Parameterized Uniformity Testing in the Huge Object Model

Clément L. Canonne*

Sayantan Sen†

Joy Qiping Yang‡

Abstract

We continue the study of distribution testing in the Huge Object Model (Goldreich and Ron, 2023), focusing on the flagship question of *parameterized uniformity testing*: given access to data in $\{0, 1\}^n$, is its underlying probability distribution uniform on a small subset of size m ? The focus here is on both the number of data points required (sample complexity) and the number of bits from these data points one needs to observe (query complexity). While previous work fully characterized both of these in the “tiny support” regime $m \ll n$, very little was understood in the “medium-to-large” support one ($n \leq m \leq 2^n$). Our work remedies this state of affairs, establishing the existence of a sample-query complexity tradeoff and showing that (in contrast to the tiny-support regime) adaptive algorithms can significantly outperform non-adaptive ones.

1 Introduction

Distribution testing, defined by Goldreich, Goldwasser, and Ron [GGR98] as a special case of property testing, and implicitly by Goldreich and Ron [GR00] in the context of graph property testing, was systematically started a field of study with the work of Batu, Fortnow, Rubinfeld, Smith, and White [BFRSW00]. Over the past decades, it has then become a vibrant area of research, witnessing a flurry of results and new techniques, and with connections to other areas of computer science and statistics.

Along the way, many different properties of distributions have been studied, some quite extensively: among those, most relevant to this work, are *uniformity* (is the data coming from the uniform distribution over the whole domain Ω ?), *generalized uniformity* [BC17] (is the underlying data distribution a uniform distribution over *some* subset of the domain?), and *parameterized uniformity* [GR23b] (is the data distribution a uniform distribution over some *fixed-size* subset of the domain?), as well as *grainedness* (is this distribution consistent with the empirical distribution over an m -size dataset?) [Gol16; GR23a].

Another fruitful line of research focused on the *data access model*, in order to capture the type of queries to the samples available to the algorithms in various settings, beyond the standard i.i.d. sampling model, as data structures and types of data evolved. One such model was recently proposed by Goldreich and Ron [GR23b], to model data distributions over very large objects – settings where a sample is only implicitly provided, and querying the specific value comes at additional cost. Seeing these objects as n -bit strings for $n \gg 1$, the algorithms now not only “pay” to obtain a sample $s \in \{0, 1\}^n$, but also incur a cost for each of the n bits of s they choose to read. (For detailed definitions, see Section 2.) Parameterized uniformity testing and grainedness were two of the flagship properties whose study was well-motivated in this new Distribution over Huge Objects (DoHO) model. Yet, in spite of this rather extensive study by Goldreich and Ron, there remains a large gap

*University of Sydney. Email: clement.canonne@sydney.edu.au.

†Centre for Quantum Technologies, National University of Singapore. Email: sayantan789@gmail.com.

‡University of Sydney. Email: qyan6238@uni.sydney.edu.au.

in our understanding of the former, as we elaborate below. Narrowing this gap is the main objective of this paper.

The gap, and our focus. For $m \leq n$, it is known that testing parameterized uniformity in the DoHO model (over a domain of size $\{0, 1\}^n$, with support size parameter m , henceforth m -uniformity) has both *sample* and *query* complexity $\tilde{O}(m)$ [GR23b] and $\Omega(\frac{m}{\log m})$ [CSY25]. However, for $m > n$, only the upper bound holds, and the best known lower bound is an $\Omega(m^{2/3})$ sample complexity lower bound lifted from the “standard” testing setting, which in particular implies this sample and query complexity lower bound in the DoHO model. In addition, Canonne, Sen, and Yang [CSY25] established an alternative query complexity upper bound of $O(m^{2/3}n)$ (for all values of m, n), which implies that the $\tilde{O}(m)$ upper bound cannot be tight in all parameter regimes (as it is suboptimal whenever $m > n^3$).

Thus, while for $m \leq n$ the sample and query complexities of m -uniformity are well understood, for $m > n$ the picture is much less clear, and the known results hint that one can somehow “trade” sample and query complexities:

- use $\tilde{O}(m)$ samples (much more than the required $\Omega(m^{2/3})$!) but only $\tilde{O}(m)$ queries; or
- use $O(m^{2/3})$ samples but at the cost of $O(m^{2/3}n)$ queries.

This intriguing state of affairs in the $m \gg n$ regime is the main focus of our work:

What are the sample and query complexities of testing m -uniformity in the DoHO model, *and can one trade one for the other?*

Interestingly, both known algorithms mentioned above are *non-adaptive*, in that they decide which coordinates to query prior to observing the samples themselves. One could consider a broader class of algorithms proceeding in rounds, first querying *some* coordinates of the samples, before, based on the information obtained, querying additional coordinates of a carefully chosen subset of samples.

Can adaptive testers achieve a better sample-query complexity tradeoff than non-adaptive ones?

We make significant progress on these questions in two aspects: first, we design an adaptive algorithm achieving a much improved upper bound, with both sample and query complexities $O(m^{2/3}n^{1/3})$ – and, moreover, able to interpolate between the two, answering the first question in the positive. Second, we complement this upper bound with a new $\Omega(\sqrt{mn})$ query complexity lower bound against *non-adaptive* testers, thereby settling the second question as well: *yes, adaptivity helps*. Finally, to complement our adaptive algorithm, we provide a more involved lower bound that holds against *adaptive* algorithms, and establishes a similar (though quantitatively weaker) statement, allowing a further tradeoff between samples and queries in the lower bound. We emphasize that these results, which hold in the $m \gg n$ regime, are quite surprising: as it was known that, for $m \leq n$, there is no such trade-off, and adaptivity does not help.

1.1 Our results and techniques

We here state our main results, before providing a high-level overview of the algorithm and proof.

Theorem 1.1 (Upper bound (Informal; see [Theorem 3.1](#))). *There is an m -uniformity adaptive tester in the DoHO model with sample and query complexities $\tilde{O}(m^{2/3}n^{1/3})$. More generally, one can achieve sample complexity $\tilde{O}(t + m^{2/3})$ with query complexity $\tilde{O}(t + m^2n/t^2)$, for any $1 \leq t \leq O(m)$.*

Theorem 1.2 (Non-adaptive lower bound (Informal; see [Theorem 4.1](#))). *Any non-adaptive tester for m -uniformity in the DoHO model must either have sample complexity $\tilde{\Omega}(m)$ or query complexity $\tilde{\Omega}(\sqrt{mn})$.*

Theorem 1.3 (Adaptive Lower bound (Informal; see [Theorem 5.1](#))). *Any (possibly adaptive) tester for m -uniformity in the DoHO model that takes $s \leq \tilde{O}(m^{3/4})$ samples, must have query complexity $\tilde{\Omega}(\frac{mn}{s})$. In particular, any algorithm taking $s = \Theta(m^{2/3})$ samples must make $\tilde{\Omega}(m^{1/3}n)$ queries.*

In particular, while the second one is for a restricted regime of parameters and is quantitatively weaker, it applies to a broader range of algorithms, including the one behind our upper bound, [Theorem 1.1](#). We emphasize that while these two lower bounds apply more generally, in light of the previous discussion, they are only meaningful for $m \geq n$.

Existing sample and query complexity bounds			
Adaptivity	Non-adaptive UB [GR23b]	Non-adaptive UB [CSY25]	Adaptive LB [GR23b]
Bounds	$q \leq \tilde{O}(m)$	$q \leq m^{2/3} \cdot n$	$q \geq \tilde{\Omega}(n)$
Conditions	$s = \tilde{O}(m)$	$s = O(m^{2/3})$	$s \leq \tilde{O}(m)$
Our sample-query complexity bounds			
Adaptivity	Adaptive UB	Non-adaptive LB	Adaptive LB
Bounds	$sq \leq O(m^{4/3}n)$	$q \geq \tilde{\Omega}(\sqrt{mn})$	$sq \geq \tilde{\Omega}(mn)$
Conditions	$O(q) \geq s \geq \Omega(m^{2/3})$	$s \leq \tilde{O}(m)$	$s \leq \tilde{O}(m^{3/4})$

Table 1: A table of the sample-query complexity bounds for $\varepsilon = \Theta(1)$, where s denotes the sample complexity, and q is the query complexity. To derive the adaptive upper bound tradeoff, we set in our upper bound $s = O(t) \geq m^{2/3}$, from which $q \leq O(s + m^2n/s^2)$, which implies $q - s \leq O(m^2n/s^2)$, so that $qs^2 \leq O(m^2n)$; and since $s \geq \Omega(m^{2/3})$, we get $qsm^{2/3} \leq O(qs^2) \leq O(m^2n)$, implying $qs \leq O(m^{4/3}n)$.

1.2 The adaptive m -uniformity tester

At a high level, our adaptive tester leverages two key insights:

- (i) *It is actually unnecessary to check elements with large masses* – we will show that these “big” elements combined (which we denote as the set $\mathcal{S} = \mathcal{S}(\tau)$, where τ is the threshold chosen for an element to be considered big¹) will only contribute little to the distance from m -uniform (see [Lemma 3.6](#)). This means that in order to be far from m -uniform in EMD, the *small elements* combined need to be far from m -uniform. More formally, we show that the *pseudo-distribution*² induced by the small elements needs to be far from m -uniform. Thus, we can ignore the heavy elements altogether and avoid fully querying them. (To do so, algorithmically, we will first identify every string in $\mathcal{S}$, by taking $\tilde{O}(1/\tau)$ samples and only querying $\tilde{O}(1/\varepsilon)$ bits from each of them.)
- (ii) *The set of small elements* (i.e., any string with mass smaller than τ , in $\overline{\mathcal{S}}$) *create few collisions*: that is, altogether induce a pseudodistribution with *small ℓ_2 -norm* (and also *small ℓ_3 -norm*;

¹For a different chosen threshold τ , our tester provides a different sample-query tradeoff.

²For a probability distribution D over a (discrete) domain Ω and a set $S \subseteq \Omega$, we define the *pseudo-distribution* D_S by $D_S(x) = D(x)$ for every $x \in S$ (note that D_S may not be normalized). See [Section 2](#).

see [fact 3.8](#) for more details). Therefore, the number of pairwise collisions when taking $O(m^{1/2})$ samples, bounded by $m \cdot \|D_{\bar{S}}\|_2^2$ in expectation, will be small. As a result, we can afford to fully query every sample from $\bar{S}$ participating in a pairwise collision, as this only costs a total of $O(m \cdot \|D_{\bar{S}}\|_2^2 \cdot n)$ queries. Which is great news for us, as obtaining good estimates of the ℓ_2 and ℓ_3 norms suffices for testing m -uniformity [BC17; CSY25] in the standard model, and we show this still holds in the DoHO one.

Accordingly, our tester has two main steps: (i) first, identify strings (clusters) with mass more than $1/\tau$ (the heavy elements), and then (ii) estimate the true ℓ_2 and ℓ_3 norm on what remains by querying all n -bit sample strings with mass smaller than $1/\tau$. This second step is the source of adaptivity in our algorithm as it depends on the first; moreover, querying every bit of the sample strings that participate in collisions in this second step guarantees that the resulting estimates are accurate (since the algorithm learns the exact “identity” of each string in this step, and thus will not not conflate two distinct strings as one, which could lead to a wrong collision count).

Algorithm 1 HIGH LEVEL OVERVIEW OF THE ADAPTIVE m -UNIFORMITY TESTER

Require: Sample and query access to an unknown distribution D , parameters m, n, τ, ε

Ensure: Return accept if D is m -uniform and reject if D is ε -far from m -uniform, with probability at least $2/3$

▷ Query $\tilde{O}(1/\varepsilon)$ bits per sample.

1: $\mathcal{S} \leftarrow \text{FINDHEAVYELEMENTS}(D, \tau)$

▷ Adaptively based on $\bar{\mathcal{S}}$: query all n bits of each sample.

2: $\|\hat{D}_{\bar{\mathcal{S}}}\|_2^2, \|\hat{D}_{\bar{\mathcal{S}}}\|_3^3 \leftarrow$ Estimate ℓ_2 and ℓ_3 over low mass region ($\bar{\mathcal{S}}$)

3: **if** $\|\hat{D}_{\bar{\mathcal{S}}}\|_2^2, \|\hat{D}_{\bar{\mathcal{S}}}\|_3^3$ looks m -uniform **then**

4: **return** accept

5: **else**

6: **return** reject

However, this is only true for the second step, and these spurious collisions *could* happen in the first step. Indeed, since in that step we are only performing $\tilde{O}(1/\varepsilon)$ queries for each sample to detect the heavy elements, a naive implementation could be tricked by “fake” collisions, for instance if two strings are equal on the majority of the n bits, but differ on some tiny fraction of the bits (which with high probability will not be detected in few queries). Consider the following example:

Example 1.4. Let D be a uniform distribution over m strings which, pairwise, only differ by $O(1)$ bits: so that every two strings’ relative Hamming distance is at most $O(1/n) \ll \varepsilon = \Theta(1)$. Suppose we take $O(\sqrt{m})$ samples (sufficient for ℓ_2 norm estimation) but only query $\tilde{O}(1/\varepsilon)$ -bits from each sample string. Based on these queries, D is indistinguishable from a distribution supported on a single string (as all samples are identical, with overwhelming probability, on the set of bits queried). As a result, the algorithm will reject, as a point mass distribution is far for m -uniform.

This issue can arise as, by querying only $\tilde{O}(1/\varepsilon)$ bits for some elements, the algorithm can be misled into believing it is observing samples from a completely different distribution than D . Unfortunately, the obvious fix (querying more bits from these samples) is not an option, as this would blow up the query complexity. And indeed, one of the key technical contributions of our algorithm is to understand and conduct inference on this “new” distribution (obtained after querying only few bits), the partial “view” of D , which we denote D' and will describe in detail below.

Key construct D' : a typical $\tilde{O}(1/\varepsilon)$ -bit view of D . We introduce a surrogate distribution D' to analyze the $\tilde{O}(1/\varepsilon)$ view of D , which (the analysis of) our algorithm will use to identify elements

with masses greater than our chosen threshold τ . Fortunately, Goldreich and Ron [GR23b] have already laid the groundwork when considering this viewpoint for m -grainedness testing. We first recall some key notions from their work, along with notions their analysis (and ours) hinges upon. These definitions are then used to construct a “clustered” version of the probability distribution D . Throughout, we denote by s the overall sample complexity.

Definition 1.5 (Heavy strings and good sequences). For any fixed ℓ -subset $J \subseteq [n]$, we say that $\sigma \in \{0, 1\}^\ell$ is J -heavy (w.r.t. D) if $\Pr_{X \sim D}[X_J = \sigma] \geq 1/(100s^2)$. For a fixed ℓ -subset $J \subseteq [n]$, a sequence of $s' = \Omega(s^2 \cdot \ell)$ strings $(w^{(1)}, \dots, w^{(s')}) \in (\{0, 1\}^n)^{s'}$ is then J -good (w.r.t. D) if its J -restrictions hit all J -heavy strings, i.e., for every J -heavy string σ , there exists $i \in [s']$ such that $w_J^{(i)} = \sigma$. Finally, a sequence $(w^{(1)}, \dots, w^{(s')})$ is good (w.r.t. D) if it is J -good for a $1 - o(1)$ fraction of all ℓ -subsets J .

Remark 1.6. Goldreich and Ron [GR23b] provided a non-adaptive query tester for m -uniformity using m -grainedness³ as a proxy, which takes $\tilde{O}(m)$ samples and queries the same (randomly chosen) $\tilde{O}(\log m/\varepsilon)$ -size subset J of coordinates for each sample. It is natural to try to bypass this proxy step of testing m -grainedness, as it costs significantly more samples than parameterized uniformity testing ($O(m/\log m)$ for the former, compared to $O(m^{2/3})$ for the latter). We also emphasize that our approach significantly differs from that of [GR23b]: namely, our approach crucially uses adaptivity while their m -grainedness tester of is intrinsically non-adaptive. However, we first briefly describe their approach, as we draw upon the insights from their work will be to obtain our improved adaptive tester.

The above definition, which underlies the analysis of the algorithm of Goldreich and Ron in the context of testing m -grainedness, captures a typical distribution that one would most likely observe when looking only at the ℓ -subset J positions of the samples. Since two strings x and y are then indistinguishable if $x_J = y_J$, we will (have to) treat them as the same string. Indeed, given a set of J -good strings $\vec{w} = (w^{(1)}, \dots, w^{(s')})$, we can construct a distribution D' from D by merging multiple strings that are the same under the J -restriction, thus creating a “clustered” version of D . In more detail, let $I = I(J)$ be a minimal set of indices $i \in [s']$ such that $\{w_J^{(i)} \mid i \in I\} = \{w_J^{(i)} \mid i \in [s']\}$. From D , we define D' by the following mapping $f_{J, \vec{w}}$:

$$D'(x) = \begin{cases} \Pr_{X \sim D}[X_J = w_J^{(i)}] & \text{if } x = w^{(i)} \text{ for } i \in I; \\ 0 & \text{if } x_J \in \{w_J^{(i)} \mid i \in I\} \text{ but } x \notin \{w^{(i)} \mid i \in I\}; \\ \Pr_{X \sim D}[X = x] & \text{if } x_J \notin \{w_J^{(i)} \mid i \in I\}. \end{cases} \quad (1)$$

Clearly, if one takes samples $X \sim D$ and $X' \sim D'$ but only queries the set of indices J , then X_J and X'_J are identical (w.h.p.). Goldreich and Ron then base the analysis of their m -grainedness tester on X' instead, showing the following: i) if D is m -grained, then so is D' (as this property is closed under mapping); ii) When D is far from m -grained, then D' is also far from m -grained (because, as they argue, D' is close to D in EMD distance). The final ingredient is then showing that, when drawing s samples, X' and X'_J will, with high probability, have the same *collision pattern* (i.e., the empirical frequencies defined by these s samples will be identical). Thus, even though the algorithm samples from D and not D' , since it restricts its attention on a subset of indices J of the s samples,

³A distribution D is called m -grained if all probability values under D are an integer multiple of the “granularity” $1/m$. As detailed in [GR23b], the property of being m -grained is closely related to that of being m -uniform – yet not equivalent. However, one observation which we will rely on is that, if $\frac{\log m}{n} \ll \varepsilon$, then any distribution ε -close to m -grained is also ε -close to m -uniform (both in EMD). This is because, as observed in [GR23b, Corollary 2.3], any m -grained distribution can be turned into m -uniform by modifying the first $O(\log m)$ -bits in each support string.

analyzing its behavior under D' is sufficient. This is crucial: If D' has a small chance of collisions in set $\overline{\mathcal{S}'}$, then D will also see fewer collisions in the corresponding area under J -restriction.

We let $\mathcal{S}'$ be the subset of “heavy” elements of the domain (defined below, as a function of the aforementioned threshold τ ; see also (2)):

$$\mathcal{S}' := \{x \in \{0, 1\}^n \mid D'(x) \geq 1/\tau\};$$

which are those we will avoid performing additional queries on (so we will only query additional bits on strings in its complement $\overline{\mathcal{S}'}$; or, as a technicality, we are in fact querying $\overline{\mathcal{S}}$ and thus we need a way to relate elements between $\overline{\mathcal{S}'}$ and $\overline{\mathcal{S}}$. This is detailed in Section 3 and illustrated in Fig. 1).

The $O(n)$ view of D . Recall that in step (ii), we will fully query a subset of samples that are unlikely to cause many collisions, i.e., the set of elements with relatively low mass. So when we choose a set of these “not much queried” samples from the original multiset of samples (those of D) and perform additional queries (we query the full n bits in these samples), the illusion that these are samples from D' will be broken. However, by doing this, we can obtain the actual norm estimates of D (over $\overline{\mathcal{S}}$) rather than D' (over $\overline{\mathcal{S}'}$) with very few queries,⁴ leading to the final distinguishing argument below.

Distinguishing with $\overline{\mathcal{S}}$ (low mass elements). Our main argument, focusing on the low-mass elements, is as follows: (1) When D is m -uniform, then $D_{\overline{\mathcal{S}}}$ will remain m -uniform as a *pseudo-distribution* on $\overline{\mathcal{S}}$ (meaning that every $x \in \overline{\mathcal{S}}$ satisfies $D(x) = \frac{1}{m}$); (2) When D is far from m -uniform, we will show two things: First, $D'_{\mathcal{S}'}$ and $D_{\mathcal{S}}$ will be close in EMD (Lemma 3.4). Second, $D'_{\mathcal{S}'}$ will be close to some m -uniform (pseudo)-distribution in EMD (Lemma 3.6). With these two points combined, we have that $D_{\mathcal{S}}$ will be close to some m -uniform (pseudo)-distribution in EMD (Lemma A.3). Thus, as D is overall far from m -uniform in EMD, $D_{\overline{\mathcal{S}}}$ must be far from any m -uniform (pseudo)-distribution (Lemma 3.7). Thereafter, we can then focus on checking farness from m -uniform of the (pseudo)-distribution, which involves estimation of ℓ_2 and ℓ_3 norms (counting pairwise and three-wise collisions suffices). A careful analysis shows that taking $O(m^{1/2}/\text{poly}(\varepsilon))$ and $O(m^{2/3}/\text{poly}(\varepsilon))$ samples to perform these two norm estimations, respectively, suffices to distinguish the two cases. One can further save on query budget by focusing on samples that collide, whose number in expectation, are bounded by the two-norm (and three-norm) of $D'_{\mathcal{S}'}$. The query complexity for this check is thus $O((m^{1/2})^2 \|D'_{\overline{\mathcal{S}}}\|_2^2 \cdot n) + O((m^{2/3})^3 \|D'_{\overline{\mathcal{S}}}\|_3^3 \cdot n)$: as discussed above, the (expected) values of the quantities $\|D'_{\overline{\mathcal{S}}}\|_2^2$ and $\|D'_{\overline{\mathcal{S}}}\|_3^3$ can be tuned by the choice of τ , leading to the sample and query complexity tradeoff in our upper bound.

1.3 A non-adaptive lower bound

To establish our non-adaptive lower bound, a natural idea is to define two families of probability distributions, one comprised of m -uniform distributions (yes-instances) and the other of distributions far from m -uniform (no-instances), before proving that distinguishing a (randomly chosen) yes-instance from a (randomly chosen) no-instance requires either many samples, or many queries. Unfortunately, this approach comes with significant hurdles, as one needs to both create *distance* between yes- and no-instances, while also arguing indistinguishability: but most natural ways to create distance in EMD lead to distributions supported on strings pairwise far in Hamming distance – then easy to distinguish given very few queries! One fix would be to start with distributions far in TV distance, then create EMD distance through the application of suitable types of error-correcting,

⁴The key idea to guarantee low collisions when querying D over $\overline{\mathcal{S}}$ is that we can upper bound the number of collisions via D' – since D has strictly fewer collisions than to D' and thus analyzing the projection on J over $\overline{\mathcal{S}'}$ would suffice.

with additional “indistinguishability properties for low-query algorithms.” Again unfortunately, finding a type of construction and codes meeting all the requirements seemed arduous, if not impossible: instead, we take another route. Namely, we define *three* families of instances: **yes**-, **no**-, and **maybe**-instances, such that (1) **yes**-instances are m -uniform, (2) **no**-instances are far from m -uniform, (3) **no**- and **maybe**-instances are indistinguishable with $\tilde{o}(m)$ samples, and (4) **yes**- and **maybe**-instances are indistinguishable with $\tilde{o}(\sqrt{mn})$ non-adaptive queries. (Proving (4) is our main technical contribution.) Now, a valid tester must be able to consistently distinguish **yes**- and **no**-instances: it could say *anything* on **maybe**-instances, but still needs to say *something*. This implies it must also be able to either distinguish **yes**- and **maybe**-instances, or distinguish **no**- and **maybe**-instances. We do not know which of the two, but this is fine: either way, this requires either high sample complexity (by (3)) or high query complexity (by (4)), yielding our lower bound.

1.4 An adaptive lower bound

Our adaptive lower bound draws from the same idea, again defining three families of distributions, **yes**-, **no**-, and **maybe**-instances, but with considerably more technical difficulties to “foil” adaptive algorithms. The first step is to again start with “superdistributions”⁵ over m well-separated clusters, and embed hard instances for m -grainedness into these cluster distributions to obtain D_{maybe} and D_{no} . However, we now modify the m -grainedness hard instance to get a new “type of cluster”, with larger total mass than the other clusters (namely, $1/t$, for some suitable parameter t), while preserving the indistinguishability between D_{maybe} and D_{no} . The reason for this type of “heavy clusters” is to *generate spurious collisions*, inspired by the way our upper bound (algorithm) works: indeed, an algorithm, to distinguish between D_{yes} and D_{maybe} , will need to find “secret” (and light) clusters in which we hide the difference between **yes**-instances (which are m -uniform, and so in which these clusters contain several very close but distinct points) and **maybe**-instances (which are only m -grained, and in which these clusters consist of a single point). To “win”, an algorithm must thus obtain collisions in one of these “secret” clusters, and query at least $\tilde{\Omega}(n)$ coordinates of the corresponding colliding samples to figure out if they are the same, or different points within that cluster. Yet, intuitively, the algorithm will not be able to *only* query a few of these “candidate secret clusters” as it will see *many* clusters with the same number of collisions thanks to these “spurious added clusters”, and cannot *a priori* know which ones to focus on: so the algorithm is forced to query $\tilde{\Omega}(n)$ bits from *many* cluster-colliding samples to have a chance to win. Of course, this intuition is quite intricate to formalize: in order to do so, we reduce the question to a problem reminiscent of list-decoding, in which we are able to prove the lower bound against a kind of (weakly) adaptive algorithms, before lifting the argument to general adaptive algorithms.

1.5 Related work.

Several new models of distribution testing, such as the *dual access model* [RS09; CR14] *conditional sampling* [CRS15; CFGM16], *subcube conditioning* [BC18], and very recently, the *Distribution over Huge Objects* (DoHO) model [GR23b] have been introduced over the past two decades. See the surveys ([Can20; Can22; BW18]) and the book by [Gol17, Chapter 11] for detailed references.

The DoHO model has raised significant interest over the past few years. Chakraborty, Fischer, Ghosh, Mishra, and Sen [CFGMS23] studied the notion of *index-invariant* properties (i.e., properties

⁵Formally, consider two clusters $C_1 = \{x_1, x_2, x_3\}$ and $C_2 = \{x_4, x_5, x_6\}$ such that C_1 and C_2 are centered at x_1 and x_4 , respectively. While the distribution is defined over, say all the points in $\{x_1, \dots, x_6\}$, due to clustering, we can think a distribution, called a *super-distribution* defined only over $\{x_1, x_2\}$, where the total mass of all the points inside a cluster is moved to its center.

which are invariant under the permutation of the indices, in contrast to the usual notion of *label-invariant* properties) and showed that any index-invariant property whose support set has bounded VC-dimension can be learned using a constant number of queries. That paper also established tight separations of adaptive and non-adaptive testers for index-invariant properties. Later, Chakraborty, Fischer, Ghosh, Levi, Mishra, and Sen [CFGLMS25] extended these results, and proved that any index-invariant property that can be tested using constant queries, can also be estimated using constant number of queries. Adar and Fischer [AF24] defined various notions of adaptivity in this model and studied separation results among them. Finally, Adar, Fischer, and Levi [AFL24] studied the problem of support size testing in the DoHO model, and explicitly ask about trade-offs between sample and query complexities as an open question in the context of this task.

2 Preliminaries

Throughout, for an integer $k \geq 1$ we denote by $[k]$ the set $\{1, \dots, k\}$. We will use the standard asymptotic notation $O(\cdot)$, $\Theta(\cdot)$, $\Omega(\cdot)$, $o(\cdot)$ as well as the (slightly) less standard $\tilde{O}(\cdot)$, $\tilde{\Omega}(\cdot)$ and $\tilde{\Theta}(\cdot)$, which omit polylogarithmic dependencies in the parameters. For a probability distribution D over a (discrete) domain Ω and a set $S \subseteq \Omega$, we write D_S for the *pseudo-distribution* where $D_S(x) = D(x)$ for every $x \in S$, and $D(S) = \sum_{x \in S} D(x) \in [0, 1]$ for its probability mass (note that D_S may not be normalized). Often, our algorithms will rely on the *Poissonization* technique, whereby instead of taking a fixed number s of samples from a distribution D , they instead take $s' \sim \text{Poi}(s)$ samples: this simplifies some of the analysis, and is essentially without loss of generality (see, e.g., [Can20, Fact D.10] for more on Poissonization). Finally, we write “w.h.p.” to denote an outcome occurring with sufficiently high (constant) probability. We first recall the definition of Earth Mover distance (EMD) with respect to Hamming distance. For an n -bit string $x \in \{0, 1\}^n$, and a subset $J \subseteq [n]$, we denote by x_J the sub-string obtained by concatenating every bit obtained from $\{x_i\}_{i \in J}$; we say that $x_J \in \{0, 1\}^{|J|}$ is a J -restriction of x .

Definition 2.1 (EMD with respect to Hamming distance). The *Earth Mover Distance (EMD)* between distributions D_1 and D_2 over $\{0, 1\}^n$ is $\text{d}_{\text{EM}}(D_1, D_2) := \inf_{T \in \mathcal{T}(D_1, D_2)} \mathbb{E}_{(x, y) \sim T} [\text{d}_{\text{H}}(x, y)]$, where $\mathcal{T}(D_1, D_2)$ denotes the set of all coupling between D_1 and D_2 , and $\text{d}_{\text{H}}(x, y)$ the (normalized) Hamming distance between $x, y \in \{0, 1\}^n$.

We will require an equivalent characterization of EMD, which extends to (*pseudo*)-distributions:

Fact 2.2. Let $\Omega = \{0, 1\}^n$ and fix $S \subseteq \Omega$, $S' \subseteq \Omega$. Given two distributions $D, D': \Omega \rightarrow [0, 1]$ and $D(S) = D'(S')$, the earth mover distance between two (*pseudo*)-distributions D_S and $D'_{S'}$ is the optimal value of the following linear program:

$$\begin{aligned} \text{d}_{\text{EM}}(D_S, D'_{S'}) = & \min \sum_{x, y \in \Omega} w_{x, y} \cdot \text{d}_{\text{H}}(x, y). \\ & \forall x \in \Omega : \sum_{y \in \Omega} w_{x, y} = D(x) \\ & \forall y \in \Omega : \sum_{x \in \Omega} w_{x, y} = D'(y) \\ & \forall x, y \in \Omega : w_{x, y} \geq 0. \end{aligned}$$

We then define the Huge Object Model, and what adaptivity means in this model.

Definition 2.3 (DoHO model). In the *Distribution over Huge Objects* (DoHO) model, the algorithm is provided access to an unknown probability distribution D over $\{0, 1\}^n$ through a *sampling* and a *query* oracles. The sampling oracle provides i.i.d. samples from D , whose bits are not revealed to the algorithm. For any such sample $x \in \{0, 1\}^n$, the algorithm can then use the query oracle which, on query $j \in [n]$, returns the bit $x_j \in \{0, 1\}$. The goal is to minimize the total sample and query complexities, i.e., the number of calls the algorithm makes to the sampling and query oracles.

Definition 2.4 (Non-adaptive DoHO algorithms). An algorithm in the DoHO model with sample complexity s is said to be *non-adaptive* if the sets of queries $Q_1, \dots, Q_s \subseteq [n]$ made to the query oracle are determined prior to seeing the result of any query; i.e., the indices queried for each of the s samples are chosen ahead of time. In contrast, an adaptive algorithm is allowed to choose the t -th query to the query oracle as a function of the $t - 1$ bits returned in the previous queries it made.

We now define our main object of study, parameterized uniformity testing. Recall that, for a given integer $m \geq 1$, a probability distribution D over a domain Ω is said to be *m -uniform* if there exists a set $S \subseteq \Omega$ of size $|S| = m$ such that $D(x) = 1/m$ for each $x \in S$.

Definition 2.5 (Parameterized uniformity testing). Let $n, m \geq 1$ be positive integers. The task of *parameterized uniformity testing* (or *m -uniformity testing*) requires, given DoHO access to an unknown distribution D over $\{0, 1\}^n$ and a parameter $\varepsilon \in (0, 1]$, to distinguish, with probability at least $2/3$, between the cases (i) D is a *m -uniform distribution* and (ii) D is ε -far in EMD distance from all m -uniform distributions D' (i.e., $d_{\text{EM}}(D, D') \geq \varepsilon$).

3 An adaptive tester for m -uniformity testing

In this section, we prove our main result, stated below:

Theorem 3.1. *For every $m \geq \tau \geq 1$, there exists a testing algorithm for m -uniformity (Algorithm 2) with sample complexity $\tilde{O}\left(\tau + \frac{m^{1/2}}{\varepsilon^{9/2}} + \frac{m^{2/3}}{\varepsilon^{13/3}}\right)$ and query complexity $\tilde{O}\left(\frac{\tau}{\varepsilon} + \frac{m^2}{\varepsilon^{13}} \cdot \frac{n}{\tau^2}\right)$, where $1 \geq \varepsilon > \Omega(\tau/m)$ is the distance parameter.*

For constant $\varepsilon > 0$, by optimizing over τ this implies the following:

Corollary 3.2. *There is a testing algorithm for m -uniformity with sample complexity $\tilde{O}(m^{2/3}n^{1/3})$ and query complexity $\tilde{O}(m^{2/3}n^{1/3})$, where $\varepsilon = \Theta(1)$ is the distance parameter.*

Moreover, as discussed in the introduction, Theorem 3.1 provides a smooth tradeoff between sample and query complexities. For instance, seeing ε as a constant for simplicity, setting $\tau = \Theta(m^{2/3})$ retrieves, up to logarithmic factors, the result of Canonne, Sen, and Yang [CSY25] ($\tilde{O}(m^{2/3})$ samples, $\tilde{O}(m^{2/3}n)$ queries), while at the other extreme, choosing $\tau = \Theta(m)$ yields the result of Goldreich and Ron [GR23b] ($\tilde{O}(m)$ samples, $\tilde{O}(m)$ queries).

To implement the strategy discussed in the introduction and analyze Algorithm 2 (thus proving Theorem 3.1), we will require a few technical lemmas, which we provide below. All proofs can be found in Section A. Our main algorithm for m -uniformity testing (Algorithm 2) uses four subroutines:

- FINDHEAVYELEMENTS (Algorithm 3), which identifies the set of heavy elements $\mathcal{S}$, based on a given threshold τ
- SANITYCHECKNORMS (Algorithm 4), which, given a set, run a sanity check to ensure that the pseudodistribution on this set has ℓ_2 and ℓ_3 norms which are not too small (and, in particular, at least not obviously inconsistent with an m -uniform distribution)
- DRAWFILTERED (Algorithm 5), which, from a multiset of samples, only keeps the small mass ones, filtering out the heavy strings; and
- REALNORMEST (Algorithm 6), which, given a set, estimates the ℓ_2 and ℓ_3 norms of the pseudodistribution on this set to high accuracy (with a query complexity related to the value of the norms being estimated).

Algorithm 2 The m -uniformity tester

Require: Sample and query access to an unknown distribution D , parameters m, n, τ, ε

Ensure: Return accept if D is m -uniform and reject if D is ε -far from m -uniform, with probability at least $2/3$.

- 1: Set $\ell \leftarrow O\left(\frac{\log(mn/\varepsilon)}{\varepsilon}\right)$, $s_1 \leftarrow O\left(\frac{\sqrt{m}}{\varepsilon^{5/2}}\right)$, $s_2 \leftarrow O\left(\frac{m^{1/2}}{\varepsilon^{9/2}} + \frac{m^{2/3}}{\varepsilon^{13/3}}\right)$.
 - 2: Choose a set of indices $J \subseteq [n]$ of size ℓ uniformly at random.
▷ Identify large mass region $\mathcal{L}$, which costs $O(\tau \log \tau)$ samples and $O(\ell \cdot \tau \log \tau)$ queries.
 - 3: $\mathcal{L} \leftarrow \text{FINDHEAVYELEMENTS}(D, J, \tau)$
 - 4: Take $O(\frac{1}{\varepsilon^4})$ samples from D to estimate $D'(\overline{\mathcal{L}})$, under the J restriction.
 - 5: **if** $\hat{D}'(\overline{\mathcal{L}}) < O(\varepsilon)$ **then return accept**
▷ Accept only if ℓ_2 and ℓ_3 norms looks m -uniform over the small mass region: $\overline{\mathcal{L}}$.
 - 6: **if** $\text{SANITYCHECKNORMS}(D, J, \mathcal{L})$ returns reject, **then return reject**.
 - 7: $\mathcal{S}_1 \leftarrow \text{DRAWFILTERED}(s_1, \mathcal{L}, J)$; $C_1 \leftarrow \text{REALNORMEST}(\mathcal{S}_1, J, 2)$
▷ Note that $\hat{D}'(\overline{\mathcal{L}}) = \hat{D}(\overline{\mathcal{L}})$.
 - 8: **if** $\left|\frac{C_1}{\hat{D}(\overline{\mathcal{L}})} - \frac{1}{m}\right| > \frac{\varepsilon}{400}$ **then return reject**.
 - 9: $\mathcal{S}_2 \leftarrow \text{DRAWFILTERED}(s_2, \mathcal{L}, J)$
 - 10: $C_2 \leftarrow \text{REALNORMEST}(\mathcal{S}_2, J, 2)$; $\tilde{C}_2 \leftarrow \text{REALNORMEST}(\mathcal{S}_2, J, 3)$
 - 11: **if** $\left|\tilde{C}_2 - \frac{C_2^2}{\hat{D}(\overline{\mathcal{L}})}\right| \geq \Omega(\varepsilon^2) \cdot \frac{C_2^2}{\hat{D}(\overline{\mathcal{L}})}$ **then return reject**.
 - 12: **else return accept**.
-

Our first claim, which shows that most of the choices of the uniformly random subset of coordinates J to inspect are good, is implicit in the proof of Goldreich and Ron [GR23b, Claim 2.2.2].

Claim 3.3. *With probability $1 - o(1) - \frac{1}{100}$ over the choice of J and randomness over the sampling of $X \sim D$, with up to s samples, X_J will have the same empirical histogram as X' .*

Similarly, by a direct adaptation from the proof of Goldreich and Ron [GR23b, Claim 2.2.1], we obtain the following **Lemma 3.4** (we defer its proof to **Section A**):

Lemma 3.4 (D and D' are close on $\mathcal{S}$). *Given a sequence of good strings $(w^{(1)}, w^{(2)}, \dots, w^{(s')})$, a random projection J (and thus $X' \sim D'$), along with a subset $\mathcal{S}' \subseteq \Omega = \{0, 1\}^n$, one can apply $f_{J, \vec{w}}^{-1}$ on every $x' \in \mathcal{S}'$ to find $\mathcal{S}$ (see (2)). Moreover, we have that with probability $1 - o(1)$ over the choice of J , $\text{d}_{\text{EM}}(D_{\mathcal{S}}, D'_{\mathcal{S}'}) \leq \frac{\varepsilon}{10}$.*

For our argument, we will assume that we have sample access to D' .⁶ First, let us define a set that would contain most of the collisions of X' , parameterized by some threshold τ . Abusing the notation, we set the default $\mathcal{S}'$ without any parameters below:

$$\mathcal{S}'_{1/\tau} := \{x \in \{0, 1\}^n \mid D'(x) \geq 1/\tau\}; \mathcal{S}' := \mathcal{S}'_{1/\tau} \text{ and } \mathcal{S} = \{f_{J, \vec{w}}^{-1}(x) \in \{0, 1\}^n \mid x \in \mathcal{S}'\}. \quad (2)$$

Mapping elements between D and D' . Recall that in applying **Algorithm 6**, we will fully query a subset of samples that are unlikely to cause many collisions, i.e., the set of elements with relatively

⁶Technically, to D' up to a permutation of the domain elements: since we are only querying an ℓ -bit subset of coordinates, the D' in question is only defined up to a permutation (which leaves the ℓ -dimensional subcube queried invariant). This does not affect the argument, however, as parameterized uniformity is a label-invariant property.

Algorithm 3 FINDHEAVYELEMENTS(D, J, τ)

Require: Sample access to D , index set $J \subseteq [n]$, threshold τ

Ensure: Find $\mathcal{L}$: elements whose restriction on J has empirical weight $\geq \frac{3}{2}\frac{1}{\tau}$; It takes $O(\tau \log \tau)$ samples and $O(\tau \log \tau \cdot |J|)$ queries.

- 1: $s_1, s_2 \leftarrow O(\tau \log \tau)$
▷ Every heavy element will appear in $\mathcal{S}_1$, w.h.p.
 - 2: Draw multisets $\mathcal{S}_1 = \{x^{(i)}\}_{i \in [s_1]}$, $\mathcal{S}_2 = \{y^{(i)}\}_{i \in [s_2]}$ from D
 - 3: **for all** $i \in \{1, \dots, s_1\}$ **do**
 - 4: $\hat{D}'(x^{(i)}) \leftarrow \frac{1}{s_2} \sum_{j=1}^{s_2} \mathbf{1}[y^{(j)}|_J = x^{(i)}|_J]$
 - 5: $\mathcal{L} \leftarrow \emptyset$
▷ Use the empirical weight of $\mathcal{S}_2$ to confirm that it is indeed heavy
 - 6: **for all** $i \in \{1, \dots, s_1\}$ **do**
 - 7: **if** $\hat{D}'(x^{(i)}|_J) \geq \frac{3}{2} \cdot \frac{1}{\tau}$ **then**
 - 8: $\mathcal{L} \leftarrow \mathcal{L} \cup \{x^{(i)}\}$
 - 9: **return** $\mathcal{L}$
-

Algorithm 4 SANITYCHECKNORMS($D, J, \mathcal{L}, \hat{D}(\mathcal{L})$)

Require: Sample access to D , index set $J \subseteq [n]$, and the empirical large mass set $\mathcal{L}$

Ensure: $\|D_{\mathcal{L}}\|_2^2$ norm is at least $\frac{D(\mathcal{L})}{1600m}$, otherwise reject; ℓ_3 norm is at least $\frac{D(\mathcal{L})}{1600m^2}$, otherwise reject.

This procedure will cost $O(\sqrt{\frac{m}{\varepsilon}} + \frac{m^{2/3}}{\varepsilon^{1/3}})$ samples and $O\left(\frac{m}{\varepsilon} \cdot \|D'_{\mathcal{L}}\|_2^2 + \frac{m^2}{\varepsilon} \cdot \|D'_{\mathcal{L}}\|_3^3\right)$ queries.

- 1: $s_1 \leftarrow O\left(\sqrt{\frac{m}{D(\mathcal{L})}}\right) = O\left(\sqrt{\frac{m}{\varepsilon}}\right)$, $s_2 \leftarrow O\left(\frac{m^{2/3}}{D(\mathcal{L})^{1/3}}\right) = O\left(\frac{m^{2/3}}{\varepsilon^{1/3}}\right)$
▷ Try to estimate ℓ_2 norm over the set $\mathcal{L}$, and reject early if ℓ_2 norm is too small.
 - 2: $\mathcal{S}_1 \leftarrow \text{DRAWFILTERED}(s_1, \mathcal{L}, J)$; $C_1 \leftarrow \text{REALNORMEST}(\mathcal{S}_1, J, 2)$
 - 3: **if** $s_1^2 \cdot C_1 < 1$, **then return reject**
▷ Try to estimate ℓ_3 norm over the set $\mathcal{L}$, and reject early if ℓ_2 norm is too small.
 - 4: $\mathcal{S}_2 \leftarrow \text{DRAWFILTERED}(s_2, \mathcal{L}, J)$; $C_2 \leftarrow \text{REALNORMEST}(\mathcal{S}_2, J, 3)$
 - 5: **if** $s_2^3 \cdot C_2 < 1$, **then return reject**.
-

low mass. So when we choose a set of these “not much queried” samples from the original multiset of samples of D and perform additional queries (we query the full n bits in these samples), the illusion that these are samples from D' will be broken. One observation is that each $x' \in \text{Supp}(D')$ will correspond to one or multiple $x \in \text{Supp}(D)$, i.e., we can reverse the mapping defined in (1) via its preimage $\{y : (x', y) = f_{J, \vec{w}}(y)\}$. Therefore, if we have a partition $\mathcal{S}'$ and $\bar{\mathcal{S}}'$ for X' , then it will correspond to a partition of $\mathcal{S}$ and $\bar{\mathcal{S}}$ for X (here, by correspondence, we mean that no $x' \in \mathcal{S}'$ will be mapped to some $x \in \bar{\mathcal{S}}$).

Here one can think of the mapping function $f_{J, \vec{w}}: \{0, 1\}^n \rightarrow \{0, 1\}^n \times \{0, 1\}^n$ in (1): it maps one input to two outputs (see Fig. 1 for an illustration). i) For the first output, if its input x 's restriction to J matches with any element in $\vec{w}$, then it will be mapped to the corresponding $w^{(i)}$ in $\vec{w}$, otherwise we simply output x (this is exact the procedure behind (1), and it reflects the $\tilde{O}(1/\varepsilon)$ view). ii) For the second output, it maps x to itself, with the caveat that this output is *hidden* to our tester (this remains hidden until we perform additional queries on x).⁷

Our next claim shows one can efficiently identify (a good proxy for) $\mathcal{S}'$, the set of heavy elements:

⁷One can easily implement $f_{J, \vec{w}}^{-1}$ by mapping the second hidden output back to itself.

Algorithm 5 DRAWFILTERED($\mathcal{S}, \mathcal{L}, J$)

Require: Multiset $\mathcal{S}$ (samples), sets $\mathcal{L} \subseteq \{0, 1\}^n$ (uninterested region) and $J \subseteq [n]$ (query positions)

Ensure: The multiset obtained from $\mathcal{S}$ by removing all strings consistent (on J) with one string in $\mathcal{L}$

▷ For every z , we will query its J positions and check against strings in $\mathcal{L}$.

```
1: for all  $z \in \mathcal{S}$  do
2:   if  $z_J \in \{x_J : x \in \mathcal{L}\}$  then
3:      $\mathcal{S} \leftarrow \mathcal{S} \setminus \{z\}$ 
4: return  $\mathcal{S}$ 
```

Algorithm 6 REALNORMEST($\mathcal{S}, J, k$)

Require: A multiset $\mathcal{S}$ (samples), J previously queried positions, integer $k \in \{2, 3\}$

Ensure: Return an estimate of $\|D_{\bar{\mathcal{L}}}\|_k^k$, ℓ_k norm of the small mass region, by querying $O\left(\binom{|\mathcal{S}|}{k} \cdot \|D'_{\bar{\mathcal{L}}}\|_k^k\right)$ bits overall.

▷ The information of $\bar{\mathcal{L}}$ is implicitly passed through $\mathcal{S}$, as we have previously filtered samples before running this estimation procedure.

```
1:  $H' \leftarrow \text{HashTable}(\{0, 1\}^\ell \rightarrow \text{poly}(s))$ 
2:  $H \leftarrow \text{HashTable}(\{0, 1\}^n \rightarrow \text{poly}(s))$ 
   ▷ Reusing queries in previous steps. So this step is query-free.
3: for all  $z \in \mathcal{S}$  do
4:    $H'(z_J) \leftarrow H'(z_J) + 1$ 
5:  $C \leftarrow 0$ 
6: for all  $z \in \mathcal{S}$  do
   ▷ Adaptive query of full bits.
7:   if  $H'(z_J) \geq k$  then
8:     Query all  $n$  bits of  $z$ 
9:      $H(z) \leftarrow H(z) + 1$ 
10: for all  $(e, v) \in H$  do
11:    $C \leftarrow C + \binom{v}{k}$ 
12: return  $C / \binom{|\mathcal{S}|}{k}$ 
```

Claim 3.5. *There is an algorithm (Algorithm 3) which takes $\tilde{O}(\tau)$ i.i.d. samples from D' and returns, with probability $1 - o(1)$, a set $\mathcal{L}'$ such that*

$$\mathcal{S}'_{3/\tau} \subseteq \mathcal{L}' \subseteq \mathcal{S}'_{1/\tau}.$$

Moreover, $\|D'_{\mathcal{L}'}\|_2^2 \leq 3D(\bar{\mathcal{L}}')/\tau$ and $\|D'_{\mathcal{L}'}\|_3^3 \leq 9D(\bar{\mathcal{L}}')/\tau^2$. Further, if $D'(\mathcal{S}'_{1/\tau}) \geq \Omega(\varepsilon)$ then $D'(\bar{\mathcal{L}}') \geq \Omega(\varepsilon)$.

Afterwards, we will draw a fresh set of $O(1/\varepsilon^4)$ samples, to get an empirical estimate $\hat{D}'(\mathcal{L})$ of the probability mass $D'(\mathcal{L})$. If $\hat{D}'(\mathcal{L}) \geq 1 - \frac{\varepsilon}{100}$, our algorithm will accept – as this implies that $D'(\bar{\mathcal{L}}) \leq O(\varepsilon)$, and thus, by Lemma 3.6 below that, D' is $O(\tau/m + \varepsilon)$ -close to m -uniform in EMD.⁸

Lemma 3.6 ($D'_{\mathcal{S}'}$ is close to m -grained). *Assume $m \geq \frac{\tau}{3}$. With probability $1 - o(1)$ over the choice of J , we can transport $W \geq D'(\mathcal{S}') - O(\frac{1}{m})$ mass of $D'_{\mathcal{S}'}$ into a (pseudo)-distribution Q that is m -grained, costing at most $O(\frac{\tau}{m})$. Moreover, Q 's mass is equal to W .*

⁸Recall that by Footnote 3, closeness to m -grained implies closeness to m -uniform in the parameter regime $1/m \ll \varepsilon$.

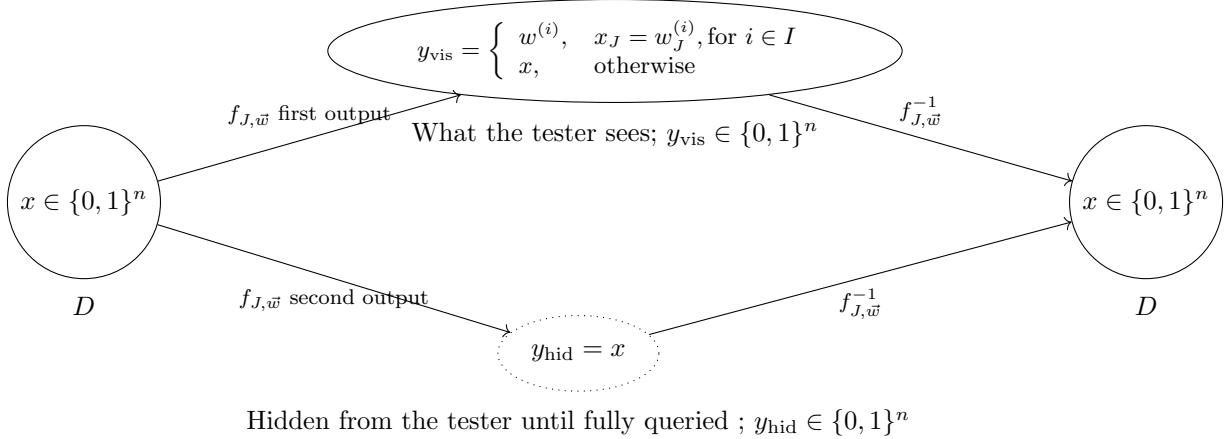


Figure 1: An illustration of the $f_{J,w}$: y_{vis} gives the clustered view (D') of the tester which queries only $\tilde{O}(1/\varepsilon)$ bits per sample (technically, it sees only the J -restriction, so $(y_{\text{vis}})_J \in \{0, 1\}^\ell$; but their empirical histogram will be the same) and y_{hid} captures the full view – what the tester might see if it later performs additional queries on x .

Lemma 3.7 (d_{EM} implies d_{TV} Farness). *Suppose $\frac{100}{m} \leq \varepsilon$. If we cannot transport at least $D(\bar{S}) - \frac{\varepsilon}{100}$ of the mass of $D_{\bar{S}}$ into any m -uniform with cost at most ε , then for any m -uniform distribution U , $\|D_{\bar{S}} - U_{\bar{S}}\|_1 \geq \frac{\varepsilon}{100}$.*

The next claim shows we can efficiently check the ℓ_2 and ℓ_3 norms, crucial to Steps 3 and 5 in Algorithm 4:

Fact 3.8 (Sanity checks). *Suppose $|\hat{D}(\bar{\mathcal{L}}) - D(\bar{\mathcal{L}})| \leq O(\varepsilon)$ and $\hat{D}(\bar{\mathcal{L}}) \geq \Omega(\varepsilon)$. Then we have the following: i) It takes $O(m^{1/2}/\varepsilon^{1/2})$ samples to distinguish between $\|D_{\bar{\mathcal{L}}}\|_2^2 < D(\bar{\mathcal{L}})(1600m)$ and $\|D_{\bar{\mathcal{L}}}\|_2^2 \geq D(\bar{\mathcal{L}})/m$. ii) It takes $O(m^{2/3}/\varepsilon^{1/3})$ samples to distinguish between $\|D_{\bar{\mathcal{L}}}\|_3^3 < D(\bar{\mathcal{L}})/(1600m^2)$ and $\|D_{\bar{\mathcal{L}}}\|_3^3 \geq D(\bar{\mathcal{L}})/m^2$.*

Building on Lemma A.1, we will also show the following, allowing us to get, in our algorithm a good estimate of the ℓ_2 norm of a pseudo-distribution:

Corollary 3.9. *Fix a set $S \subseteq \Omega$. Suppose $D(S) \geq \Omega(\varepsilon)$, $\|D_S\|_2^2 \geq \frac{1}{1600} \cdot \frac{D(S)}{m}$ and we are given $\hat{D}(S)$ such that $|\hat{D}(S) - D(S)| \leq O(\varepsilon)$. Then, with $O(\sqrt{m}/\varepsilon^{5/2})$ samples from D , one can return an estimate $\hat{D}$ such that $|\|\hat{D}_S\|_2^2 - \|D_S\|_2^2| \leq O(\varepsilon) \cdot \|D_S\|_2^2$. Moreover, $\left| \frac{\|\hat{D}_S\|_2^2}{\hat{D}(S)} - \frac{\|D_S\|_2^2}{D(S)} \right| \leq \frac{\varepsilon}{400}$.*

The next two results focus on the relation between ℓ_2 and ℓ_3 norms of pseudodistributions, relating them to closeness to uniformity, and providing an operational way to check this relation. While this was known for *bona fide* distributions, the extension to pseudo-distributions (and its “right” formulation) turns out to be quite involved, and is an important ingredient of our analysis:

Lemma 3.10. *Fix any set S such that $D(S) \geq \Omega(\varepsilon)$, $\|D_S\|_2^2 \geq \frac{1}{1600} \cdot \frac{D(S)}{m}$ and $\|D_S\|_3^3 \geq \frac{1}{1600} \cdot \frac{D(S)}{m^2}$. Assume $|\hat{D}(S) - D(S)| \leq O(\varepsilon^2 \cdot D(S))$. Then, given $s = O(m^{1/2}/\varepsilon^{9/2} + m^{2/3}/\varepsilon^{13/3})$ samples from D , we can distinguish w.h.p. between*

$$\|D_S\|_3^3 = \frac{\|D_S\|_2^4}{D(S)} \quad \text{and} \quad \|D_S\|_3^3 - \frac{\|D_S\|_2^4}{D(S)} \geq \Omega\left(\varepsilon^2 \frac{\|D_S\|_2^4}{D^2(S)}\right).$$

Claim 3.11. Fix any subset $S \subseteq \{0, 1\}^n$. Then: (i) If D is m -uniform, then

$$\frac{\|D_S\|_2^2}{D(S)} = \frac{1}{m} \text{ and } \|D_S\|_3^3 - \frac{\|D_S\|_2^4}{D(S)} = 0.$$

(ii) If $\min_U \min_{m\text{-uniform}} \|D_S - U_S\|_1 \geq \frac{\varepsilon}{100}$ and $\frac{\|D_S\|_2^2}{D(S)} \in [\frac{1 \pm \varepsilon/200}{m}]$, then

$$\|D_S\|_3^3 - \frac{\|D_S\|_2^4}{D(S)} \geq \Omega\left(\varepsilon^2 \frac{\|D_S\|_2^4}{(D(S))^2}\right).$$

Our final technical lemma will allow us to relate the distance of D from m -uniformity to that of $D_{\bar{S}}$, thus “localizing” enough distance in the pseudo-distribution considered by the algorithm:

Lemma 3.12. Suppose $\frac{100}{m} \leq \varepsilon$ and $100\frac{\tau}{m} \leq \varepsilon$. If D is ε -far from any m -uniform distribution in EMD, then, with probability $1 - o(1)$ over the choice of J , one cannot transport W' mass from $D_{\bar{S}}$ to any m -uniform pseudo-distribution $\tilde{Q}$ such that $Q(\bar{S}') = W'$ without a cost at least $\varepsilon/2$, where $W' = \frac{k}{m} \geq D(\bar{S}) - \varepsilon/100$ for some integer k .

With all these building blocks in hand, we are ready to prove our main theorem:

Proof of Theorem 3.1. Throughout the analysis, we will assume that the random query projection J is J -good for some $\vec{w} = (w^{(1)}, \dots, w^{(s')})$, which happens with probability $1 - o(1)$ and gives us a distribution D' . In view of Claim 3.3, we will further assume sample access (up to relabeling) to D' via X_J , where $X \sim D$.

Completeness. Assume D is m -uniform: note that, then, $\|D_{\bar{\mathcal{L}}}\|_2^2 = \frac{|\bar{\mathcal{L}}|}{m^2} = \frac{D(\bar{\mathcal{L}})}{m}$:

1. If $D(\bar{\mathcal{L}}) = D'(\bar{\mathcal{L}}') \leq O(\varepsilon)$, then Algorithm 2 will accept w.h.p. in Step 5.

2. If $D(\bar{\mathcal{L}}) = D'(\bar{\mathcal{L}}') > \Omega(\varepsilon)$, then we obtain an estimate of $D'(\bar{\mathcal{L}}')$ s.t. $\left|\hat{D}'(\bar{\mathcal{L}}') - D'(\bar{\mathcal{L}}')\right| \leq O(\varepsilon)$. By Fact 3.8, if $\|D_{\bar{\mathcal{L}}}\|_2^2 = \frac{D(\bar{\mathcal{L}})}{m}$, w.h.p. we will not reject in Step 3 in Algorithm 4. Furthermore, by Corollary 3.9, we have that $\left|\frac{\|\hat{D}_{\bar{\mathcal{L}}}\|_2^2}{\hat{D}(\bar{\mathcal{L}})} - \frac{\|D_{\bar{\mathcal{L}}}\|_2^2}{D(\bar{\mathcal{L}})}\right| \leq \frac{\varepsilon}{400}$ and thus $\left|\frac{\|\hat{D}_{\bar{\mathcal{L}}}\|_2^2}{\hat{D}(\bar{\mathcal{L}})} - \frac{1}{m}\right| \leq \frac{\varepsilon}{200}$, which means that w.h.p. we will not reject in Step 8 of Algorithm 2. For the final check in Step 11 of Algorithm 2, by Claim 3.11 item i) and Claim 3.10, w.h.p., we will not reject. By a union bound, over the events, we have that w.h.p., we will not reject and thus the tester will return accept.

Soundness. Suppose D is ε -far from m -uniform, which implies that D' is $\Omega(\varepsilon)$ -far from m -uniform by [GR23b][Claim 2.2.1]. By Lemma 3.6, D' is $O\left(\frac{\tau}{m} + \frac{1}{m}\right)$ -close to m -uniform on S' . Suppose $D'(\bar{S}') \leq O(\varepsilon)$ for contradiction, then we will have that D' is $O\left(\frac{\tau}{m} + \frac{1}{m} + \varepsilon\right)$ -close to m -uniform in EMD, which contradicts the premise: thus, $D'(\bar{S}') \geq \Omega(\varepsilon)$.

By Claim 3.5, we will locate a set $\mathcal{L}'$ such that $S'_{3/\tau} \subseteq \mathcal{L}' \subseteq S'$; and we have that $D'(\bar{\mathcal{L}}') \geq D'(\bar{S}') \geq \Omega(\varepsilon)$. This implies that w.h.p. Algorithm 2 will not accept in Step 5.

After obtaining $\mathcal{L}'$, the algorithm fully queries (all n bits of) the samples that fall in $\bar{\mathcal{L}}'$. This allows it to compute the corresponding set $\bar{\mathcal{L}} = \{f_{J, \vec{w}}^{-1}(x) \mid x \in \bar{\mathcal{L}}'\}$. Since $\bar{\mathcal{L}}' \supseteq \bar{S}'$, we have that $\bar{\mathcal{L}} \supseteq \bar{S}$. Now by Lemma 3.12 and Lemma 3.7 (note that as $\bar{\mathcal{L}}$ is a superset of $\bar{S}$, it is costlier to transport from $D_{\bar{\mathcal{L}}}$ to any m -uniform pseudo-distribution Q), for any m -uniform distribution U , $\|D_{\bar{\mathcal{L}}} - U_{\bar{\mathcal{L}}}\|_1 \geq \|D_{\bar{S}} - U_{\bar{S}}\|_1 \geq \frac{\varepsilon}{100}$. We then consider two cases.

1. If $\left| \frac{\|D_{\mathcal{L}}\|_2^2}{D(\mathcal{L})} - \frac{1}{m} \right| > \frac{\varepsilon}{200}$, then $\left| \frac{\|\hat{D}_{\mathcal{L}}\|_2^2}{\hat{D}(\mathcal{L})} - \frac{1}{m} \right| \geq \left| \frac{D_{\mathcal{L}}\|_2^2}{D(\mathcal{L})} - \frac{1}{m} \right| - \left| \frac{\hat{D}_{\mathcal{L}}\|_2^2}{\hat{D}(\mathcal{L})} - \frac{D_{\mathcal{L}}\|_2^2}{D(\mathcal{L})} \right| > \frac{\varepsilon}{400}$ by Corollary 3.9. Thus Algorithm 2 will reject w.h.p. in Step 8 (if it has not done so already at Step 3 in Algorithm 4).
2. If $\left| \frac{\|D_{\mathcal{L}}\|_2^2}{D(\mathcal{L})} - \frac{1}{m} \right| \leq \frac{\varepsilon}{200}$, then by Claim 3.11, we have that $\|D_{\mathcal{L}}\|_3^3 - \frac{\|D_{\mathcal{L}}\|_2^4}{D(\mathcal{L})} \geq \Omega\left(\varepsilon^2 \cdot \frac{\|D_{\mathcal{L}}\|_2^4}{D^2(\mathcal{L})}\right)$. In this case, Algorithm 2 in Step 11 will reject w.h.p. via Claim 3.10.

We conclude via a union bound that the algorithm will overall output **reject** w.h.p.

Sample complexity. We can compute the sample complexity by summing samples used in Algorithms 2 to 4, which would lead to a total sample complexity of

$$O\left(\tau \log \tau + \frac{m^{1/2}}{\varepsilon^{9/2}} + \frac{m^{2/3}}{\varepsilon^{13/3}}\right).$$

Query complexity. For every sample, we will always query on J , which costs

$$O\left(\left(\tau \log \tau + \frac{m^{1/2}}{\varepsilon^{9/2}} + \frac{m^{2/3}}{\varepsilon^{13/3}}\right) \cdot \frac{\log(mn/\varepsilon)}{\varepsilon}\right)$$

queries. On the other hand, when we compute the actual 2-norm and 3-norm estimates as done by Algorithm 6, we will pay $O(n)$ per sample that is involved in 2-wise and 3-wise collisions and in expectation there are $\binom{s'}{2} \cdot \|D'_{\mathcal{L}}\|_2^2$ and $\binom{s'}{3} \cdot \|D'_{\mathcal{L}}\|_3^3$ number of samples that fits the criteria (respectively), where s' denotes the number of samples we use in that batch. By Claim 3.5, we can upper bound the 2-norm and 3-norm on $\overline{\mathcal{L}'}$, i.e., the number of collisions would be limited even for the “clustered” version D' . Combining all these, we can compute an upper bound for the expected number of queries (before, by Markov inequality, converting it to a worst-case query complexity, at the cost of a small additional failure probability) for samples used for estimating 2-wise collisions:

$$O\left(\left(\binom{s_4}{2} + \binom{s_6}{2} + \left(\frac{m^{1/2}}{\varepsilon^{9/2}}\right)^2\right) \cdot \|D_{\mathcal{L}}\|_2^2 \cdot n\right) \leq O\left(\frac{m}{\varepsilon^9} \cdot \frac{n}{\tau}\right).$$

Similarly, for 3-wise collisions, the number of queries is

$$O\left(\left(\binom{s_5}{2} + \left(\frac{m^{2/3}}{\varepsilon^{13/3}}\right)^3\right) \cdot \|D_{\mathcal{L}}\|_3^3 \cdot n\right) \leq O\left(\frac{m^2}{\varepsilon^{13}} \cdot \frac{n}{\tau^2}\right).$$

Combining all, we have that the query complexity is bounded by

$$O\left(\left(\tau \log \tau + \frac{m^{1/2}}{\varepsilon^{9/2}} + \frac{m^{2/3}}{\varepsilon^{13/3}}\right) \cdot \frac{\log(mn/\varepsilon)}{\varepsilon} + \frac{m^2}{\varepsilon^{13}} \cdot \frac{n}{\tau^2}\right),$$

where we restrict $\tau \leq \tilde{O}(m)$ (otherwise the stated sample and query complexities are worse than that of [GR23b, Corollary 2.3]), which implies $\frac{m^2}{\varepsilon^{13}} \cdot \frac{n}{\tau^2} \geq \frac{m}{\varepsilon^9} \cdot \frac{n}{\tau}$. $\square$

4 A non-adaptive lower bound for m -uniformity testing

In this section, we prove the following theorem, which shows that non-adaptive testers for m -uniformity either must take many samples, or make many queries—and, in light of Corollary 3.2 establishes a separation between adaptive and non-adaptive testers:

Theorem 4.1 (Non-adaptive lower bound). *Let $m, n \geq 1$ be sufficiently large integers with $m \leq n^{O(1)}$. Any non-adaptive ε_0 -tester for m -uniformity in the DoHO model must either have sample complexity $s \geq c \frac{m}{\log m}$ or query complexity $q \geq c' \frac{\sqrt{mn}}{\log^2 m}$, where $c, c', \varepsilon_0 > 0$ are absolute constants.*

Proof. By a standard application of (the easy direction of) Yao’s minimax principle, it is enough to design two distributions $\mathcal{D}_{\text{yes}}, \mathcal{D}_{\text{no}}$ over instances such that (1) $\mathcal{D}_{\text{yes}}$ is supported on m -uniform distributions, (2) $\mathcal{D}_{\text{no}}$ is supported on distributions ε_0 -far from m -uniform (in EMD distance), and (3) any deterministic non-adaptive tester which takes $s = o\left(\frac{m}{\log m}\right)$ samples must make $\Omega(\sqrt{mn})$ queries. As discussed in [Section 1.1](#), the key idea of our proof is to actually define *three* distributions of instances, $\mathcal{D}_{\text{yes}}, \mathcal{D}_{\text{no}}$, and $\mathcal{D}_{\text{maybe}}$, such that

1. $\mathcal{D}_{\text{yes}}$ is supported on m -uniform distributions;
2. $\mathcal{D}_{\text{no}}$ is supported on distributions ε_0 -far from m -uniform;
3. $\mathcal{D}_{\text{maybe}}$ is supported on *some* distributions, not guaranteed to be m -uniform; yet
4. $\mathcal{D}_{\text{no}}$ and $\mathcal{D}_{\text{maybe}}$ are hard to distinguish given $o\left(\frac{m}{\log m}\right)$ samples, and
5. $\mathcal{D}_{\text{yes}}$ and $\mathcal{D}_{\text{maybe}}$ are hard to distinguish given $o\left(\frac{\sqrt{mn}}{\log^2 m}\right)$ non-adaptive queries.

Assuming the above, fix any *honest-to-goodness* non-adaptive tester $\mathcal{A}$ for m -uniformity. By definition, $\mathcal{A}$ needs to output `accept` on (distributions from) $\mathcal{D}_{\text{yes}}$ with probability at least $2/3$, and `reject` on (distributions from) $\mathcal{D}_{\text{no}}$ with probability at least $2/3$; it is, of course, completely off the hook on (distributions from) $\mathcal{D}_{\text{maybe}}$, since they do not necessarily satisfy either the uniformity nor fairness promise. Yet $\mathcal{A}$ does need to output *something* on these as well: denote by p the probability it outputs `accept` when run on a distribution from $\mathcal{D}_{\text{maybe}}$: $p = \Pr_{D \sim \mathcal{D}_{\text{maybe}}}[\mathcal{A}^D = \text{accept}]$. Clearly, we must either have $p \geq 1/2$ or $p < 1/2$.

- If $p \geq 1/2$, then $\mathcal{A}$ is a valid distinguisher between $\mathcal{D}_{\text{no}}$ and $\mathcal{D}_{\text{maybe}}$: and so it must have sample complexity $s = \Omega\left(\frac{m}{\log m}\right)$;
- If $p < 1/2$, then $\mathcal{A}$ is a valid distinguisher between $\mathcal{D}_{\text{yes}}$ and $\mathcal{D}_{\text{maybe}}$: and so it must have query complexity $q = \Omega\left(\frac{\sqrt{mn}}{\log^2 m}\right)$;

which then establishes the conclusion of [Theorem 4.1](#). So our goal is to design distributions $\mathcal{D}_{\text{yes}}, \mathcal{D}_{\text{no}}$, and $\mathcal{D}_{\text{maybe}}$ over probability distributions satisfying the desired properties. To do so, we will rely on a result of Canonne, Sen, and Yang [[CSY25](#)] on testing m -grainedness, restated below:

Theorem 4.2 (Canonne, Sen, and Yang [[CSY25](#), Theorem 1.1]). *Let N be a sufficiently large integer, and $m = O(N)$. Then, there exists a constant $\varepsilon_0 \in (0, 1)$ such that ε_0 -testing (in the standard setting) m -grainedness of distributions over $[N]$ requires $\Omega\left(\frac{m}{\log m}\right)$ samples. Specifically, there exist two families $\mathcal{D}_1, \mathcal{D}_2$ of probability distributions over $[N]$ such that (1) distributions in $\mathcal{D}_1$ are m -grained, (2) distributions in $\mathcal{D}_2$ are ε_0 -far from m -grained in TV distance; and (3) any algorithm distinguishing a randomly chosen distribution from $\mathcal{D}_1$ from a randomly chosen one from $\mathcal{D}_2$ with probability of success at least .51 must take $c \cdot \frac{m}{\log m}$, where $c > 0$ is an absolute constant.*

(Moreover, as a technical point, by inspection of their proof one can check that probability distributions in $\mathcal{D}_2$ have support size at most $2m$, while these in $\mathcal{D}_1$ have by definition support size at most m . Further, $\|D_1\|_\infty, \|D_2\|_\infty \leq C' \cdot \frac{\log^2 m}{m}$ for every D_1, D_2 from $\mathcal{D}_1$ and $\mathcal{D}_2$, where $C' > 0$ is an absolute constant.) We will take $N = 2^n$ there, and embed the families $\mathcal{D}_1, \mathcal{D}_2$ into $\{0, 1\}^n$ in a way which “creates EMD distance” from TV distance. Specifically, choosing any good linear-distance

error-correcting code $\mathcal{C} \subseteq \{0, 1\}^n$ ⁹, we fix a set $\Delta = \{x^{(1)}, x^{(2)}, \dots, x^{(2m)}\} \subseteq \mathcal{C}$ of $2m$ points at constant pairwise distance: $d_H(x^{(i)}, x^{(j)}) \geq 1/4$ for all $1 \leq i < j \leq 2m$.¹⁰ Then, our distribution $\mathcal{D}_{\text{no}}$ is defined by drawing a distribution D_2 (over $[N]$) from $\mathcal{D}_2$, and mapping its support to Δ ; while our distribution $\mathcal{D}_{\text{maybe}}$ is similarly obtained by drawing a distribution D_1 (over $[N]$) from $\mathcal{D}_1$, and mapping its support to Δ . (Nearly) finally, $\mathcal{D}_{\text{yes}}$ is defined similarly, but with an extra step:

1. draw an m -grained distribution D_0 (over $[N]$) from $\mathcal{D}_1$;
2. map its support to Δ , obtaining an m -grained distribution D'_0 over Δ , such that, for each $1 \leq i \leq 2m$, $D'_0(x^{(i)}) = L_i/m$ for some integer $L_i \in \{0, 1, \dots, m\}$;
3. for each $1 \leq i \leq 2m$, define D''_0 by dividing the probability mass of $x^{(i)}$ equally over L_i distinct points $x^{(i,1)}, \dots, x^{(i,L_i)}$, all within small Hamming distance of $x^{(i)}$: that is,

$$D''_0(x^{(i,1)}) = \dots = D''_0(x^{(i,L_i)}) = \frac{1}{m}$$

and $d_H(x^{(i)}, x^{(i,\ell)}) \leq O(1)/n$ for all $1 \leq i \leq 2m$ and $1 \leq \ell \leq L_i$ (this is possible as $m \leq n^{O(1)}$).

Finally, in both the **yes** and **maybe** cases we apply a uniformly random permutation to each cluster, that is, for every $1 \leq i \leq 2m$, we independently apply a uniformly random partition to $\{x^{(i)}, x^{(i,1)}, \dots, x^{(i,L_i)}\}$ (this shuffling is so that no algorithm has *a priori* knowledge of which point is the cluster center). It is easy to see that, by the above, (1) $\mathcal{D}_{\text{no}}$ is supported on distributions that are $\varepsilon_0/4$ -far from m -grained in EMD (for the value of ε_0 from [Theorem 4.2](#)) and thus *a fortiori* $\varepsilon_0/4$ -far from m -uniform; (2) $\mathcal{D}_{\text{yes}}$ is supported on m -uniform distributions, and (3) $\mathcal{D}_{\text{maybe}}$ is supported on m -uniform distributions, *and inherits the lower bound from Theorem 4.2*: it takes $c \cdot \frac{m}{\log m}$ to distinguish $\mathcal{D}_{\text{maybe}}$ from $\mathcal{D}_{\text{no}}$ with success probability at least .51.

This gives [Items 1 to 4](#) of our above desiderata: to conclude, it only remains to establish [Item 5](#).

Item 5: distinguishing $\mathcal{D}_{\text{yes}}$ from $\mathcal{D}_{\text{maybe}}$. First, we fix some notation: we refer to the points in $\Delta = \{x^{(1)}, x^{(2)}, \dots, x^{(2m)}\} \subseteq \{0, 1\}^n$ as *cluster centers*, and, for a given $1 \leq i \leq 2m$, to the (up to $m+1$) points $x^{(i)}, x^{(i,1)}, \dots, x^{(i,L_i)}$ as a *cluster C_i* . We will also implicitly use a coupling argument when considering $D_{\text{yes}} \sim \mathcal{D}_{\text{yes}}$ and $D_{\text{maybe}} \sim \mathcal{D}_{\text{maybe}}$, specifically drawing $D_{\text{maybe}} \sim \mathcal{D}_{\text{maybe}}$ and constructing (as defined above) the corresponding D_{yes} , so that our instances are generated by pairs $(D_{\text{yes}}, D_{\text{maybe}})$. This allows us to argue about the overall probability mass of each cluster under each such D_{yes} and D_{maybe} : i.e., $D_{\text{yes}}(C_i) = D_{\text{maybe}}(C_i)$ for all $1 \leq i \leq 2m$.

Fix any deterministic non-adaptive algorithm $\mathcal{A}$ with sample complexity s and query complexity $q \leq c' \frac{\sqrt{mn}}{\log^2 m}$. Observe that we can assume $s \geq C \cdot \sqrt{m}$ for some (sufficiently small) absolute constant $C > 0$. Indeed, otherwise $\mathcal{A}$ will, with probability at least 9/10 over the s samples it obtains from the distribution D , “hit” each of the $2m$ clusters at most once, regardless of whether $D = D_{\text{yes}}$ or $D = D_{\text{maybe}}$. But since by construction each cluster has exactly the same total probability under both and the uniformly random permutations implies that elements within a given cluster all appear interchangeable, the algorithm is unable to distinguish between the two.

Define $r = s/(C\sqrt{m}) \geq 1$. The average number of queries $\mathcal{A}$ makes to each of the s samples is then $q/s \leq c' \cdot (\sqrt{mn}/\log^2 m)/(Cr\sqrt{m}) \leq n/(100r \log^2 m)$, the last inequality holds for a suitable choice of $c' > 0$. An averaging argument then shows that, for at least $s(1 - \frac{1}{10r \log m})$ of the samples it obtains, $\mathcal{A}$ queries at most $\frac{n}{10 \log m}$ coordinates: denote the subset of these “little-queried samples”

⁹Existence of such codes is well known in the literature: see, e.g., [\[BSGHSV06\]](#).

¹⁰This is always possible as long as, say, $m \leq 2^{0.9n}$.

$L \subseteq [s]$, and the “heavily-queried” ones $H = [s] \setminus L$.¹¹ For the samples indexed by L , by a union bound over the $s \leq O(m)$ samples and given the $O(1/n)$ Hamming distance between same-cluster points, we get that, with probability at least 9/10, $\mathcal{A}$ is not able to detect whether samples falling within a same cluster are the same point (as in D_{yes}) or distinct (as in D_{maybe}).

The heavily-queried set of samples H *could* allow $\mathcal{A}$ to distinguish between D_{yes} and D_{maybe} , if they included a collision (i.e., if at least two samples from H fell in the same cluster). But the expected number of cluster collisions within H can be bounded as

$$\binom{|H|}{2} \frac{L_{\max}}{m} \leq \frac{s^2}{200r^2 \log^2 m} \cdot \frac{C' \log^2 m}{m} = \frac{C^2 \cdot C'}{200} \leq \frac{1}{10},$$

where $L_{\max} = \max_{1 \leq i \leq 2m} L_i$ (with the notation above), which is at most $C' \log^2 m$ by the discussion after [Theorem 4.2](#); and the last inequality follows from a suitable choice of the small constant $C > 0$. By Markov’s inequality, this in turn implies that the probability to observe *any* cluster collision among the samples in H is at most 1/10. However, by a union bound, with probability at least 8/10 we have that (1) samples from L reveal no information, as even collisions within clusters are not detected; and (2) samples from H involve no collisions within clusters, and so also reveal no information. This implies that $\mathcal{A}$ cannot distinguish between D_{yes} and D_{maybe} , as claimed. This concludes the proof of [Item 5](#) and, by the foregoing discussion, of [Theorem 4.1](#). $\square$

5 An adaptive lower bound for m -uniformity testing

The goal of this section is to prove the following adaptive lower bound for m -uniformity testing:

Theorem 5.1 (Full adaptive lower bound). *Consider any full adaptive algorithm $\mathcal{A}$ for m -uniformity testing that takes s samples and performs q queries, and fix $\varepsilon = \Omega(1)$. If $s \leq O\left(\frac{m^{3/4}}{\text{polylog } m}\right)$, then $\mathcal{A}$ must perform $q = \Omega\left(\frac{mn}{s \text{ polylog } m}\right)$ queries.*

For our lower bound, we will restrict our attention to *cluster distributions* (i.e., the support of the distribution can be clustered into m Hamming balls of diameter $O(\log m/n)$, where balls are at pairwise distance $\varepsilon = \Omega(1)$ from each other). Furthermore, we will first focus on an adaptive model that is more restrictive, which we term the *weakly-adaptive* setting, then later we will generalize it to adaptive algorithms without any restriction. An algorithm is given s samples from a distribution D , and, for each sample, the algorithm is provided “for free” with some extra information, namely the index i of the cluster the sample belongs to.¹² Based on this cluster information, the algorithm can build a frequency histogram for the s samples. We restrict the adaptiveness of the algorithm as follows: the algorithm $\mathcal{A}$ must decide which strings (and how many bits) to query, upon seeing this frequency histogram of the cluster distribution. This is captured by the following definition:

Definition 5.2 (weakly-adaptive for cluster distributions). An algorithm is called a *weakly-adaptive* algorithm if it fixes its query set after seeing the frequency histogram $(Y_1, \dots, Y_m)$ of cluster distribution given by the cluster oracle, where Y_i is the number of samples from cluster i .

¹¹It is worth highlighting that this is where the assumption that $\mathcal{A}$ is non-adaptive comes in play, as L and H then are fully specified by the index of the samples themselves. An adaptive algorithm, in contrast, could define L and H “on-the-fly” based on the values of earlier samples: and indeed, it is not hard to see that an adaptive algorithm can distinguish D_{yes} from D_{maybe} with only $O(\sqrt{m} + n)$ queries.

¹²This is to capture the fact, upon $\tilde{O}\left(\frac{1}{\varepsilon}\right)$ queries to each sample (and this is constant if ε is constant), the algorithm will know whether each pair of sample strings are far or not.

Note that this is exactly the adaptiveness used to design our upper bound algorithm. Also note that we restrict the number of clusters in the cluster distribution to be m , and thus one can think of the frequency histogram as being generated by a discrete distribution over $[m]$.

Now we are ready to state our lower bound theorem for the weakly adaptive setting.

Lemma 5.3 (Lower bound for weakly adaptive setting). *Consider any weakly adaptive algorithm $\mathcal{A}$ for m -uniformity testing that takes s samples and performs q queries and $\varepsilon = \Omega(1)$. If $s \leq O\left(\frac{m^{3/4}}{\text{polylog } m}\right)$, then $\mathcal{A}$ must perform $q = \Omega\left(\frac{mn}{s \text{polylog } m}\right)$ queries.*

Assume $m \geq n$, the regime of interest for this lower bound. Let $s = o(m/\log m)$ (otherwise, we have an upper bound), and $s = \Omega(m^{2/3})$, as otherwise we already have a lower bound (regardless of the number of queries q).

We first construct a new pair of hard instances variables $((X_i)_{i \in [m]}, (X'_i)_{i \in [m]})$ obtained by suitable modification to the m -grainedness hard instances $((U_i)_{i \in [m]}, (V_i)_{i \in [m]})$ over $[m]$: for $i \in [m]$, recall that

$$U_i \in \begin{cases} \{1/m\} & \text{w.p. } \Omega(1) \\ \{0\} & \text{w.p. } \Omega(1) \\ \left\{\frac{2}{m}, \frac{3}{m}, \dots, \frac{O(\log^2 m)}{m}\right\} & \text{w.p. } \Omega(1) \end{cases}, V_i \in \begin{cases} \{1/(2m)\} & \text{w.p. } \Omega(1) \\ \left\{\frac{2}{m}, \frac{3}{m}, \dots, \frac{O(\log^2 m)}{m}\right\} & \text{w.p. } \Omega(1) \end{cases}.$$

Then, let

$$X_i = \begin{cases} U_i & \text{w.p. } 2/3 \\ 1/t & \text{w.p. } t/(3m) \end{cases}, \quad X'_i = \begin{cases} V_i & \text{w.p. } 2/3 \\ 1/t & \text{w.p. } t/(3m) \end{cases}$$

where $m \geq t \geq 1$ is an integer dividing m , which we will choose momentarily (see discussion after the proof of [Lemma 5.7](#)). We assume that $s = O(t)$ throughout.

We first show that $((X_i)_{i \in [m]}, (X'_i)_{i \in [m]})$ is a hard instance for m -grainedness testing. We follow the moment-matching technique¹³, similar to [\[CSY25\]](#).

Claim 5.4. *$((X_i)_{i \in [m]}, (X'_i)_{i \in [m]})$ remains a hard instance of testing m -grainedness and has lower bound of $\Omega\left(\frac{m}{\log m}\right)$ samples.*

Proof. There are two things to check: $\mathbb{E}[X_i], \mathbb{E}[X'_i] \leq O\left(\frac{1}{m}\right)$, and the first L moments match like they did for U_i, V_i , i.e., $\mathbb{E}[X_i^\ell] = \mathbb{E}[(X'_i)^\ell]$ for $1 \leq \ell \leq L$. The first item can be verified by noting that $\mathbb{E}[X_i] = \frac{2}{3} \cdot \mathbb{E}[U_i] + \frac{1}{t} \cdot \frac{t}{3m} \leq O\left(\frac{1}{m}\right)$. As for the second item, we have that, for every $1 \leq \ell \leq L$

$$\mathbb{E}[X_i^\ell] = \frac{2}{3} \cdot \mathbb{E}[U_i^\ell] + \frac{t}{3m} \cdot \left(\frac{1}{t}\right)^\ell = \frac{2}{3} \cdot \mathbb{E}[V_i^\ell] + \frac{t}{3m} \cdot \left(\frac{1}{t}\right)^\ell = \mathbb{E}[(X'_i)^\ell].$$

□

Similar to our non-adaptive lower bound, we have a distribution over triples of distributions $(D_{\text{yes}}, D_{\text{maybe}}, D_{\text{no}})$, each defined over m well-separated clusters in $\{0, 1\}^n$ obtained from a modified hard instance $((X_i)_{i \in [m]}, (X'_i)_{i \in [m]})$ for m -grainedness on $[m]$ as follows:

1. Pick $x_1, \dots, x_m$ uniformly at random from $\{0, 1\}^n$; with high probability all their pairwise (relative) Hamming distances are $\Theta(1)$;

¹³The moment-matching technique roughly says that if the moments of two instances (one yes-instance and one no-instance) match up to a suitable threshold, the instances are hard to distinguish w.r.t. a number of i.i.d. samples, which depends on the threshold.

2. D_{yes} and D_{maybe} share the same cluster distribution over $[m]$, with cluster probabilities defined by $(X_i)_{i \in [m]}$.
3. D_{no} has cluster distribution defined by $(X'_i)_{i \in [m]}$.
4. For every cluster i such that $X_i = \frac{1}{t}$ (resp. $X'_i = \frac{1}{t}$), divide the total probability mass $\frac{1}{t}$ under D_{yes} and D_{maybe} (resp. D_{no}) of the cluster over $\frac{m}{t}$ (uniformly random) points within radius $O(\log(\frac{m}{t})/n)$ of x_i .
5. For every cluster i such that $X_i \neq \frac{1}{t}$, set $D_{\text{maybe}}(x_i) = X_i$ (no dividing needed), and
 - If $X_i = \frac{1}{m}$, set $D_{\text{yes}}(x_i) = \frac{1}{m}$ (no dividing needed);
 - If $X_i = \frac{k}{m}$ for some $2 \leq k \leq O(\log^2 m)$, divide equally under D_{yes} the probability mass X_i of the cluster over k (uniformly random) points within radius $O(\log \log m/n)$ of x_i .
6. For every cluster i such that $X'_i \neq \frac{1}{t}$, set $D_{\text{no}}(x_i) = X'_i$ (no dividing needed).

It is easy to verify that these three distributions satisfy the following:

- D_{yes} is m -uniform;
- D_{maybe} is not m -uniform but m -grained;
- D_{no} is far from m -uniform.

Moreover, we note that $\Omega(\frac{m}{\log m})$ samples are still required to distinguish between D_{maybe} and D_{no} : the changes made to $1/t$ masses are the same on all three distributions (D_{yes} , D_{maybe} and D_{no}) and we did not modify any other masses; their moments remain matched following the proof idea of [claim 5.4](#). Since $\mathcal{A}$ takes $s = o(m/\log m)$ samples, it thus cannot distinguish between the two with constant advantage; thus, by the same argument as in [Theorem 4.1](#), it must be the case that it distinguishes between D_{yes} and D_{maybe} .

Outline of the proof. We thus turn to the distinguishability between D_{yes} and D_{maybe} : by our construction, the only way to tell the difference between D_{yes} and D_{maybe} is to first find a cluster collision (i.e., $Y_i \geq 2$) in a cluster with probability mass in $\{\frac{2}{m}, \dots, \frac{\log^2 m}{m}\}$ – as this is the only segment where they differ. We will refer to them as the *secret clusters*.

The outline of the argument is as follows: after looking over the frequency histogram of the cluster distribution, the algorithm is given some number T of chances to retrieve an index i such that $X_i \in \{k/m\}_{k \in \{2, \dots, \log^2 m\}}$. One can see this as a list-decoding task, where the algorithm must return a small list of candidates containing at least one witnessing a “hidden secret”. We will argue that it can only “win” if $X_i \in \{k/m\}_{k \in \{2, \dots, \log^2 m\}}$ is among the T choices it selected, *and* it queries each of the T corresponding samples in at least $\Omega(\frac{n}{\text{polylog}(m)})$ coordinates.

Indeed, suppose the algorithm only spends $o(\frac{n}{\text{polylog}(m)})$ queries on the potential candidate strings cluster, then with high probability (via a multiplicative Chernoff bound and a union bound), it cannot find the “hidden secret” (here, the coordinates witnessing whether the string is x_i or a nearby point in the Hamming ball) even if there is a secret to find; and thus cannot distinguish between the D_{yes} and D_{maybe} . On the other hand, we will show later that unless $T = \Omega(\text{poly}(m))$ (by setting $s = \text{poly}(m)$ and $t = \text{poly}(m)$), the algorithm cannot find a secret cluster. While the algorithm can choose not to query some candidate strings cluster, that would imply that those strings clusters can

be removed from the candidate set, and thus suggesting that strictly less than T candidates suffice for the list-decoding task, contradicting the lower bound we will establish later.

Therefore, it has to spend $\Omega\left(\frac{n}{\text{polylog } m}\right)$ queries per sample, and for T samples, this gives a query bound of $\Omega\left(\frac{n}{\text{polylog } m} \cdot T\right) = \Omega(\text{poly}(m) \cdot n)$.

Remark 5.5 (Weak and full adaptivity). *While this lower bound setting is not fully adaptive, it is “nearly there,” considering the invariant that we could maintain for any fully adaptive algorithm: suppose we have a lower bound on this weaker adaptive setting, we consider every bit that we query on each string from one cluster, they will be random strings generated independently of strings from other clusters. Thus, intuitively, information gathered from queries of one cluster will not help in identifying (secret) bits hidden in other clusters. Therefore, we obtain no further useful information (conditioned on the histogram already obtained) that can be used adaptively to detect them.*

Formalizing the argument. To simplify the argument, we will use the standard Poissonization technique, and assume the algorithm, instead of s samples, is given $\text{Poi}(s)$ samples. Below, we will look at the task of finding secret collisions given the frequency histogram. Under the Poissonized setting, the hard instances’ generation for the frequency histogram is a two-step process: first draw $(X_i)_{i \in [m]}$; then return $Y_i \sim \text{Poi}(s \cdot X_i)$ for $i \in [m]$.

Definition 5.6 (Two truncated Poissons). Let p_0 be the truncated Poisson $\text{Poi}(s/t)$ at count 2 and $p_{1,k}$ be the truncated Poisson $\text{Poi}(\frac{sk}{m})$ at count 2, i.e., $p_0(j) = \Pr[Y_i = j \mid Y_i \geq 2 \text{ and } Y_i \sim \text{Poi}(s/t)]$; further, $p_{1,k}(j) = \Pr[Y_i = j \mid Y_i \geq 2 \text{ and } Y_i \sim \text{Poi}(sk/m)]$ for some $k \in \{2, \dots, \log^2 m\}$. We will abbreviate $p_{1,k}$ as p_1 in the remainder of the section (as our proof will show that p_0 and $p_{1,k}$ are indistinguishable for any $k \in \{2, \dots, \log^2 m\}$).

The next lemma then bounds the number of clusters of each type in which at least 2 distinct samples fall (i.e., witnessing at least one collision among the s samples).

Lemma 5.7 (Much more p_0 than p_1). *Given $\text{Poi}(s)$ samples over the pseudo-distributions formed by $(X_i)_{i \in [m]}$, there are at least $\Omega(s^2/t)$ truncated Poissons from p_0 and at most $O(s^2 \text{polylog}(m)/m)$ truncated Poissons from p_1 , with probability at least 99/100.*

Proof. From the description, we note that with probability $t/(3m)$ the mass will be $1/t$. By the independence of $(X_i)_{i \in [m]}$ and an application of Chernoff bound, we have that w.h.p. $\Theta(t)$ out of the m positions will have mass $1/t$.

On the other hand, we also have that with probability $\Theta(1)$, the mass will be one in $\{\frac{2}{m}, \dots, \frac{k}{m}\}$. By independence over $(X_i)_{i \in [m]}$ and applying Chernoff bound, we have that w.h.p. $\Theta(m)$ out of the m positions will have mass from $\{\frac{2}{m}, \dots, \frac{k}{m}\}$.

Conditioned on the above event happening (otherwise we simply say that the algorithm wins by being lucky, which happens with small probability), we look at the expected number of random variables that have a collision, i.e., having $Y_i \geq 2$ where $Y_i \sim \text{Poi}(s \cdot X_i)$. Recall that we are in the Poissonized setting, so we look at the following.

Conditioned on $X_i = \frac{1}{t}$, we have that for $Y_i \sim \text{Poi}(s \cdot X_i)$; $\lambda = s/t$,

$$\Pr[Y_i = 2] = e^{-\lambda} \cdot \frac{\lambda^2}{2} \asymp (1 - o(1)) \cdot \frac{\lambda^2}{2} = \Theta\left(\frac{s^2}{t^2}\right).$$

And since there are at least $\Omega(t)$ elements with $1/t$ masses, the expected number of Poisson random variables with count at least 2 resulting from $1/t$ masses is at least

$$\sum_{i=1}^m \mathbb{E} [\mathbb{1}_{\{Y_i \geq 2 \text{ and } Y_i \sim \text{Poi}(s \cdot X_i)\}}] \geq \sum_{i=1}^{\Omega(t)} \mathbb{E} [\mathbb{1}_{\{Y_i \geq 2\}}] = \Omega(t)(1 - e^{-\lambda} \cdot (1 + \lambda)) = \Omega(t \cdot \lambda^2) = \Omega\left(\frac{s^2}{t}\right).$$

On the other hand, let $X_i = k/m$ for some $k \in \{2, \dots, O(\log^2 m)\}$. Let $\lambda = sk/m$. There are at most $O(m)$ elements with masses from $\left\{\frac{2}{m}, \dots, \frac{O(\log^2 m)}{m}\right\}$, the expected number of Poisson random variables with count at least 2, resulting from masses in $\left\{\frac{2}{m}, \dots, \frac{O(\log^2 m)}{m}\right\}$ is at most

$$\sum_{i=1}^m \mathbb{E} [\mathbb{1}_{\{Y_i \geq 2 \text{ and } Y_i \sim \text{Poi}(sX_i)\}}] \leq \sum_{i=1}^{O(m)} \mathbb{E} [\mathbb{1}_{\{Y_i \geq 2\}}] = O(m) \cdot (1 - e^{-\lambda} \cdot (1 + \lambda)) = O\left(\frac{s^2 \log^4 m}{m}\right).$$

By a Chernoff bound we conclude that w.h.p., there are at least $\Omega(s^2/t)$ Poissons generated from $\text{Poi}(s/t)$ and at most $O\left(\frac{s^2 \text{polylog}(m)}{m}\right)$ Poissons generated from $\left\{(\text{Poi}\left(\frac{sk}{m}\right))_{k \in \left\{\frac{2}{m}, \dots, \frac{\log^2 m}{m}\right\}}\right\}$. $\square$

We now define the list-decoding task alluded to in the overview, setting $K = \Omega(s^2/t)$, $\eta = O\left(\frac{t \text{polylog } m}{m}\right)$:

Definition 5.8 ((η, K)-List decodable index identification ((η, K)-LDII)). Given at least K samples from p_0 and at most $K \cdot \eta$ from p_1 (and applying a uniformly random permutation to them), as well as an integer T , the LDII task asks to return a list of at most T candidate indices such that at least one of the indices i is from p_1 .

We have the following guarantee about (η, K)-LDII.

Lemma 5.9. Assume that $1/\eta^3 \geq \Omega(K)$. No algorithm can solve the LDII task from [Definition 5.8](#) with probability at least $2/3$, unless $T = \Omega(\min(K\eta, 1/\eta))$.

Now we argue that no algorithm can find collisions resulted from the secret clusters

$$\left\{ \text{Poi}\left(\frac{sk}{m}\right)_{k \in \left\{\frac{2}{m}, \dots, \frac{\log^2 m}{m}\right\}} \right\}$$

with T trials and we do so via a reduction: the proof of [Lemma 5.9](#) follows from [Lemma 5.14](#) described below. We construct the following problem:

Definition 5.10 ((η, K)-List decodable index identification from sampling). The (η, K)-LDIIS problem is defined as follows: with probability η , draw a sample from p_1 , otherwise, draw a sample from p_0 . Suppose an algorithm is given K number of draws and needs to identify one draw that comes from p_1 . After looking at all samples, the algorithm can decide to keep a subset of the samples, and wins if, among these samples, at least one was drawn from p_1 . The complexity of the algorithm is the number T' of samples it keeps.

Remark 5.11. While the proof below is in the more restrictive setting where each decision is made only on the current sample, the algorithm “intuitively” cannot really use adaptivity to increase its chances: every sample is drawn independently from the same mixture of two truncated Poissons. So looking at other samples will not increase the distinguishing power, and thus cannot reduce the length of T' . We make this “intuition” formal in [Section B](#).

Lemma 5.12. Assume that $s = O(t)$. When distinguishing between p_0 and p_1 , let α be the Type I error (i.e., returning 1 when seeing a sample from p_0), and β be the Type II error (returning 0 when seeing a sample from p_1). Then, either $\alpha \geq \Omega(1)$ or $\alpha + \beta \geq 1$.

Proof. Let A be an algorithm for distinguishing the truncated Poisson distributions p_0 and p_1 as defined in [Definition 5.6](#). Denote by $S_{A,R}$ the accepting region, i.e., the algorithm A will return 1 if $Y \in S_{A,R}$, where R is the internal randomness of the algorithm. With this notation, we have $\alpha = p_0(S_{A,R})$ and $\beta = 1 - p_1(S_{A,R})$.

Suppose $S_{A,R}$ is not an empty set, otherwise $p_1(S_{A,R}) = 0$ (the algorithm does not accept anything, and so $\beta = 1$ (implying the conclusion of the lemma)). We show that if $2 \in S_{A,R}$ then $\alpha \geq \Omega(1)$, and otherwise $\alpha + \beta \geq 1$.

To do so, we first observe that, if $S_{A,R}$ to be $\{2\}$, i.e., the algorithm accepts if $X = 2$, then

$$\begin{aligned} \alpha &= \mathbb{E}_{p_0,R}[A(X, R) = 1] = p_0(S_{A,R}) = \Pr_{X \sim p_0}(X = 2) = \frac{e^{-\lambda_0} \cdot \frac{\lambda_0^2}{2}}{(1 - e^{-\lambda_0}) \cdot (1 + \lambda_0)} \\ &= \frac{\lambda_0^2/2}{e^{\lambda_0} - 1 - \lambda_0} = 1 - \Theta(\lambda_0) = \Omega(1), \end{aligned}$$

the last two equalities for $\lambda_0 = O(1)$, i.e., for $t = \Omega(s)$. Thus, if $2 \in S_{A,R}$, then the above is a lower bound on α , and so $\alpha = \Omega(1)$. Turning to the case when $2 \notin S_{A,R}$, observe that for $j > 2$ we have that $p_0(X = j) \geq p_1(X = j)$. This means that if $2 \notin S_{A,R}$ then $p_0(S_{A,R}) \geq p_1(S_{A,R})$, and since $\alpha = p_0(S_{A,R})$ and $\beta = 1 - p_1(S_{A,R})$, we get $\alpha \geq 1 - \beta$, i.e., $\alpha + \beta \geq 1$. This concludes the proof. $\square$

Lemma 5.13. *Assuming that $s = O(m/\log^2 m)$ and $s = O(t)$, the (η, K) -LDIIS ([Definition 5.10](#)) problem needs to retain at least $T' \geq \Omega(\min(K\eta, 1/\eta))$ (assuming that $1/\eta^3 \geq O(K)$) in order to succeed with probability at least $2/3$.*

Sketch proof. We consider for each of the K samples, a separate distinguisher $(A_i)_{i \in [K]}$. Using [Lemma 5.12](#), we have two cases to consider: i) suppose that at least half of them satisfy that $\alpha_i \geq \Omega(1)$, where α_i is the Type I error of A_i . In this case, in expectation we will retain at least $\alpha_i \cdot K \cdot \frac{1}{2} \geq \Omega(K)$ in our list. ii) Suppose we are in the opposite case, then at least half of them satisfy that $\alpha_i + \beta_i \geq 1$. In this case, we will either retain $\alpha_i \cdot K$ samples or since there are $K \cdot \eta$ samples from p_1 , we will keep $(1 - \beta_i) \cdot K \cdot \eta$ samples from p_1 . Now, we need that $\alpha_i \cdot K \leq T'$ and $(1 - \beta_i) \cdot K \cdot \eta \geq 0.5$ in order to break the lower bound. In other words, we want

$$\alpha_i \leq T'/K \text{ and } 0.5 \leq (1 - \beta_i) \cdot K \cdot \eta \leq \alpha_i \cdot K \cdot \eta.$$

Solving inequalities gives us that $T' \geq 0.5/\eta$. Over both cases, $T' \geq \Omega(\min(K\eta, 1/\eta))$.¹⁴ $\square$

Lemma 5.14. *The LDIIS ([Definition 5.10](#)) problem with T retaining samples can be reduced to the LDII ([Definition 5.8](#)) problem.*

Proof. We describe the reduction. Suppose we have an algorithm $\mathcal{A}$ that can pick up one sample from p_1 in a candidate set of size at most T , given at least $a = \frac{s^2}{t}$ samples from p_0 and at most $b = \frac{s^2 \text{polylog } m}{m} = a\eta$ samples from p_1 , then we can solve the previous problem. The reduction is straightforward: first, draw $\Theta(a)$ samples from the $(\frac{b}{a}, \Theta(a))$ -LDIIS problem (setting $K = \Theta(a), \eta = \frac{b}{a}$); with high probability, there will be at least a samples of p_0 and at most b samples from p_1 . Let T' be the lower bound of the $(\frac{b}{a}, \Theta(a))$ -LDIIS problem. Now algorithm $\mathcal{A}$ will choose a set of samples with size at most $T = T'$ as the retained set of sampling problem. If no algorithm for LDIIS can find a sample from p_1 if the retained set has size at most T , no algorithm for LDII can identify any sample from p_1 if $T = T'$. $\square$

¹⁴See formal proof in [Section B](#) for more details (regarding the limitation of $1/\eta^3 \ll O(K)$ and the $\Omega(K\eta)$ lower bound instead of $\Omega(K)$).

Remark 5.15 (On the parameter t). *At a high level, the hardness of our hard instance construction parameterized by t comes first from the fact that no algorithm with the frequency histogram of the clusters can reliably distinguish between the histograms of the clusters with mass $1/t$ vs. that of $\{\frac{k}{m}\}_{k \in \{2, \dots, \log^2 m\}}$ (made formal by Lemma 5.12 and the proof in Section B). On the other hand, to “win”, any algorithm needs to locate samples from (k/m) -mass clusters that have a collision. But the majority of collisions are produced by samples from elements with mass $1/t$, rather than from elements with mass k/m , and so the algorithm has to guess a lot before reaching one sample from k/m . Now, tuning t allows us to obtain a trade-off: we set $s = O(t)$ as this is the extremal point: with more samples than this, we can distinguish $1/t$ and k/m just by the frequency histogram. Setting t higher means that we can use more samples; as $1/t$ becomes closer to k/m , the $(1/t)$ -mass clusters will produce fewer collisions compared to the k/m ones, and thus the algorithm needs fewer guesses to take over all collisions.*

Given the hardness result of our list decoding problem from a mixture of p_0 and p_1 , we can prove the lower bound on the weakly-adaptive setting:

Proof of Lemma 5.3. We first recall our regime of interest: $\Omega(m^{2/3}) \leq s \leq O\left(\frac{m}{\log m}\right)$. We set $K = \Omega(s^2/t)$ and $\eta = O\left(\frac{t \text{polylog } m}{m}\right)$ and apply Lemma 5.9, which requires the following condition:

$$\frac{1}{\eta^3} \geq \Omega(K) \Leftrightarrow \left(\frac{m}{t \text{polylog } m}\right)^3 \geq \Omega\left(\frac{s^2}{t}\right) \Leftrightarrow s \leq O\left(\frac{m^{3/2}}{t \text{polylog } m}\right). \quad (3)$$

We collect all the constraints for the setting of s, t :

$$m^{2/3} \leq s \leq O(t) \leq O(m), s \leq O\left(\frac{m}{\text{polylog } m}\right), s \cdot t \leq O\left(\frac{m^{3/2}}{\text{polylog } m}\right). \quad (4)$$

From Lemma 5.9 and the fact that any weakly-adaptive algorithm needs to query each of its guesses at least $\Omega(n/\text{polylog } m)$, we obtain that any weakly-adaptive algorithm has query complexity to distinguish D_{yes} and D_{maybe} :

$$\begin{aligned} q &\geq \Omega\left(\min\left\{K \cdot \eta, \frac{1}{\eta}\right\}\right) \cdot \frac{n}{\text{polylog}(m)} \\ &= \Omega\left(\min\left\{\frac{s^2 \text{polylog } m}{m}, \frac{m}{t \text{polylog } m}\right\} \frac{n}{\text{polylog}(m)}\right) \\ &\geq \Omega\left(\frac{m}{t} \frac{n}{\text{polylog}(m)}\right), \end{aligned}$$

where the right hand side of the min is always smaller: $s^2 t \geq s^3 \geq \Omega(m^2)$, and thus $\frac{s^2 \text{polylog } m}{m} \geq \frac{m}{t \text{polylog } m} \Leftrightarrow s^2 t \geq \Omega(m^2) \geq m^2/\text{polylog } m$.

This means that for any $s \leq O\left(\frac{m^{3/2}}{t \text{polylog } m}\right)$, there exists a construction of hard instances by setting $t = \Theta(s)$ such that this setting of s, t satisfy the constraint in (4), and thus

$$q \geq \Omega\left(\frac{m}{s} \cdot \frac{n}{\text{polylog } m}\right).$$

To conclude, as $s = \Theta(\sqrt{st}) \leq O\left(\frac{m^{3/4}}{\text{polylog } m}\right)$, no algorithm can distinguish between D_{maybe} and D_{no} ; and also the algorithm needs $q \geq \Omega\left(\frac{m}{s} \cdot \frac{n}{\text{polylog } m}\right)$ queries to distinguish D_{yes} and D_{maybe} , and so the algorithm cannot distinguish D_{yes} and D_{no} with probability at least $2/3$. $\square$

We are now able to conclude with the main theorem of this section:

Proof of Theorem 5.1. To generalize the lower bound from weakly-adaptive to full-adaptive, we prove by reducing the full-adaptive setting to the LDII task over the course of an algorithm’s adaptive state. We first describe the view of the adaptive algorithm $\mathcal{A}$ at every step.

In the beginning, it sees the frequency histogram and picks a guess to query. In the i -th step, $\mathcal{A}$ sees the frequency histogram, and also knows the value of $X_1, \dots, X_{i-1}$, its previous queries. $\mathcal{A}$ then needs to choose the i -th X_i to query. For each query, $\mathcal{A}$ sees the value of X_i , given the hardness of LDII, most likely, $X_i = \frac{1}{m}$ or $X_i = \frac{1}{t}$. After knowing the value of X_i , $\mathcal{A}$ can simulate the query string on its own (after paying $O\left(\frac{n}{\text{polylog}(m)}\right)$ bits to figure out the value of X_i): if it is $X_i = \frac{1}{m}$, then generate a uniform random strings; otherwise, generate a uniform random string with the $\log\left(\frac{m}{t}\right)$ secret bits. Clearly, these extra strings will not help in reducing the complexity of solving LDII (otherwise any algorithm for LDII can generate these random strings: generate sufficiently many of them, and upon seeing X_i ’s value, choose an appropriate random string; and it can break the lower bound with the extra randomness).

Finally, we note that in the hardness proof of LDIIS (see Remark B.1), the lower bound holds even if you know values of other X_j for $j \neq i$ when making decision for the i -th sample. Therefore, knowing the configurations of $X_1, \dots, X_{i-1}$, will not break the lower bound of LDII (as LDIIS reduces to LDII). The full adaptive lower bound then follows from the proof of Lemma 5.3. $\square$

Acknowledgment The authors would like to thank anonymous reviewers from COLT 2026 and SODA 2027 for their suggestions, which considerably improved the presentation of the paper. SS’s research is supported by the NRF Investigatorship award (NRF-NRFI10-2024-0006) and the CQT Young Researcher Career Development Grant (25-YRCDG-SS). JY was supported by the JD Technology Scholarship for Postgraduate Research in Artificial intelligence and the Faculty of Engineering Career Advancement Award from The University of Sydney.

AI disclosure. All ideas, proofs, writing, and typos were generated solely by humans (the listed authors).

References

- [AF24] Tomer Adar and Eldar Fischer. Refining the Adaptivity Notion in the Huge Object Model. In: *APPROX/RANDOM*. Vol. 317. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2024, 45:1–45:16.
- [AFL24] Tomer Adar, Eldar Fischer, and Amit Levi. Support Testing in the Huge Object Model. In: *APPROX/RANDOM*. Vol. 317. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2024, 46:1–46:16.
- [BC17] Tugkan Batu and Clément L Canonne. Generalized uniformity testing. In: *Symposium on Foundations of Computer Science (FOCS)*. 2017, pp. 880–889.
- [BC18] Rishiraj Bhattacharyya and Sourav Chakraborty. Property Testing of Joint Distributions using Conditional Samples. In: *ACM Trans. Comput. Theory* 10.4 (2018), 16:1–16:20.
- [BFRSW00] 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)*. 2000, pp. 259–269.

- [BSGHSV06] Eli Ben-Sasson, Oded Goldreich, Prahladh Harsha, Madhu Sudan, and Salil Vadhan. Robust PCPs of Proximity, Shorter PCPs, and Applications to Coding. In: *SIAM Journal on Computing* 36.4 (2006), pp. 889–974.
- [BW18] Sivaraman Balakrishnan and Larry Wasserman. Hypothesis testing for high-dimensional multinomials: A selective review. In: *The Annals of Applied Statistics* (2018). DOI: [10.1214/18-AOAS1155SF](https://doi.org/10.1214/18-AOAS1155SF). URL: <https://doi.org/10.1214/18-AOAS1155SF>.
- [Can20] Clément L Canonne. A survey on distribution testing: Your data is big. But is it blue? In: *Theory of Computing* (2020), pp. 1–100.
- [Can22] Clément L Canonne. Topics and techniques in distribution testing: A biased but representative sample. In: *Foundations and Trends® in Communications and Information Theory* (2022), pp. 1032–1198.
- [CFGLMS25] Sourav Chakraborty, Eldar Fischer, Arijit Ghosh, Amit Levi, Gopinath Mishra, and Sayantan Sen. Testing vs Estimation for Index-Invariant Properties in the Huge Object Model. In: *STOC*. ACM, 2025, pp. 1007–1018.
- [CFGM16] Sourav Chakraborty, Eldar Fischer, Yonatan Goldhirsh, and Arie Matsliah. On the Power of Conditional Samples in Distribution Testing. In: *SIAM J. Comput.* 45.4 (2016), pp. 1261–1296.
- [CFGMS23] Sourav Chakraborty, Eldar Fischer, Arijit Ghosh, Gopinath Mishra, and Sayantan Sen. Testing of Index-Invariant Properties in the Huge Object Model. In: *COLT*. Vol. 195. Proceedings of Machine Learning Research. PMLR, 2023, pp. 3065–3136.
- [CR14] Clément L. Canonne and Ronitt Rubinfeld. Testing Probability Distributions Underlying Aggregated Data. In: *ICALP (1)*. Vol. 8572. Lecture Notes in Computer Science. Springer, 2014, pp. 283–295.
- [CRS15] Clément L. Canonne, Dana Ron, and Rocco A. Servedio. Testing Probability Distributions using Conditional Samples. In: *SIAM J. Comput.* 44.3 (2015), pp. 540–616.
- [CSY25] 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. In: *ITCS*. Vol. 325. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2025, 26:1–26:19.
- [DKS18] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Sharp bounds for generalized uniformity testing. In: *Advances in Neural Information Processing Systems (NeurIPS)* (2018).
- [GGR98] Oded Goldreich, Shafi Goldwasser, and Dana Ron. Property Testing and its Connection to Learning and Approximation. In: *J. ACM* 45.4 (1998), pp. 653–750.
- [Gol16] Oded Goldreich. The uniform distribution is complete with respect to testing identity to a fixed distribution. In: *Electronic Colloquium on Computational Complexity (ECCC)*. 2016, p. 1.
- [Gol17] Oded Goldreich. *Introduction to property testing*. Cambridge University Press, 2017.
- [GR00] Oded Goldreich and Dana Ron. On Testing Expansion in Bounded-Degree Graphs. In: *Electron. Colloquium Comput. Complex.* TR00-020 (2000).
- [GR23a] Oded Goldreich and Dana Ron. A Lower Bound on the Complexity of Testing Grained Distributions. In: *Computational Complexity (CC)* (2023), p. 11.

- [GR23b] Oded Goldreich and Dana Ron. Testing Distributions of Huge Objects. In: *TheoretCS*. 2023, p. 78.
- [RS09] Ronitt Rubinfeld and Rocco A. Servedio. Testing monotone high-dimensional distributions. In: *Random Struct. Algorithms* 34.1 (2009), pp. 24–44.

A Deferred proofs from Section 3

We will rely on the following result on estimating moments of (pseudo)-distributions:

Lemma A.1 (Adapted from [DKS18, Lemma 2.2]). *Let $r \geq 1$ and fix $S \subseteq \Omega$. We have the following: i) There is an algorithm that, given a parameter $0 < \delta < 1$ and sample access to a distribution D on Ω , draws $O(1/(\delta^2 \|D_S\|_r))$ samples in expectation and outputs an estimate $\hat{\gamma}_r$ such that, w.h.p., $|\hat{\gamma}_r - \|D_S\|_r^r| \leq \delta \cdot \|D_S\|_r^r$. ii) For any $c > 1$, there is an algorithm which, given $\text{Poi}(O(s))$ samples from a distribution D on Ω , distinguishes w.h.p. between $s^r \cdot \|D_S\|_r^r \geq 20c$ and $s^r \cdot \|D_S\|_r^r \leq c/20$.*

Fact 3.8 (Sanity checks). *Suppose $|\hat{D}(\bar{\mathcal{L}}) - D(\bar{\mathcal{L}})| \leq O(\varepsilon)$ and $\hat{D}(\bar{\mathcal{L}}) \geq \Omega(\varepsilon)$. Then we have the following: i) It takes $O(m^{1/2}/\varepsilon^{1/2})$ samples to distinguish between $\|D_{\bar{\mathcal{L}}}\|_2^2 < D(\bar{\mathcal{L}})(1600m)$ and $\|D_{\bar{\mathcal{L}}}\|_2^2 \geq D(\bar{\mathcal{L}})/m$. ii) It takes $O(m^{2/3}/\varepsilon^{1/3})$ samples to distinguish between $\|D_{\bar{\mathcal{L}}}\|_3^3 < D(\bar{\mathcal{L}})/(1600m^2)$ and $\|D_{\bar{\mathcal{L}}}\|_3^3 \geq D(\bar{\mathcal{L}})/m^2$.*

Proof of Fact 3.8 item i). We use the same algorithm as in Lemma A.1 in [DKS18], i.e., draw $\text{Poi}(O(s))$ samples and check if $s^2 \|\hat{D}_{\bar{\mathcal{L}}}\|_2^2 \geq c$. Let $s^2 = 40c \cdot \frac{m}{D(\bar{\mathcal{L}})}$, note that $20c \cdot \frac{m}{D(\bar{\mathcal{L}})} \leq s^2 \leq 80c \cdot \frac{m}{D(\bar{\mathcal{L}})}$. By Lemma A.1 item 2, we look at the two cases: when $\|D_{\bar{\mathcal{L}}}\|_2^2 \leq \frac{D(\bar{\mathcal{L}})}{1600m}$, then we have that

$$s^2 \cdot \|D_{\bar{\mathcal{L}}}\|_2^2 \leq 80c \cdot \frac{m}{D(\bar{\mathcal{L}})} \cdot \frac{1}{1600} \cdot \frac{D(\bar{\mathcal{L}})}{m} \leq \frac{c}{20}.$$

When $\|D_{\bar{\mathcal{L}}}\|_2^2 \geq \frac{D(\bar{\mathcal{L}})}{m}$, then we have that

$$s^2 \cdot \|D_{\bar{\mathcal{L}}}\|_2^2 \geq 20c \cdot \frac{m}{D(\bar{\mathcal{L}})} \cdot \frac{D(\bar{\mathcal{L}})}{m} = 20c.$$

And $s \leq O(\sqrt{\frac{m}{\varepsilon}})$. □

Proof of Fact 3.8 item ii). We use the same algorithm as in Lemma A.1 in [DKS18], i.e., draw $\text{Poi}(O(s))$ samples and check if $s^3 \|\hat{D}_{\bar{\mathcal{L}}}\|_3^3 \geq c$. Let $s^3 = 40c \cdot \frac{m^2}{D(\bar{\mathcal{L}})}$, we have that $20c \cdot \frac{m^2}{D(\bar{\mathcal{L}})} \leq s^3 \leq 80c \cdot \frac{m^2}{D(\bar{\mathcal{L}})}$

Suppose $\|D_{\bar{\mathcal{L}}}\|_3^3 \leq \frac{1}{1600} \cdot \frac{D(\bar{\mathcal{L}})}{m^2}$, then

$$s^3 \cdot \|D_{\bar{\mathcal{L}}}\|_3^3 \leq 80c \cdot \frac{m^2}{D(\bar{\mathcal{L}})} \cdot \frac{1}{1600} \cdot \frac{D(\bar{\mathcal{L}})}{m^2} = \frac{c}{20}.$$

Suppose $\|D_{\bar{\mathcal{L}}}\|_3^3 \geq \frac{D(\bar{\mathcal{L}})}{m^2}$, then

$$s^3 \cdot \|D_{\bar{\mathcal{L}}}\|_3^3 \geq 20c \cdot \frac{m^2}{D(\bar{\mathcal{L}})} \cdot \frac{D(\bar{\mathcal{L}})}{m^2} = 20c.$$

And $s = \left(40c \cdot \frac{m^2}{D(\bar{\mathcal{L}})}\right)^{1/3} \leq O\left(\frac{m^{2/3}}{\varepsilon^{1/3}}\right)$. □

Corollary 3.9. *Fix a set $S \subseteq \Omega$. Suppose $D(S) \geq \Omega(\varepsilon)$, $\|D_S\|_2^2 \geq \frac{1}{1600} \cdot \frac{D(S)}{m}$ and we are given $\hat{D}(S)$ such that $|\hat{D}(S) - D(S)| \leq O(\varepsilon)$. Then, with $O(\sqrt{m}/\varepsilon^{5/2})$ samples from D , one can return an estimate $\hat{D}$ such that $|\|\hat{D}_S\|_2^2 - \|D_S\|_2^2| \leq O(\varepsilon) \cdot \|D_S\|_2^2$. Moreover, $\left|\frac{\|\hat{D}_S\|_2^2}{\hat{D}(S)} - \frac{\|D_S\|_2^2}{D(S)}\right| \leq \frac{\varepsilon}{400}$.*

Proof of Corollary 3.9. By Lemma A.1, setting $\delta = O(\varepsilon)$, then we have that $|\|\hat{D}_S\|_2^2 - \|D_S\|_2^2| \leq O(\varepsilon) \cdot \|D_S\|_2^2$ using $O\left(\frac{1}{\varepsilon^2 \|D_S\|_2}\right) \leq O\left(\frac{\sqrt{m}}{\varepsilon^2 \sqrt{D(S)}}\right) \leq O\left(\frac{\sqrt{m}}{\varepsilon^{5/2}}\right)$ samples. Also $|\hat{D}(S) - D(S)| \leq O(\varepsilon)$ holds from premise. This implies that

$$|\|\hat{D}_S\|_2^2 / \|D_S\|_2^2 - 1| \leq O(\varepsilon) \text{ and } |\hat{D}(S)/D(S) - 1| \leq O(\varepsilon).$$

and therefore,

$$1 - O(\varepsilon) = \frac{1 - O(\varepsilon)}{1 + O(\varepsilon)} \leq \frac{\|\hat{D}_S\|_2^2 / \|D_S\|_2^2}{\hat{D}(S)/D(S)} \leq \frac{1 + O(\varepsilon)}{1 - O(\varepsilon)} = 1 + O(\varepsilon).$$

Noting that,

$$\begin{aligned} \left| \frac{\|\hat{D}_S\|_2^2}{\hat{D}(S)} - \frac{\|D_S\|_2^2}{D(S)} \right| &= \left| \frac{\|\hat{D}_S\|_2^2 / \|D_S\|_2^2}{\hat{D}(S)/D(S)} - 1 \right| \cdot \frac{\|D_S\|_2^2}{D(S)} \\ &\leq O(\varepsilon) \frac{\|D_S\|_2^2}{D(S)} \leq O(\varepsilon), \end{aligned}$$

where the last inequality we use $\|D_S\|_2 \leq \|D_S\|_1 \leq 1$. $\square$

Lemma 3.10. Fix any set S such that $D(S) \geq \Omega(\varepsilon)$, $\|D_S\|_2^2 \geq \frac{1}{1600} \cdot \frac{D(S)}{m}$ and $\|D_S\|_3^3 \geq \frac{1}{1600} \cdot \frac{D(S)}{m^2}$. Assume $|\hat{D}(S) - D(S)| \leq O(\varepsilon^2 \cdot D(S))$. Then, given $s = O\left(m^{1/2}/\varepsilon^{9/2} + m^{2/3}/\varepsilon^{13/3}\right)$ samples from D , we can distinguish w.h.p. between

$$\|D_S\|_3^3 = \frac{\|D_S\|_2^4}{D(S)} \quad \text{and} \quad \|D_S\|_3^3 - \frac{\|D_S\|_2^4}{D(S)} \geq \Omega\left(\varepsilon^2 \frac{\|D_S\|_2^4}{D^2(S)}\right).$$

Proof of Lemma 3.10. Given the lower bounds: $\|D_S\|_2^2 \geq \frac{1}{1600} \cdot \frac{D(S)}{m}$ and $\|D_S\|_3^3 \geq \frac{1}{1600} \cdot \frac{D(S)}{m^2}$, we can sample $s_1 = O\left(\frac{\sqrt{m}}{\varepsilon^2 \sqrt{D(S)}}\right) \leq O\left(\frac{m^{1/2}}{\varepsilon^2 \cdot \varepsilon^{1/2}}\right)$ and $s_2 = O\left(\frac{m^{2/3}}{\varepsilon^2 (D(S))^{1/3}}\right) \leq O\left(\frac{m^{2/3}}{\varepsilon^2 \cdot \varepsilon^{1/3}}\right)$ samples from D (respectively) to obtain

$$|\|\hat{D}_S\|_2^2 - \|D_S\|_2^2| \leq \delta \|D_S\|_2^2 \text{ and } |\|\hat{D}_S\|_3^3 - \|D_S\|_3^3| \leq \delta \|D_S\|_3^3,$$

respectively by Lemma A.1. We will set δ later in the proof.

Completeness. When $\|D_S\|_3^3 = \frac{\|D_S\|_2^4}{D(S)}$, then

$$\begin{aligned} \frac{\|\hat{D}_S\|_3^3 - \frac{\|\hat{D}_S\|_2^4}{\hat{D}(S)}}{\frac{\|\hat{D}_S\|_2^4}{\hat{D}(S)}} &= \frac{\|\hat{D}_S\|_3^3}{\|\hat{D}_S\|_2^4 / \hat{D}(S)} - 1 \\ &\leq \frac{1 + \delta}{(1 - \delta)/(1 + O(\varepsilon^2))} \frac{\|D_S\|_3^3}{\|D_S\|_2^4 / D(S)} - 1 \\ &= \frac{1 + \delta}{(1 - \delta)/(1 + O(\varepsilon^2))} \cdot 1 - 1 \\ &\leq (1 + 2\delta)(1 + O(\varepsilon^2)) - 1 \\ &\leq 2\delta + O(\varepsilon^2) \leq O(\varepsilon^2) \end{aligned}$$

where we use the $\|D_S\|_3^3 = \frac{\|D_S\|_2^4}{D(S)}$ for the third line and we set $\delta = O(\varepsilon^2)$ for the last inequality.

Soundness. When $\|D_S\|_3^3 - \frac{\|D_S\|_2^4}{D(S)} \geq \Omega\left(\varepsilon^2 \frac{\|D_S\|_2^4}{D^2(S)}\right)$, then we have

$$\frac{\|D_S\|_3^3 - \frac{\|D_S\|_2^4}{D(S)}}{\|D_S\|_2^4/D(S)} = \frac{\|D_S\|_3^3}{\|D_S\|_2^4/D(S)} - 1 \geq \Omega\left(\varepsilon^2 \frac{1}{D(S)}\right) \quad (5)$$

we look at its empirical estimates:

$$\begin{aligned} \frac{\|\hat{D}_S\|_3^3 - \frac{\|\hat{D}_S\|_2^4}{\hat{D}(S)}}{\frac{\|\hat{D}_S\|_2^4}{\hat{D}(S)}} &= \frac{\|\hat{D}_S\|_3^3}{\|\hat{D}_S\|_2^4/\hat{D}(S)} - 1 \\ &\geq \frac{1 - \delta}{(1 + \delta)/(1 - O(\varepsilon^2))} \frac{\|D_S\|_3^3}{\|D_S\|_2^4/D(S)} - 1 \\ &\geq \frac{(1 - \delta)(1 - O(\varepsilon^2))}{(1 + \delta)} (1 + \Omega(\varepsilon^2)) - 1 \\ &\geq (1 - 2\delta)(1 - O(\varepsilon^2))(1 + \Omega(\varepsilon^2)) - 1 \\ &\geq 1 - 2\delta - O(\varepsilon^2) + \Omega(\varepsilon^2) - 1 \\ &\geq \Omega(\varepsilon^2), \end{aligned}$$

where the third line uses (5) and we set $\delta = O(\varepsilon^2)$ for the last inequality.

We conclude that with $O\left(\frac{m^{1/2}}{\varepsilon^{9/2}} + \frac{m^{2/3}}{\varepsilon^{13/3}}\right)$ samples from D , we can distinguish the two cases with high constant probability. $\square$

Lemma 3.4 (D and D' are close on $\mathcal{S}$). *Given a sequence of good strings $(w^{(1)}, w^{(2)}, \dots, w^{(s')})$, a random projection J (and thus $X' \sim D'$), along with a subset $\mathcal{S}' \subseteq \Omega = \{0, 1\}^n$, one can apply $f_{J, \vec{w}}^{-1}$ on every $x' \in \mathcal{S}'$ to find $\mathcal{S}$ (see (2)). Moreover, we have that with probability $1 - o(1)$ over the choice of J , $d_{\text{EM}}(D_S, D'_{\mathcal{S}'}) \leq \frac{\varepsilon}{10}$.*

Proof of Lemma 3.4. We will adapt our proof from [GR23b, Claim 2.2.1]. The issue with the original proof is that it does not specify the transport scheme they would use is exactly as $X \rightarrow X'$ (where $X \sim D$ and $X' \sim D'$). Note that for our proof to work, we need a “local” transport scheme, as specified by how we construct X' . Our transport scheme will follow how we construct X' via $f_{J, \vec{w}}$: We move x to $w^{(i)}$ if $w_J^{(i)} = x_J$ otherwise x is unchanged. So the only case that costs us is when $x \neq w^{(i)}$ but their J -restriction are equal. Under this transport scheme, any string in $\mathcal{S}$ of X will only be moved to $\mathcal{S}'$ of X' .

We will first upper bound the probability that x is $(\varepsilon/20)$ -far in Hamming distance from $w^{(i)}$ in this case. Let $\text{Bad}_x(J)$ denote the event that $\exists w^{(i)}$ where $w_J^{(i)} = x_J$ but $d_H(w^{(i)}, x) \geq \frac{\varepsilon}{20}$. In addition, let $F_\varepsilon(S) = \{x \in \Omega \mid \forall y \in S, d_H(x, y) \geq \varepsilon\}$ denote the set of strings that are far from all

strings in S . We look at the following:

$$\begin{aligned}
\Pr_{x \sim X, J} [\text{Bad}_x(J)] &= \mathbb{E}_{x \sim X} [\Pr_J [\text{Bad}_x(J)]] \\
&\leq \mathbb{E}_{x \sim X} \left[\sum_{i \in [s']} \Pr_J \left[x_J = w_J^{(i)} \wedge d_H(x, w^{(i)}) \geq \frac{\varepsilon}{20} \right] \right] \\
&= \sum_{i \in [s']} \mathbb{E}_{x \sim X} \left[\Pr_J \left[x_J = w_J^{(i)} \wedge d_H(x, w^{(i)}) \geq \frac{\varepsilon}{20} \right] \right] \\
&\leq \sum_{i \in [s']} \max_{x \in F_{\varepsilon/20}(\{w^{(i)}\})} [\Pr_J [x_J = w_J^{(i)}]] \\
&\leq s' \cdot (1 - (\varepsilon/20))^\ell.
\end{aligned}$$

This is upper bounded by $o(\varepsilon)$ when $\ell = O\left(\frac{\log(s/\varepsilon)}{\varepsilon}\right)$ as $s' = \text{Poly}(s/\varepsilon)$. For the purpose of contradiction, we assume that

$$\Pr_{J \in \binom{[n]}{\ell}} \left[\Pr_{x \sim X} [\text{Bad}_x(J)] \geq \frac{\varepsilon}{20} \right] \geq \Omega(1).$$

This would imply that $\Pr_{x \sim X, J \in \binom{[n]}{\ell}} [\text{Bad}_x(J)] \geq \Omega(\varepsilon)$, which contradicts the probability upper bound of $o(\varepsilon)$. So we have that

$$\Pr_{J \in \binom{[n]}{\ell}} \left[\Pr_{x \sim X} [\text{Bad}_x(J)] \leq \frac{\varepsilon}{20} \right] = 1 - o(1).$$

With the above probability, we can upper bound the EMD between X and X' : The strings that are moved at most $\varepsilon/20$ distance will contribute at most $\varepsilon/20$ in EMD (worst case we move all of them by $\varepsilon/20$ in Hamming distance, which has probability mass 1); with probability $1 - o(1)$ over J , the probability mass that requires transporting more than $\varepsilon/20$ in Hamming distance is bounded by $\varepsilon/20$ (worst case we move them by 1 in Hamming distance, but the probability mass is at most $\varepsilon/20$). Therefore, it takes at most $\varepsilon/10$ to transport from D_S to $D'_{S'}$, thus concluding the proof. $\square$

Claim 3.5. *There is an algorithm ([Algorithm 3](#)) which takes $\tilde{O}(\tau)$ i.i.d. samples from D' and returns, with probability $1 - o(1)$, a set $\mathcal{L}'$ such that*

$$\mathcal{S}'_{3/\tau} \subseteq \mathcal{L}' \subseteq \mathcal{S}'_{1/\tau}.$$

Moreover, $\|D'_{\mathcal{L}'}\|_2^2 \leq 3D(\overline{\mathcal{L}'})/\tau$ and $\|D'_{\mathcal{L}'}\|_3^3 \leq 9D(\overline{\mathcal{L}'})/\tau^2$. Further, if $D'(\overline{\mathcal{S}'_{1/\tau}}) \geq \Omega(\varepsilon)$ then $D'(\overline{\mathcal{L}'}) \geq \Omega(\varepsilon)$.

Proof of Claim 3.5. There are at most $\tau/3$ elements in $\mathcal{S}'_{3/\tau}$. Draw $s_1 = O(\tau \log \tau)$ samples from D' and denote them $\mathcal{X}_1 = \{x_1, \dots, x_{s_1}\}$.

$$\Pr \left[x \in \mathcal{S}'_{3/\tau} \text{ and } x \notin \mathcal{X}_1 \right] \leq \sum_{x \in \mathcal{S}'_{3/\tau}} \left(1 - \frac{3}{\tau} \right)^{s_1} \leq \frac{\tau}{3} \exp \left(-\frac{3}{\tau} s_1 \right) = o(1).$$

Draw another multi-set of $s_2 = O(\tau \log s_1) = O(\tau \log \tau)$ samples (denote as $\mathcal{X}_2$) from D' . Then check every sample in $\mathcal{X}_2$ against every $x \in \mathcal{X}_1$ to estimate its probability mass. Note that we can estimate every $|\hat{D}'(x) - D'(x)| \leq \frac{1}{10\tau}$ accuracy (by a Chernoff bound followed by a union bound over $\mathcal{X}_1$). If

$\hat{D}'(x) \geq \frac{1.5}{\tau}$, then we add it in a list $\mathcal{L}$. Otherwise, it goes to $\bar{\mathcal{L}}$ (implicitly). We have the following high probability events (after a union bound over $\mathcal{X}_1$):

- (i) If $x \in \mathcal{X}_1$ and $D'(x) \geq \frac{3}{\tau}$ then $\hat{D}'(\text{Ball}(x, \varepsilon)) \geq \frac{2.9}{\tau}$, and thus $x \in \mathcal{L}$ ¹⁵.
 - (ii) Any $x \in \mathcal{X}_1$, and $D'(x) < \frac{1}{\tau}$, then $\hat{D}'(\text{Ball}(x, \varepsilon)) < \frac{1.1}{\tau}$ and therefore, $x \notin \mathcal{L}$.
- By $\mathcal{S}'_{3/\tau} \subset \mathcal{X}_1$ and item (i), we have that $\mathcal{S}'_{3/\tau} \subset \mathcal{L}$.

Because of item (ii), any $x \in \mathcal{X}_1$ but $D'(x) < \frac{1}{\tau}$ will be excluded from the target set $\mathcal{L}$. This means that if $x \in \mathcal{L}$ then $D'(x) \geq \frac{1}{\tau}$, which implies that $\mathcal{L} \subseteq \mathcal{S}'_{1/\tau}$. Also, if $x \notin \mathcal{L}$, due to item (i), it would mean that $D'(x) \leq \frac{3}{\tau}$, i.e., $\|D'_{\bar{\mathcal{L}}}\|_{\infty} \leq \frac{3}{\tau}$.

Combining the above, we can conclude that with probability $1 - o(1)$, $\mathcal{S}'_{3/\tau} \subseteq \mathcal{L} \subseteq \mathcal{S}'_{1/\tau}$. We note that

$$\|D'_{\bar{\mathcal{L}}}\|_2^2 \leq \|D'_{\bar{\mathcal{L}}}\|_{\infty} \cdot \|D'_{\bar{\mathcal{L}}}\|_1 \leq \frac{3 \cdot D'(\bar{\mathcal{L}})}{\tau} \text{ and } \|D'_{\bar{\mathcal{L}}}\|_3^3 \leq \|D'_{\bar{\mathcal{L}}}\|_{\infty} \cdot \|D'_{\bar{\mathcal{L}}}\|_2^2 \leq \frac{9 \cdot D'(\bar{\mathcal{L}})}{\tau^2}. \quad (6)$$

Finally, by $\mathcal{L} \subseteq \mathcal{S}'_{1/\tau}$, we have that $D'(\mathcal{L}) \leq D'(\mathcal{S}'_{1/\tau})$ which implies that $D'(\bar{\mathcal{L}}) \geq D'(\bar{\mathcal{S}}'_{1/\tau})$. Thus if $D'(\bar{\mathcal{S}}'_{1/\tau}) \geq \Omega(\varepsilon)$ then $D'(\bar{\mathcal{L}}) \geq \Omega(\varepsilon)$ holds. $\square$

Claim A.2. Fix a set of interest S and a distribution D . Suppose for any m -uniform distribution U , $\|D_S - U_S\|_1 \geq \varepsilon/100$ holds. Then $\forall x \in \left[\frac{1-\varepsilon/200}{m}, \frac{1+\varepsilon/200}{m}\right]$, we have that

$$\sum_{i \in S} \min\{D(i), |x - D(i)|\} \geq \varepsilon/200.$$

Proof. Denote H the set of i such that $D(i) \geq x/2$. Then $\delta = \sum_{i \in S} \min\{D(i), |x - D(i)|\} = \|D_S - c_{x,H}\|_1$, where $c_{x,H}$ is the (pseudo)-distribution that assigns x to elements on H and 0 otherwise. Let Q_H be the (pseudo)-distribution that assigns $1/m$ to a set of strings H . We have that

$$\|Q_H - c_{x,H}\|_1 = \sum_{i \in H} \left| \frac{1}{m} - x \right| \leq \frac{\varepsilon}{200},$$

where we use $x = \frac{1 \pm \varepsilon/200}{m}$ in the premise for the last inequality. Since $\|D_S - Q_H\|_1 \geq \varepsilon/100$ from our premise, we have

$$\|Q_H - c_{x,H}\|_1 + \|D_S - c_{x,H}\|_1 \geq \|D_S - Q_H\|_1 \Rightarrow \delta = \|D_S - c_{x,H}\|_1 \geq \|D_S - Q_H\|_1 - \frac{\varepsilon}{200} \geq \varepsilon/200.$$

$\square$

Lemma 3.6 ($D'_{\mathcal{S}'}$ is close to m -grained). Assume $m \geq \frac{\tau}{3}$. With probability $1 - o(1)$ over the choice of J , we can transport $W \geq D'(\mathcal{S}') - O(\frac{1}{m})$ mass of $D'_{\mathcal{S}'}$ into a (pseudo)-distribution Q that is m -grained, costing at most $O(\frac{\tau}{m})$. Moreover, Q 's mass is equal to W .

Proof of Lemma 3.6. If $\mathcal{S}' = \emptyset$, then it trivially holds. Suppose $D'(\mathcal{S}')$ is m -grained and $\mathcal{S}' \neq \emptyset$, otherwise, we can remove at most $1/m$ of its mass to make it $W = k/m$ for some integer k . In other words, we will not consider that $1/m$ mass in the following transport scheme.

Since $\forall x \in \mathcal{S}', D'(x) \geq \frac{3}{\tau}$, we know that $\ell = |\text{Supp}(D') \cap \mathcal{S}'| \leq \tau/3$. Suppose $\text{Supp}(D') \cap \mathcal{S}' = \{x_1, \dots, x_s\}$, then for every $D'(x_i)$, one can move at most $\frac{1}{m}$ mass to make it m -grained (since from the premise: $\frac{1}{m} \leq \frac{3}{\tau}$). We fix some arbitrary point x' , and move every such residual mass from x_i to

¹⁵For a sample $x \in \{0, 1\}^n$ and a parameter ε , a ball around x or radius ε is defined as $\text{Ball}(x, \varepsilon) = \{y \in \{0, 1\}^n : d_H(x, y) \leq \varepsilon\}$.

x' , costing us $d_H(x_i, x') \leq 1$. Indeed, this procedure will give us a new (pseudo)-distribution Q that is m -grained, i.e., Q is supported on $\{x_1, \dots, x_s, x'\}$, where $Q(x_i)$ is m -grained for $i \in [s]$ and as a result $Q(x') = \left(\frac{k}{m} - \sum_{i \in [s]} D'(x_i)\right)$ must also be m -grained. In total, we moved W mass of $D'_{S'}$ into Q , with a cost of $\leq \frac{\tau}{3m} \times 1 = O(\tau/m)$. $\square$

Lemma A.3. *Assume $m \geq \frac{\tau}{3}$. With probability $1 - o(1)$ over the choice of J , we can transport at least $(D(\mathcal{S}) - O(\frac{1}{m}))$ mass of $D_{\mathcal{S}}$ into a (pseudo)-distribution Q that is m -grained costing at most $O(\frac{\tau}{m}) + \frac{\varepsilon}{10}$.*

Proof of Lemma A.3. By Lemma 3.4, we can transport from $D_{\mathcal{S}}$ into $D'_{S'}$ with a cost at most $\varepsilon/10$. By Lemma 3.6, we can transport at least $D'(\mathcal{S}') - \frac{1}{m} = D(\mathcal{S}) - \frac{1}{m}$ mass of $D'_{S'}$ into a pseudo-distribution Q costing at most $O(\tau/m)$. Chaining the two transports, we can transport $D(\mathcal{S}) - \frac{1}{m}$ mass of D into Q costing at most $O(\tau/m) + \varepsilon/10$. $\square$

Lemma 3.12. *Suppose $\frac{100}{m} \leq \varepsilon$ and $100\frac{\tau}{m} \leq \varepsilon$. If D is ε -far from any m -uniform distribution in EMD, then, with probability $1 - o(1)$ over the choice of J , one cannot transport W' mass from $D_{\bar{\mathcal{S}}}$ to any m -uniform pseudo-distribution $\tilde{Q}$ such that $Q(\bar{\mathcal{S}}') = W'$ without a cost at least $\varepsilon/2$, where $W' = \frac{k}{m} \geq D(\bar{\mathcal{S}}) - \varepsilon/100$ for some integer k .*

Proof of Lemma 3.12. Suppose we are in the regime where m -uniform is equivalent to m -grained (to figure out the corner case later). By Lemma A.3, we can transport at least $(D(\mathcal{S}) - \frac{1}{m})$ mass of $D_{\mathcal{S}}$ into a (pseudo)-distribution Q that is m -grained paying at most $O(\frac{\tau}{m} + \varepsilon)$. Fix a reverse string $x_0 \in \Omega = \{0, 1\}^n$. For the sake of contradiction, we assume that we can also move mass $(D(\bar{\mathcal{S}}) - \varepsilon/100)$ of $D_{\bar{\mathcal{S}}}$ into a pseudo-distribution $\tilde{Q}$ that is m -grained paying at most $\varepsilon/2$. Then we could combine the two transport schemes by moving the last $\leq O(\frac{2}{m})$ mass into x_0 , with a combined cost of $\frac{\tau}{m} + \frac{2}{m} + \varepsilon/10$. Note that since Q and $\tilde{Q}$ are both m -grained, the remaining $O(2/m)$ moved into x_0 must also be m -grained. In effect, we obtain a transport scheme that serves as evidence that the EMD distance between D and some m -uniform is at most $\frac{\tau}{m} + \frac{2}{m} + \frac{\varepsilon}{10} + \frac{\varepsilon}{2} \leq \varepsilon$. This will contradict with our premise that D is ε -far from any m -uniform distribution in EMD. Therefore, it will at least cost us $\varepsilon/2$ to move most mass of $D_{\bar{\mathcal{S}}}$ to a m -uniform distribution. $\square$

Lemma 3.7 (d_{EM} implies d_{TV} Farness). *Suppose $\frac{100}{m} \leq \varepsilon$. If we cannot transport at least $D(\bar{\mathcal{S}}) - \frac{\varepsilon}{100}$ of the mass of $D_{\bar{\mathcal{S}}}$ into any m -uniform with cost at most ε , then for any m -uniform distribution U , $\|D_{\bar{\mathcal{S}}} - U_{\bar{\mathcal{S}}}\|_1 \geq \frac{\varepsilon}{100}$.*

Proof of Lemma 3.7. Suppose for the sake of contradiction, there exists a m -uniform distribution U such that

$$\|D_{\bar{\mathcal{S}}} - U_{\bar{\mathcal{S}}}\|_1 \leq \varepsilon/100.$$

For our proof, we want to find a way to make $D(\bar{\mathcal{S}})$ and $U(\bar{\mathcal{S}})$ “equal” in mass.

Suppose we are in the case $U(\bar{\mathcal{S}}) \leq D(\bar{\mathcal{S}}) \leq U(\bar{\mathcal{S}}) + \frac{\varepsilon}{100}$. Let $D(x) = \tilde{D}(x) + \Delta(x)$, where $\tilde{D}(\bar{\mathcal{S}}) \geq D(\bar{\mathcal{S}}) - \frac{\varepsilon}{100}$ and $\tilde{D}(\bar{\mathcal{S}}) = U(\bar{\mathcal{S}}) = \frac{k}{m}$ for some integer k ; note that $\Delta(x) \geq 0$ and $\sum_{x \in \bar{\mathcal{S}}} \Delta(x) \leq \frac{\varepsilon}{100}$.¹⁶

¹⁶The tricky thing is that we might not find a set where $D(L) = \frac{k}{m}$ for some integer k ; and there is a solution for this program since $|D(\bar{\mathcal{S}}) - U(\bar{\mathcal{S}})| \leq \|D_{\bar{\mathcal{S}}} - U_{\bar{\mathcal{S}}}\|_1 \leq \frac{\varepsilon}{100}$.

Fix a $\tilde{D}$ and Δ satisfying the constraints. We have that,

$$\begin{aligned}\|\tilde{D} - U_{\bar{S}}\|_1 &= \sum_{x \in \bar{S}} |\tilde{D}(x) + \Delta(x) - \Delta(x) - U_{\bar{S}}(x)| \\ &\leq \sum_{x \in \bar{S}} |D(x) - U_{\bar{S}}(x)| + \sum_{x \in \bar{S}} \Delta(x) \\ &= \|D_{\bar{S}} - U_{\bar{S}}\|_1 + \frac{\varepsilon}{100}.\end{aligned}$$

Since $\tilde{D}(\bar{S})$ contains a mass of at least $D(\bar{S}) - \frac{\varepsilon}{100}$ of $D(\bar{S})$, we can applying Lemma 3.12, and we have that

$$d_{\text{EM}}(\tilde{D}, U_{\bar{S}}) \geq \Omega(\varepsilon).$$

We can scale $\tilde{D}$ and $U_{\bar{S}}$ such that their sum is 1, i.e., let $Q_1(x) = \tilde{D}(x)/\tilde{D}(\bar{S})$ and $Q_2(x) = U(x)/U(\bar{S}) = \frac{1}{k}$. We note that now both Q_1 and Q_2 are bona fide distributions (albeit they might not be m -grained any more):

$$\Omega\left(\frac{\varepsilon}{D(\bar{S})}\right) \leq \frac{d_{\text{EM}}(\tilde{D}, U_{\bar{S}})}{D(\bar{S})} = d_{\text{EM}}(Q_1, Q_2) \leq \frac{1}{2}\|Q_1 - Q_2\|_1.$$

Scaling back everything, we can show that

$$\varepsilon \leq D(\bar{S}) \frac{1}{2} \|Q_1 - Q_2\|_1 = \frac{1}{2} \|\tilde{D} - U_{\bar{S}}\|_1.$$

Therefore, $\|D_{\bar{S}} - U_{\bar{S}}\|_1 + \frac{\varepsilon}{100} \geq \|\tilde{D} - U_{\bar{S}}\|_1 \geq 2\varepsilon \Rightarrow \|D_{\bar{S}} - U_{\bar{S}}\|_1 \geq 2\varepsilon - \frac{\varepsilon}{100} \geq \frac{\varepsilon}{100}$.

If $U(\bar{S}) - \frac{\varepsilon}{100} \leq D(\bar{S}) \leq U(\bar{S})$, we can solve a program $U(x) = \tilde{U}(x) + \Delta_1(x)$, where $\Delta_1(x) \geq 0$, $\tilde{U}(\bar{S}) \geq D(\bar{S})$ is $\frac{k'}{m}$ for some integer k' and $\sum_{x \in \bar{S}} \Delta(x) \leq \frac{\varepsilon}{100} + \frac{1}{m}$. Note that we still have

$$\begin{aligned}\|\tilde{U}_{\bar{S}} - D_{\bar{S}}\|_1 &= \sum_{x \in \bar{S}} |\tilde{U}_{\bar{S}}(x) + \Delta(x) - \Delta(x) - D_{\bar{S}}(x)| \\ &\leq \sum_{x \in \bar{S}} |U(x) - D_{\bar{S}}(x)| + \sum_{x \in \bar{S}} \Delta(x) \\ &\leq \|D_{\bar{S}} - U_{\bar{S}}\|_1 + \frac{\varepsilon}{100} + \frac{1}{m} \leq \frac{3\varepsilon}{100}.\end{aligned}$$

Now $\tilde{U}(\bar{S}) \leq D(\bar{S}) \leq \tilde{U}(\bar{S}) + \frac{1}{m}$. We apply the argument derived from the other case: Find a $\tilde{D}_1$ such that it contains more than $(D(\bar{S}) - \frac{1}{m})$ mass of D (with equal mass to $\tilde{D}_1(\bar{S}) = \tilde{U}(\bar{S})$), which implies that

$$\|\tilde{D}_1 - \tilde{U}\|_1 \geq 2\varepsilon \Rightarrow \|D_{\bar{S}} - \tilde{U}\|_1 \geq 2\varepsilon - \frac{1}{m} \geq \left(2 - \frac{1}{100}\right)\varepsilon,$$

which contradicts with previous derivation that $\|\tilde{U}_{\bar{S}} - D_{\bar{S}}\|_1 \leq \frac{3\varepsilon}{100}$. $\square$

Claim 3.11. Fix any subset $S \subseteq \{0, 1\}^n$. Then: (i) If D is m -uniform, then

$$\frac{\|D_S\|_2^2}{D(S)} = \frac{1}{m} \text{ and } \|D_S\|_3^3 - \frac{\|D_S\|_2^4}{D(S)} = 0.$$

(ii) If $\min_{U \text{ } m\text{-uniform}} \|D_S - U_S\|_1 \geq \frac{\varepsilon}{100}$ and $\frac{\|D_S\|_2^2}{D(S)} \in \left[\frac{1 \pm \varepsilon/200}{m}\right]$, then

$$\|D_S\|_3^3 - \frac{\|D_S\|_2^4}{D(S)} \geq \Omega\left(\varepsilon^2 \frac{\|D_S\|_2^4}{(D(S))^2}\right).$$

Proof of Claim 3.11. We have

$$\begin{aligned} \sum_{i \in S} D(i) \left(D(i) - \frac{\|D_S\|_2^2}{D(S)} \right)^2 &= \sum_{i \in S} D(i)^3 - 2D(i)^2 \frac{\|D_S\|_2^2}{D(S)} + D(i) \left(\frac{\|D_S\|_2^2}{D(S)} \right)^2 \\ &= \|D_S\|_3^3 - \frac{\|D_S\|_2^4}{D(S)}. \end{aligned}$$

(i) Note that when $\|D_S - U_S\|_1 = 0$, which means that when $D(i) = \frac{1}{m}$ for every $i \in S$, then indeed,

$$\|D_S\|_3^3 = \frac{|S|}{m^3} = \frac{\|D_S\|_2^4}{D(S)} = \frac{|S|^2/m^4}{|S|/m} = \frac{|S|}{m^3}.$$

Thus $\|D_S\|_3^3 - \frac{\|D_S\|_2^4}{D(S)} = 0$ holds.

(ii) When $\|D_S - U_S\|_1 \geq \varepsilon/100$ and $\frac{1-\varepsilon/200}{m} \leq \frac{\|D_S\|_2^2}{D(S)} \leq \frac{1+\varepsilon/200}{m}$, by Claim A.2, we have that

$$\sum_{i \in S} \min \left\{ D(i), \left| D(i) - \frac{\|D_S\|_2^2}{D(S)} \right| \right\} \geq \varepsilon/200.$$

Denote $S_l := \left\{ i \in [S] \mid D(i) \leq \frac{\|D_S\|_2^2}{D(S)}/2 \right\}$ and $S_h := S \setminus S_l$. We know one of the sum (over S_l or S_h) will be greater than $\varepsilon/400$. Suppose

$$\sum_{i \in S_l} \min \left\{ D(i), \left| D(i) - \frac{\|D_S\|_2^2}{D(S)} \right| \right\} \geq \varepsilon/400,$$

which implies that $\sum_{i \in S_l} D(i) \geq \varepsilon/200$. We have that

$$\sum_{i \in S_l} D(i) \left(D(i) - \frac{\|D_S\|_2^2}{D(S)} \right)^2 \geq \sum_{i \in S_l} D(i) \cdot \left(\frac{\|D_S\|_2^2}{2D(S)} \right)^2 \geq \frac{\varepsilon}{400} \left(\frac{\|D_S\|_2^2}{2D(S)} \right)^2.$$

Suppose $\sum_{i \in S_h} \min \left\{ D(i), \left| D(i) - \frac{\|D_S\|_2^2}{D(S)} \right| \right\} \geq \varepsilon/400$, we have that

$$2D(S) \geq 2 \sum_{i \in S_h} D(i) > \sum_{i \in S_h} \frac{\|D_S\|_2^2}{D(S)} = |S_h| \frac{\|D_S\|_2^2}{D(S)} \Rightarrow \frac{1}{|S_h|} \geq \frac{1}{2} \frac{\|D_S\|_2^2}{D^2(S)}. \quad (7)$$

Moreover,

$$\begin{aligned} \sum_{i \in S_h} D(i) \left(D(i) - \frac{\|D_S\|_2^2}{D(S)} \right)^2 &\geq \sum_{i \in S_h} \left(\frac{\|D_S\|_2^2}{D(S)} / 2 \right) \left(D(i) - \frac{\|D_S\|_2^2}{D(S)} \right)^2 \\ &\geq \left(\frac{\|D_S\|_2^2}{D(S)} / 2 \right) \frac{\left(\sum_{i \in S_h} \left(D(i) - \frac{\|D_S\|_2^2}{D(S)} \right) \right)^2}{|S_h|} \\ &\geq \left(\frac{\|D_S\|_2^2}{D(S)} / 2 \right) \left(\sum_{i \in S_h} \left(D(i) - \frac{\|D_S\|_2^2}{D(S)} \right) \right)^2 \frac{1}{2} \frac{\|D_S\|_2^2}{D^2(S)} \\ &= \left(\frac{\|D_S\|_2^2}{D(S)} / 2 \right) \frac{\|D_S\|_2^2}{2D^2(S)} \cdot \frac{\varepsilon^2}{(400)^2} \\ &= \frac{\varepsilon^2}{64 \times 10^4} \frac{\|D_S\|_2^4}{D^3(S)}, \end{aligned}$$

where we use Cauchy-Schwarz inequality for the second inequality and (7) for the third inequality. Overall, we have that

$$\|D_S\|_3^3 - \frac{\|D_S\|_2^4}{D(S)} \geq \min \left(\frac{\varepsilon}{400} \left(\frac{\|D_S\|_2^2}{2D(S)} \right)^2, \frac{\varepsilon^2}{64 \times 10^4} \frac{\|D_S\|_2^4}{D^3(S)} \right) \geq \frac{\varepsilon^2}{64 \times 10^4} \frac{\|D_S\|_2^4}{D^2(S)}.$$

□

B Deferred proofs from Section 5

Lemma 5.13. *Assuming that $s = O(m/\log^2 m)$ and $s = O(t)$, the (η, K) -LDIIS (Definition 5.10) problem needs to retain at least $T' \geq \Omega(\min(K\eta, 1/\eta))$ (assuming that $1/\eta^3 \geq O(K)$) in order to succeed with probability at least $2/3$.*

Full proof for Lemma 5.13. Assume that $s = O(m/\log^2 m)$ and $s = O(t)$. Fix any fully adaptive algorithm for keeping or dropping one sample: it can look at all $\{Y_1, \dots, Y_K\}$ at the same time, where K is the total number of samples. Let us focus on the i -th sample. Denote by $S_{A,R}^{(1)} \otimes S_{A,R}^{(2)} \otimes \dots \otimes S_{A,R}^{(K)}$ the accept region for the i -th sample, i.e., if $Y_j \in S_{A,R}^{(j)}$ for all $j \in [K]$, then we say that the algorithm will accept the i -th sample. We calculate α , the Type I error:

$$\begin{aligned} \alpha &= \mathbb{E}_{Y_1, \dots, Y_i \sim p_0, Y_K, R} [A(Y_1, \dots, Y_K, R) = 1] \\ &= \mathbb{E}_{Y_1, \dots, Y_i \sim p_0, \dots, Y_K, R} [Y_1 \in S_{A,R}^{(1)} \wedge \dots \wedge Y_K \in S_{A,R}^{(K)}] \\ &= \mathbb{E}_R [Y_i \in S_{A,R}^{(i)}] \cdot \prod_{j \neq i} \mathbb{E} [Y_j \in S_{A,R}^{(j)}] \\ &= p_0(S_{A,R}^{(i)}) \cdot \sigma_i \end{aligned} \tag{8}$$

where $\sigma_i = \prod_{j \neq i} \mathbb{E} [Y_j \in S_{A,R}^{(j)}]$. For Type II, we have

$$\begin{aligned} \beta &= \mathbb{E}_{Y_1, \dots, Y_i \sim p_1, Y_K, R} [A(Y_1, \dots, Y_K, R) = 0] \\ &= \mathbb{E}_{Y_1, \dots, Y_i \sim p_0, \dots, Y_K} [Y_1 \notin S_{A,R}^{(1)} \vee \dots \vee Y_K \notin S_{A,R}^{(K)}] \\ &= \left(1 - \mathbb{E} [Y_i \in S_{A,R}^{(i)}] \cdot \prod_{j \neq i} \mathbb{E} [Y_j \in S_{A,R}^{(j)}] \right) \\ &= 1 - \mathbb{E} [Y_i \in S_{A,R}^{(i)}] \cdot \sigma_i \\ &= 1 - p_1(S_{A,R}^{(i)}) \cdot \sigma_i. \end{aligned} \tag{9}$$

As in the proof of Lemma 5.12, we will consider two cases: case i), first, we consider the case where $\{2\} = S_{A,R}^{(i)}$. Following the proof of Lemma 5.12, we have that $p_0(S_{A,R}^{(i)}) \cdot \sigma_i \geq \Omega(\sigma_i)$ and recalling that p_1 has parameter $\frac{sk}{m}$, similarly $p_1(S_{A,R}^{(i)}) = 1 - \Theta(\frac{sk}{m}) = \Theta(1)$, as long as $sk = O(m)$, from which

$$1 - p_1(S_{A,R}^{(i)}) \cdot \sigma_i = 1 - (1 - \Theta(\frac{sk}{m})) \cdot \sigma_i = 1 - \Theta(\sigma_i)$$

Suppose that $\sigma_i \geq \frac{1}{100} \cdot \eta$, then $\alpha \geq \Omega(\eta)$. If $\sigma_i < \frac{1}{100} \cdot \eta$, then $\alpha + \beta \geq \Omega(\sigma_i) + 1 - \Theta(\sigma_i) \geq 1 - \frac{1}{50} \cdot \eta$. For the general case that $2 \in S_{A,R}^{(i)}$, we first note that $\alpha \geq \Omega(\sigma_i)$ as accepting more means that false

positive will only increase. So for $\sigma_i \geq \frac{1}{100} \cdot \eta$, $\alpha_i \geq \Omega(\eta)$. We then look at the case that $\sigma_i < \frac{1}{100} \cdot \eta$. Given that $p_0(j) \geq p_1(j)$, when $j > 2$. Using $p_0(S_{A,R}^{(i)} \setminus \{2\}) \geq p_1(S_{A,R}^{(i)} \setminus \{2\})$, we have that

$$\begin{aligned} \alpha_i + \beta_i &\geq 1 + (p_0(S_{A,R}^{(i)}) - p_1(S_{A,R}^{(i)})) \cdot \sigma_i \\ &= 1 + (p_0(2) + p_0(S_{A,R}^{(i)} \setminus \{2\})) \cdot \sigma_i - (p_1(2) + p_1(S_{A,R}^{(i)} \setminus \{2\})) \cdot \sigma_i \\ &\geq 1 + (p_0(2) - p_1(2)) \cdot \sigma_i. \end{aligned}$$

This thus reduces back to the case when $\{2\} = S_{A,R}^{(i)}$, and thus we conclude that either $\alpha_i \geq \Omega(\eta)$ or $\alpha_i + \beta_i \geq 1 - \frac{1}{50} \cdot \eta$ for case i).

In case ii), where $2 \notin S_{A,R}^{(i)}$, we have that

$$\alpha + \beta = p_0(S_{A,R}^{(i)}) \cdot \sigma_i + 1 - p_1(S_{A,R}^{(i)}) \cdot \sigma_i \geq p_1(S_{A,R}^{(i)}) \cdot \sigma_i + 1 - p_1(S_{A,R}^{(i)}) \cdot \sigma_i = 1.$$

Overall, we have that for our i -th distinguisher A_i : either $\alpha_i \geq \Omega(\eta)$ or $\alpha_i + \beta_i \geq 1 - \frac{1}{50} \eta$. If at least half of them satisfy that $\alpha_i \geq \Omega(\eta)$, then over all A_i , we at least retain $\alpha_i \cdot K \cdot \frac{1}{2} \geq \Omega(K \cdot \eta)$ samples. In the opposite case, if at least half of them satisfy that $\alpha_i + \beta_i \geq 1 - \Theta(\eta)$, then we need $\alpha_i \cdot K \leq T'$ and $(1 - \beta_i) \cdot K \cdot \eta \geq 0.25$ to break the lower bound. Solving the inequalities, we have

$$\alpha_i \leq \frac{T'}{K} \text{ and } 0.25 \leq (1 - \beta_i) \cdot K \cdot \eta \leq \left(\alpha_i + \frac{1}{50} \cdot \eta \right) \cdot K \cdot \eta.$$

Solving the inequalities, we have that $\frac{T'}{K} \geq \alpha_i \geq \frac{0.25}{K\eta} - \frac{1}{50} \cdot \eta^2 \Rightarrow T' \geq \frac{0.25}{\eta} - \frac{1}{50} \cdot K \cdot \eta^2$. Over both cases, we have that $T' \geq \min(K \cdot \eta, \frac{1}{\eta})$ if $1/\eta \gg K \cdot \eta^2$.

From expectation to high probability: we first note that one can use Markov's inequality for the false negatives: if $\mathbb{E}[(1 - \beta_i) \cdot K \cdot \eta] \leq 0.25$, then with high constant probability, the algorithm will return strictly less than 1 sample from p_1 .

Second, for the false positives, we note that even conditioning on the extra information (observations of $(Y_j)_{j \neq i}$ or even labels of $(X_j)_{j \neq i}$ (see [Remark B.1](#))), the probability of error remains bounded below: $\alpha_i \geq \min\left\{\frac{T'}{K}, \eta\right\}$. Therefore, we can couple each execution of A_i over p_0 , i.e., $\text{Bern}(\alpha_i)$ with the output of $\min\left\{\text{Bern}\left(\frac{T'}{K}\right), \text{Bern}(\eta)\right\}$, noting that the latter Bernoullis are fully independent across i . An application of Chernoff bound suffices to give high probability error bound. $\square$

Remark B.1. Notice in (8) and (9), we can replace the accept region to be, e.g., $X_j = 1/t$ or $X_j = 1/m$. Using the same independence argument, one can still get the “same” σ_i in the following equality. In other words, even if one is provided the extra information regarding other configurations, when making a decision for the i -th sample, the complexity lower bound still applies.