Testing Probability Distributions Underlying Aggregated Data

Citation: Canonne, Clement, and Ronitt Rubinfeld. "Testing Probability Distributions Underlying Aggregated Data." Lecture Notes in Computer Science (2014): 283–295.

As Published: http://dx.doi.org/10.1007/978-3-662-43948-7_24

Publisher: Springer-Verlag

Persistent URL: http://hdl.handle.net/1721.1/101001

Version: Original manuscript: author's manuscript prior to formal peer review

Testing probability distributions underlying aggregated data

Clément Canonne[*] Ronitt Rubinfeld[†]

February 15, 2014

Abstract

In this paper, we analyze and study a hybrid model for testing and learning probability distributions. Here, in addition to samples, the testing algorithm is provided with one of two different types of oracles to the unknown distribution D over $[n]$. More precisely, we define both the *dual* and *cumulative dual access models*, in which the algorithm A can both sample from D and respectively, for any $i \in [n]$,

- query the probability mass $D(i)$ *(query access)*; or
- get the total mass of $\{1, \dots, i\}$, i.e. $\sum_{j=1}^{i} D(j)$ *(cumulative access)*

These two models, by generalizing the previously studied sampling and query oracle models, allow us to bypass the strong lower bounds established for a number of problems in these settings, while capturing several interesting aspects of these problems – and providing new insight on the limitations of the models. Finally, we show that while the testing algorithms can be in most cases strictly more efficient, some tasks remain hard even with this additional power.

1 Introduction

1.1 Background

Given data sampled from a population or an experiment, understanding the distribution from which it has been drawn is a fundamental problem in statistics, and one which has been extensively studied for decades. However, it is only rather recently that these questions have been considered when the distribution is over a *very large* domain (see for instance [4, 18, 21]). In this case, the usual techniques in statistics and learning theory become impractical, motivating the search for better algorithms, in particular by relaxing the goals so that learning is not required. This is useful in many real-world applications where only a particular aspect of the distribution is investigated, such as estimating the entropy or the distance between two distributions. In these examples, as well as many others, one *can* achieve sublinear sample complexity. However, strong lower bounds show that the complexity of these tasks is still daunting, as it has polynomial, and often nearly linear, dependence on the size of the support of the distribution. To address this difficulty, new lines of research have emerged. One approach is to obtain more efficient algorithms for special classes of distributions. For instance, improved algorithms whose sample complexity is polylogarithmic in the domain size can be achieved by requiring it to satisfy specific smoothness assumptions, or to be of a convenient shape (monotone, unimodal, or a "k-histogram" [6, 20, 14]). A different approach applies to general distributions, but gives the algorithm more power in form of more flexible access to the distribution: as in many applications the data has already been collected and aggregated, it may be reasonable to assume that the testing algorithm can make other limited queries to the probability density function. For example, the algorithm may be provided with query access to the probability density function of the distribution [24], or samples from conditional distributions induced by the original distribution [12, 10, 11].

[*] Columbia University. Email: ccanonne@cs.columbia.edu

[†] CSAIL, MIT and the Blavatnik School of Computer Science, Tel Aviv University. Email: ronitt@csail.mit.edu. Research supported by NSF grants CCF-1217423 and CCF-1065125.

1.2 Our model: dual and cumulative dual oracles

In this work, we consider the power of two natural oracles. The first is a *dual oracle*, which combines the standard model for distributions and the familiar one commonly assumed for testing Boolean and real-valued functions. In more detail, the testing algorithm is granted access to the unknown distribution D through two independent oracles, one providing samples of the distribution, while the other, on query i in the domain of the distribution, provides the value of the probability density function at i. [1]

Definition 1 (Dual access model)**.** Let D be a fixed distribution over $[n]= \{1, \dots, n\}$. A *dual oracle for D* is a pair of oracles $(\mathsf{SAMP}_D, \mathsf{EVAL}_D)$ defined as follows: when queried, the *sampling* oracle SAMP_D returns an element $i \in [n]$, where the probability that i is returned is $D(i)$ independently of all previous calls to any oracle; while the *evaluation* oracle EVAL_D takes as input a query element $j \in [n]$, and returns the probability weight $D(j)$ that the distribution puts on j.

It is worth noting that this type of dual access to a distribution has been considered (under the name *combined oracle*) in [7] and [19], where they address the task of estimating (multiplicatively) the entropy of the distribution, or the f-divergence between two of them (see Sect. 4 for a discussion of their results).

The second oracle that we consider provides samples of the distribution as well as queries to the *cumulative distribution function* (cdf) at any point in the domain[2].

Definition 2 (Cumulative Dual access model)**.** Let D be a fixed distribution over $[n]$. A *cumulative dual oracle for D* is a pair of oracles $(\mathsf{SAMP}_D, \mathsf{CEVAL}_D)$ defined as follows: the *sampling* oracle SAMP_D behaves as before, while the *evaluation* oracle CEVAL_D takes as input a query element $j \in [n]$, and returns the probability weight that the distribution puts on $[j]$, that is $D([j]) = \sum_{i=1}^{j} D(i)$.

1.3 Motivation and discussion

As a first motivation to this hybrid model, consider the following scenario: There is a huge and freely available dataset, which a computationally-limited party – call it Arthur – needs to process. Albeit all the data is public and Arthur can view any element of his choosing, extracting further information from the dataset (such as the number of occurrences of a particular element) takes too much time. However, a third-party, Merlin, has already spent resources in preprocessing this dataset and is willing to disclose such information – yet at a price. This leaves Arthur with the following question: *how can he get his work done as quickly as possible, paying as little as possible?* This type of question is captured by our new model, and can be analyzed in this framework. For instance, if the samples are stored in sorted order, implementing either of our oracles becomes possible with only a logarithmic overhead per query. It is worth noting that Google has published their N-gram models, which describe their distribution model on 5-word sequences in the English language. In addition, they have made available the texts on which their model was constructed. Thus, samples of the distribution in addition to query access to probabilities of specific domain elements may be extracted from the Google model.

A second and entirely theoretical motivation for studying distribution testing in these two dual oracle settings arises from attempting to understand the limitations and underlying difficulties of the standard sampling model. Indeed, by circumventing the lower bound, one may get a better grasp on the core issues whence the hardness stemmed in the first place.

Another motivation arises from data privacy, when a curator administers a database of highly sensitive records (e.g, healthcare information, or financial records). Differential privacy [15, 17, 16] studies mechanisms which allow the curator to release relevant information about its database without without jeopardizing the privacy of the individual records. In particular, mechanisms have been considered that enable the curator to *release* a sanitized approximation $\tilde{D}$ of its database D, which "behaves" essentially the same for all queries

[1] Note that in both definitions, one can decide to disregard the corresponding evaluation oracle, which in effect amounts to falling back to the standard sampling model; moreover, for our domain $[n]$, any EVAL_D query can be simulated by (at most) two queries to a CEVAL_D oracle – in other terms, the cumulative dual model is at least as powerful as the dual one.

[2] We observe that such a cumulative evaluation oracle CEVAL appears in [6] (Sect. 8).

of a certain type – such as *counting* or *interval queries*[3] [9]. Specifically, if the user needs to test a property of a database, it is sufficient to test whether the sanitized database has the property, using now both samples and interval (i.e., CEVAL) or counting (EVAL) queries. As long as the tester has some tolerance (in that it accepts databases that are close to having the property), it is then possible to decide whether the true database itself is close to having the property of interest.

Finally, a further motivation is the tight connection between the dual access model and the *data-stream model*, as shown by Guha et al. ([19], Theorem 25): more precisely, they prove that any (multiplicative) approximation algorithm for a large class of functions of the distribution (functions that are invariant by relabeling of any two elements of the support) in the dual access model yields a space-efficient, $O(1)$-pass approximation algorithm for the same function in the data-stream model.

1.4 Our results and techniques

We focus here on four fundamental and pervasive problems in distribution testing, which are testing *uniformity*, *identity* to a known distribution D^*, *closeness* between two (unknown) distributions D_1, D_2, and finally *entropy and support size*. As usual in the distribution testing literature, the notion of distance we use is the *total variation distance* (or statistical distance), which is essentially the ℓ_1 distance between the probability distributions (see Sect. 2 for the formal definition). Testing closeness is thus the problem of deciding if two distributions are equal or far from each other in total variation distance; while tolerant testing aims at deciding whether they are sufficiently close versus far from each other.

As shown in Table 1, which summarizes our results and compares them to the corresponding bounds for the standard sampling-only (SAMP), evaluation-only (EVAL) and conditional sampling (COND) models, we indeed manage to bypass the aforementioned limitations of the sampling model, and give (often tight) algorithms with sample complexity either constant (with relation to n) or logarithmic, where a polynomial dependence was required in the standard setting.

Our main finding overall is that *both dual models allow testing algorithms to significantly outperform both* SAMP *and* COND *algorithms*, either with relation to the dependence on n or, for the latter, in $1/\varepsilon$; further, these testing algorithms are *significantly simpler*, both conceptually and in their analysis, and can often be made robust to some multiplicative noise in the evaluation oracle. Another key observation is that this new flexibility not only allows us to tell whether two distributions are close or far, but also to efficiently estimate their distance[4].

In more detail, we show that for the problem of testing equivalence between distributions, both our models allow to get rid of any dependence on n, with a (tight) sample complexity of $\Theta(1/\varepsilon)$. The upper bound is achieved by adapting an EVAL-only algorithm of [24] (for identity testing) to our setting, while the lower bound is obtained by designing a far-from-uniform instance which “defeats” simultaneously both oracles of our models. Turning to tolerant testing of equivalence, we describe algorithms whose sample complexity is again independent of n, in sharp contrast with the $n^{1-o(1)}$ lower bound of the standard sampling model. Moreover, we are able to show that, at least in the Dual access model, our quadratic dependence on ε is optimal. The same notable improvements apply to the query complexity of estimating the support size of the distribution, which becomes constant (with relation to n) in both of our access models – versus quasilinear if one only allows sampling.

As for the task of (additively) estimating the entropy of an arbitrary distribution, we give an algorithm whose sample complexity is only polylogarithmic in n, and show that this is tight in the Dual access model, up to the exponent of the logarithm. Once more, this is to be compared to the $n^{1-o(1)}$ lower bound for sampling.

While it is not clear, looking at these problems, whether the additional flexibility that the Cumulative Dual access grants over the Dual one can *unconditionally* yield strictly more sample-efficient testing algorithms, we do provide a separation between the two models in Sect. 4.2 by showing an exponential improvement in the query complexity for estimating the entropy of a distribution given the promise that

[3] A counting query is of the form “how many records in the database satisfy predicate χ?” – or, equivalently, “what is the probability that a random record drawn from the database satisfies χ?”.

[4] For details on the equivalence between tolerant testing and distance estimation, the reader is referred to [23].

Problem	SAMP	COND [10, 11]	EVAL	Dual	Cumulative Dual
Testing uniformity	$\Theta\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ [18, 5, 22]	$\tilde{O}\left(\frac{1}{\varepsilon^2}\right)$, $\Omega\left(\frac{1}{\varepsilon^2}\right)$	$O\left(\frac{1}{\varepsilon}\right)$ [24], $\Omega\left(\frac{1}{\varepsilon}\right)^*$	$\Theta\left(\frac{1}{\varepsilon}\right)$ (†)	$\Theta\left(\frac{1}{\varepsilon}\right)$ (†)
Testing $\equiv D^*$	$\tilde{\Theta}\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ [3, 22]	$\tilde{O}\left(\frac{1}{\varepsilon^4}\right)$			
Testing $D_1 \equiv D_2$	$\Theta\left((\max\left(\frac{N^{2/3}}{\varepsilon^{4/3}}, \frac{\sqrt{N}}{\varepsilon^2}\right)\right)$ [5, 28, 13]	$\tilde{O}\left(\frac{\log^5 n}{\varepsilon^4}\right)$	$\Omega\left(\frac{1}{\varepsilon}\right)^*$		
Tolerant uniformity	$O\left(\frac{1}{(\varepsilon_2-\varepsilon_1)^2}\frac{n}{\log n}\right)$ [27, 26] $\Omega\left(\frac{n}{\log n}\right)$ [27, 25]	$\tilde{O}\left(\frac{1}{(\varepsilon_2-\varepsilon_1)^{20}}\right)$	$\Omega\left(\frac{1}{(\varepsilon_2-\varepsilon_1)^2}\right)^*$	$\Theta\left(\frac{1}{(\varepsilon_2-\varepsilon_1)^2}\right)$ (†)	$O\left(\frac{1}{(\varepsilon_2-\varepsilon_1)^2}\right)$ (†)
Tolerant D^*	$\Omega\left(\frac{n}{\log n}\right)$ [27, 25]				
Tolerant D_1, D_2					
Estimating entropy to $\pm\Delta$	$\Theta\left(\frac{n}{\log n}\right)$ [27, 25]			$O\left(\frac{\log^2 \frac{n}{\Delta}}{\Delta^2}\right)$ (†), $\Omega(\log n)$	$O\left(\frac{\log^2 \frac{n}{\Delta}}{\Delta^2}\right)$ (†)
Estimating support size to $\pm\varepsilon n$	$\Theta\left(\frac{n}{\log n}\right)$ [27, 25]			$\Theta\left(\frac{1}{\varepsilon^2}\right)$	$O\left(\frac{1}{\varepsilon^2}\right)$

Table 1: Summary of results. (†) stands for "robust to multiplicative noise". The bounds with an asterisk are those which, in spite of being for different models, derive from the results of the last two columns.

the latter is (close to) monotone. This leads us to suspect that for the task of testing monotonicity (for which we have preliminary results), under a structural assumption on the distribution, or more generally for properties intrinsically related to the underlying total order of the domain, such a speedup holds. Moreover, we stress out the fact that our $\Omega\big(1/(\varepsilon_2-\varepsilon_1)^2\big)$ lower bound for tolerant identity testing does not apply to the Cumulative Dual setting.

One of the main techniques we use for algorithms in the dual model is a general approach[5] for estimating very efficiently any quantity of the form $\mathbb{E}_{i\sim D}\left[\Phi(i, D(i))\right]$, for any *bounded* function Φ. In particular, in the light of our lower bounds, this technique is both an intrinsic and defining feature of the Dual model, as it gives essentially tight upper bounds for the problems we consider.

On the other hand, for the task of proving lower bounds, we no longer can take advantage of the systematic characterizations known for the sampling model (see e.g. [2], Sect. 2.4.1). For this reason, we have to rely on reductions from known-to-be-hard problems (such as estimating the bias of a coin), or prove indistinguishability in a *customized* fashion.

1.5 Organization

After the relevant definitions and preliminaries in Sect. 2, we pursue by considering the first three problems of testing equivalence of distributions in Sect. 3, where we describe our testing upper and lower bounds. We then turn to the harder problem of *tolerant* testing. Finally, we tackle in Sect. 4 the task of performing entropy and support size estimation, and give for the latter matching upper and lower bounds.

2 Preliminaries

We consider discrete probability distributions over the subset of integers $[n] = \{1, \dots, n\}$. As aforementioned, the notion of distance we use between distributions D_1, D_2 is their *total variation distance*, defined as

$$\mathrm{d_{TV}}(D_1, D_2) \overset{\text{def}}{=} \max_{S\subseteq[n]} \left(D_1(S) - D_2(S)\right) = \frac{1}{2}\sum_{i\in[n]} \left|D_1(i) - D_2(i)\right| .$$

Recall that any property $\mathcal{P}$ can equivalently be seen as the subset of distributions that satisfy it; in particular, the distance $\mathrm{d_{TV}}(D, \mathcal{P})$ from some D to $\mathcal{P}$ is the minimum distance to any distribution in this subset, $\min_{D'\in\mathcal{P}} \mathrm{d_{TV}}(D, D')$.

Testing algorithms for distributions over $[n]$ are defined as follows[6]:

[5] We note that a similar method was utilized in [7], albeit in a less systematic way.

Definition 3. Fix any property $\mathcal{P}$ of distributions, and let ORACLE_D be an oracle providing some type of access to D. A *q-query testing algorithm for* $\mathcal{P}$ is a randomized algorithm $\mathcal{T}$ which takes as input n, $\varepsilon \in (0,1]$, as well as access to ORACLE_D. After making at most $q(\varepsilon, n)$ calls to the oracle, $\mathcal{T}$ outputs either ACCEPT or REJECT, such that the following holds:

- if $D \in \mathcal{P}$, $\mathcal{T}$ outputs ACCEPT with probability at least $2/3$;
- if $\mathrm{d_{TV}}(D, \mathcal{P}) \geq \varepsilon$, $\mathcal{T}$ outputs REJECT with probability at least $2/3$.

We shall also be interested in *tolerant* testers – roughly, algorithms robust to a relaxation of the first item above:

Definition 4. Fix property $\mathcal{P}$ and ORACLE_D as above. A *q-query tolerant testing algorithm for* $\mathcal{P}$ is a randomized algorithm $\mathcal{T}$ which takes as input n, $0 \leq \varepsilon_1 < \varepsilon_2 \leq 1$, as well as access to ORACLE_D. After making at most $q(\varepsilon_1, \varepsilon_2, n)$ calls to the oracle, $\mathcal{T}$ outputs either ACCEPT or REJECT, such that the following holds:

- if $\mathrm{d_{TV}}(D, \mathcal{P}) \leq \varepsilon_1$, $\mathcal{T}$ outputs ACCEPT with probability at least $2/3$;
- if $\mathrm{d_{TV}}(D, \mathcal{P}) \geq \varepsilon_2$, $\mathcal{T}$ outputs REJECT with probability at least $2/3$.

Observe in particular that if $\mathrm{d_{TV}}(D, \mathcal{P}) \in (0, \varepsilon)$ (resp. $\mathrm{d_{TV}}(D, \mathcal{P}) \in (\varepsilon_1, \varepsilon_2)$), the tester's output can be arbitrary. Furthermore, we stress that the two definitions above only deal with the query complexity, and not the running time. However, it is worth noting that while our lower bounds hold even for such computationally unbounded algorithms, all our upper bounds are achieved by testing algorithms whose running time is polynomial in the number of queries they make.

Remark 1. We will sometimes refer as a *multiplicatively noisy* EVAL_D (or similarly for CEVAL_D) to an evaluation oracle with takes an additional input parameter $\tau > 0$ and returns a value $\hat{d}_i$ within a multiplicative factor $(1+\tau)$ of the true $D(i)$. Note however that this notion of noisy oracle does not preserve the two-query simulation of a dual oracle by a cumulative dual one.

Finally, recall the following well-known result on distinguishing biased coins (which can for instance be derived from Eq. (2.15) and (2.16) of [1]), that shall come in handy in proving our lower bounds:

Fact 1. *Let $p \in [\eta, 1-\eta]$ for some fixed $\eta > 0$, and suppose $m \leq \frac{c_\eta}{\varepsilon^2}$, with c_η a sufficiently small constant and $\varepsilon < \eta$. Then,*

$$\mathrm{d_{TV}}(\mathrm{Bin}(m,p), \mathrm{Bin}(m,p+\varepsilon)) < \frac{1}{3}.$$

We shall make extensive use of Chernoff bounds; for completeness, we state them in Appendix A.

3 Uniformity and identity of distributions

3.1 Testing

In this section, we consider the three following testing problems, each of them a generalization of the previous:

Uniformity testing: given oracle access to D, decide whether $D = \mathcal{U}$ (the uniform distribution on $[n]$) or is far from it;

Identity testing: given oracle access to D and the full description of a fixed D^*, decide whether $D = D^*$ or is far from it;

Closeness testing: given independent oracle accesses to D_1, D_2 (both unknown), decide whether $D_1 = D_2$ or D_1, D_2 are far from each other.

[6] Note that, as standard in property testing, the threshold 2/3 is arbitrary: any $1-\delta$ confidence can be achieved at the cost of a multiplicative factor $\log(1/\delta)$ in the query complexity, by repeating the test and outputting the majority vote.

We begin by stating here two results from the literature that transpose straighforwardly in our setting. Observe that since the problem of testing closeness between two unknown distributions D_1, D_2 in particular encompasses the identity to known D^* testing (and a fortiori the uniformity testing) one, this upper bound automatically applies to these as well.

Theorem 1 (Theorem 24 from [24])**.** *In the query access model, there exists a tester for identity to a known distribution D^* with query complexity $O\big(\frac{1}{\varepsilon}\big)$.*

Note that the tester given in [24] is neither tolerant nor robust; however, it only uses query access. [10] later adapt this algorithm to give a tester for closeness between two unknown distributions, in a setting which can be seen as "relaxed" dual access model[7]:

Theorem 2 (Theorem 12 from [10])**.** *In the dual access model, there exists a tester for closeness between two unknown distributions D_1, D_2 with sample complexity $O\big(\frac{1}{\varepsilon}\big)$.*

It is worth noting that the algorithm in question is conceptually very simple – namely, it consists in drawing samples from both distributions and then querying the respective probability mass both distributions put on them, hoping to detect a violation.

Remark 2. As mentioned, the setting of the theorem is slightly more general than stated – indeed, it only assumes "approximate" query access to D_1, D_2 (in their terminology, this refers to an evaluation oracle that outputs, on query $x \in [n]$, a good *multiplicative* estimate of $D_i(x)$, for *most* of the points x).

Lower bound Getting more efficient testing seems unlikely – the dependence on $1/\varepsilon$ being "as good as it gets". The following result formalizes this, showing that indeed both Theorems 1 and 2 are tight, even for the least challenging task of testing uniformity:

Theorem 3 (Lower bound for dual oracle testers)**.** *In the dual access model, any tester for uniformity must have query complexity $\Omega\big(\frac{1}{\varepsilon}\big)$.*

Albeit the lower bound above applies only to the dual model, one can slightly adapt the proof to get the following improvement:

Theorem 4 (Lower bound for cumulative dual oracle testers)**.** *In the cumulative dual access model, any tester for uniformity must have sample complexity $\Omega\big(\frac{1}{\varepsilon}\big)$.*

Albeit the lower bound above applies only to the dual model, one can slightly adapt the construction to get the following improvement:

Theorem 5 (Lower bound for cumulative dual oracle testers)**.** *In the cumulative dual access model, any tester for uniformity must have sample complexity $\Omega\big(\frac{1}{\varepsilon}\big)$.*

Sketch. Theorem 5 directly implies Theorem 3, so we focus on the former. The high-level idea is to trick the algorithm by somehow "disabling" the additional flexibility coming from the oracles.

To do so, we start with a distribution that is far from uniform, but easy to recognize when given evaluation queries. We then shuffle its support randomly in such a way that (a) sampling will not, with overwhelming probability, reveal anything, while (b) evaluation queries essentially need to find a needle in a haystack. Note that the choice of the shuffling must be done carefully, as the tester has access to the cumulative distribution function of any no-instance D: in particular, using a random permutation will not work. Indeed, it is crucial for the cumulative distribution function to be as close as the linear function $x \in [n] \mapsto \frac{x}{n}$ as possible; meaning that the set of elements on which D differs from $\mathcal{U}$ had better be a consecutive "chunk" (otherwise, looking at the value of the cdf at a uniformly random point would give away the difference with uniform with non-negligible probability: such a point x is likely to have at least a "perturbed point" before *and* after it, so that $\sum_{i\leq x} D(x) \neq \frac{x}{n}$).

[7] In the sense that the evaluation oracle, being simulated via another type of oracle, is not only noisy but also allowed to err on a small set of points.

Fix any $\varepsilon \in (0, \frac{1}{2}]$; for $n \geq \frac{1}{\varepsilon}$, set $m \stackrel{\text{def}}{=} (1-\varepsilon)n - 1$, and consider testing a distribution D on $[n]$ which is either (a) the uniform distribution or (b) chosen uniformly at random amongst the family of distributions $(D_r)_{0\leq r\leq m}$, defined this way: for any offset $0 \leq r < m$, D_r is obtained as follows:

1. Set $D(1) = \varepsilon + \frac{1}{n}$, $D(2) = \cdots = D(\varepsilon n + 1) = 0$, and $D(k) = \frac{1}{n}$ for the remaining $m = (1-\varepsilon)n - 1$ points;
2. Shift the whole support (modulo n) by adding r.

At a high-level, what this does is keeping the "chunk" on which the cdf of the no-instance grouped together, and just place it at a uniformly random position; outside this interval, the cdf's are exactly the same, and the only way to detect a difference with CEVAL is to make a query in the "chunk". Furthermore, it is not hard to see that any no-instance distribution will be exactly ε-far from uniform, so that any tester $\mathcal{T}$ must distinguish between cases (a) and (b) with probability at least 2/3.

Suppose by contradiction that there exists a tester $\mathcal{T}$ making $q = o\big(\frac{1}{\varepsilon}\big)$ queries (without loss of generality, we can further assume $\mathcal{T}$ makes exactly q queries; and that for any SAMP query, the tester also gets "for free" the result of an evaluation query on the sample). Given dual access to a $D = D_r$ generated as in case (b), observe first that, since the outputs of the sample queries are independent of the results of the evaluation queries, one can assume that some evaluation queries are performed first, followed by some sample queries, before further evaluation queries (where the evaluation points may depend arbitrarily on the sample query results) are made. That is, we subdivide the queries in 3: first, q_1 consecutive EVAL queries, then a sequence of q_2 SAMP queries, and finally q_3 EVAL queries. Define the following "bad" events:

- E_1: one of the first q_1 evaluation queries falls outside the set $G \stackrel{\text{def}}{=} \{\varepsilon n + 2 + r, \ldots, n + r\} \mod n$;
- E_2: one of the q_2 sampling queries returns a sample outside G, conditioned on $\overline{E_1}$;
- E_3: one of the q_3 evaluation queries is on a point outside G, conditioned on $\overline{E_1} \cap \overline{E_2}$.

It is clear that, conditioned on $\overline{E_1} \cap \overline{E_2} \cap \overline{E_3}$, all the tester sees is exactly what its view would have been in case (a) (probabilities equal to $\frac{1}{n}$ for any EVAL query, and uniform sample from G for any SAMP one). It is thus sufficient to show that $\Pr\big[\, \overline{E_1} \cap \overline{E_2} \cap \overline{E_3} \,\big] = 1 - o(1)$.

- As r is chosen uniformly at random, $\Pr[\, E_1 \,] \leq q_1 \frac{n-m}{n} = q_1(\varepsilon + \frac{1}{n})$;
- since $D(G) = \frac{m}{n} = 1 - \varepsilon - \frac{1}{n} \geq 1 - 2\varepsilon$, $\Pr[\, E_2 \,] \leq 1 - (1-2\varepsilon)^{q_2}$;
- finally, $\Pr[\, E_3 \,] \leq q_3(\varepsilon + \frac{1}{n})$;

we therefore have $\Pr[\, E_1 \cup E_2 \cup E_3 \,] \leq (q_1 + q_3)(\varepsilon + \frac{1}{n}) + 1 - (1-2\varepsilon)^{q_2} = O(q\varepsilon) = o(1)$, as claimed. □

3.2 Tolerant testing

In this section, we describe tolerant testing algorithms for the three problems of uniformity, identity and closeness; note that by a standard reduction (see Parnas et al. ([23], Section 3.1), this is equivalent to estimating the distance between the corresponding distributions. As hinted in the introduction, our algorithm relies on a general estimation approach that will be illustrated further in Section 4, and which constitutes a fundamental feature of the dual oracle: namely, the ability to estimate cheaply quantities of the form $\mathbb{E}_{i\sim D}\left[\Phi(i, D(i))\right]$ for any *bounded* function Φ.

Theorem 6. *In the dual access model, there exists a tolerant tester for uniformity with query complexity* $O\Big(\frac{1}{(\varepsilon_2-\varepsilon_1)^2}\Big)$.

Proof. We describe such a tester $\mathcal{T}$; as it will start by estimating the quantity $2\mathrm{d}_{\mathrm{TV}}(D,\mathcal{U})$ up to some additive $\gamma \stackrel{\text{def}}{=} \varepsilon_2 - \varepsilon_1$ (and then accept if and only if its estimate $\hat{d}$ is at most $2\varepsilon_1 + \gamma = \varepsilon_1 + \varepsilon_2$).

In order to approximate this quantity, observe that[8]

$$\mathrm{d_{TV}}(D,\mathcal{U}) = \frac{1}{2}\sum_{i\in[n]}\left|D(i)-\frac{1}{n}\right| = \sum_{i:D(i)>\frac{1}{n}}\left(D(i)-\frac{1}{n}\right) = \sum_{i:D(i)>\frac{1}{n}}\left(1-\frac{1}{nD(i)}\right)\cdot D(i)$$

$$= \mathbb{E}_{i\sim D}\left[\left(1-\frac{1}{nD(i)}\right)\mathbb{1}_{\left\{D(i)>\frac{1}{n}\right\}}\right] \tag{1}$$

where $\mathbb{1}_E$ for the indicator function of set (or event) E; thus, $\mathcal{T}$ only has to do get an empirical estimate of this expected value, which can be done by taking $m = O\big(1/(\varepsilon_2-\varepsilon_1)^2\big)$ samples s_i from D, querying $D(s_i)$ and computing $X_i = \left(1-\frac{1}{nD(s_i)}\right)\mathbb{1}_{\left\{D(s_i)>\frac{1}{n}\right\}}$ (cf. Alg. 1).

Algorithm 1 Tester $\mathcal{T}$: Estimate-L_1

```
Require: SAMP_D and EVAL_D oracle access, parameters $0 \leq \varepsilon_1 < \varepsilon_2$
  Set $m \overset{\text{def}}{=} \Theta\left(\frac{1}{\gamma^2}\right)$, where $\gamma \overset{\text{def}}{=} \frac{\varepsilon_2-\varepsilon_1}{2}$.
  Draw $s_1, \dots, s_m$ from $D$
  for $i = 1$ to $m$ do
      With EVAL, get $X_i \overset{\text{def}}{=} \left(1-\frac{1}{nD(s_i)}\right)\mathbb{1}_{\left\{D(s_i)>\frac{1}{n}\right\}}$
  end for
  Compute $\hat{d} \overset{\text{def}}{=} \frac{1}{m}\sum_{i=1}^m X_i$.
  if $\hat{d} \leq \frac{\varepsilon_1+\varepsilon_2}{2}$ then
      return ACCEPT
  else
      return REJECT
  end if
```

Analysis Define the random variable X_i as above; from Eq.(1), we can write its expectation as

$$\mathbb{E}[X_i] = \sum_{k=1}^n D(k)\left|1-\frac{1}{nD(k)}\right|\mathbb{1}_{\left\{D(k)>\frac{1}{n}\right\}} = \mathrm{d_{TV}}(D,\mathcal{U}). \tag{2}$$

Since the X_i's are independent and take value in $[0,1]$, an additive Chernoff bound ensures that

$$\Pr\left[\left|\hat{d}-\mathrm{d_{TV}}(D,\mathcal{U})\right| \geq \gamma\right] \leq 2e^{-2\gamma^2 m} \tag{3}$$

which is at most $1/3$ by our choice of m. Conditioning from now on on the event $\left|\hat{d}-\mathrm{d_{TV}}(D,\mathcal{U})\right| < \gamma$:

- if $\mathrm{d_{TV}}(D,\mathcal{U}) \leq \varepsilon_1$, then $\hat{d} \leq \varepsilon_1+\gamma$, and $\mathcal{T}$ outputs ACCEPT;
- if $\mathrm{d_{TV}}(D,\mathcal{U}) > \varepsilon_2$, then $\hat{d} > \varepsilon_2-\gamma$, and $\mathcal{T}$ outputs REJECT.

Furthermore, the algorithm makes m SAMP queries, and m EVAL queries. □

Remark 3. Note that we can also do it with EVAL queries only (same query complexity), by internally drawing uniform samples: indeed,

$$2\mathrm{d_{TV}}(D,\mathcal{U}) = \sum_{i\in[n]}\left|D(i)-\frac{1}{n}\right| = \sum_{i\in[n]}\left|nD(i)-1\right|\cdot\frac{1}{n} = 2\mathbb{E}_{x\sim\mathcal{U}}\left[\left|nD(x)-1\right|\mathbb{1}_{\left\{\frac{1}{n}>D(x)\right\}}\right]$$

This also applies to the first corollary below, as long as the known distribution is efficiently samplable by the algorithm.

[8] Note that dividing by $D(i)$ is "legal", since if $D(i) = 0$ for some $i \in [n]$, this point will never be sampled, and thus no division by 0 will ever occur.

Indeed, the proof above can be easily extended to other distributions than uniform, and even to the case of two unknown distributions:

Corollary 1. *In the dual access model, there exists a tolerant tester for identity to a known distribution with query complexity* $O\Big(\frac{1}{(\varepsilon_2-\varepsilon_1)^2}\Big)$.

Corollary 2. *In the dual access model, there exists a tolerant tester for closeness between two unknown distributions with query complexity* $O\Big(\frac{1}{(\varepsilon_2-\varepsilon_1)^2}\Big)$. *As noted in the next subsection, this is optimal (up to constants).*

Interestingly, this tester can be made robust to multiplicative noise, i.e. can be shown to work even when the answers to the EVAL queries are only accurate up to a factor $(1+\gamma)$ for $\gamma > 0$: it suffices to set $\gamma = \varepsilon/2$, getting on each point $\hat{D}(i) \in [(1+\gamma)^{-1}, 1+\gamma]D(i)$, and work with $X_i = \Big(1 - D^*(s_i)/\hat{D}(s_i)\Big)\mathbb{1}_{\{\hat{D}(s_i)>D^*(s_i)\}}$ and estimate the expectation up to $\pm\gamma$ (or, for closeness between two unknown distributions, setting $\gamma = \varepsilon/4$).

3.2.1 Lower bound

In this subsection, we show that the upper bounds of Lemma 6, Corollaries 1 and 2 are tight.

Theorem 7. *In the dual access model, performing* $(\varepsilon_1, \varepsilon_2)$*-testing for uniformity requires sample complexity* $\Omega\Big(\frac{1}{(\varepsilon_2-\varepsilon_1)^2}\Big)$ *(the bound holds even when only asking* ε_1 *to be* $\Omega(1)$*).*

Proof. The overall idea lies on a reduction from distinguishing between two types of biased coins to tolerant testing for uniformity. In more detail, given access to samples from a fixed coin (promised to be of one of these two types), we define a probability distribution as follows: the domain $[n]$ is randomly partitioned into $K = 1/\varepsilon^2$ pairs of buckets, each bucket having same number of elements; the distribution is uniform within each bucket, and the two buckets of each pair are balanced to have total weight $2/K$. However, within each pair of buckets (A, B), the probability mass is divided according to a coin toss (performed "on-the-fly" when a query is made by the tolerant tester), so that either (a) $D(A) = (1+\alpha)/K$ and $D(B) = (1-\alpha)/K$, or (b) $D(A) = D(B) = 1/K$. Depending on whether the coin used for this choice is fair or $(\frac{1}{2}+\varepsilon)$)biased, the resulting distribution will (with high probability) have different distance from uniformity – sufficiently for a tolerant tester to distinguish between the two cases.

Construction We start by defining the instances of distributions we shall consider. Fix any $\varepsilon \in (0, \frac{1}{100})$; without loss of generality, assume n is even, and $n \gg 1/\varepsilon$. Define $\alpha = 2/(1+\varepsilon)$, $K = 1/\varepsilon^2$, $p^+ = (1+\varepsilon)/2$ and $p^- = (1+20\varepsilon)/2$, and consider the family of distributions $\mathcal{D}^+$ (resp. $\mathcal{D}^-$) defined by the following construction:

- pick uniformly at random a partition[9] of $[n]$ in $2K$ sets of size $n/(2K)$ $A_1, \dots, A_K, B_1, \dots, B_K$;
- for all $k \in [K]$, draw independently at random $X_k \sim \mathrm{Bern}(p^+)$ (resp. $X_k \sim \mathrm{Bern}(p^-)$), and set for all $x \in A_k$, $y \in B_k$

$$D^+(x) = \begin{cases} \frac{1+\alpha}{n} & \text{if } X_i = 1 \\ \frac{1}{n} & \text{o.w.} \end{cases} \qquad \text{and} \qquad D^+(y) = \begin{cases} \frac{1-\alpha}{n} & \text{if } X_i = 1 \\ \frac{1}{n} & \text{o.w.} \end{cases}$$

(the pairing between A_k and B_k ensures the final measure indeed sums to one). Regardless of the choice of the initial partition, but with fluctuations over the random coin tosses $X_1, \dots, X_k$, we have that the total variation distance between a distribution $D^+ \in \mathcal{D}^+$ (resp. $D^- \in \mathcal{D}^-$) and uniform is on expectation what

[9] For convenience, it will be easier to think of the A_i's and B_i's as consecutive intervals, the first ones covering $[\frac{n}{2}]$ while the former cover $[n] \setminus [\frac{n}{2}]$ (see Fig. 1).

we aimed for:

$$\mathbb{E}\big[\mathrm{d_{TV}}\big(D^+,\mathcal{U}\big)\big] = \frac{1}{2}\cdot 2\cdot\sum_{k=1}^{K}\frac{n}{2K}\cdot\frac{\alpha}{n}p^+ = \frac{1}{2}\alpha p^+ = \frac{1}{2}$$
$$\mathbb{E}\big[\mathrm{d_{TV}}\big(D^-,\mathcal{U}\big)\big] = \frac{1}{2}p^-\alpha = \frac{1+20\varepsilon}{1+\varepsilon}\cdot\frac{1}{2} > \frac{1}{2}+7\varepsilon$$

and with an additive Chernoff bound on the sum of $1/\varepsilon^2$ i.i.d. choices for the X_k's, we have that for (D^+,D^-): for any choice of the initial partition $\Pi = (A_k,B_k)_{k\in[K]}$, with probability at least 99/100,

$$\mathrm{d_{TV}}\big(D^+_\Pi,\mathcal{U}\big) < \frac{1}{2}+3\varepsilon$$
$$\mathrm{d_{TV}}\big(D^-_\Pi,\mathcal{U}\big) > \frac{1}{2}+4\varepsilon$$

where by $D^\pm_\Pi$ we denote the distribution defined as above, but fixing the partition for the initial step to be Π. We will further implicitly condition on this event happening; any tester for uniformity called with $(\varepsilon',\varepsilon'+c\varepsilon)$ must therefore distinguish between D^+ and D^-. Suppose we have such a tester $\mathcal{T}$, with (without loss of generality) exact sample complexity $q = q(\varepsilon) = o\big(\frac{1}{\varepsilon^2}\big)$.

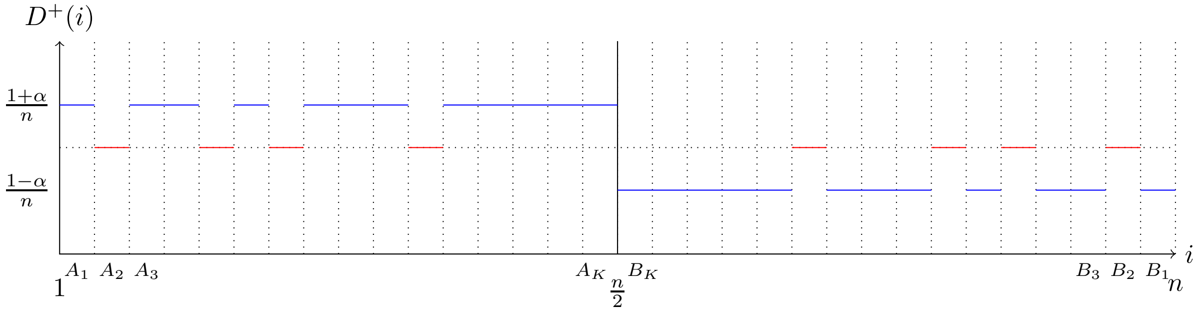

Figure 1: The yes-instance D^+ (for a fixed Π, taken to be consecutive intervals).

Reduction We will reduce the problem of distinguishing between (a) a $\frac{p^+\alpha}{2}$ and (b) a $\frac{p^-\alpha}{2}$ biased coin to telling D^+ and D^- apart.

Given $\mathsf{SAMP}_{\text{coin}}$ access to i.i.d. coin tosses coming either from one of those two situations, define a distinguisher $\mathcal{A}$ as follows:

- choose uniformly at random a partition $\Pi = (A^0_k, A^1_k)_{k\in[K]}$ of $[n]$; for convenience, for any $i\in[n]$, we shall write $\pi(i)$ for the index $k\in[K]$ such that $i\in A^0_k\cup A^1_k$, and $\varsigma(i)\in\{0,1\}$ for the part in which it belongs – so that $i\in A^{\varsigma(i)}_{\pi(i)}$ for all i;
- run $\mathcal{T}$, maintaining a set C of triples[10] (k, D^0_k, D^1_k) (initially empty), containing the information about the (A^0_k, A^1_k) for which the probabilities have already be decided;
- EVAL: whenever asked an evaluation query on some $i\in[n]$:
  - if $\pi(i)\in C$, return $D^{\varsigma(i)}_{\pi(i)}$;
  - otherwise, let $k=\pi(i)$; ask a fresh sample b_k from $\mathsf{SAMP}_{\text{coin}}$ and draw a uniform random bit b'_k; set

$$(D^0_k, D^1_k) = \begin{cases} (\frac{1}{n},\frac{1}{n}) & \text{if } b_k = 0\\ (\frac{1+\alpha}{n},\frac{1-\alpha}{n}) & \text{if } b_k=1, b'_k=1\\ (\frac{1-\alpha}{n},\frac{1+\alpha}{n}) & \text{if } b_k=1, b'_k=0\end{cases} \qquad \text{(“Choosing the profile”)}$$

[10] Abusing the notation, we will sometimes write “$k\in C$” for “there is a triple in C with first component k”.

then add (k, D_k^0, D_k^1) to C; and return $D_k^{c(i)}$.

- SAMP: whenever asked a sample: let $\gamma = \frac{n}{2K}\sum_{k\in C} d_k$ the current probability mass of the "committed points"; observe that the distribution D_C induced by the d_k's on $\{\, i \in [n] \,:\, \pi(i) \in C \,\}$ is fully known by $\mathcal{A}$;
  - with probability γ, $\mathcal{A}$ draws $i \sim D_C$ and returns it;
  - otherwise, $\mathcal{A}$ draws $k \sim \mathcal{U}([K] \setminus C)$. As before, it gets b_k from $\mathsf{SAMP}_{\text{coin}}$ and a uniform random bit b'_k; gets (D_k^0, D_k^1) as in the EVAL case, commits to it as above by (k, D_k^0, D_k^1) to C. Finally, it draws a random sample i from the piecewise constant distribution induced by (D_k^0, D_k^1) on $A_k^0 \cup A_k^1$, where each $j \in A_k^0$ (resp. A_k^1) has equal probability mass $D_k^0 \cdot \frac{n}{2K}$ (resp. $D_k^1 \cdot \frac{n}{2K}$), and returns i.

Observe that $\mathcal{A}$ makes at most q queries to $\mathsf{SAMP}_{\text{coin}}$; provided we can argue that $\mathcal{A}$ answers $\mathcal{T}$'s queries consistently to what a corresponding $D^{\pm}$ (depending on whether we are in case (a) or (b)) would look like, we can conclude.

This is the case, as (i) $\mathcal{A}$ is always consistent with what its previous answers induce on the distribution (because of the maintaining of the set C); (ii) any EVAL query on a new point exactly simulates the "on-the-fly" construction of a $D^{\pm}$; and any SAMP query is either consistent with the part of $D^{\pm}$ already built, or in case of a new point gets a sample exactly distributed according to the $D^{\pm}$ built "on-the-fly"; this is because in any $D^{\pm}$, every $A_k \cup B_k$ has same probability mass $1/(2K)$; therefore, in order to get one sample, tossing K i.i.d. coins to decide the "profiles" of *every* $A_k \cup B_k$ before sampling from the overall support $[n]$ is equivalent to first choosing uniformly at random a particular $S = A_k \cup B_k$, tossing one coin to decide *only its particular profile*, and then drawing a point accordingly from S.

In other terms, $\mathcal{A}$ will distinguish, with only $o(1/\varepsilon^2)$ i.i.d. samples, between cases (a) ($\frac{1}{2}$-biased coin) and (b) ($\frac{1}{2} + \Omega(\varepsilon)$)-biased coin with probability at least $6/10$ – task which, for ε sufficiently small, is known to require $\Omega(1/\varepsilon^2)$ samples (cf. Fact 1), thus leading to a contradiction. □

4 Entropy and support size

4.1 Additive and multiplicative estimations of entropy

In this section, we describe simple algorithms to perform additive and multiplicative estimation (which in turns straightforwardly implies tolerant testing) of the *entropy* $H(D)$ of the unknown distribution D, defined as

$$H(D) \overset{\text{def}}{=} -\sum_{i\in[n]} D(i)\log D(i) \in [0, \log n]$$

We remark that Batu et al. ([7], Theorem 14) gives a similar algorithm, based on essentially the same approach but relying on a Chebyshev bound, yielding a $(1+\gamma)$-multiplicative approximation algorithm for entropy with sample complexity $O\big((1+\gamma)^2\log^2 n/\gamma^2 h^2\big)$, given a lower bound $h > 0$ on $H(D)$.

Guha et al. ([19], Theorem 5.2) then refined their result, using as above a threshold for the estimation along with a multiplicative Chernoff bound to get the sample complexity down to $O\big(\log n/\gamma^2 h\big)$ – thus matching the $\Omega(\log n/\gamma(2+\gamma)h)$ lower bound of [7] (Theorem 18); we recall their results for multiplicative estimation of the entropy below[11].

Theorem 8 (Upper bound [[19], Theorem 5.2]). *Fix $\gamma > 0$. In the dual access model, there exists an algorithm that, given a parameter $h > 0$ and the promise that $H(D) \geq h$, estimates the entropy within a multiplicative $(1+\gamma)$ factor, with sample complexity $\Theta\Big(\frac{\log n}{\gamma^2 h}\Big)$.*

[11] In particular, note that translating their lower bound for additive estimation implies that the dependence on n of our algorithm is tight.

Theorem 9 (Lower bound [[7], Theorem 18])**.** *Fix $\gamma > 0$. In the dual access model, any algorithm that, given a parameter $h > 0$ and the promise that $H(D) = \Omega(h)$, estimates the entropy within a multiplicative $(1+\gamma)$ factor must have sample complexity $\Omega\left(\frac{\log n}{\gamma(2+\gamma)h}\right)$.*

Observe that the additive bound we give (based on a different cutoff threshold), however, still performs better in many cases, e.g. $\Delta = \gamma h > 1$ and $h > 1$; and does not require any *a priori* knowledge on a lower bound $h > 0$. Moreover, we believe that this constitutes a good illustration of the more general technique used, and a good example of what the dual model allows: approximation of quantities of the form $\mathbb{E}_{i\sim D}\left[\Phi(i, D(i))\right]$, where Φ is any *bounded* function of both an element of the domain and its probability mass under the distribution D.

Additive estimate The overall idea is to observe that for a distribution D, the entropy $H(D)$ can be rewritten as

$$H(D) = \sum_{x\in[n]} D(x)\log\frac{1}{D(x)} = \mathbb{E}_{x\sim D}\left[\log\frac{1}{D(x)}\right] \tag{4}$$

The quantity $\log\frac{1}{D(x)}$ cannot be easily upperbounded, which we need for concentration results. However, recalling that the function $x \mapsto x\log(1/x)$ is increasing for $x \in (0, \frac{1}{e})$ (and has limit 0 when $x \to 0^+$), one can refine the above identity as follows: for any *cutoff threshold* $\tau \in (0, \frac{1}{e})$, write

$$H(D) = \sum_{x:D(x)\geq\tau} D(x)\log\frac{1}{D(x)} + \sum_{x:D(x)<\tau} D(x)\log\frac{1}{D(x)} \tag{5}$$

so that

$$\begin{aligned}
H(D) \geq \sum_{x:D(x)\geq\tau} D(x)\log\frac{1}{D(x)} &\geq H(D) - \sum_{x:D(x)<\tau} D(x)\log\frac{1}{D(x)} \\
&\geq H(D) - n\cdot\tau\log\frac{1}{\tau}
\end{aligned}$$

Without loss of generality, assume $\frac{\Delta}{n} < \frac{1}{2}$. Fix $\tau \overset{\text{def}}{=} \frac{\frac{\Delta}{n}}{10\log\frac{n}{\Delta}}$, so that $n\cdot\tau\log\frac{1}{\tau} \leq \frac{\Delta}{2}$; and set

$$\varphi\colon y \mapsto \log\frac{1}{y}\mathbb{1}_{\{y\geq\tau\}}$$

Then, the above discussion gives us

$$H(D) \geq \mathbb{E}_{x\sim D}[\varphi(D(x))] \geq H(D) - \frac{\Delta}{2} \tag{6}$$

and getting an additive $\Delta/2$-approximation of $\mathbb{E}_{x\sim D}[\varphi(D(x))]$ is enough for estimating $H(D)$ within $\pm\Delta$; further, we now have

$$0 \leq \varphi(D(x)) \leq \log\frac{1}{\tau} \sim \log\frac{n}{\Delta} \text{ a.s.} \tag{7}$$

so using an additive Chernoff bound, taking $m = \Theta\left(\frac{\log^2\frac{n}{\Delta}}{\Delta^2}\right)$ samples $x_1, \dots, x_m$ from SAMP_D and computing the quantities $\varphi(D(x_i))$ using EVAL_D implies

$$\Pr\left[\left|\frac{1}{m}\sum_{i=1}^{m}\varphi(D(x_i)) - \mathbb{E}_{x\sim D}[\varphi(D(x))\right|] \geq \frac{\Delta}{2}\right] \leq 2e^{-\frac{\Delta^2 m}{\log^2\frac{1}{\tau}}} \leq \frac{1}{3}$$

This leads to the following theorem:

Theorem 10. *In the dual access model, there exists an algorithm estimating the entropy up to an additive Δ, with sample complexity $\Theta\Big(\frac{\log^2 \frac{n}{\Delta}}{\Delta^2}\Big)$.*

or, in terms of tolerant testing:

Corollary 3. *In the dual access model, there exists an (Δ_1, Δ_2)-tolerant tester for entropy with sample complexity $\tilde{\Theta}\Big(\frac{\log^2 n}{(\Delta_1-\Delta_2)^2}\Big)$.*

Proof. We describe such a $\mathcal{T}$ in Algorithm 2; the claimed query complexity is straighforward. □

Algorithm 2 Tester $\mathcal{T}$: Estimate-Entropy

```
Require: $\mathsf{SAMP}_D$ and $\mathsf{EVAL}_D$ oracle access, parameters $0 \leq \Delta \leq \frac{n}{2}$
Ensure: Outputs $\hat{H}$ s.t. w.p. at least 2/3, $\hat{H} \in [H(D) - \Delta, H(D) + \Delta/2]$
  Set $\tau \overset{\text{def}}{=} \frac{\frac{\Delta}{n}}{10 \log \frac{n}{\Delta}}$ and $m = \lceil \frac{\ln 6}{\Delta^2} \log^2 \frac{1}{\tau} \rceil$.
  Draw $s_1, \dots, s_m$ from $D$
  for $i = 1$ to $m$ do
      With $\mathsf{EVAL}$, get $X_i \overset{\text{def}}{=} \log \frac{1}{D(s_i)} \mathbb{1}_{\{D(s_i) \geq \tau\}}$
  end for
  return $\hat{H} \overset{\text{def}}{=} \frac{1}{m} \sum_{i=1}^m X_i$
```

Remark 4. The tester above can easily be adapted to be made multiplicatively robust; indeed, it is enough that the EVAL oracle only provides $(1+\gamma)$-accurate estimates $\hat{D}(i)$ of the probabilities $D(i)$, where γ is chosen to be $\gamma \overset{\text{def}}{=} \min(2^{\Delta/3} - 1, 1)$ so that the algorithm will output with high probability an additive $(\Delta/2)$-estimate of a quantity

$$H(D) \geq \mathbb{E}_{x \sim D}\left[\widehat{\varphi}(x)\right] \geq \sum_{x: D(x) \geq (1+\gamma)\tau} D(x) \log \frac{1}{D(x)} - \log(1+\gamma) \geq H(D) + n \cdot \underbrace{(1+\gamma)\tau \log(1+\gamma)\tau}_{\geq -2\tau \log \frac{1}{2\tau}} - \frac{\Delta}{3}$$

and taking for instance $\tau \overset{\text{def}}{=} \frac{\frac{\Delta}{n}}{30 \log \frac{n}{\Delta}}$ ensures the right-hand-side is at least $H(D) - \frac{\Delta}{6} - \frac{\Delta}{3} = H(D) - \frac{\Delta}{2}$.

4.2 Additive estimation of entropy for monotone distributions

In the previous section, we saw how to obtain an additive estimate of the entropy of the unknown distribution, using essentially $O\big(\log^2 n\big)$ sampling and evaluation queries; moreover, this dependence on n is optimal. However, one may wonder if, by taking advantage of *cumulative* queries, it becomes possible to obtain a better query complexity. We partially answer this question, focusing on a particular class of distributions for which the cumulative dual query access seems particularly well-suited: namely the class of *monotone* distributions[12].

Before describing how this assumption can be leveraged to obtain an exponential improvement in the sample complexity for cumulative dual query algorithms, we first show that given only *dual* access to a distribution promised to be $o(1)$-close to monotone, no such speedup can hold. By establishing (see Remark 6) that the savings obtained for (close to) monotone distributions are only possible with cumulative dual access, this will yield a separation between the two oracles, proving the latter is strictly more powerful.

4.2.1 Lower bound for dual oracles

Theorem 11. *In the dual access model, any algorithm that estimates the entropy of distributions $O(1/\log n)$-close to monotone even to an additive constant must make $\Omega(\log n)$ queries to the oracle.*

[12] Recall that a distribution D over a totally ordered domain is said to be monotone if for all $i \in [n-1]$ $D(i) \geq D(i+1)$

Proof. We will define two families of distributions, $\mathcal{D}_1$ and $\mathcal{D}_2$, such that for any two D_1, D_2 drawn uniformly at random from $\mathcal{D}_1$ and $\mathcal{D}_2$:

1. D_1 and D_2 are $(2/\log n)$-close to monotone;
2. $|H(D_1) - H(D_2)| = 1/4$;
3. no algorithm making $o(\log n)$ queries to a dual oracle can distinguish between D_1 and D_2 with constant probability.

In more detail, the families are defined by the following process: for $K_n \overset{\text{def}}{=} n^{1/4}$, $\ell_n \overset{\text{def}}{=} \log n$ and $\gamma_n \overset{\text{def}}{=} 1/\log n$,

- Draw a subset $S \subset \{2, \dots, n\}$ of size ℓ_n uniformly at random;
- Set $D_1(1) = 1 - \gamma_n$, and $D_1(i) = \gamma_n/\ell_n = 1/\log^2 n$ for all $i \in S$.

(D_2 is obtained similarly, but with a subset S of size $K_n\ell_n = n^{1/4}\log n$ and $D_2(i) = \gamma_n/(\ell_n K_n)$) Roughly, both distributions have a very heavy first element (whose role is to "disable" sampling queries by hogging them with high probability), and then a random subset of size respectively logarithmic or polynomial, on which they are uniform. To determine whether a distribution is drawn from $\mathcal{D}_1$ or $\mathcal{D}_2$, intuitively a testing algorithm has to find a point $i > 1$ with non-zero mass – and making a query on this point then gives away the type of distribution. However, since sampling queries will almost always return the very first element, finding such a $i > 1$ amounts to finding a needle in a haystack (without sampling) or to sampling many times (to get a non-trivial element) – and thus requires many queries. Before formalizing this intuition, we prove the first two items of the above claims:

Distance to monotonicity By moving all elements of S at the beginning of the support (points $2, \dots, |S|+1$), the distribution would be monotone; so in particular

$$\mathrm{d_{TV}}(D_i, \textsc{Monotone}) \leq \frac{1}{2} \cdot 2\,|S| \cdot \frac{\gamma_n}{|S|} = 2\gamma_n = \frac{2}{\log n}, \qquad i \in \{1,2\}$$

Difference of entropy By their definition, for any two D_1, D_2, we have

$$|H(D_1) - H(D_2)| = \left| \sum_{i=2}^{n} D_1(i)\log D_1(i) - \sum_{i=2}^{n} D_2(i)\log D_2(i) \right| = \gamma_n \log K_n = \frac{1}{4}.$$

We now turn to the main item, the indistinguishability:

Telling D_1 and D_2 apart Assume we have an algorithm $\mathcal{T}$, which can estimate entropy of distributions that are $O(1/\log n)$-close to monotone up to an additive $1/3$ making $q(n) = o(\log n)$ queries; we claim that $\mathcal{T}$ cannot be correct with probability $2/3$. As argued before, we can further assume without loss of generality that $\mathcal{T}$ makes exactly $2q$ queries, q sampling queries and q evaluation ones; and that for any SAMP query, it gets "for free" the result of an evaluation query on the sample. Finally, and as the sampling queries are by definition non-adaptive, this also allows us to assume that $\mathcal{T}$ starts by making its q SAMP queries.

Let B_1 be the event that one of the q first queries results in sampling an element $i > 1$ (that is, B_1 is the event that the "hogging element" fails its role). Clearly, B_1 has same probability no matter with of the two families the unknown distribution belongs to, and

$$\Pr[\, B_1 \,] = 1 - (1-\gamma_n)^q = 1 - 2^{q\log(1-1/\log n)} \leq 1 - 2^{-2q/\log n} = O(q/\log n) = o(1) \tag{8}$$

so with probability $1 - o(1)$, $\bar{B}_1$ holds. We further condition on this: i.e., the testing algorithm only saw the first element (which does not convey any information) after the sampling stage.

The situation is now as follows: unless one of its queries hits one of the relevant points in the uniform set S (call this event B_2), the algorithm will see in both case the same thing – a sequence of points with probability zero. But by construction, in both cases, the probability over the (uniform) choice of the support S to hit a relevant point with one query is either $\ell_n/(n-1) = \log n/(n-1)$ or $K_n\ell_n/(n-1) = n^{1/4}\log n/(n-1)$; so that the probability of finding such a point in n queries is at most

$$\Pr[\,B_2\,] \leq 1 - \left(1 - \frac{K_n\ell_n}{n-1}\right)^q = O\left(\frac{q\log n}{n^{3/4}}\right) = o(1) \tag{9}$$

Conditioning on $\bar{B}_1 \cup \bar{B}_2$, we get that $\mathcal{T}$ sees exactly the same transcript if the distribution is drawn from $\mathcal{D}_1$ or $\mathcal{D}_2$; so overall, with probability $1 - o(1)$ it cannot distinguish between the two cases – contradicting the assumption. □

4.2.2 Upper bound: exponential speedup for cumulative dual oracles

We now establish the positive result in the case of algorithms given cumulative dual query access. Note that Batu et al. [7] already consider the problem of getting a (multiplicative) estimate of the entropy of D, under the assumption that the distribution is monotone; and describe (both in the evaluation-only and sample-only models) polylog(n)-query algorithms for this task, which work by recursively splitting the domain in a suitable fashion to get a partition into near uniform and negligible intervals.

The main insight here (in addition to the mere fact that we allow ourself a stronger type of access to D) is to use, instead of an *ad hoc* partition of the domain, a specific one tailored for monotone distributions, introduced by Birgé [8] – and which crucially *does not depend on the distribution itself.*

Definition 5 (Oblivious decomposition)**.** Given a parameter $\varepsilon > 0$, the corresponding *oblivious decomposition of* $[n]$ is the partition $\mathcal{I}_\varepsilon = (I_1, \dots, I_\ell)$, where $\ell = \left\lceil \frac{\log(\varepsilon n+1)}{\varepsilon} \right\rceil = \Theta\left(\frac{\log n}{\varepsilon}\right)$ and $|I_{k+1}| = (1+\varepsilon)\,|I_k|$, $1 \leq k < \ell$.

For a distribution D and parameter ε, define $\bar{D}_\varepsilon$ to be the *flattened distribution* with relation to the oblivious decomposition $\mathcal{I}_\varepsilon$:

$$\forall k \in [\ell], \forall i \in I_k, \quad \bar{D}_\varepsilon(i) = \frac{D(I_k)}{|I_k|}$$

We insist that while $\bar{D}_\varepsilon$ (obviously) depends on D, the partition $\mathcal{I}_\varepsilon$ itself does not; in particular, it can be computed prior to getting any sample or information about D.

Theorem 12 ([8])**.** *If D is monotone non-increasing, then* $\mathrm{d_{TV}}\big(D, \bar{D}_\varepsilon\big) \leq \varepsilon$.

Remark 5. A proof of this theorem, self-contained and phrased in terms of discrete distributions (whereas the original paper by Birgé is primarily intended for continuous ones) can be found in [14] – Theorem 3.1.

Corollary 4. *Suppose D is ε-close to monotone non-increasing. Then* $\mathrm{d_{TV}}\big(D, \bar{D}_\varepsilon\big) \leq 3\varepsilon$*; furthermore, $\bar{D}_\varepsilon$ is also ε-close to monotone non-increasing.*

Finally, we shall also need the following well-known result relating total variation distance and difference of entropies (see e.g. [29], Eq. (4)):

Fact 2 (Total variation and Entropy)**.** *Let D_1, D_2 be two distributions on $[n]$ such that* $\mathrm{d_{TV}}(D_1, D_2) \leq \alpha$, *for $\alpha \in [0,1]$. Then* $|H(D_1) - H(D_2)| \leq \alpha\log(n-1) + h_2(\alpha) \leq \alpha\log\frac{n}{\alpha} + (1-\alpha)\log\frac{1}{1-\alpha}$*, where h_2 is the binary entropy function*[13]*.*

High-level idea Suppose we use the oblivious decomposition from Definition 5, with small parameter α (to be determined later), to reduce the domain into $\ell = o(n)$ intervals. Then, we can set out to approximate the entropy of the induced *flat* distribution – that we can efficiently simulate from the cumulative dual oracles, roughly reducing the complexity parameter from n to ℓ; it only remains to use the previous approach, slightly adapted, on this flat distribution. Of course, we have to be careful not to incur too much a loss at each step, where we first approximate $H(D)$ by $H(\bar{D})$, and then specify our cutoff threshold to only consider significant contributions to $H(\bar{D})$.

[13] That is, $h_2(p) = -p\log p - (1-p)\log(1-p)$ is the entropy of a Bernoulli random variable with parameter p.

Details Consider the Birgé decomposition of $[n]$ into $\ell = \Theta(\log(n\alpha)/\alpha)$ intervals (for α to be defined shortly). Theorem 12 ensures the corresponding (unknown) flattened distribution $\bar{D}$ is α-close to D; which, by the fact above, implies that

$$\left|H(\bar{D}) - H(D)\right| \leq \alpha \left(\log \frac{n}{\alpha} + 2\right) \tag{10}$$

Taking $\alpha \overset{\text{def}}{=} \Theta(\Delta/\log n)$, the right-hand-side is at most $\Delta/2$; so that it is now sufficient to estimate $H(\bar{D})$ to $\pm\Delta/2$, where both sampling and evaluation access to $\bar{D}$ can easily be simulated from the CEVAL_D and SAMP_D oracles. But although $\bar{D}$ is a distribution on $[n]$, its "actual" support is morally only the $\ell = \tilde{\Theta}\big(\log^2 n/\Delta\big)$. Indeed, we may write the entropy of $\bar{D}$ as

$$H(\bar{D}) = \sum_{k=1}^{\ell}\sum_{x\in I_k} \bar{D}(x)\log\frac{1}{\bar{D}(x)} = \sum_{k=1}^{\ell}\sum_{x\in I_k} \frac{D(I_k)}{|I_k|}\log\frac{|I_k|}{D(I_k)} = \sum_{k=1}^{\ell} D(I_k)\log\frac{|I_k|}{D(I_k)} = \mathbb{E}_{k\sim\bar{D}}\left[\log\frac{1}{d_k}\right]$$

where $d_k = \frac{D(I_k)}{|I_k|} \approx (1+\alpha)^{-k} D(I_k)$.

As in the previous section, we can then define a cutoff threshold τ (for d_k) and only estimate $\mathbb{E}_{k\sim\bar{D}}\left[\log\frac{1}{d_k}\mathbb{1}_{\{d_k\geq\tau\}}\right]$, for this purpose, we need $\ell\cdot\tau\log 1/\tau$ to be at most $\Delta/4$, i.e.

$$\tau \overset{\text{def}}{=} \Theta\left(\frac{\Delta/\ell}{\log\Delta/\ell}\right) = \tilde{\Theta}\left(\frac{\Delta^2}{\log^2 n}\right)$$

and to get with high probability a $\Delta/4$-approximation, it is as before sufficient to make $m = O\big(\Delta^2/\log^2(1/\tau)\big) = \tilde{O}\Big(\frac{\log^2\frac{\log n}{\Delta}}{\Delta^2}\Big)$ queries.

Theorem 13. *In the cumulative dual access model, there exists an algorithm for monotone distributions estimating the entropy up to an additive* Δ*, with sample complexity* $\tilde{O}\Big(\log^2\frac{\log n}{\Delta}/\Delta^2\Big)$.

Remark 6. We remark that the above result and algorithm (after some minor changes in the constants) still applies if D is only guaranteed to be $O(1/\log n)$-*close* to monotone; indeed, as stated in Corollary 4, the oblivious decomposition is (crucially) robust, and $\bar{D}$ will still be $O(\varepsilon)$-close to D.

4.3 Additive estimation of support size

We now turn to the task of estimating the effective support size of the distribution: given the promise that D puts on every element of the domain either no weight or at least some minimum probability mass $1/n > 0$, the goal is to output a good estimate (up to $\pm\varepsilon n$) of the number of elements in the latter situation.

Theorem 14. *In the dual access model, there exists an algorithm* Estimate-Support *that, on input a threshold* $n\in\mathbb{N}^*$ *and a parameter* $\varepsilon > 0$*, and given access to a distribution* D *(over an arbitrary set) satisfying*

$$\min_{x\in\operatorname{supp}(D)} D(x) \geq \frac{1}{n}$$

estimates the support size $|\operatorname{supp}(D)|$ *up to an additive* εn*, with query complexity* $O\big(\frac{1}{\varepsilon^2}\big)$.

Proof. Write $k \overset{\text{def}}{=} |\operatorname{supp}(D)|$. We describe Estimate-Support which outputs (w.p. at least 2/3) an estimate as required:

If $\varepsilon > \frac{2}{\sqrt{n\ln 3n}}$**:** The algorithm will draw $m = \left\lceil\frac{4}{\varepsilon^2}\right\rceil$ samples $x_1,\dots,x_m$ from D, query their probability mass $D(x_i)$, and output $\hat{k} = \lceil Y\rceil$, where

$$Y \overset{\text{def}}{=} \frac{1}{m}\sum_{i=1}^{m}\frac{\mathbb{1}_{\left\{D(x_i)\geq\frac{1}{n}\right\}}}{D(x_i)}$$

If $\varepsilon \leq \frac{2}{\sqrt{n\ln 3n}}$: in this case, Estimate-Support just draws $m = n\ln 3n = O\big(\frac{1}{\varepsilon^2}\big)$ samples $x_1,\dots,x_m$ from D, and returns the number $\hat{k}$ of distincts elements it got (no query access is needed in this case).

Analysis In the first (and interesting) case, let ϕ be the function defined over the coset of D by $\phi(x) = \frac{1}{D(x)} \cdot \mathbb{1}_{\left\{D(x)\geq \frac{1}{n}\right\}}$, so that $\mathbb{E}_{x\sim D}[\phi(x)] = \sum_{x:D(x)>\frac{1}{n}} D(x)\cdot \frac{1}{D(x)} = \left|\left\{\ x \ :\ D(x) > \frac{1}{n}\ \right\}\right| = k$; and as the r.v. $\phi(x_1),\dots,\phi(x_m)$ are i.i.d and taking value in $[0,n]$, an additive Chernoff bound yields

$$\Pr\Big[\,|Y-k| > \frac{\varepsilon n}{2}\,\Big] \leq 2e^{-\frac{\varepsilon^2 m}{2}} < \frac{1}{3}$$

Conditioned on this not happening, $k + \frac{\varepsilon}{2}n \leq Y \leq \hat{k} \leq Y+1 \leq k+\frac{\varepsilon}{2}n+1 \leq k+\varepsilon n$ (as $\varepsilon > \frac{2}{n}$), and $\hat{k}$ is as stated.

Turning now to the second case, observe first that the promise on D implies that $1 \leq k \leq n$. It is sufficient to bound the probability that an element of the support is *never* seen during the m draws – let F denote this event. By a union bound,

$$\Pr[\,F\,] \leq k\cdot\left(1-\frac{1}{n}\right)^m \leq ne^{n\ln(3n)\ln(1-\frac{1}{n})} \leq ne^{-\ln 3n} = \frac{1}{3}$$

so w.p. at least $2/3$, every element of the support is drawn, and Estimate-Support returns (exactly) k. □

4.3.1 Lower bound

In this subsection, we show that the upper bound of Theorem 14 is tight.

Theorem 15. *In the dual access model, ε-additively estimating support size requires query complexity $\Omega\big(\frac{1}{\varepsilon^2}\big)$.*

Proof. Without loss of generality, suppose n is even, and let $k = \frac{n}{2}$. For any $p \in [0,1]$, consider the following process Φ_p, which yields a random distribution D_p on $[n]$ (See Fig.2):

- draw k i.i.d. random variables $X_1,\dots,X_k \sim \mathrm{Bern}(p)$;
- for $i \in [k]$, set $D(i) = \frac{1}{n}(1+X_i)$ and $D(n-i) = \frac{1}{n}(1-X_i)$

Note that by construction $D(i) + D(n-i) = \frac{2}{n}$ for all $i \in [k]$.

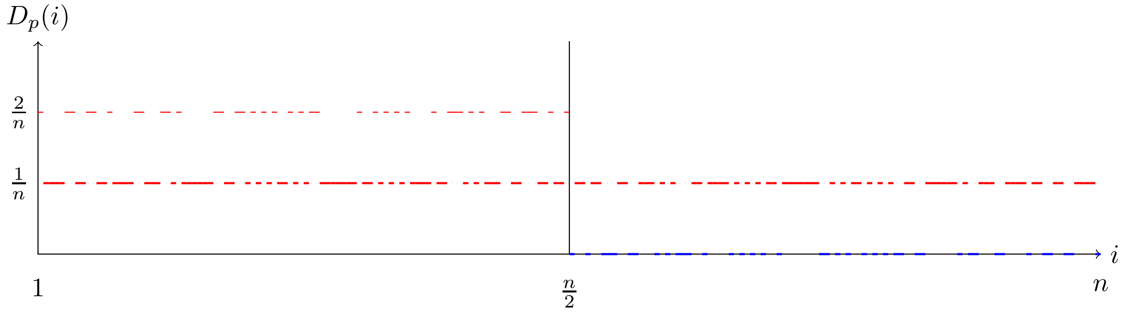

Figure 2: An instance of distribution D_p with $p = 4/10$.

Define now, for any $\varepsilon \in (0,1/6)$, the families of distributions $\mathcal{D}^+$ and $\mathcal{D}^-$ induced the above construction, taking p to be respectively $p^+ \overset{\text{def}}{=} \frac{1}{2}$ and $p^- \overset{\text{def}}{=} \frac{1}{2}-6\varepsilon$. Hereafter, by D^+ (resp. D^-), we refer to a distribution from $\mathcal{D}^+$ (resp. $\mathcal{D}^-$) generated randomly as above (we assume further, without loss of generality, that

$n \gg 1/\varepsilon^2$):

$$\mathbb{E}\big[\operatorname{supp}\big(D^+\big)\big] = n - kp^+ = n\left(1 - \frac{p^+}{2}\right) = \frac{3}{4}n$$
$$\mathbb{E}\big[\operatorname{supp}\big(D^-\big)\big] = n - kp^- = n\left(1 - \frac{p^-}{2}\right) = \left(\frac{3}{4} + 3\varepsilon\right)n$$

and, with an additive Chernoff bound,

$$\Pr\left[\operatorname{supp}\big(D^+\big) \geq \frac{3}{4}n + \frac{\varepsilon}{2}n\right] \leq e^{-\frac{\varepsilon^2 n}{2}} < \frac{1}{100}$$
$$\Pr\left[\operatorname{supp}\big(D^-\big) \leq \frac{3}{4}n + \frac{5\varepsilon}{2}n\right] \leq e^{-\frac{\varepsilon^2 n}{2}} < \frac{1}{100}$$

We hereafter condition on these events E^+ and E^- every time we consider a given D^+ or D^-, and set for convenience $s^+ \stackrel{\text{def}}{=} \frac{3}{4}(n + 2\varepsilon)$, $s^- \stackrel{\text{def}}{=} \frac{3}{4}(n + 10\varepsilon)$.

Reduction We shall once again reduce the problem of distinguishing between (a) a fair coin and (b) an $(\frac{1}{2} - 6\varepsilon)$-biased coin to the problem of approximating the support size: suppose by contradiction we have a tester $\mathcal{T}$ for the latter problem, making $q = o\big(\frac{1}{\varepsilon^2}\big)$ queries on input ε.

Given parameter $\varepsilon \in (0, 1/100)$ and $\mathsf{SAMP}_{\text{coin}}$ access to i.i.d. coin tosses coming from one of those two situations (($p^+ = \frac{1}{2}$, or $p^- = \frac{1}{2} - 6\varepsilon$), define a distinguisher $\mathcal{A}$ as follows:

- after picking an even integer $n \gg 1/\varepsilon^2$, $\mathcal{A}$ will maintain a set $C \subseteq [n] \times \{0, \frac{1}{n}, \frac{2}{n}\}$ (initially empty), and run $\mathcal{T}$ as a subroutine with parameter ε;
- EVAL: when $\mathcal{T}$ makes an evaluation query on a point $i \in [n]$
  - if i has already been committed to (there is a pair (i, d_i) in C), it returns d_i;
  - otherwise, it asks for a sample b from $\mathsf{SAMP}_{\text{coin}}$, and sets

    $$d_i = \begin{cases} \frac{1}{n} & \text{if } b = 0 \\ \frac{2}{n} & \text{if } b = 1 \text{ and } i \in [k] \\ 0 & \text{if } b = 1 \text{ and } i \in [n] \setminus [k] \end{cases}$$

    before adding (i, d_i) and $(n - i, \frac{2}{n} - d_i)$ to C and returning d_i.
- SAMP: when $\mathcal{T}$ makes an sampling query, $\mathcal{A}$ draws u.a.r. $i \sim [k]$, and then proceeds as in the EVAL case to get d_i and d_{n-i} (that is, if they are not in C, it first generates them from a $\mathsf{SAMP}_{\text{coin}}$ query and commits to them); and then, it returns i w.p. $(nd_i)/2$, and $n - i$ w.p. $(nd_{n-i})/2$.

It is easy to see that the process above exactly simulates dual access to a distribution D generated either according to Φ_{p^+} or Φ_{p^-} – in particular, this is true of the sampling queries because each pair $(i, n - i)$ has same total mass $\frac{2}{n}$ under any such distribution, so drawing from D is equivalent to drawing uniformly $i \in [k]$, and then returning at random i or $n - i$ according to the conditional distribution of D on $\{i, n - i\}$.

Furthermore, the number of queries to $\mathsf{SAMP}_{\text{coin}}$ is at most the number of queries made by $\mathcal{T}$ to $\mathcal{A}$, that is $o\big(\frac{1}{\varepsilon^2}\big)$. Conditioning on E^+ (or E^-, depending on whether we are in case (a) or (b)), the distribution D has support size at most s^+ (resp. at least s^-). As the estimate $\hat{s}$ that $\mathcal{T}$ will output will, with probability at least 2/3, be εn-close to the real support size, and as $s^- - s^+ = 2\varepsilon n$, $\mathcal{A}$ will distinguish between cases (a) and (b) with probability at least $2/3 - 2/100 > 6/10$ – contradicting the fact that $\Omega\big(1/\varepsilon^2\big)$ samples are required to distinguish between a fair and a $(\frac{1}{2} - 6\varepsilon)$-biased coin with this probability. □

References

[1] J. A. Adell and P. Jodra. Exact Kolmogorov and total variation distances between some familiar discrete distributions. *J. of Ineq. App.*, 2006(1):64307, 2006.

[2] Ziv Bar-Yossef. *The Complexity of Massive Data Set Computations*. PhD thesis, UC Berkeley, 2002. Christos Papadimitriou.

[3] T. Batu, E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White. Testing random variables for independence and identity. In *Proceedings of FOCS*, pages 442–451, 2001.

[4] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing that distributions are close. In *Proceedings of FOCS*, pages 189–197, 2000.

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

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

[7] Tugkan Batu, Sanjoy Dasgupta, Ravi Kumar, and Ronitt Rubinfeld. The complexity of approximating the entropy. *SIAM Journal on Computing*, 35(1):132–150, 2005.

[8] Lucien Birgé. On the risk of histograms for estimating decreasing densities. *The Annals of Statistics*, 15(3):1013–1022, 1987.

[9] Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to noninteractive database privacy. *J. ACM*, 60(2):12, 2013.

[10] Clément Canonne, Dana Ron, and Rocco A. Servedio. Testing probability distributions using conditional samples. Technical Report abs/1211.2664, ArXiV, November 2012.

[11] Clément Canonne, Dana Ron, and Rocco A. Servedio. Testing equivalence between distributions using conditional samples. In *Proceedings of SODA*, 2014.

[12] Sourav Chakraborty, Eldar Fischer, Yonatan Goldhirsh, and Arie Matsliah. On the power of conditional samples in distribution testing. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, ITCS '13, pages 561–580, New York, NY, USA, 2013. ACM.

[13] S.-O. Chan, I. Diakonikolas, G. Valiant, and P. Valiant. Optimal Algorithms for Testing Closeness of Discrete Distributions. In *Proceedings of SODA*, 2014.

[14] C. Daskalakis, I. Diakonikolas, R. Servedio, G. Valiant, and P. Valiant. Testing k-modal distributions: Optimal algorithms via reductions. In *Proceedings of SODA*, 2013.

[15] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 202–210. ACM, 2003.

[16] Cynthia Dwork. Differential privacy: A survey of results. In *Theory and Applications of Models of Computation*, volume 4978, pages 1–19. Springer, 2008.

[17] Cynthia Dwork and Kobbi Nissim. Privacy-preserving datamining on vertically partitioned databases. In *Advances in Cryptology–CRYPTO 2004*, pages 528–544. Springer, 2004.

[18] O. Goldreich and D. Ron. On testing expansion in bounded-degree graphs. Technical Report TR00-020, ECCC, 2000.

[19] Sudipto Guha, Andrew McGregor, and Suresh Venkatasubramanian. Streaming and sublinear approximation of entropy and information distances. *CoRR*, abs/cs/0508122, 2005.

[20] P. Indyk, R. Levi, and R. Rubinfeld. Approximating and Testing k-Histogram Distributions in Sub-linear Time. In *Proceedings of PODS*, pages 15–22, 2012.

[21] Shang-Keng Ma. Calculation of entropy from data of motion. *Journal of Statistical Physics*, 26(2):221–240, 1981.

[22] L. Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE-IT*, 54(10):4750–4755, 2008.

[23] Michal Parnas, Dana Ron, and Ronitt Rubinfeld. Tolerant property testing and distance approximation. *J. Comput. Syst. Sci.*, 72(6):1012–1042, September 2006.

[24] R. Rubinfeld and R. A. Servedio. Testing monotone high-dimensional distributions. *RSA*, 34(1):24–44, January 2009.

[25] G. Valiant and P. Valiant. A CLT and tight lower bounds for estimating entropy. Technical Report TR10-179, ECCC, 2010.

[26] G. Valiant and P. Valiant. Estimating the unseen: A sublinear-sample canonical estimator of distributions. Technical Report TR10-180, ECCC, 2010.

[27] G. Valiant and P. Valiant. Estimating the unseen: an $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs. In *Proceedings of STOC*, pages 685–694, 2011. See also [25] and [26].

[28] P. Valiant. Testing symmetric properties of distributions. *SICOMP*, 40(6):1927–1968, 2011.

[29] Zhengmin Zhang. Estimating mutual information via Kolmogorov distance. *Information Theory, IEEE Transactions on*, 53(9):3280–3282, 2007.

A Chernoff Bounds

Theorem 16. *Let $Y_1, \ldots, Y_m$ be m independent random variables that take on values in $[0,1]$, where $\mathbb{E}[Y_i] = p_i$, and $\sum_{i=1}^m p_i = P$. For any $\gamma \in (0,1]$ we have*

$$\textit{(additive bound)} \qquad \Pr\left[\sum_{i=1}^m Y_i > P + \gamma m\right],\ \Pr\left[\sum_{i=1}^m Y_i < P - \gamma m\right] \leq \exp(-2\gamma^2 m) \tag{11}$$

$$\textit{(multiplicative bound)} \qquad \Pr\left[\sum_{i=1}^m Y_i > (1+\gamma)P\right] < \exp(-\gamma^2 P/3) \tag{12}$$

and

$$\textit{(multiplicative bound)} \qquad \Pr\left[\sum_{i=1}^m Y_i < (1-\gamma)P\right] < \exp(-\gamma^2 P/2). \tag{13}$$

The following extension of the multiplicative bound is useful when only upper and/or lower bounds on P are known:

Corollary 5. *In the setting of Theorem 16 suppose that* $P_L \le P \le P_H$. *Then for any* $\gamma \in (0, 1]$, *we have*

$$\Pr\left[\sum_{i=1}^{m} Y_i > (1+\gamma)P_H\right] < \exp(-\gamma^2 P_H/3) \tag{14}$$

$$\Pr\left[\sum_{i=1}^{m} Y_i < (1-\gamma)P_L\right] < \exp(-\gamma^2 P_L/2) \tag{15}$$