Electronic Colloquium on Computational Complexity, Revision 1 of Report No. 155 (2012)

Testing probability distributions using conditional samples

Clément L. Canonne* Dana Ron† Rocco A. Servedio‡

January 16, 2015

Abstract

We study a new framework for property testing of probability distributions, by considering distribution testing algorithms that have access to a *conditional sampling oracle*. This is an oracle that takes as input a subset $S \subseteq [N]$ of the domain $[N]$ of the unknown probability distribution D and returns a draw from the conditional probability distribution D restricted to S. This new model allows considerable flexibility in the design of distribution testing algorithms; in particular, testing algorithms in this model can be adaptive.

We study a wide range of natural distribution testing problems in this new framework and some of its variants, giving both upper and lower bounds on query complexity. These problems include testing whether D is the uniform distribution $\mathcal{U}$; testing whether $D = D^*$ for an explicitly provided D^*; testing whether two unknown distributions D_1 and D_2 are equivalent; and estimating the variation distance between D and the uniform distribution. At a high level our main finding is that the new conditional sampling framework we consider is a powerful one: while all the problems mentioned above have $\Omega(\sqrt{N})$ sample complexity in the standard model (and in some cases the complexity must be almost linear in N), we give $\mathrm{poly}(\log N, 1/\epsilon)$-query algorithms (and in some cases $\mathrm{poly}(1/\epsilon)$-query algorithms independent of N) for all these problems in our conditional sampling setting.

*ccanonne@cs.columbia.edu, Columbia University. Supported by NSF grants CCF-1115703 and CCF-1319788.
†danaron@post.tau.ac.il, Tel Aviv University. Supported by ISF grants 246/08 and 671/13.
‡rocco@cs.columbia.edu, Columbia University. Supported by NSF grants CCF-0915929 and CCF-1115703.

ISSN 1433-8092

Contents

1 Introduction **2**

1.1 Background: Distribution testing in the standard model . . . 2

1.2 The conditional sampling model . . . 3

1.3 Our results . . . 4

1.4 The work of Chakraborty et al. [CFGM13] . . . 8

2 Preliminaries **10**

2.1 Definitions . . . 10

2.2 Useful tools . . . 10

3 Some useful procedures **12**

3.1 The procedure Compare . . . 12

3.2 The procedure Estimate-Neighborhood . . . 14

3.3 The procedure Approx-Eval-Simulator . . . 16

4 Algorithms and lower bounds for testing uniformity **27**

4.1 A $\tilde{O}(1/\epsilon^2)$-query PCOND algorithm for testing uniformity . . . 27

4.2 An $\Omega(1/\epsilon^2)$ lower bound for COND_D algorithms that test uniformity . . . 31

5 Testing equivalence to a known distribution D^* **33**

5.1 A $\mathrm{poly}(\log n, 1/\epsilon)$-query PCOND_D algorithm . . . 33

5.2 A $(\log N)^{\Omega(1)}$ lower bound for PCOND_D . . . 37

5.3 A $\mathrm{poly}(1/\epsilon)$-query COND_D algorithm . . . 41

6 Testing equality between two unknown distributions **51**

6.1 An approach based on PCOND queries . . . 51

6.2 An approach based on simulating EVAL . . . 57

7 An algorithm for estimating the distance to uniformity **59**

7.1 Finding a reference point . . . 62

8 A $\tilde{O}\left(\frac{\log^3 N}{\epsilon^3}\right)$-query ICOND_D algorithm for testing uniformity **65**

9 An $\Omega(\log N/\log\log N)$ lower bound for ICOND_D algorithms that test uniformity **68**

9.1 A lower bound against non-adaptive algorithms . . . 70

9.2 A lower bound against adaptive algorithms: Outline of the proof of Theorem 16 . . . 75

9.3 Extended transcripts and drawing $D \sim \mathcal{P}_{\mathrm{No}}$ on the fly. . . . 76

9.4 Bounding $d_{\mathrm{TV}}\left(\mathfrak{A}_{\mathrm{otf}}^{(k),\mathrm{N}}, \mathfrak{A}_{\mathrm{otf}}^{(k+1),\mathrm{N}}\right)$ . . . 79

10 Conclusion **84**

1 Introduction

1.1 Background: Distribution testing in the standard model

One of the most fundamental problem paradigms in statistics is that of inferring some information about an unknown probability distribution D given access to independent samples drawn from it. More than a decade ago, Batu et al. [BFR+00][1] initiated the study of problems of this type from within the framework of *property testing* [RS96, GGR98]. In a property testing problem there is an unknown "massive object" that an algorithm can access only by making a small number of local inspections of the object, and the goal is to determine whether the object has a particular property. The algorithm must output ACCEPT if the object has the desired property and output REJECT if the object is far from every object with the property. (See [Fis01, Ron08, Ron10, Gol10] for detailed surveys and overviews of the broad field of property testing; we give precise definitions tailored to our setting in Section 2.)

In distribution property testing the "massive object" is an unknown probability distribution D over an N-element set, and the algorithm accesses the distribution by drawing independent samples from it. A wide range of different properties of probability distributions have been investigated in this setting, and upper and lower bounds on the number of samples required have by now been obtained for many problems. These include testing whether D is uniform [GR00, BFR+10, Pan08], testing whether D is identical to a given known distribution D^* [BFF+01], testing whether two distributions D_1, D_2 (both available via sample access) are identical [BFR+00, Val11], and testing whether D has a monotonically increasing probability mass function [BFRV11], as well as related problems such as estimating the entropy of D [BDKR05, VV11], and estimating its support size [RRSS09, Val11, VV11]. Similar problems have also been studied by researchers in other communities, see e.g., [Ma81, Pan04, Pan08].

One broad insight that has emerged from this past decade of work is that while sublinear-sample algorithms do exist for many distribution testing problems, the number of samples required is in general quite large. Even the basic problem of testing whether D is the uniform distribution $\mathcal{U}$ over $[N] = \{1, \ldots, N\}$ versus ϵ-far from uniform requires $\Omega(\sqrt{N})$ samples[2] for constant ϵ, and the other problems mentioned above have sample complexities at least this high, and in some cases *almost linear in* N [RRSS09, Val11, VV11]. Since such sample complexities could be prohibitively high in real-world settings where N can be extremely large, it is natural to explore problem variants where it may be possible for algorithms to succeed using fewer samples. Indeed, researchers have studied distribution testing in settings where the unknown distribution is guaranteed to have some special structure, such as being monotone, k-modal or a "k-histogram" over $[N]$ [BKR04, DDS+13, ILR12], or being monotone over $\{0,1\}^n$ [RS09] or over other posets [BFRV11], and have obtained significantly more sample-efficient algorithms using these additional assumptions.

[1] There is a more recent full version of this work [BFR+10] and we henceforth reference this recent version.

[2] To verify this, consider the family of all distributions that are uniform over half of the domain, and 0 elsewhere. Each distribution in this family is $\Theta(1)$-far from the uniform distribution. However, it is not possible to distinguish with sufficiently high probability between the uniform distribution and a distribution selected randomly from this family, given a sample of size $\sqrt{N}/c$ (for a sufficiently large constant $c > 1$). This is the case because for the uniform distribution as well as each distribution in this family, the probability of observing the same element more than once is very small. Conditioned on such a collision event not occurring, the samples are distributed identically.

1.2 The conditional sampling model

In this work we pursue a different line of investigation: rather than restricting the class of probability distributions under consideration, we consider testing algorithms that may use a more powerful form of access to the unknown distribution D. This is a *conditional sampling oracle*, which allows the algorithm to obtain a draw from D_S, the conditional distribution of D restricted to a subset S of the domain (where S is specified by the algorithm). More precisely, we have:

Definition 1 *Fix a distribution D over $[N]$. A* COND *oracle for D, denoted* COND_D*, is defined as follows: The oracle is given as input a* query set $S \subseteq [N]$ *, chosen by the algorithm, that has $D(S) > 0$. The oracle returns an element $i \in S$, where the probability that element i is returned is $D_S(i) = D(i)/D(S)$, independently of all previous calls to the oracle.*[3]

As mentioned earlier, a recent work of Chakraborty et al. [CFGM13] introduced a very similar conditional model; we discuss their results and how they relate to our results in Subsection 1.4. For compatibility with our COND_D notation we will write SAMP_D to denote an oracle that takes no input and, each time it is invoked, returns an element from $[N]$ drawn according to D independently from all previous draws. This is the sample access to D that is used in the standard model of testing distributions, and this is of course the same as a call to $\mathsf{COND}_D([N])$.

Motivation and Discussion. One purely theoretical motivation for the study of the COND model is that it may further our understanding regarding what forms of information (beyond standard sampling) can be helpful for testing properties of distributions. In both learning and property testing it is generally interesting to understand how much power algorithms can gain by making queries, and COND queries are a natural type of query to investigate in the context of distributions. As we discuss in more detail below, in several of our results we actually consider restricted versions of COND queries that do not require the full power of obtaining conditional samples from arbitrary sets.

A second attractive feature of the COND model is that it enables a new level of richness for algorithms that deal with probability distributions. In the standard model where only access to SAMP_D is provided, all algorithms must necessarily be non-adaptive, with the same initial step of simply drawing a sample of points from SAMP_D, and the difference between two algorithms comes only from how they process their samples. In contrast, the essence of the COND model is to allow algorithms to *adaptively* determine later query sets S based on the outcomes of earlier queries.

A natural question about the COND model is its plausibility: are there settings in which an investigator could actually make conditional samples from a distribution of interest? We feel that the COND framework provides a reasonable first approximation for scenarios that arise in application areas (e.g., in biology or chemistry) where the parameters of an experiment can be adjusted so

[3]Note that as described above the behavior of $\mathsf{COND}_D(S)$ is undefined if $D(S) = 0$, i.e., the set S has zero probability under D. While various definitional choices could be made to deal with this, we shall assume that in such a case, the oracle (and hence the algorithm) outputs "failure" and terminates. This will not be a problem for us throughout this paper, as (a) our lower bounds deal only with distributions that have $D(i) > 0$ for all $i \in [N]$, and (b) in our algorithms $\mathsf{COND}_D(S)$ will only ever be called on sets S which are "guaranteed" to have $D(S) > 0$. (More precisely, each time an algorithm calls $\mathsf{COND}_D(S)$ it will either be on the set $S = [N]$, or will be on a set S which contains an element i which has been returned as the output of an earlier call to COND_D.)

as to restrict the range of possible outcomes. For example, a scientist growing bacteria or yeast cells in a controlled environment may be able to deliberately introduce environmental factors that allow only cells with certain desired characteristics to survive, thus restricting the distribution of all experimental outcomes to a pre-specified subset. We further note that techniques which are broadly reminiscent of COND sampling have long been employed in statistics and polling design under the name of "stratified sampling" (see e.g. [Wik13, Ney34]). We thus feel that the study of distribution testing in the COND model is well motivated both by theoretical and practical considerations.

Given the above motivations, the central question is whether the COND model enables significantly more efficient algorithms than are possible in the weaker SAMP model. Our results (see Subsection 1.3) show that this is indeed the case.

Before detailing our results, we note that several of them will in fact deal with a weaker variant of the COND model, which we now describe. In designing COND-model algorithms it is obviously desirable to have algorithms that only invoke the COND oracle on query sets S which are "simple" in some sense. Of course there are many possible notions of simplicity; in this work we consider the size of a set as a measure of its simplicity, and consider algorithms which only query small sets. More precisely, we consider the following restriction of the general COND model:

PCOND **oracle:** We define a PCOND (short for "pair-cond") *oracle for* D is a restricted version of COND_D that only accepts input sets S which are either $S = [N]$ (thus providing the power of a SAMP_D oracle) or $S = \{i, j\}$ for some $i, j \in [N]$, i.e. sets of size two. The PCOND oracle may be viewed as a minimalist variant of COND that essentially permits an algorithm to compare the relative weights of two items under D (and to draw random samples from D, by setting $S = [N]$).

ICOND **oracle:** We define an ICOND (short for "interval-cond") *oracle for* D as a restricted version of COND_D that only accepts input sets S which are intervals $S = [a, b] = \{a, a + 1, \ldots, b\}$ for some $a \leq b \in [N]$ (note that taking $a = 1$, $b = N$ this provides the power of a SAMP_D oracle). This is a natural restriction on COND queries in settings where the N points are endowed with a total order.

To motivate the PCOND model (which essentially gives the ability to compare two elements), one may consider a setting in which a human domain expert can provide an estimate of the relative likelihood of two distinct outcomes in a limited-information prediction scenario.

1.3 Our results

We give a detailed study of a range of natural distribution testing problems in the COND model and its variants described above, establishing both upper and lower bounds on their query complexity. Our results show that the ability to do conditional sampling provides a significant amount of power to property testers, enabling polylog(N)-query, or even constant-query, algorithms for problems whose sample complexities in the standard model are $N^{\Omega(1)}$; see Table 1. While we have considered a variety of distribution testing problems in the COND model, our results are certainly not exhaustive, and many directions remain to be explored; we discuss some of these in Section 10.

Problem	Our results		Standard model
Is D uniform?	COND_D	$\Omega\left(\frac{1}{\epsilon^2}\right)$	$\Theta\left(\frac{\sqrt{N}}{\epsilon^2}\right)$ [GR00, BFR+10, Pan08]
	PCOND_D	$\tilde{O}\left(\frac{1}{\epsilon^2}\right)$	
	ICOND_D	$\tilde{O}\left(\frac{\log^3 N}{\epsilon^3}\right)$ $\Omega\left(\frac{\log N}{\log\log N}\right)$	
Is $D = D^*$ for a known D^*?	COND_D	$\tilde{O}\left(\frac{1}{\epsilon^4}\right)$	$\Theta\left(\frac{\sqrt{N}}{\epsilon^2}\right)$ [BFF+01, Pan08, VV14]
	PCOND_D	$\tilde{O}\left(\frac{\log^4 N}{\epsilon^4}\right)$ $\Omega\left(\sqrt{\frac{\log N}{\log\log N}}\right)$	
Are D_1, D_2 (both unknown) equivalent?	COND_{D_1,D_2}	$\tilde{O}\left(\frac{\log^5 N}{\epsilon^4}\right)$	$\Theta\left(\max\left(\frac{N^{2/3}}{\epsilon^{4/3}}, \frac{\sqrt{N}}{\epsilon^2}\right)\right)$ [BFR+10, Val11, CDVV14]
	PCOND_{D_1,D_2}	$\tilde{O}\left(\frac{\log^6 N}{\epsilon^{21}}\right)$	
How far is D from uniform?	PCOND_D	$\tilde{O}\left(\frac{1}{\epsilon^{20}}\right)$	$O\left(\frac{1}{\epsilon^2}\frac{N}{\log N}\right)$ [VV11, VV10b] $\Omega\left(\frac{N}{\log N}\right)$ [VV11, VV10a]

Table 1: Comparison between the COND model and the standard model on a variety of distribution testing problems over $[N]$. The upper bounds for the first three problems are for testing whether the property holds (i.e. $d_{\mathrm{TV}} = 0$) versus $d_{\mathrm{TV}} \geq \epsilon$, and for the last problem the upper bound is for estimating the distance to uniformity to within an additive $\pm\epsilon$.

1.3.1 Testing distributions over unstructured domains

In this early work on the COND model our main focus has been on the simplest (and, we think, most fundamental) problems in distribution testing, such as testing whether D is the uniform distribution $\mathcal{U}$; testing whether $D = D^*$ for an explicitly provided D^*; testing whether $D_1 = D_2$ given COND_{D_1} and COND_{D_2} oracles; and estimating the variation distance between D and the uniform distribution. In what follows d_{TV} denotes the variation distance.

Testing uniformity. We give a PCOND_D algorithm that tests whether $D = \mathcal{U}$ versus $d_{\mathrm{TV}}(D,\mathcal{U}) \geq \epsilon$ using $\tilde{O}(1/\epsilon^2)$ calls to PCOND_D, independent of N. We show that this PCOND_D algorithm is nearly optimal by proving that any COND_D tester (which may use arbitrary subsets $S \subseteq [N]$ as its query sets) requires $\Omega(1/\epsilon^2)$ queries for this testing problem.

Testing equivalence to a known distribution. As described above, for the simple problem of testing uniformity we have an essentially optimal PCOND testing algorithm and a matching lower bound. A more general and challenging problem is that of testing whether D (accessible via a PCOND or COND oracle) is equivalent to D^*, where D^* is an arbitrary "known" distribution over $[N]$ that is explicitly provided to the testing algorithm at no cost (say as a vector $(D^*(1), \ldots, D^*(N))$ of probability values). For this "known D^*" problem, we give a PCOND_D algorithm testing whether $D = D^*$ versus $d_{\mathrm{TV}}(D, D^*) \geq \epsilon$ using $\tilde{O}((\log N)^4/\epsilon^4)$ queries. We further show that the $(\log N)^{\Omega(1)}$ query complexity of our PCOND_D algorithm is inherent in the problem, by proving that any PCOND_D algorithm for this problem must use $\sqrt{\log(N)/\log\log(N)}$ queries for constant ϵ.

Given these $(\log N)^{\Theta(1)}$ upper and lower bounds on the query complexity of PCOND_D-testing equivalence to a known distribution, it is natural to ask whether the full COND_D oracle provides more power for this problem. We show that this is indeed the case, by giving a $\tilde{O}(1/\epsilon^4)$-query algorithm (independent of N) that uses unrestricted COND_D queries.

Testing equivalence between two unknown distributions. We next consider the more challenging problem of testing whether two unknown distributions D_1, D_2 over $[N]$ (available via COND_{D_1} and COND_{D_2} oracles) are identical versus ϵ-far. We give two very different algorithms for this problem. The first uses PCOND oracles and has query complexity $\tilde{O}((\log N)^6/\epsilon^{21})$, while the second uses COND oracles and has query complexity $\tilde{O}((\log N)^5/\epsilon^4)$. We believe that the proof technique of the second algorithm is of independent interest, since it shows how a COND_D oracle can efficiently simulate an "approximate EVAL_D oracle." (An EVAL_D oracle takes as input a point $i \in [N]$ and outputs the probability mass $D(i)$ that D puts on i; we briefly explain our notion of approximating such an oracle in Subsection 1.3.3.)

Estimating the distance to uniformity. We also consider the problem of estimating the variation distance between D and the uniform distribution $\mathcal{U}$ over $[N]$, to within an additive error of $\pm\epsilon$. In the standard SAMP_D model this is known to be a very difficult problem, with an $\Omega(N/\log N)$ lower bound established in [VV11, VV10a]. In contrast, we give a PCOND_D algorithm that makes only $\tilde{O}(1/\epsilon^{20})$ queries, independent of N.

1.3.2 Testing distributions over structured domains

In the final portion of the paper we view the domain $[N]$ as an ordered set $1 \leq \cdots \leq N$. (Note that in all the testing problems and results described previously, the domain could just as well have been viewed as an unstructured set of abstract points $x_1, \ldots, x_N$.) With this perspective it is natural to consider an additional oracle. We define an ICOND (short for "interval-cond") *oracle for* D as a restricted version of COND_D, which only accepts input sets S that are intervals $S = [a, b] = \{a, a+1, \ldots, b\}$ for some $a \leq b \in [N]$ (note that taking $a = 1$, $b = N$ this provides the power of a SAMP_D oracle).

We give an $\tilde{O}((\log N)^3/\epsilon^3)$-query ICOND_D algorithm for testing whether D is uniform versus ϵ-far from uniform. We show that a $(\log N)^{\Omega(1)}$ query complexity is inherent for uniformity testing using ICOND_D, by proving an $\Omega(\log N/\log\log N)$-query ICOND_D lower bound.

Along the way to establishing our main testing results described above, we develop several powerful tools for analyzing distributions in the COND and PCOND models, which we believe may be of independent interest and utility in subsequent work on the COND and PCOND models. These include as mentioned above a procedure for approximately simulating an "evaluation oracle", as well as a procedure for estimating the weight of the "neighborhood" of a given point in the domain of the distribution. (See further discussion of these tools in Subsection 1.3.3.)

1.3.3 A high-level discussion of our algorithms

To maintain focus here we describe only the ideas behind our algorithms; intuition for each of our lower bounds can be found in an informal discussion preceding the formal proof, see the beginnings

of Sections 4.2, 5.2, and 9. As can be seen in the following discussion, our algorithms share some common themes, though each has its own unique idea/technique, which we emphasize below.

Our simplest testing algorithm is the algorithm for **testing whether D is uniform** over $[N]$ (using PCOND_D queries). The algorithm is based on the observation that if a distribution is ϵ-far from uniform, then the total weight (according to D) of points $y \in [N]$ for which $D(y) \geq (1+\Omega(\epsilon))/N$ is $\Omega(\epsilon)$, and the fraction of points $x \in [N]$ for which $D(x) \leq (1-\Omega(\epsilon))/N$ is $\Omega(\epsilon)$. If we obtain such a pair of points (x, y), then we can detect this deviation from uniformity by performing $\Theta(1/\epsilon^2)$ PCOND_D queries on the pair. Such a pair can be obtained with high probability by making $\Theta(1/\epsilon)$ SAMP_D queries (so as to obtain y) as well as selecting $\Theta(1/\epsilon)$ points uniformly (so as to obtain x). This approach yields an algorithm whose complexity grows like $1/\epsilon^4$. To actually get an algorithm with query complexity $\tilde{O}(1/\epsilon^2)$ (which, as our lower bound shows, is tight), a slightly more refined approach is applied.

When we take the next step to **testing equality to an arbitrary (but fully specified) distribution** D^*, the abovementioned observation generalizes so as to imply that if we sample $\Theta(1/\epsilon)$ points from D and $\Theta(1/\epsilon)$ from D^*, then with high probability we shall obtain a pair of points (x, y) such that $D(x)/D(y)$ differs by at least $(1 \pm \Omega(\epsilon))$ from $D^*(x)/D^*(y)$. Unfortunately, this cannot necessarily be detected by a small number of PCOND_D queries since (as opposed to the uniform case), $D^*(x)/D^*(y)$ may be very large or very small. However, we show that by sampling from both D and D^* and allowing the number of samples to grow with $\log N$, with high probability we either obtain a pair of points as described above for which $D^*(x)/D^*(y)$ is a constant, or we detect that for some set of points B we have that $|D(B) - D^*(B)|$ is relatively large.[4]

As noted previously, we prove a lower bound showing that a polynomial dependence on $\log N$ is unavoidable if only PCOND_D queries (in addition to standard sampling) are allowed. To obtain our more efficient $\mathrm{poly}(1/\epsilon)$-queries algorithm, which uses more general COND_D queries, we extend the observation from the uniform case in a different way. Specifically, rather than comparing the relative weight of pairs of points, we compare the relative weight of pairs in which one element is a point and the other is a subset of points. Roughly speaking, we show how points can be paired with subsets of points of comparable weight (according to D^*) such that the following holds. If D is far from D^*, then by taking $\tilde{O}(1/\epsilon)$ samples from D and selecting subsets of points in an appropriate manner (depending on D^*), we can obtain (with high probability) a point x and a subset Y such that $D(x)/D(Y)$ differs significantly from $D^*(x)/D^*(Y)$ *and* $D^*(x)/D^*(Y)$ is a constant.

In our next step, to **testing equality between two unknown distributions D_1 and D_2**, we need to cope with the fact that we no longer "have a hold" on a known distribution. Our PCOND algorithm can be viewed as creating such a hold in the following sense. By sampling from D_1 we obtain (with high probability) a (relatively small) set of points R that *cover* the distribution D_1. By "covering" we mean that except for a subset having small weight according to D_1, all points y in $[N]$ have a *representative* $r \in R$, i.e. a point r such that $D_1(y)$ is close to $D_1(r)$. We then show that if D_2 is far from D_1, then one of the following must hold: (1) There is relatively large weight, either according to D_1 or according to D_2, on points y such that for some $r \in R$ we have that $D_1(y)$ is close to $D_1(r)$ but $D_2(y)$ is not sufficiently close to $D_2(r)$; (2) There exists a point $r \in R$ such

[4] Here we use B for "Bucket", as we consider a bucketing of the points in $[N]$ based on their weight according to D^*. We note that bucketing has been used extensively in the context of testing properties of distributions, see e.g. [BFR+10, BFF+01].

that the set of points y for which $D_1(y)$ is close to $D_1(r)$ has significantly different weight according to D_2 as compared to D_1. We note that this algorithm can be viewed as a variant of the PCOND algorithm for the case when one of the distributions is known (where the "buckets" B, which were defined by D^* in that algorithm (and were disjoint), are now defined by the points in R (and are not necessarily disjoint)).

As noted previously, our (general) COND algorithm for testing the equality of two (unknown) distributions is based on a subroutine that estimates $D(x)$ (to within $(1 \pm O(\epsilon))$) for a given point x given access to COND_D. Obtaining such an estimate for *every* $x \in [N]$ cannot be done efficiently for some distributions.[5] However, we show that if we allow the algorithm to output UNKNOWN on some subset of points with total weight $O(\epsilon)$, then the relaxed task can be performed using $\mathrm{poly}(\log N, 1/\epsilon)$ queries, by performing a kind of randomized binary search "with exceptions". This relaxed version, which we refer to as an *approximate* EVAL *oracle*, suffices for our needs in distinguishing between the case that D_1 and D_2 are the same distribution and the case in which they are far from each other. It is possible that this procedure will be useful for other tasks as well.

The algorithm for **estimating the distance to uniformity** (which uses $\mathrm{poly}(1/\epsilon)$ PCOND_D queries) is based on a subroutine for finding a *reference point* x together with an estimate $\widehat{D}(x)$ of $D(x)$. A reference point should be such that $D(x)$ is relatively close to $1/N$ (if such a point cannot be found then it is evidence that D is very far from uniform). Given a reference point x (together with $\widehat{D}(x)$) it is possible to estimate the distance to uniformity by obtaining (using PCOND queries) estimates of the ratio between $D(x)$ and $D(y)$ for $\mathrm{poly}(1/\epsilon)$ uniformly selected points y. The procedure for finding a reference point x together with $\widehat{D}(x)$ is based on estimating both the weight and the size of a subset of points y such that $D(y)$ is close to $D(x)$. The procedure shares a common subroutine, ESTIMATE-NEIGHBORHOOD, with the PCOND algorithm for testing equivalence between two unknown distributions.

Finally, the ICOND_D algorithm for testing uniformity is based on a version of the approximate EVAL oracle mentioned previously, which on one hand uses only ICOND_D (rather than general COND_D) queries, and on the other hand exploits the fact that we are dealing with the uniform distribution rather than an arbitrary distribution.

1.4 The work of Chakraborty et al. [CFGM13]

Chakraborty et al. [CFGM13] proposed essentially the same COND model that we study, differing only in what happens on query sets S such that $D(S) = 0$. In our model such a query causes the COND oracle and algorithm to return FAIL, while in their model such a query returns a uniform random $i \in S$.

Related to testing equality of distributions, [CFGM13] provides an (adaptive) algorithm for testing whether D is equivalent to a specified distribution D^* using $\mathrm{poly}(\log^* N, 1/\epsilon)$ COND queries. Recall that we give an algorithm for this problem that performs $\tilde{O}(1/\epsilon^4)$ COND queries. [CFGM13] also gives a *non-adaptive* algorithm for this problem that performs $\mathrm{poly}(\log N, 1/\epsilon)$ COND queries.[6]

[5] As an extreme case consider a distribution D for which $D(1) = 1 - \phi$ and $D(2) = \cdots = D(N) = \phi/(N-1)$ for some very small ϕ (which in particular may depend on N), and for which we are interested in estimating $D(2)$. This requires $\Omega(1/\phi)$ queries.

[6] We note that it is only possible for them to give a non-adaptive algorithm because their model is more permissive

Testing equivalence between two unknown distributions is not considered in [CFGM13], and the same is true for testing in the PCOND model.

[CFGM13] also presents additional results for a range of other problems, which we discuss below:

- An (adaptive) algorithm for testing uniformity that performs $\text{poly}(1/\epsilon)$ queries.[7] The sets on which the algorithms performs COND queries are of size linear in $1/\epsilon$. Recall that our algorithm for this problem performs $\tilde{O}(1/\epsilon^2)$ PCOND queries and that we show that every algorithm must perform $\Omega(1/\epsilon^2)$ queries (when there is no restriction on the types of queries). We note that their analysis uses the same observation that ours does regarding distributions that are far from uniform (see the discussion in Subsection 1.3.3), but exploits it in a different manner.

  They also give a non-adaptive algorithm for this problem that performs $\text{poly}(\log N, 1/\epsilon)$ COND queries and show that $\Omega(\log \log N)$ is a lower bound on the necessary number of queries for non-adaptive algorithms.

- An (adaptive) algorithm for testing whether D is equivalent to a specified distribution D^* using $\text{poly}(\log^* N, 1/\epsilon)$ COND queries. Recall that we give an algorithm for this problem that performs $\tilde{O}(1/\epsilon^4)$ COND queries.

  They also give a non-adaptive algorithm for this problem that performs $\text{poly}(\log N, 1/\epsilon)$ COND queries.

- An (adaptive) algorithm for testing any label-invariant (i.e., invariant under permutations of the domain) property that performs $\text{poly}(\log N, 1/\epsilon)$ COND queries. As noted in [CFGM13], this in particular implies an algorithm with this complexity for estimating the distance to uniformity. Recall that we give an algorithm for this estimation problem that performs $\text{poly}(1/\epsilon)$ PCOND queries.

  The algorithm for testing any label-invariant property is based on learning a certain approximation of the distribution D and in this process defining some sort of approximate EVAL oracle. To the best of our understanding, our notion of an approximate EVAL oracle (which is used to obtain one or our results for testing equivalence between two unknown distributions) is quite different.

  They also show that there exists a label-invariant property for which any adaptive algorithm must perform $\Omega(\sqrt{\log \log N})$ COND queries.

- Finally they show that there exist general properties that require $\Omega(N)$ COND queries.

than ours (if a query set S is proposed for which $D(S) = 0$, their model returns a uniform random element of S while our model returns FAIL). In our stricter model, any non-adaptive algorithm which queries a proper subset $S \subsetneq N$ would output FAIL on some distribution D.

[7] The precise polynomial is not specified – we believe it is roughly $1/\epsilon^4$ as it follows from an application of the identity tester of [BFF+01] with distance $\Theta(\epsilon^2)$ on a domain of size $O(1/\epsilon)$.

2 Preliminaries

2.1 Definitions

Throughout the paper we shall work with discrete distributions over an N-element set whose elements are denoted $\{1, \ldots, N\}$; we write $[N]$ to denote $\{1, \ldots, N\}$ and $[a, b]$ to denote $\{a, \ldots, b\}$. For a distribution D over $[N]$ we write $D(i)$ to denote the probability of i under D, and for $S \subseteq [N]$ we write $D(S)$ to denote $\sum_{i \in S} D(i)$. For $S \subseteq [N]$ such that $D(S) > 0$ we write D_S to denote the conditional distribution of D restricted to S, so $D_S(i) = \frac{D(i)}{D(S)}$ for $i \in S$ and $D_S(i) = 0$ for $i \notin S$.

As is standard in property testing of distributions, throughout this work we measure the distance between two distributions D_1 and D_2 using the *total variation distance*:

$$d_{\mathrm{TV}}(D_1, D_2) \stackrel{\text{def}}{=} \frac{1}{2}\|D_1 - D_2\|_1 = \frac{1}{2}\sum_{i \in [N]} |D_1(i) - D_2(i)| = \max_{S \subseteq [N]} |D_1(S) - D_2(S)|.$$

We may view a *property* $\mathcal{P}$ of distributions over $[N]$ as a subset of all distributions over $[N]$ (consisting of all distributions that have the property). The distance from D to a property $\mathcal{P}$, denoted $d_{\mathrm{TV}}(D, \mathcal{P})$, is defined as $\inf_{D' \in \mathcal{P}}\{d_{\mathrm{TV}}(D, D')\}$.

We define testing algorithms for properties of distributions over $[N]$ as follows:

Definition 2 *Let $\mathcal{P}$ be a property of distributions over $[N]$. Let ORACLE_D be some type of oracle which provides access to D. A $q(\epsilon, N)$-query ORACLE* testing algorithm for *$\mathcal{P}$ is an algorithm T which is given ϵ, N as input parameters and oracle access to an ORACLE_D oracle. For any distribution D over $[N]$ algorithm T makes at most $q(\epsilon, N)$ calls to ORACLE_D, and:*

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

This definition can easily be extended to cover situations in which there are two "unknown" distributions D_1, D_2 that are accessible via ORACLE_{D_1} and ORACLE_{D_2} oracles. In particular we shall consider algorithms for testing whether $D_1 = D_2$ versus $d_{\mathrm{TV}}(D_1, D_2)$ in such a setting. We sometimes write T^{ORACLE_D} to indicate that T has access to ORACLE_D.

2.2 Useful tools

On several occasions we will use the *data processing inequality for variation distance*. This fundamental result says that for any two distributions D, D', applying any (possibly randomized) function to D and D' can never increase their statistical distance; see e.g. part (iv) of Lemma 2 of [Rey11] for a proof of this lemma.

Lemma 1 (Data Processing Inequality for Total Variation Distance) *Let D, D' be two distributions over a domain Ω. Fix any randomized function[8] F on Ω, and let $F(D)$ be the distribution such that a draw from $F(D)$ is obtained by drawing independently x from D and f from F and then outputting $f(x)$ (likewise for $F(D')$). Then we have*

$$d_{\mathrm{TV}}(F(D), F(D')) \leq d_{\mathrm{TV}}(D, D').$$

We next give several variants of Chernoff bounds (see e.g. Chapter 4 of [MR95]).

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

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

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

and

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

The bound in Equation (2) is derived from the following more general bound, which holds from any $\gamma > 0$:

$$\Pr\left[\sum_{i=1}^m Y_i > (1+\gamma)P\right] \leq \left(\frac{e^\gamma}{(1+\gamma)^{1+\gamma}}\right)^P, \tag{4}$$

and which also implies that for any $B > 2eP$,

$$\Pr\left[\sum_{i=1}^m Y_i > B\right] \leq 2^{-B}. \tag{5}$$

The following extension of the multiplicative bound is useful when we only have upper and/or lower bounds on P (see Exercise 1.1 of [DP09]):

Corollary 2 *In the setting of Theorem 1 suppose that $P_L \leq P \leq P_H$. Then for any $\gamma \in (0,1]$, we have*

$$\Pr\left[\sum_{i=1}^m Y_i > (1+\gamma)P_H\right] < \exp(-\gamma^2 P_H/3) \tag{6}$$

$$\Pr\left[\sum_{i=1}^m Y_i < (1-\gamma)P_L\right] < \exp(-\gamma^2 P_L/2) \tag{7}$$

[8] Which can be seen as a distribution over functions over Ω.

We will also use the following corollary of Theorem 1:

Corollary 3 *Let $0 \leq w_1, \ldots, w_m \in \mathbb{R}$ be such that $w_i \leq \kappa$ for all $i \in [m]$ where $\kappa \in (0,1]$. Let $X_1, \ldots, X_m$ be i.i.d. Bernoulli random variables with $\Pr[X_i = 1] = 1/2$ for all i, and let $X = \sum_{i=1}^m w_i X_i$ and $W = \sum_{i=1}^m w_i$. For any $\gamma \in (0,1]$,*

$$\Pr\left[X > (1+\gamma)\frac{W}{2}\right] < \exp\left(-\gamma^2 \frac{W}{6\kappa}\right) \quad \textit{and} \quad \Pr\left[X < (1-\gamma)\frac{W}{2}\right] < \exp\left(-\gamma^2 \frac{W}{4\kappa}\right) ,$$

and for any $B > e \cdot W$,

$$\Pr[X > B] < 2^{-B/\kappa} .$$

Proof: Let $w'_i = w_i/\kappa$ (so that $w'_i \in [0,1]$), let $W' = \sum_{i=1}^m w'_i = W/\kappa$, and for each $i \in [m]$ let $Y_i = w'_i X_i$, so that Y_i takes on values in $[0,1]$ and $\mathrm{E}[Y_i] = w'_i/2$. Let $X' = \sum_{i=1}^m w'_i X_i = \sum_{i=1}^m Y_i$, so that $\mathrm{E}[X'] = W'/2$. By the definitions of W' and X' and by Equation (2), for any $\gamma \in (0,1]$,

$$\Pr\left[X > (1+\gamma)\frac{W}{2}\right] = \Pr\left[X' > (1+\gamma)\frac{W'}{2}\right] < \exp\left(-\gamma^2 \frac{W'}{6}\right) = \exp\left(-\gamma^2 \frac{W}{6\kappa}\right), \tag{8}$$

and similarly by Equation (3)

$$\Pr\left[X < (1-\gamma)\frac{W}{2}\right] < \exp\left(-\gamma^2 \frac{W}{4\kappa}\right) . \tag{9}$$

For $B > e \cdot W = 2e \cdot W/2$ we apply Equation (5) and get

$$\Pr\left[X > B\right] = \Pr\left[X' > B/\kappa\right] < 2^{-B/\kappa}, \tag{10}$$

as claimed. ■

3 Some useful procedures

In this section we describe some procedures that will be used by our algorithms. On a first pass the reader may wish to focus on the explanatory prose and performance guarantees of these procedures (i.e., the statements of Lemma 2 and Lemma 3, as well as Definition 3 and Theorem 4) and otherwise skip to p.27; the internal details of the proofs are not necessary for the subsequent sections that use these procedures.

3.1 The procedure Compare

We start by describing a procedure that estimates the ratio between the weights of two disjoint sets of points by performing COND queries on the union of the sets. More precisely, it estimates the ratio (to within $1 \pm \eta$) if the ratio is not too high and not too low. Otherwise, it may output high or low, accordingly. In the special case when each set is of size one, the queries performed are PCOND queries.

Algorithm 1: COMPARE

Input: COND query access to a distribution D over $[N]$, disjoint subsets $X, Y \subset [N]$, parameters $\eta \in (0,1]$, $K \geq 1$, and $\delta \in (0, 1/2]$.

1. Perform $\Theta\left(\frac{K \log(1/\delta)}{\eta^2}\right)$ COND_D queries on the set $S = X \cup Y$, and let $\hat{\mu}$ be the fraction of times that a point $y \in Y$ is returned.
2. If $\hat{\mu} < \frac{2}{3} \cdot \frac{1}{K+1}$, then return Low.
3. Else, if $1 - \hat{\mu} < \frac{2}{3} \cdot \frac{1}{K+1}$, then return High.
4. Else return $\rho = \frac{\hat{\mu}}{1-\hat{\mu}}$.

Lemma 2 *Given as input two disjoint subsets of points X, Y together with parameters $\eta \in (0,1]$, $K \geq 1$, and $\delta \in (0, 1/2]$, as well as* COND *query access to a distribution D, the procedure* COMPARE *(Algorithm 1) either outputs a value $\rho > 0$ or outputs* High *or* Low*, and satisfies the following:*

1. *If $D(X)/K \leq D(Y) \leq K \cdot D(X)$ then with probability at least $1 - \delta$ the procedure outputs a value $\rho \in [1-\eta, 1+\eta] D(Y)/D(X)$;*
2. *If $D(Y) > K \cdot D(X)$ then with probability at least $1 - \delta$ the procedure outputs either* High *or a value $\rho \in [1-\eta, 1+\eta] D(Y)/D(X)$;*
3. *If $D(Y) < D(X)/K$ then with probability at least $1 - \delta$ the procedure outputs either* Low *or a value $\rho \in [1-\eta, 1+\eta] D(Y)/D(X)$.*

The procedure performs $O\left(\frac{K \log(1/\delta)}{\eta^2}\right)$ COND *queries on the set $X \cup Y$.*

Proof: The bound on the number of queries performed by the algorithm follows directly from the description of the algorithm, and hence we turn to establish its correctness.

Let $w(X) = \frac{D(X)}{D(X)+D(Y)}$ and let $w(Y) = \frac{D(Y)}{D(X)+D(Y)}$. Observe that $\frac{w(Y)}{w(X)} = \frac{D(Y)}{D(X)}$ and that for $\hat{\mu}$ as defined in Line 1 of the algorithm, $\mathrm{E}[\hat{\mu}] = w(Y)$ and $\mathrm{E}[1 - \hat{\mu}] = w(X)$. Also observe that for any $B \geq 1$, if $D(Y) \geq D(X)/B$, then $w(Y) \geq \frac{1}{B+1}$ and if $D(Y) \leq B \cdot D(X)$, then $w(X) \geq \frac{1}{B+1}$.

Let E_1 be the event that $\hat{\mu} \in [1 - \eta/3, 1 + \eta/3] w(Y)$ and let E_2 be the event that $(1 - \hat{\mu}) \in [1 - \eta/3, 1 + \eta/3] w(X)$. Given the number of COND queries performed on the set $X \cup Y$, by applying a multiplicative Chernoff bound (see Theorem 1), if $w(Y) \geq \frac{1}{4K}$ then with probability at least $1 - \delta/2$ the event E_1 holds, and if $w(X) \geq \frac{1}{4K}$, then with probability at least $1 - \delta/2$ the event E_2 holds. We next consider the three cases in the lemma statement.

1. If $D(X)/K \leq D(Y) \leq KD(X)$, then by the discussion above, $w(Y) \geq \frac{1}{K+1}$, $w(X) \geq \frac{1}{K+1}$, and with probability at least $1 - \delta$ we have that $\hat{\mu} \in [1 - \eta/3, 1 + \eta/3] w(Y)$ and $(1 - \hat{\mu}) \in [1 - \eta/3, 1 + \eta/3] w(X)$. Conditioned on these bounds holding,

$$\hat{\mu} \geq \frac{1 - \eta/3}{K+1} \geq \frac{2}{3} \cdot \frac{1}{K+1} \quad \text{and} \quad 1 - \hat{\mu} \geq \frac{2}{3} \cdot \frac{1}{K+1} .$$

It follows that the procedure outputs a value $\rho = \frac{\hat{\mu}}{1-\hat{\mu}} \in [1-\eta, 1+\eta]\frac{w(Y)}{w(X)}$ as required by Item 1.

2. If $D(Y) > K \cdot D(X)$, then we consider two subcases.

   (a) If $D(Y) > 3K \cdot D(X)$, then $w(X) < \frac{1}{3K+1}$, so that by a multiplicative Chernoff bound (stated in Corollary 2), with probability at least $1-\delta$ we have that

   $$1 - \hat{\mu} < \frac{1+\eta/3}{3K+1} \leq \frac{4}{3} \cdot \frac{1}{3K+1} \leq \frac{2}{3} \cdot \frac{1}{K+1},$$

   causing the algorithm to output High. Thus Item 2 is established for this subcase.

   (b) If $K \cdot D(X) < D(Y) \leq 3K \cdot D(X)$, then $w(X) \geq \frac{1}{3K+1}$ and $w(Y) \geq \frac{1}{2}$, so that the events E_1 and E_2 both hold with probability at least $1-\delta$. Assume that these events in fact hold. This implies that $\hat{\mu} \geq \frac{1-\eta/3}{2} \geq \frac{2}{3} \cdot \frac{1}{K+1}$, and the algorithm either outputs High or outputs $\rho = \frac{\hat{\mu}}{1-\hat{\mu}} \in [1-\eta, 1+\eta]\frac{w(Y)}{w(X)}$, so Item 2 is established for this subcase as well.

3. If $D(Y) < D(X)/K$, so that $D(X) > K \cdot D(Y)$, then the exact same arguments are applied as in the previous case, just switching the roles of Y and X and the roles of $\hat{\mu}$ and $1-\hat{\mu}$ so as to establish Item 3.

We have thus established all items in the lemma. ■

3.2 The procedure Estimate-Neighborhood

In this subsection we describe a procedure that, given a point x, provides an estimate of the weight of a set of points y such that $D(y)$ is similar to $D(x)$. In order to specify the behavior of the procedure more precisely, we introduce the following notation. For a distribution D over $[N]$, a point $x \in [N]$ and a parameter $\gamma \in [0,1]$, let

$$U_\gamma^D(x) \stackrel{\text{def}}{=} \left\{ y \in [N] : \frac{1}{1+\gamma} D(x) \leq D(y) \leq (1+\gamma) D(x) \right\} \tag{11}$$

denote the set of points whose weight is "γ-close" to the weight of x. If we take a sample of points distributed according to D, then the expected fraction of these points that belong to $U_\gamma^D(x)$ is $D(U_\gamma^D(x))$. If this value is not too small, then the actual fraction in the sample is close to the expected value. Hence, if we could efficiently determine for any given point y whether or not it belongs to $U_\gamma^D(x)$, then we could obtain a good estimate of $D(U_\gamma^D(x))$. The difficulty is that it is not possible to perform this task efficiently for "boundary" points y such that $D(y)$ is very close to $(1+\gamma)D(x)$ or to $\frac{1}{1+\gamma}D(x)$. However, for our purposes, it is not important that we obtain the weight and size of $U_\gamma^D(x)$ for a specific γ, but rather it suffices to do so for γ in a given range, as stated in the next lemma. The parameter β in the lemma is the threshold above which we expect the algorithm to provide an estimate of the weight, while $[\kappa, 2\kappa)$ is the range in which γ is permitted to lie; finally, η is the desired (multiplicative) accuracy of the estimate, while δ is a bound on the probability of error allowed to the subroutine.

Lemma 3 *Given as input a point x together with parameters $\kappa, \beta, \eta, \delta \in (0, 1/2]$ as well as* PCOND *query access to a distribution D, the procedure* Estimate-Neighborhood *(Algorithm 2) outputs a pair $(\hat{w}, \alpha) \in [0,1] \times (\kappa, 2\kappa)$ such that α is uniformly distributed in $\{\kappa + i\theta\}_{i=0}^{\kappa/\theta - 1}$ for $\theta = \frac{\kappa\eta\beta\delta}{64}$, and such that the following holds:*

1. *If $D(U_\alpha^D(x)) \geq \beta$, then with probability at least $1 - \delta$ we have $\hat{w} \in [1-\eta, 1+\eta] \cdot D(U_\alpha^D(x))$, and $D(U_{\alpha+\theta}^D(x) \setminus U_\alpha^D(x)) \leq \eta\beta/16$;*
2. *If $D(U_\alpha^D(x)) < \beta$, then with probability at least $1 - \delta$ we have $\hat{w} \leq (1+\eta) \cdot \beta$, and $D(U_{\alpha+\theta}^D(x) \setminus U_\alpha^D(x)) \leq \eta\beta/16$.*

The number of PCOND *queries performed by the procedure is $O\left(\frac{\log(1/\delta) \cdot \log(\log(1/\delta)/(\delta\beta\eta^2))}{\kappa^2\eta^4\beta^3\delta^2}\right)$.*

Algorithm 2: Estimate-Neighborhood

Input: PCOND query access to a distribution D over $[N]$, a point $x \in [N]$ and parameters $\kappa, \beta, \eta, \delta \in (0, 1/2]$

1: Set $\theta = \frac{\kappa\eta\beta\delta}{64}$ and $r = \frac{\kappa}{\theta} = \frac{64}{\eta\beta\delta}$.
2: Select a value $\alpha \in \{\kappa + i\theta\}_{i=0}^{r-1}$ uniformly at random.
3: Call the SAMP_D oracle $\Theta(\log(1/\delta)/(\beta\eta^2))$ times and let S be the set of points obtained.
4: For each point y in S call $\textsc{Compare}_D(\{x\}, \{y\}, \theta/4, 4, \delta/(4|S|))$ (if a point y appears more than once in S, then Compare is called only once on y).
5: Let $\hat{w}$ be the fraction of occurrences of points y in S for which Compare returned a value $\rho(y) \in [1/(1+\alpha+\theta/2), (1+\alpha+\theta/2)]$. (That is, S is viewed as a multiset.)
6: Return $(\hat{w}, \alpha)$.

Proof of Lemma 3: The number of PCOND queries performed by Estimate-Neighborhood is the size of S times the number of PCOND queries performed in each call to Compare. By the setting of the parameters in the calls to Compare, the total number of PCOND queries is $O\left(\frac{(|S|)\cdot\log|S|/\delta)}{\theta^2}\right) = O\left(\frac{\log(1/\delta)\cdot\log(\log(1/\delta)/(\delta\beta\eta^2))}{\kappa^2\eta^4\beta^3\delta^2}\right)$. We now turn to establishing the correctness of the procedure.

Since D and x are fixed, in what follows we shall use the shorthand U_γ for $U_\gamma^D(x)$. For $\alpha \in \{\kappa + i\theta\}_{i=0}^{r-1}$, let $\Delta_\alpha \stackrel{\text{def}}{=} U_{\alpha+\theta} \setminus U_\alpha$. We next define several "desirable" events. In all that follows we view S as a multiset.

1. Let E_1 be the event that $D(\Delta_\alpha) \leq 4/(\delta r)$. Since there are r disjoint sets Δ_α for $\alpha \in \{\kappa + i\theta\}_{i=0}^{r-1}$, the probability that E_1 occurs (taken over the uniform choice of α) is at least $1 - \delta/4$. From this point on we fix α and assume E_1 holds.
2. The event E_2 is that $|S \cap \Delta_\alpha|/|S| \leq 8/(\delta r)$ (that is, at most twice the upper bound on the expected value). By applying the multiplicative Chernoff bound using the fact that $|S| = \Theta(\log(1/\delta)/(\beta\eta^2)) = \Omega(\log(1/\delta) \cdot (\delta r))$, we have that $\Pr_S[E_2] \geq 1 - \delta/4$.

3. The event E_3 is defined as follows: If $D(U_\alpha) \geq \beta$, then $|S \cap U_\alpha|/|S| \in [1-\eta/2, 1+\eta/2] \cdot D(U_\alpha)$, and if $D(U_\alpha) < \beta$, then $|S \cap U_\alpha|/|S| < (1+\eta/2) \cdot \beta$. Once again applying the multiplicative Chernoff bound (for both cases) and using that fact that $|S| = \Theta(\log(1/\delta)/(\beta\eta^2))$, we have that $\Pr_S[E_3] \geq 1 - \delta/4$.

4. Let E_4 be the event that all calls to COMPARE return an output as specified in Lemma 2. Given the setting of the confidence parameter in the calls to COMPARE we have that $\Pr[E_4] \geq 1 - \delta/4$ as well.

Assume from this point on that events E_1 through E_4 all hold where this occurs with probability at least $1 - \delta$. By the definition of Δ_α and E_1 we have that $D(U_{\alpha+\theta} \setminus U_\alpha) \leq 4/(\delta r) = \eta\beta/16$, as required (in both items of the lemma). Let T be the (multi-)subset of points y in S for which COMPARE returned a value $\rho(y) \in [1/(1+\alpha+\theta/2), (1+\alpha+\theta/2)]$ (so that $\hat{w}$, as defined in the algorithm, equals $|T|/|S|$). Note first that conditioned on E_4 we have that for every $y \in U_{2\kappa}$ it holds that the output of COMPARE when called on $\{x\}$ and $\{y\}$, denoted $\rho(y)$, satisfies $\rho(y) \in [1-\theta/4, 1+\theta/4](D(y)/D(x))$, while for $y \notin U_{2\kappa}$ either COMPARE outputs High or Low or it outputs a value $\rho(y) \in [1-\theta/4, 1+\theta/4](D(y)/D(x))$. This implies that if $y \in U_\alpha$, then $\rho(y) \leq (1+\alpha) \cdot (1+\theta/4) \leq 1+\alpha+\theta/2$ and $\rho(y) \geq (1+\alpha)^{-1} \cdot (1-\theta/4) \geq (1+\alpha+\theta/2)^{-1}$, so that $S \cap U_\alpha \subseteq T$. On the other hand, if $y \notin U_{\alpha+\theta}$ then either $\rho(y) > (1+\alpha+\theta) \cdot (1-\theta/4) \geq 1+\alpha+\theta/2$ or $\rho(y) < (1+\alpha+\theta)^{-1} \cdot (1+\theta/4) \leq (1+\alpha+\theta/2)^{-1}$ so that $T \subseteq S \cap U_{\alpha+\theta}$. Combining the two we have:

$$S \cap U_\alpha \;\subseteq\; T \;\subseteq\; S \cap U_{\alpha+\theta}\,. \tag{12}$$

Recalling that $\hat{w} = \frac{|T|}{|S|}$, the left-hand side of Equation (12) implies that

$$\hat{w} \geq \frac{|S \cap U_\alpha|}{|S|}\,, \tag{13}$$

and by E_1 and E_2, the right-hand-side of Equation (12) implies that

$$\hat{w} \;\leq\; \frac{|S \cap U_\alpha)|}{|S|} + \frac{8}{\delta r} \;\leq\; \frac{|S \cap U_\alpha|}{|S|} + \frac{\beta\eta}{8}\,. \tag{14}$$

We consider the two cases stated in the lemma:

1. If $D(U_\alpha) \geq \beta$, then by Equation (13), Equation (14) and (the first part of) E_3, we have that $\hat{w} \in [1-\eta, 1+\eta] \cdot D(U_\alpha)$.

2. If $D(U_\alpha) < \beta$, then by Equation (14) and (the second part of) E_3, we have that $\hat{w} \leq (1+\eta)\beta$.

The lemma is thus established. ■

3.3 The procedure APPROX-EVAL-SIMULATOR

3.3.1 Approximate EVAL oracles.

We begin by defining the notion of an "approximate EVAL oracle" that we will use. Intuitively this is an oracle which gives a multiplicatively $(1 \pm \epsilon)$-accurate estimate of the value of $D(i)$ for all i

in a fixed set of probability weight at least $1-\epsilon$ under D. More precisely, we have the following definition:

Definition 3 *Let D be a distribution over $[N]$. An (ϵ,δ)-approximate EVAL_D simulator is a randomized procedure ORACLE with the following property: For each $0<\epsilon<1$, there is a fixed set $S^{(\epsilon,D)} \subsetneq [N]$ with $D(S^{(\epsilon,D)}) < \epsilon$ for which the following holds. Given as input an element $i^* \in [N]$, the procedure ORACLE either outputs a value $\alpha \in [0,1]$ or outputs $\mathsf{UNKNOWN}$ or FAIL. The following holds for all $i^* \in [N]$:*

(i) If $i^* \notin S^{(\epsilon,D)}$ then with probability at least $1-\delta$ the output of ORACLE on input i^* is a value $\alpha \in [0,1]$ such that $\alpha \in [1-\epsilon, 1+\epsilon]D(i^*)$;

(i) If $i^* \in S^{(\epsilon,D)}$ then with probability at least $1-\delta$ the procedure either outputs $\mathsf{UNKNOWN}$ or outputs a value $\alpha \in [0,1]$ such that $\alpha \in [1-\epsilon, 1+\epsilon]D(i^*)$.

We note that according to the above definition, it may be the case that different calls to ORACLE on the same input element $i^* \in [N]$ may return different values. However, the "low-weight" set $S^{(\epsilon,D)}$ is an *a priori* fixed set that does not depend in any way on the input point i^* given to the algorithm. The key property of an (ϵ,δ)-approximate EVAL_D oracle is that it reliably gives a multiplicatively $(1\pm\epsilon)$-accurate estimate of the value of $D(i)$ for all i in some fixed set of probability weight at least $1-\epsilon$ under D.

3.3.2 Constructing an approximate EVAL_D simulator using COND_D

In this subsection we show that a COND_D oracle can be used to obtain an approximate EVAL simulator:

Theorem 4 *Let D be any distribution over $[N]$ and let $0<\epsilon,\delta<1$. The algorithm* Approx-Eval-Simulator *has the following properties: It uses*

$$\tilde{O}\left(\frac{(\log N)^5 \cdot (\log(1/\delta))^2}{\epsilon^3}\right)$$

calls to COND_D and it is an (ϵ,δ)-approximate EVAL_D simulator.

A few notes: First, in the proof we give below of Theorem 4 we assume throughout that $0<\epsilon\le 1/40$. This incurs no loss of generality because if the desired ϵ parameter is in $(1/40,1)$ then the parameter can simply be set to $1/40$. We further note that in keeping with our requirement on a COND_D algorithm, the algorithm Approx-Eval-Simulator only ever calls the COND_D oracle on sets S which are either $S=[N]$ or else contain at least one element i that has been returned as the output of an earlier call to COND_D. To see this, note that Line 6 is the only line when COND_D queries are performed. In the first execution of the outer "For" loop clearly all COND queries are on set $S_0=[N]$. In subsequent stages the only way a set S_j is formed is if either (i) S_j is set to $\{i^*\}$ in Line 10, in which case clearly i^* was previously received as the response of a $\mathsf{COND}_D(S_{j-1})$

query, or else (ii) a nonzero fraction of elements $i_1, \ldots, i_m$ received as responses to $\mathsf{COND}_D(S_{j-1})$ queries belong to S_j (see Line 19).

A preliminary simplification. Fix a distribution D over $[N]$. Let Z denote $\mathrm{supp}(D)$, i.e. $Z = \{i \in [N] : D(i) > 0\}$. We first claim that in proving Theorem 4 we may assume without loss of generality that no two distinct elements $i, j \in Z$ have $D(i) = D(j)$ – in other words, we shall prove the theorem under this assumption on D, and we claim that this implies the general result. To see this, observe that if Z contains elements $i \neq j$ with $D(i) = D(j)$, then for any arbitrarily small $\xi > 0$ and any arbitrarily large M we can perturb the weights of elements in Z to obtain a distribution D' supported on Z such that (i) no two elements of Z have the same probability under D', and (ii) for every $S \subseteq [N]$, $S \cap Z \neq \emptyset$ we have $d_{\mathrm{TV}}(D_S, D'_S) \leq \xi/M$. Since the variation distance between D'_S and D_S is at most ξ/M for an arbitrarily small ξ, the variation distance between (the execution of any M-query COND algorithm run on D) and (the execution of any M-query COND algorithm run on D') will be at most ξ. Since ξ can be made arbitrarily small this means that indeed without loss of generality we may work with D' in what follows.

Thus, we henceforth assume that the distribution D has no two elements in $\mathrm{supp}(D)$ with the same weight. For such a distribution we can explicitly describe the set $S^{(\epsilon,D)}$ from Definition 3 that our analysis will deal with. Let $\pi : \{1, \ldots, |Z|\} \to Z$ be the bijection such that $D(\pi(1)) > \cdots > D(\pi(|Z|))$ (note that the bijection π is uniquely defined by the assumption that $D(i) \neq D(j)$ for all distinct $i, j \in Z$). Given a value $0 < \tau < 1$ we define the set $L_{\tau,D}$ to be $([N] \setminus Z) \cup \{\pi(s), \ldots, \pi(|Z|)\}$ where s is the smallest index in $\{1, \ldots, |Z|\}$ such that $\sum_{j=s}^{|Z|} D(\pi(j)) < \tau$ (if $D(\pi(|Z|))$ itself is at least τ then we define $L_{\tau,D} = [N] \setminus Z$). Thus intuitively $L_{\tau,D}$ contains the τ fraction (w.r.t. D) of $[N]$ consisting of the lightest elements. The desired set $S^{(\epsilon,D)}$ is precisely $L_{\epsilon,D}$.

Intuition for the algorithm. The high-level idea of the EVAL_D simulation is the following: Let $i^* \in [N]$ be the input element given to the EVAL_D simulator. The algorithm works in a sequence of stages. Before performing the j-th stage it maintains a set S_{j-1} that contains i^*, and it has a high-accuracy estimate $\hat{D}(S_{j-1})$ of the value of $D(S_{j-1})$. (The initial set S_0 is simply $[N]$ and the initial estimate $\hat{D}(S_0)$ is of course 1.) In the j-th stage the algorithm attempts to construct a subset S_j of S_{j-1} in such a way that (i) $i^* \in S_j$, and (ii) it is possible to obtain a high-accuracy estimate of $D(S_j)/D(S_{j-1})$ (and thus a high-accuracy estimate of $D(S_j)$). If the algorithm cannot construct such a set S_j then it outputs $\mathsf{UNKNOWN}$; otherwise, after at most (essentially) $O(\log N)$ stages, it reaches a situation where $S_j = \{i^*\}$ and so the high-accuracy estimate of $D(S_j) = D(i^*)$ is the desired value.

A natural first idea towards implementing this high-level plan is simply to split S_{j-1} randomly into two pieces and use one of them as S_j. However this simple approach may not work; for example, if S_{j-1} has one or more elements which are very heavy compared to i^*, then with a random split it may not be possible to efficiently estimate $D(S_j)/D(S_{j-1})$ as required in (ii) above. Thus we follow a more careful approach which first identifies and removes "heavy" elements from S_{j-1} in each stage.

In more detail, during the j-th stage, the algorithm first performs COND_D queries on the set S_{j-1} to identify a set $H_j \subseteq S_{j-1}$ of "heavy" elements; this set essentially consists of all elements which individually each contribute at least a κ fraction of the total mass $D(S_{j-1})$. (Here κ is a "not-too-small" quantity but it is significantly less than ϵ.) Next, the algorithm performs

additional COND_D queries to estimate $D(i^*)/D(S_{j-1})$. If this fraction exceeds $\kappa/20$ then it is straightforward to estimate $D(i^*)/D(S_{j-1})$ to high accuracy, so using $\hat{D}(S_{j-1})$ it is possible to obtain a high-quality estimate of $D(i^*)$ and the algorithm can conclude. However, the typical case is that $D(i^*)/D(S_{j-1}) < \kappa/20$. In this case, the algorithm next estimates $D(H_j)/D(S_{j-1})$. If this is larger than $1-\epsilon/10$ then the algorithm outputs $\mathsf{UNKNOWN}$ (see below for more discussion of this). If $D(H_j)/D(S_{j-1})$ is less than $1-\epsilon/10$ then $D(S_{j-1} \setminus H_j)/D(S_{j-1}) \geq \epsilon/10$ (and so $D(S_{j-1} \setminus H_j)/D(S_{j-1})$ can be efficiently estimated to high accuracy), but each element k of $S_{j-1} \setminus H_j$ has $D(k)/D(S_{j-1}) \leq \kappa \ll \epsilon/10 \leq D(S_{j-1} \setminus H_j)/D(S_{j-1})$. Thus it must be the case that the weight under D of $S_{j-1} \setminus H_j$ is "spread out" over many "light" elements.

Given that this is the situation, the algorithm next chooses S'_j to be a random subset of $S_{j-1} \setminus (H_j \cup \{i^*\})$, and sets S_j to be $S'_j \cup \{i^*\}$. It can be shown that with high probability (over the random choice of S_j) it will be the case that $D(S_j) \geq \frac{1}{3} D(S_{j-1} \setminus H_j)$ (this relies crucially on the fact that the weight under D of $S_{j-1} \setminus H_j$ is "spread out" over many "light" elements). This makes it possible to efficiently estimate $D(S_j)/D(S_{j-1} \setminus H_j)$; together with the high-accuracy estimate of $D(S_{j-1} \setminus H_j)/D(S_{j-1})$ noted above, and the high-accuracy estimate $\hat{D}(S_{j-1})$ of $D(S_{j-1})$, this means it is possible to efficiently estimate $D(S_j)$ to high accuracy as required for the next stage. (We note that after defining S_j but before proceeding to the next stage, the algorithm actually checks to be sure that S_j contains at least one point that was returned from the $\mathsf{COND}_D(S_{j-1})$ calls made in the past stage. This check ensures that whenever the algorithm calls $\mathsf{COND}_D(S)$ on a set S, it is guaranteed that $D(S) > 0$ as required by our COND_D model. Our analysis shows that doing this check does not affect correctness of the algorithm since with high probability the check always passes.)

Intuition for the analysis. We require some definitions to give the intuition for the analysis establishing correctness. Fix a nonempty subset $S \subseteq [N]$. Let π_S be the bijection mapping $\{1, \ldots, |S|\}$ to S in such a way that $D_S(\pi_S(1)) > \cdots > D_S(\pi_S(|S|))$, i.e. $\pi_S(1), \ldots, \pi_S(|S|)$ is a listing of the elements of S in order from heaviest under D_S to lightest under D_S. Given $j \in S$, we define the *S-rank of j*, denoted $\mathrm{rank}_S(j)$, to be the value $\sum_{i: D_S(\pi(i)) \leq D_S(j)} D_S(\pi(i))$, i.e. $\mathrm{rank}_S(j)$ is the sum of the weights (under D_S) of all the elements in S that are no heavier than j under D_S. Note that having $i^* \notin L_{\epsilon,N}$ implies that $\mathrm{rank}_{[N]}(i^*) \geq \epsilon$.

We first sketch the argument for correctness. (It is easy to show that the algorithm only outputs FAIL with very small probability so we ignore this possibility below.) Suppose first that $i^* \notin L_{\epsilon,D}$. A key lemma shows that if $i^* \notin L_{\epsilon,D}$ (and hence $\mathrm{rank}_{[N]}(i^*) \geq \epsilon$), then with high probability every set S_{j-1} constructed by the algorithm is such that $\mathrm{rank}_{S_{j-1}}(i^*) \geq \epsilon/2$. (In other words, if i^* is not initially among the ϵ-fraction (under D) of lightest elements, then it never "falls too far" to become part of the $\epsilon/2$-fraction (under $D_{S_{j-1}}$) of lightest elements for S_{j-1}, for any j.). Given that (with high probability) i^* always has $\mathrm{rank}_{S_{j-1}}(i^*) \geq \epsilon/2$, though, then it must be the case that (with high probability) the procedure does not output $\mathsf{UNKNOWN}$ (and hence it must with high probability output a numerical value). This is because there are only two places where the procedure can output $\mathsf{UNKNOWN}$, in Lines 14 and 19; we consider both cases below.

1. In order for the procedure to output $\mathsf{UNKNOWN}$ in Line 14, it must be the case that the elements of H_j – each of which individually has weight at least $\kappa/2$ under $D_{S_{j-1}}$ – collectively have weight at least $1 - 3\epsilon/20$ under $D_{S_{j-1}}$ by Line 13. But i^* has weight at most $3\kappa/40$ under

$D_{S_{j-1}}$ (because the procedure did not go to Line 2 in Line 10), and thus i^* would need to be in the bottom $3\epsilon/20$ of the lightest elements, i.e. it would need to have $\text{rank}_{S_{j-1}}(i^*) \leq 3\epsilon/20$; but this contradicts $\text{rank}_{S_{j-1}}(i^*) \geq \epsilon/2$.

2. In order for the procedure to output UNKNOWN in Line 19, it must be the case that all elements $i_1, \ldots, i_m$ drawn in Line 6 are not chosen for inclusion in S_j. In order for the algorithm to reach Line 19, though, it must be the case that at least $(\epsilon/10 - \kappa/20)m$ of these draws do not belong to $H_j \cup \{i^*\}$; since these draws do not belong to H_j each one occurs only a small number of times among the m draws, so there must be many distinct values, and hence the probability that none of these distinct values is chosen for inclusion in S'_j is very low.

Thus we have seen that if $i^* \notin L_{\epsilon,D}$, then with high probability the procedure outputs a numerical value; it remains to show that with high probability this value is a high-accuracy estimate of $D(i^*)$. However, this follows easily from the fact that we inductively maintain a high-quality estimate of $D(S_{j-1})$ and the fact that the algorithm ultimately constructs its estimate of $\hat{D}(i^*)$ only when it additionally has a high-quality estimate of $D(i^*)/D(S_{j-1})$. This fact also handles the case in which $i^* \in L_{\epsilon,D}$ – in such a case it is allowable for the algorithm to output UNKNOWN, so since the algorithm with high probability outputs a high-accuracy estimate when it outputs a numerical value, this means the algorithm performs as required in Case (ii) of Definition 3.

We now sketch the argument for query complexity. We will show that the heavy elements can be identified in each stage using $\text{poly}(\log N, 1/\epsilon)$ queries. Since the algorithm constructs S_j by taking a random subset of S_{j-1} (together with i^*) at each stage, the number of stages is easily bounded by (essentially) $O(\log N)$. Since the final probability estimate for $D(i^*)$ is a product of $O(\log N)$ conditional probabilities, it suffices to estimate each of these conditional probabilities to within a multiplicative factor of $(1 \pm O\left(\frac{\epsilon}{\log N}\right))$. We show that each conditional probability estimate can be carried out to this required precision using only $\text{poly}(\log N, 1/\epsilon)$ calls to COND_D; given this, the overall $\text{poly}(\log N, 1/\epsilon)$ query bound follows straightforwardly.

Now we enter into the actual proof. We begin our analysis with a simple but useful lemma about the "heavy" elements identified in Line 7.

Lemma 4 *With probability at least $1 - \delta/9$, every set H_j that is ever constructed in Line 7 satisfies the following for all $\ell \in S_{j-1}$:*

(i) If $D(\ell)/D(S_{j-1}) > \kappa$, then $\ell \in H_j$;

(ii) If $D(\ell)/D(S_{j-1}) < \kappa/2$ then $\ell \notin H_j$.

Proof: Fix an iteration j. By Line 7 in the algorithm, a point ℓ is included in H_j if it appears at least $\frac{3}{4}\kappa m$ times among $i_1, \ldots, i_m$ (which are the output of COND_D queries on S_{j-1}). For the first item, fix an element ℓ such that $D(\ell)/D(S_{j-1}) > \kappa$. Recall that $m = \Omega(M^2 \log(M/\delta)/(\epsilon^2\kappa)) = \Omega(\log(MN/\delta)/\kappa)$ (since $M = \Omega(\log(N))$). By a multiplicative Chernoff bound, the probability (over the choice of $i_1, \ldots, i_m$ in S_{j-1}) that ℓ appears less than $\frac{3}{4}\kappa m$ times among $i_1, \ldots, i_m$ (that is, less than 3/4 times the lower bound on the expected value) is at most $\delta/(9MN)$ (for an appropriate constant in the setting of m). On the other hand, for each fixed ℓ such that $D(\ell)/D(S_{j-1}) < \kappa/2$,

Algorithm 3: APPROX-EVAL-SIMULATOR

Input: access to COND_D; parameters $0 < \epsilon, \delta < 1$; input element $i^* \in [N]$

Set $S_0 = [N]$ and $\hat{D}(S_0) = 1$.Set $M = \log N + \log(9/\delta) + 1$. Set $\kappa = \Theta(\epsilon/(M^2 \log(M/\delta)))$.
for $j = 1$ to M **do**
  if $|S_{j-1}| = 1$ **then**
    return $\hat{D}(S_{j-1})$ (and exit)
  end if
  Perform $m = \Theta(\max\{M^2 \log(M/\delta)/(\epsilon^2\kappa), \log(M/(\delta\kappa))/\kappa^2\})$ COND_D queries on S_{j-1} to obtain points $i_1, \ldots, i_m \in S_{j-1}$.
  Let $H_j = \{k \in [N] : k$ appears at least $\frac{3}{4}\kappa m$ times in the list $i_1, \ldots, i_m\}$
  Let $\hat{D}_{S_{j-1}}(i^*)$ denote the fraction of times that i^* appears in $i_1, \ldots, i_m$
  if $\hat{D}_{S_{j-1}}(i^*) \geq \frac{\kappa}{20}$ **then**
    Set $S_j = \{i^*\}$, set $\hat{D}(S_j) = \hat{D}_{S_{j-1}}(i^*) \cdot \hat{D}(S_{j-1})$, increment j, and go to Line 2.
  end if
  Let $\hat{D}_{S_{j-1}}(H_j)$ denote the fraction of elements among $i_1, \ldots, i_m$ that belong to H_j.
  if $\hat{D}_{S_{j-1}}(H_j) > 1 - \epsilon/10$ **then**
    return $\mathsf{UNKNOWN}$ (and exit)
  end if
  Set S'_j to be a uniform random subset of $S_{j-1} \setminus (H_j \cup \{i^*\})$ and set S_j to be $S'_j \cup \{i^*\}$.
  Let $\hat{D}_{S_{j-1}}(S_j)$ denote the fraction of elements among $i_1, \ldots, i_m$ that belong to S_j
  if $\hat{D}_{S_{j-1}}(S_j) = 0$ **then**
    return $\mathsf{UNKNOWN}$ (and exit)
  end if
  Set $\hat{D}(S_j) = \hat{D}_{S_{j-1}}(S_j) \cdot \hat{D}(S_{j-1})$
end for
Output FAIL.

the probability that ℓ appears at least $\frac{3}{4}\kappa m$ times (that is, at least 3/2 times the upper bound on the expected value) is at most $\delta/(9MN)$ as well. The lemma follows by taking a union bound over all (at most N) points considered above and over all M settings of j. ■

Next we show that with high probability Algorithm Approx-Eval-Simulator returns either UNKNOWN or a numerical value (as opposed to outputting FAIL in Line 23):

Lemma 5 *For any D, ϵ, δ and i^*, Algorithm* Approx-Eval-Simulator *outputs* FAIL *with probability at most $\delta/9$.*

Proof: Fix any element $i \neq i^*$. The probability (taken only over the choice of the random subset in each execution of Line 16) that i is placed in S'_j in each of the first $\log N + \log(9/\delta)$ executions of Line 16 is at most $\frac{\delta}{9N}$. Taking a union bound over all $N-1$ points $i \neq i^*$, the probability that any point other than i^* remains in S_{j-1} through all of the first $\log N + \log(9/\delta)$ executions of the outer "for" loop is at most $\frac{\delta}{9}$. Assuming that this holds, then in the execution of the outer "for" loop when $j = \log N + \log(9/\delta) + 1$, the algorithm will return $\hat{D}(S_{j-1}) = \hat{D}(i^*)$ in Line 4. ■

For the rest of the analysis it will be helpful for us to define several "desirable" events and show that they all hold with high probability:

1. Let E_1 denote the event that every set H_j that is ever constructed in Line 7 satisfies both properties (i) and (ii) stated in Lemma 4. By Lemma 4 the event E_1 holds with probability at least $1 - \delta/9$.

2. Let E_2 denote the event that in every execution of Line 9, the estimate $\hat{D}_{S_{j-1}}(i^*)$ is within an additive $\pm\frac{\kappa}{40}$ of the true value of $D(i^*)/D(S_{j-1})$. By the choice of m in Line 6 (i.e., using $m = \Omega(\log(M/\delta)/\kappa^2)$), an additive Chernoff bound, and a union bound over all iterations, the event E_2 holds with probability at least $1 - \delta/9$.

3. Let E_3 denote the event that if Line 10 is executed, the resulting value $\hat{D}_{S_{j-1}}(i^*)$ lies in $[1 - \frac{\epsilon}{2M}, 1 + \frac{\epsilon}{2M}]D(i^*)/D(S_{j-1})$. Assuming that event E_2 holds, if Line 10 is reached then the true value of $D(i^*)/D(S_{j-1})$ must be at least $\kappa/40$, and consequently a multiplicative Chernoff bound and the choice of m (i.e. using $m = \Omega(M^2 \log(M/\delta)/(\epsilon^2\kappa))$) together imply that $\hat{D}_{S_{j-1}}(i^*)$ lies in $[1 - \frac{\epsilon}{2M}, 1 + \frac{\epsilon}{2M}]D(i^*)/D(S_{j-1})$ except with failure probability at most $\delta/9$.

4. Let E_4 denote the event that in every execution of Line 12, the estimate $\hat{D}_{S_{j-1}}(H_j)$ is within an additive error of $\pm\frac{\epsilon}{20}$ from the true value of $D(H_j)/D(S_{j-1})$. By the choice of m in Line 6 (i.e., using $m = \Omega(\log(M/\delta)/\epsilon^2)$) and an additive Chernoff bound, the event E_4 holds with probability at least $1 - \delta/9$.

The above arguments show that E_1, E_2, E_3 and E_4 all hold with probability at least $1 - 4\delta/9$.

Let E_5 denote the event that in every execution of Line 16, the set S'_j which is drawn satisfies $D(S'_j)/D(S_{j-1} \setminus (H_j \cup \{i^*\})) \geq 1/3$. The following lemma says that conditioned on E_1 through E_4 all holding, event E_5 holds with high probability:

Lemma 6 *Conditioned on E_1 through E_4 the probability that E_5 holds is at least $1-\delta/9$.*

Proof: Fix a value of j and consider the j-th iteration of Line 16. Since events E_2 and E_4 hold, it must be the case that $D(S_{j-1} \setminus (H_j \cup \{i^*\}))/D(S_{j-1}) \geq \epsilon/40$. Since event E_1 holds, it must be the case that every $i \in (S_{j-1} \setminus (H_j \cup \{i^*\}))$ has $D(i)/D(S_{j-1}) \leq \kappa$. Now since S'_j is chosen by independently including each element of $S_{j-1} \setminus (H_j \cup \{i^*\})$ with probability $1/2$, we can apply the first part of Corollary 3 and get

$$\Pr\left[D(S'_j) < \frac{1}{3}D(S_{j-1} \setminus (H_j \cup \{i^*\}))\right] \leq \epsilon^{-4\epsilon/(40\cdot 9\cdot 4\kappa)} < \frac{\delta}{9M}\,,$$

where the last inequality follows by the setting of $\kappa = \Omega(\epsilon/(M^2\log(1/\delta)))$. ■

Thus we have established that E_1 through E_5 all hold with probability at least $1-5\delta/9$.

Next, let E_6 denote the event that the algorithm never returns UNKNOWN and exits in Line 19. Our next lemma shows that conditioned on events E_1 through E_5, the probability of E_6 is at least $1-\delta/9$:

Lemma 7 *Conditioned on E_1 through E_5 the probability that E_6 holds is at least $1-\delta/9$.*

Proof: Fix any iteration j of the outer "For" loop. In order for the algorithm to reach Line 18 in this iteration, it must be the case (by Lines 9 and 13) that at least $(\epsilon/10-\kappa/20)m > (\epsilon/20)m$ points in $i_1,\ldots,i_m$ do not belong to $H_j \cup \{i^*\}$. Since each point not in H_j appears at most $\frac{3}{4}\kappa m$ times in the list $i_1,\ldots,i_m$, there must be at least $\frac{\epsilon}{15\kappa}$ distinct such values. Hence the probability that none of these values is selected to belong to S'_j is at most $1/2^{\epsilon/(15\kappa)} < \delta/(9M)$. A union bound over all (at most M) values of j gives that the probability the algorithm ever returns UNKNOWN and exits in Line 19 is at most $\delta/9$, so the lemma is proved. ■

Now let E_7 denote the event that in every execution of Line 17, the estimate $\hat{D}_{S_{j-1}}(S_j)$ lies in $[1-\frac{\epsilon}{2M}, 1+\frac{\epsilon}{2M}]D(S_j)/D(S_{j-1})$. The following lemma says that conditioned on E_1 through E_5, event E_7 holds with probability at least $1-\delta/9$:

Lemma 8 *Conditioned on E_1 through E_5, the probability that E_7 holds is at least $1-\delta/9$.*

Proof: Fix a value of j and consider the j-th iteration of Line 17.The expected value of $\hat{D}_{S_{j-1}}(S_j)$ is precisely

$$\frac{D(S_j)}{D(S_{j-1})} = \frac{D(S_j)}{D(S_{j-1} \setminus (H_j \cup \{i^*\}))} \cdot \frac{D(S_{j-1} \setminus (H_j \cup \{i^*\}))}{D(S_{j-1})}. \tag{15}$$

Since events E_2 and E_4 hold we have that $\frac{D(S_{j-1}\setminus(H_j\cup\{i^*\}))}{D(S_{j-1})} \geq \epsilon/40$, and since event E_5 holds we have that $\frac{D(S_j)}{D(S_{j-1}\setminus(H_j\cup\{i^*\}))} \geq 1/3$ (note that $D(S_j) \geq D(S'_j)$). Thus we have that (15) is at least $\epsilon/120$. Recalling the value of m (i.e., using $m = \Omega(M^2\log(M/\delta)/\epsilon^2\kappa) = \Omega(M^2\log(M/\delta)/\epsilon^3)$) a multiplicative Chernoff bound gives that indeed $\hat{D}_{S_{j-1}}(S_j) \in [1-\frac{\epsilon}{2M}, 1+\frac{\epsilon}{2M}]D(S_j)/D(S_{j-1})$ with failure probability at most $\delta/(9M)$. A union bound over all M possible values of j finishes the proof. ■

At this point we have established that events E_1 through E_7 all hold with probability at least $1 - 7\delta/9$.

We can now argue that each estimate $\hat{D}(S_j)$ is indeed a high-accuracy estimate of the true value $D(S_j)$:

Lemma 9 *With probability at least* $1 - 7\delta/9$ *each estimate* $\hat{D}(S_j)$ *constructed by* Approx-Eval-Simulator *lies in* $[(1 - \frac{\epsilon}{2M})^j, (1 + \frac{\epsilon}{2M})^j]D(S_j)$.

Proof: We prove the lemma by showing that if all events E_1 through E_7 hold, then the following claim (denoted (*)) holds: each estimate $\hat{D}(S_j)$ constructed by Approx-Eval-Simulator lies in $[(1 - \frac{\epsilon}{2M})^j, (1 + \frac{\epsilon}{2M})^j]D(S_j)$. Thus for the rest of the proof we assume that indeed all events E_1 through E_7 hold.

The claim (*) is clearly true for $j = 0$. We prove (*) by induction on j assuming it holds for $j - 1$. The only places in the algorithm where $\hat{D}(S_j)$ may be set are Lines 10 and 21. If $\hat{D}(S_j)$ is set in Line 21 then (*) follows from the inductive claim for $j - 1$ and Lemma 8. If $\hat{D}(S_j)$ is set in Line 10, then (*) follows from the inductive claim for $j - 1$ and the fact that event E_3 holds. This concludes the proof of the lemma. ■

Finally, we require the following crucial lemma which establishes that if $i^* \notin L_{\epsilon,N}$ (and hence the initial rank $\text{rank}_{[N]}$ of i^* is at least ϵ), then with very high probability the rank of i^* never becomes too low during the execution of the algorithm:

Lemma 10 *Suppose* $i^* \notin L_{\epsilon,N}$. *Then with probability at least* $1 - \delta/9$, *every set* S_{j-1} *constructed by the algorithm has* $\text{rank}_{S_{j-1}}(i^*) \geq \epsilon/2$.

We prove Lemma 10 in Section 3.3.3 below.

With these pieces in place we are ready to prove Theorem 4.

Proof of Theorem 4: It is straightforward to verify that algorithm Approx-Eval-Simulator has the claimed query complexity. We now argue that Approx-Eval-Simulator meets the two requirements (i) and (ii) of Definition 3. Throughout the discussion below we assume that all the "favorable events" in the above analysis (i.e. events E_1 through E_7, Lemma 5, and Lemma 10) indeed hold as desired (incurring an overall failure probability of at most δ).

Suppose first that $i^* \notin L_{\epsilon,D}$. We claim that by Lemma 10 it must be the case that the algorithm does not return UNKNOWN in Line 14. To verify this, observe that in order to reach Line 14 it would need to be the case that $D(i^*)/D(S_{j-1}) \leq 3\kappa/40$ (so the algorithm does not instead go to Line 22 in Line 10). Since by Lemma 4 every element k in H_j satisfies $D(k)/D(S_{j-1}) \geq \kappa/2$, this means that i^* does not belong to H_j. In order to reach Line 14, by event E_4 we must have $D(H_j)/D(S_{j-1}) \geq 1 - 3\epsilon/20$. Since every element of H_j has more mass under D (at least $\kappa/2$) than i^* (which has at most $3\kappa/40$), this would imply that $\text{rank}_{S_{j-1}}(i^*) \leq 3\epsilon/20$, contradicting Lemma 10. Furthermore, by Lemma 7 it must be the case that the algorithm does not return UNKNOWN in Line 19. Thus the algorithm terminates by returning an estimate $\hat{D}(S_j) = \hat{D}(i^*)$ which, by Lemma 9, lies in $[(1 - \frac{\epsilon}{2M})^j, (1 + \frac{\epsilon}{2M})^j]D(i^*)$. Since $j \leq M$ this estimate lies in $[1 - \epsilon, 1 + \epsilon]D(i^*)$ as required.

Now suppose that $i^* \in L_{\epsilon,D}$. By Lemma 5 we may assume that the algorithm either outputs UNKNOWN or a numerical value. As above, Lemma 9 implies that if the algorithm outputs a numerical value then the value lies in $[1-\epsilon, 1+\epsilon]D(i^*)$ as desired. This concludes the proof of Theorem 4. ■

3.3.3 Proof of Lemma 10.

The key to proving Lemma 10 will be proving the next lemma. (In the following, for S a set of real numbers we write $\mathrm{sum}(S)$ to denote $\sum_{\alpha \in S} \alpha$.)

Lemma 11 *Fix $0 < \epsilon \leq 1/40$. Set $\kappa = \Omega(\epsilon/(M^2 \log(1/\delta)))$. Let $T = \{\alpha_1, \ldots, \alpha_n\}$ be a set of values $\alpha_1 < \cdots < \alpha_n$ such that $\mathrm{sum}(T) = 1$. Fix $\ell \in [N]$ and let $T_L = \{\alpha_1, \ldots, \alpha_\ell\}$ and let $T_R = \{\alpha_{\ell+1}, \ldots, \alpha_n\}$, so $T_L \cup T_R = T$. Assume that $\mathrm{sum}(T_L) \geq \epsilon/2$ and that $\alpha_\ell \leq \kappa/10$.*

Fix H to be any subset of T satisfying the following two properties: (i) H includes every α_j such that $\alpha_j \geq \kappa$; and (ii) H includes no α_j such that $\alpha_j < \kappa/2$. (Note that consequently H does not intersect T_L.)

Let T' be a subset of $(T \setminus (H \cup \{\alpha_\ell\})$ selected uniformly at random. Let $T'_L = T' \cap T_L$ and let $T'_R = T' \cap T_R$.

Then we have the following:

1. *If $\mathrm{sum}(T_L) > 20\epsilon$, then with probability at least $1 - \delta/M$ (over the random choice of T') it holds that*
$$\frac{\mathrm{sum}(T'_L \cup \{\alpha_\ell\})}{\mathrm{sum}(T' \cup \{\alpha_\ell\})} \geq 9\epsilon;$$
2. *If $\epsilon/2 \leq \mathrm{sum}(T_L) < 20\epsilon$, then with probability at least $1 - \delta/M$ (over the random choice of T') it holds that*
$$\frac{\mathrm{sum}(T'_L \cup \{\alpha_\ell\})}{\mathrm{sum}(T' \cup \{\alpha_\ell\})} \geq \mathrm{sum}(T_L)\,(1-\rho)\,,$$
where $\rho = \frac{\ln 2}{M}$.

Proof of Lemma 10 using Lemma 11: We apply Lemma 11 repeatedly at each iteration j of the outer "For" loop. The set H of Lemma 11 corresponds to the set H_j of "heavy" elements that are removed at a given iteration, the set of values T corresponds to the values $D(i)/D(S_{j-1})$ for $i \in S_{j-1}$, and the element α_ℓ of Lemma 11 corresponds to $D(i^*)/D(S_{j-1})$. The value $\mathrm{sum}(T_L)$ corresponds to $\mathrm{rank}_{S_{j-1}}(i^*)$ and the value

$$\frac{\mathrm{sum}(T'_L \cup \{\alpha_\ell\})}{\mathrm{sum}(T' \cup \{\alpha_\ell\})}$$

corresponds to $\mathrm{rank}_{S_j}(i^*)$. Observe that since $i^* \notin L_{\epsilon,N}$ we know that initially $\mathrm{rank}_{[N]}(i^*) \geq \epsilon$, which means that the first time we apply Lemma 11 (with $T = \{D(i) : i \in [N]\}$) we have $\mathrm{sum}(T_L) \geq \epsilon$.

By Lemma 11 the probability of failure in any of the (at most M) iterations is at most $\delta/9$, so we assume that there is never a failure. Consequently for all j we have that if $\text{rank}_{S_{j-1}}(i^*) \geq 20\epsilon$ then $\text{rank}_{S_j}(i^*) \geq 9\epsilon$, and if $\epsilon/2 \leq \text{rank}_{S_{j-1}}(i^*) < 20\epsilon$ then $\text{rank}_{S_j}(i^*) \geq \text{rank}_{S_j}(i^*) \cdot (1-\rho)$. Since $\text{rank}_{S_0}(i^*) \geq \epsilon$, it follows that for all $j \leq M$ we have $\text{rank}_{S_j}(i^*) \geq \epsilon \cdot (1-\rho)^M > \epsilon/2$. ■

Proof of Lemma 11. We begin with the following claim:

Claim 12 *With probability at least $1 - \delta/(2M)$ (over the random choice of T') it holds that $\text{sum}(T'_L) \geq \frac{1}{2} \cdot \text{sum}(T_L) \cdot (1 - \rho/2)$.*

Proof: Recall from the setup that every element $\alpha_i \in T_L$ satisfies $\alpha_i \leq \kappa/10$, and $\text{sum}(T_L) \geq \epsilon/2$. Also recall that $\kappa = \Omega(\epsilon/(M^2 \log(1/\delta)))$ and that $\rho = \frac{\ln 2}{M}$, so that $\rho^2\epsilon/(6\kappa) \geq \ln(2M/\delta)$. The claim follows by applying the first part of Corollary 3 (with $\gamma = \rho/2$). ■

Part (1) of Lemma 11 is an immediate consequence of Claim 12, since in part (1) we have

$$\frac{\text{sum}(T'_L \cup \{\alpha_\ell\})}{\text{sum}(T' \cup \{\alpha_\ell\})} \geq \text{sum}(T'_L) \geq \frac{1}{2} \cdot \text{sum}(T_L) \cdot \left(1 - \frac{\rho}{2}\right) \geq \frac{1}{2} \cdot 20\epsilon \cdot \left(1 - \frac{\rho}{2}\right) \geq 9\epsilon.$$

It remains to prove Part (2) of the lemma. We will do this using the following claim:

Claim 13 *Suppose $\epsilon/2 \leq \text{sum}(T_L) \leq 20\epsilon$. Then with probability at least $1 - \delta/(2M)$ (over the random choice of T') it holds that $\text{sum}(T'_R) \leq \frac{1}{2}\text{sum}(T_R) \cdot (1 + \rho/2)$.*

Proof: Observe first that $\alpha_i < \kappa$ for each $\alpha_i \in T_R \setminus H$. We consider two cases.

If $\text{sum}(T_R \setminus H) \geq 4\epsilon$, then we apply the first part of Corollary 3 to the α_i's in $T_R \setminus H$ and get that

$$\begin{aligned} \Pr\left[\text{sum}(T'_R) > \frac{1}{2}\text{sum}(T_R) \cdot (1+\rho/2)\right] &\leq \Pr\left[\text{sum}(T'_R) > \frac{1}{2}\text{sum}(T_R \setminus H) \cdot (1+\rho/2)\right] \\ &< \exp(-\rho^2 \text{sum}(T_R \setminus H)/24\kappa) \qquad (16) \\ &\leq \exp(-\rho^2\epsilon/(6\kappa)) \leq \frac{\delta}{2M} \qquad (17) \end{aligned}$$

(recall from the proof of Claim 12 that $\rho^2\epsilon/(6\kappa) \geq \ln(2M/\delta)$).

If $\text{sum}(T_R \setminus H) < 4\epsilon$, (so that the expected value of $\text{sum}(T'_R)$ is less than 2ϵ) then we can apply the second part of Corollary 3 as we explain next. Observe that by the premise of the lemma, $\text{sum}(T_R) \geq 1 - 20\epsilon$ which is at least $1/2$ (recalling that ϵ is at most $1/40$). Consequently, the event "$\text{sum}(T'_R) \geq \frac{1}{2} \cdot \text{sum}(T_R) \cdot (1+\rho/2)$" implies the event "$\text{sum}(T'_R) \geq \frac{1}{4}$", and by applying the second part of Corollary 3 we get

$$\Pr\left[\text{sum}(T'_R) > \frac{1}{2}\text{sum}(T_R) \cdot (1+\rho/2)\right] \leq \Pr\left[\text{sum}(T'_R) > \frac{1}{4}\right] < 2^{-1/4\kappa} < \frac{\delta}{2M}, \qquad (18)$$

as required. ■

Now we can prove Lemma 11. Using Claims 12 and 13 we have that with probability at least $1-\delta/M$,

$$\mathrm{sum}(T'_L) \geq \frac{1}{2}\cdot \mathrm{sum}(T_L)\cdot(1-\rho/2) \qquad \text{and} \qquad \mathrm{sum}(T'_R) \leq \frac{1}{2}\mathrm{sum}(T_R)\cdot(1+\rho/2);$$

we assume that both these inequalities hold going forth. Since

$$\frac{\mathrm{sum}(T'_L \cup \{\alpha_\ell\})}{\mathrm{sum}(T' \cup \{\alpha_\ell\})} = \frac{\mathrm{sum}(T'_L)+\alpha_\ell}{\mathrm{sum}(T')+\alpha_\ell} > \frac{\mathrm{sum}(T'_L)}{\mathrm{sum}(T')},$$

it is sufficient to show that $\frac{\mathrm{sum}(T'_L)}{\mathrm{sum}(T')} \geq \mathrm{sum}(T_L)(1-\rho)$; we now show this. As $\mathrm{sum}(T') = \mathrm{sum}(T'_L)+\mathrm{sum}(T'_R)$,

$$\begin{aligned}
\frac{\mathrm{sum}(T'_L)}{\mathrm{sum}(T')} &= \frac{\mathrm{sum}(T'_L)}{\mathrm{sum}(T'_L)+\mathrm{sum}(T'_R)} = \frac{1}{1+\frac{\mathrm{sum}(T'_R)}{\mathrm{sum}(T'_L)}} \\
&\geq \frac{1}{1+\frac{(1/2)\cdot\mathrm{sum}(T_R)\cdot(1+\rho/2)}{(1/2)\cdot\mathrm{sum}(T_L)\cdot(1-\rho/2)}} \\
&= \frac{\mathrm{sum}(T_L)\cdot(1-\rho/2)}{\mathrm{sum}(T_L)\cdot(1-\rho/2)+\mathrm{sum}(T_R)\cdot(1+\rho/2)} \\
&\geq \frac{\mathrm{sum}(T_L)\cdot(1-\rho/2)}{\mathrm{sum}(T_L)\cdot(1+\rho/2)+\mathrm{sum}(T_R)\cdot(1+\rho/2)} \\
&= \mathrm{sum}(T_L)\cdot\frac{1-\rho/2}{1+\rho/2} > \mathrm{sum}(T_L)\cdot(1-\rho).
\end{aligned}$$

This concludes the proof of Lemma 11. ■

4 Algorithms and lower bounds for testing uniformity

4.1 A $\tilde{O}(1/\epsilon^2)$-query PCOND algorithm for testing uniformity

In this subsection we present an algorithm PCOND_D-Test-Uniform and prove the following theorem:

Theorem 5 PCOND_D-Test-Uniform *is a* $\tilde{O}(1/\epsilon^2)$*-query* PCOND_D *testing algorithm for uniformity, i.e. it outputs* ACCEPT *with probability at least* $2/3$ *if* $D = \mathcal{U}$ *and outputs* REJECT *with probability at least* $2/3$ *if* $d_{\mathrm{TV}}(D,\mathcal{U}) \geq \epsilon$.

Intuition. For the sake of intuition we first describe a simpler approach that yields a $\tilde{O}(1/\epsilon^4)$-query algorithm, and then build on those ideas to obtain our real algorithm with its improved $\tilde{O}(1/\epsilon^2)$ bound. Fix D to be a distribution over $[N]$ that is ϵ-far from uniform. Let

$$H = \left\{ h \in [N] \;\middle|\; D(h) \geq \frac{1}{N} \right\} \text{ and } L = \left\{ \ell \in [N] \;\middle|\; D(\ell) < \frac{1}{N} \right\}.$$

It is easy to see that since D is ϵ-far from uniform, we have

$$\sum_{h \in H} \left(D(h) - \frac{1}{N} \right) = \sum_{\ell \in L} \left(\frac{1}{N} - D(\ell) \right) \geq \frac{\epsilon}{2} \,. \tag{19}$$

From this it is not hard to show that

(i) many elements of $[N]$ must be "significantly light" in the following sense: Define $L' \subseteq L$ to be $L' = \left\{ \ell \in [N] \;\middle|\; D(\ell) < \frac{1}{N} - \frac{\epsilon}{4N} \right\}$. Then it must be the case that $|L'| \geq (\epsilon/4)N$.

(ii) D places significant weight on elements that are "significantly heavy" in the following sense: Define $H' \subseteq H$ to be $H' = \left\{ h \in [N] \;\middle|\; D(h) \geq \frac{1}{N} + \frac{\epsilon}{4N} \right\}$. Then it must be the case that $D(H') \geq (\epsilon/4)$.

Using (i) and (ii) it is fairly straightforward to give a $O(1/\epsilon^4)$-query PCOND_D testing algorithm as follows: we can get a point in L' with high probability by randomly sampling $O(1/\epsilon)$ points uniformly at random from $[N]$, and we can get a point in H' with high probability by drawing $O(1/\epsilon)$ points from SAMP_D. Then at least one of the $O(1/\epsilon^2)$ pairs that have one point from the first sample and one point from the second will have a multiplicative factor difference of $1 + \Omega(\epsilon)$ between the weight under D of the two points, and this can be detected by calling the procedure Compare (see Subsection 3.1). Since there are $O(1/\epsilon^2)$ pairs and for each one the invocation of Compare uses $\tilde{O}(1/\epsilon^2)$ queries, the overall sample complexity of this simple approach is $\tilde{O}(1/\epsilon^4)$.

Our actual algorithm PCOND_D-Test-Uniform for testing uniformity extends the above ideas to get a $\tilde{O}(1/\epsilon^2)$-query algorithm. More precisely, the algorithm works as follows: it first draws a "reference sample" of $O(1)$ points uniformly from $[N]$. Next, repeatedly for $O\big(\log \frac{1}{\epsilon}\big)$ iterations, the algorithm draws two other samples, one uniformly from $[N]$ and the other from SAMP_D. (These samples have different sizes at different iterations; intuitively, each iteration is meant to deal with a different "scale" of probability mass that points could have under D.) At each iteration it then uses Compare to do comparisons between pairs of elements, one from the reference sample and the other from one of the two other samples. If D is ϵ-far from uniform, then with high probability at some iteration the algorithm will either draw a point from SAMP_D that has "very big" mass under D, or draw a point from the uniform distribution over $[N]$ that has "very small" mass under D, and this will be detected by the comparisons to the reference points. Choosing the sample sizes and parameters for the Compare calls carefully at each iteration yields the improved query bound.

Let m_j denote the number of PCOND_D queries used to run $\textsc{Compare}_D$ in a given execution of Line 7 during the j-th iteration of the outer loop. By the setting of the parameters in each such call and Lemma 2, $m_j = O\big(\frac{t}{\epsilon^2 2^{2j}}\big)$. It is easy to see that the algorithm only performs PCOND_D queries and that the total number of queries that the algorithm performs is

$$O\left(\sum_{j=1}^{t} q \cdot s_j \cdot m_j \right) = O\left(\sum_{j=1}^{t} 2^j \log\left(\frac{1}{\epsilon}\right) \cdot \frac{\log(\frac{1}{\epsilon})}{\epsilon^2 2^{2j}} \right) = O\left(\frac{(\log(\frac{1}{\epsilon}))^2}{\epsilon^2} \right) .$$

We prove Theorem 5 by arguing completeness and soundness below.

Completeness: Suppose that D is the uniform distribution. Then for any fixed pair of points (x, y), Lemma 2 implies that the call to Compare on $\{x\}, \{y\}$ in Line 7 causes the algorithm to

Algorithm 4: PCOND_D-TEST-UNIFORM

Input: error parameter $\epsilon > 0$; query access to PCOND_D oracle

1: Set $t = \log(\frac{4}{\epsilon}) + 1$.
2: Select $q = \Theta(1)$ points $i_1, \ldots, i_q$ independently and uniformly from $[N]$.
3: **for** $j = 1$ to t **do**
4: Call the SAMP_D oracle $s_j = \Theta\big(2^j \cdot t\big)$ times to obtain points $h_1, \ldots, h_{s_j}$ distributed according to D.
5: Select s_j points $\ell_1, \ldots, \ell_{s_j}$ independently and uniformly from $[N]$.
6: **for all** pairs $(x, y) = (i_r, h_{r'})$ and $(x, y) = (i_r, \ell_{r'})$ (where $1 \le r \le q$, $1 \le r' \le s_j$) **do**
7: Call $\textsc{Compare}_D(\{x\}, \{y\}, \Theta(\epsilon 2^j), 2, \exp(-\Theta(t)))$.
8: **if** the $\textsc{Compare}$ call does not return a value in $[1 - 2^{j-5}\frac{\epsilon}{4}, 1 + 2^{j-5}\frac{\epsilon}{4}]$ **then**
9: output REJECT (and exit).
10: **end if**
11: **end for**
12: **end for**
13: Output ACCEPT

output REJECT in Line 9 with probability at most $e^{-\Theta(t)} = \mathrm{poly}(\epsilon)$. By taking a union bound over all $\mathrm{poly}(1/\epsilon)$ pairs of points considered by the algorithm, the algorithm will accept with probability at least $2/3$, as required.

Soundness: Now suppose that D is ϵ-far from uniform (we assume throughout the analysis that $\epsilon = 1/2^k$ for some integer k, which is clearly without loss of generality). We define H, L as above and further partition H and L into "buckets" as follows: for $j = 1, \ldots, t-1 = \log(\frac{4}{\epsilon})$, let

$$H_j \stackrel{\text{def}}{=} \left\{ h \;\middle|\; \left(1 + 2^{j-1} \cdot \frac{\epsilon}{4}\right) \cdot \frac{1}{N} \le D(h) < \left(1 + 2^j \cdot \frac{\epsilon}{4}\right) \cdot \frac{1}{N} \right\} ,$$

and for $j = 1, \ldots, t-2$ let

$$L_j \stackrel{\text{def}}{=} \left\{ \ell \;\middle|\; \left(1 - 2^j \cdot \frac{\epsilon}{4}\right) \cdot \frac{1}{N} < D(\ell) \le \left(1 - 2^{j-1} \cdot \frac{\epsilon}{4}\right) \cdot \frac{1}{N} \right\} .$$

Also define

$$H_0 \stackrel{\text{def}}{=} \left\{ h \;\middle|\; \frac{1}{N} \le D(h) < \left(1 + \frac{\epsilon}{4}\right) \cdot \frac{1}{N} \right\} , \quad L_0 \stackrel{\text{def}}{=} \left\{ \ell \;\middle|\; \left(1 - \frac{\epsilon}{4}\right) \cdot \frac{1}{N} < D(\ell) < \frac{1}{N} \right\} ,$$

and

$$H_t \stackrel{\text{def}}{=} \left\{ h \;\middle|\; D(h) \ge \frac{2}{N} \right\} , \quad L_{t-1} \stackrel{\text{def}}{=} \left\{ \ell \;\middle|\; D(\ell) \le \frac{1}{2N} \right\} .$$

First observe that by the definition of H_0 and L_0, we have

$$\sum_{h \in H_0} \left(D(h) - \frac{1}{N}\right) \le \frac{\epsilon}{4} \quad \text{and} \quad \sum_{\ell \in L_0} \left(\frac{1}{N} - D(\ell)\right) \le \frac{\epsilon}{4} .$$

Therefore (by Equation (19)) we have

$$\sum_{j=1}^{t}\sum_{h\in H_j}\left(D(h)-\frac{1}{N}\right)\geq\frac{\epsilon}{4}\quad\text{and}\quad\sum_{j=1}^{t-1}\sum_{\ell\in L_j}\left(\frac{1}{N}-D(\ell)\right)\geq\frac{\epsilon}{4}\,.$$

This implies that for some $1\leq j(H)\leq t$, and some $1\leq j(L)\leq t-1$, we have

$$\sum_{h\in H_{j(H)}}\left(D(h)-\frac{1}{N}\right)\geq\frac{\epsilon}{4t}\quad\text{and}\quad\sum_{\ell\in L_{j(L)}}\left(\frac{1}{N}-D(\ell)\right)\geq\frac{\epsilon}{4t}\,.\tag{20}$$

The rest of the analysis is divided into two cases depending on whether $|L|\geq\frac{N}{2}$ or $|H|>\frac{N}{2}$.

Case 1: $|L|\geq\frac{N}{2}$. In this case, with probability at least 99/100, in Line 2 the algorithm will select at least one point $i_r\in L$. We consider two subcases: $j(H)=t$, and $j(H)\leq t-1$.

- $j(H)=t$: In this subcase, by Equation (20) we have that $\sum_{h\in H_{j(H)}}D(h)\geq\frac{\epsilon}{4t}$. This implies that when $j=j(H)=t=\log(\frac{4}{\epsilon})+1$, so that $s_j=s_t=\Theta\left(\frac{t}{\epsilon}\right)$, with probability at least 99/100 the algorithm selects a point $h_{r'}\in H_t$ in Line 4. Assume that indeed such a point $h_{r'}$ is selected. Since $D(h_{r'})\geq\frac{2}{N}$, while $D(i_r)<\frac{1}{N}$, Lemma 2 implies that with probability at least $1-\text{poly}(\epsilon)$ the Compare call in Line 7 outputs either High or a value that is at least $\frac{7}{12}=\frac{1}{2}+\frac{1}{12}$. Since $\frac{7}{12}>\frac{1}{2}+2^{j-5}\frac{\epsilon}{4}$ for $j=t$, the algorithm will output REJECT in Line 9.
- $j(H)<t$: By Equation (20) and the definition of the buckets, we have

  $$\sum_{h\in H_{j(H)}}\left(\left(1+2^{j(H)}\frac{\epsilon}{4}\right)\frac{1}{N}-\frac{1}{N}\right)\geq\frac{\epsilon}{4t}\,,$$

  implying that $\left|H_{j(H)}\right|\geq\frac{N}{2^{j(H)}t}$ so that $D(H_{j(H)})\geq\frac{1}{2^{j(H)}t}$. Therefore, when $j=j(H)$ so that $s_j=\Theta\left(2^{j(H)}t\right)$, with probability at least 99/100 the algorithm will get a point $h_{r'}\in H_{j(H)}$ in Line 4. Assume that indeed such a point $h_{r'}$ is selected. Since $D(h_{r'})\geq\left(1+2^{j(H)-1}\frac{\epsilon}{4}\right)\frac{1}{N}$, while $D(i_r)\leq\frac{1}{N}$, for $\alpha_{j(H)}=2^{j(H)-1}\frac{\epsilon}{4}$, we have

  $$\frac{D(h_{r'})}{D(i_r)}\geq 1+\alpha_{j(H)}\,.$$

  Since Compare is called in Line 7 on the pair $\{i_r\},\{h_{r'}\}$ with the "δ" parameter set to $\Theta(\epsilon 2^j)$, with probability $1-\text{poly}(\epsilon)$ the algorithm outputs REJECT as a result of this Compare call.

Case 2: $|H|>\frac{N}{2}$. This proceeds similarly to Case 1. In this case we have that with high constant probability the algorithm selects a point $i_r\in H$ in Line 2. Here we consider the subcases $j(L)=t-1$ and $j(L)\leq t-2$. In the first subcase we have that $\sum_{\ell\in L_t}\frac{1}{N}\geq\frac{\epsilon}{4t}$, so that $|L_t|\geq(\frac{\epsilon}{4t})N$, and in the second case we have that $\sum_{\ell\in L_{j(L)}}(2^{j(L)}\frac{\epsilon}{4})\frac{1}{N}\geq\frac{\epsilon}{4t}$, so that $\left|L_{j(L)}\right|\geq\frac{N}{2^{j(L)}t}$. The analysis of each subcase is similar to Case 1. This concludes the proof of Theorem 5. ■

4.2 An $\Omega(1/\epsilon^2)$ lower bound for COND_D algorithms that test uniformity

In this subsection we give a lower bound showing that the query complexity of the PCOND_D algorithm of the previous subsection is essentially optimal, even for algorithms that may make general COND_D queries:

Theorem 6 *Any COND_D algorithm for testing whether $D = \mathcal{U}$ versus $d_{\mathrm{TV}}(D,\mathcal{U}) \geq \epsilon$ must make $\Omega(1/\epsilon^2)$ queries.*

The high-level idea behind Theorem 6 is to reduce it to the well-known fact that distinguishing a fair coin from a $(\frac{1}{2}+4\epsilon)$-biased coin requires $\Omega\left(\frac{1}{\epsilon^2}\right)$ coin tosses. We show that any q-query algorithm COND_D testing algorithm A can be transformed into an algorithm A' that successfully distinguishes q tosses of a fair coin from q tosses of a $(\frac{1}{2}+4\epsilon)$-biased coin.

Proof of Theorem 6: First note that we may assume without loss of generality that $0 < \epsilon \leq 1/8$. Let A be any q-query algorithm that makes COND_D queries and tests whether $D = \mathcal{U}$ versus $d_{\mathrm{TV}}(D,\mathcal{U}) \geq \epsilon$. We may assume without loss of generality that in every possible execution algorithm A makes precisely q queries (this will be convenient later).

Let D_{No} be the distribution that has $D_{\mathrm{No}}(i) = \frac{1+2\epsilon}{N}$ for each $i \in \left[1, \frac{N}{2}\right]$ and has $D_{\mathrm{No}}(i) = \frac{1-2\epsilon}{N}$ for each $i \in \left[\frac{N}{2}+1, N\right]$. (This is the "no"-distribution for our lower bound; it is ϵ-far in variation distance from the uniform distribution $\mathcal{U}$.) By Definition 2, it must be the case that

$$Z := \left|\Pr\left[A^{\mathsf{COND}_{D_{\mathrm{No}}}} \text{ outputs } \mathsf{ACCEPT}\right] - \Pr\left[A^{\mathsf{COND}_{\mathcal{U}}} \text{ outputs } \mathsf{ACCEPT}\right]\right| \geq 1/3.$$

The proof works by showing that given A as described above, there must exist an algorithm A' with the following properties: A' is given as input a q-bit string $(b_1,\ldots,b_q) \in \{0,1\}^q$. Let D_0 denote the uniform distribution over $\{0,1\}^q$ and let $D_{4\epsilon}$ denote the distribution over $\{0,1\}^q$ in which each coordinate is independently set to 1 with probability $1/2+4\epsilon$. Then algorithm A' has

$$\left|\Pr_{b\sim D_0}[A'(b) \text{ outputs } \mathsf{ACCEPT}] - |\Pr_{b\sim D_{4\epsilon}}[A'(b) \text{ outputs } \mathsf{ACCEPT}]\right| = Z. \tag{21}$$

Given (21), by the data processing inequality for total variation distance (Lemma 1) we have that $Z \leq d_{\mathrm{TV}}(D_0, D_{4\epsilon})$. It is easy to see that $d_{\mathrm{TV}}(D_0, D_{4\epsilon})$ is precisely equal to the variation distance $d_{\mathrm{TV}}(\mathrm{Bin}(q,1/2), \mathrm{Bin}(q,1/2+4\epsilon))$. However, in order for the variation distance between these two binomial distributions to be as large as $1/3$ it must be the case that $q \geq \Omega(1/\epsilon^2)$:

Fact 14 (Distinguishing Fair from Biased Coin) *Suppose $m \leq \frac{c}{\epsilon^2}$, with c a sufficiently small constant and $\epsilon \leq 1/8$. Then,*

$$d_{\mathrm{TV}}\left(\mathrm{Bin}\left(m,\frac{1}{2}\right), \mathrm{Bin}\left(m,\frac{1}{2}+4\epsilon\right)\right) \leq \frac{1}{3}.$$

(Fact 14 is well known; it follows, for example, as an immediate consequence of Equations (2.15) and (2.16) of [AJ06].) Thus to prove Theorem 6 it remains only to describe algorithm A' and prove Equation (21).

As suggested above, algorithm A' uses algorithm A; in order to do this, it must perfectly simulate the COND_D oracle that A requires, both in the case when $D = \mathcal{U}$ and in the case when $D = D_{\mathrm{No}}$. We show below that when its input b is drawn from D_0 then A' can perfectly simulate the execution of A when it is run on the $\mathsf{COND}_{\mathcal{U}}$ oracle, and when b is drawn from $D_{4\epsilon}$ then A' can perfectly simulate the execution of A when it is run on the $\mathsf{COND}_{D_{\mathrm{No}}}$ oracle.

Fix any step $1 \leq t \leq q$. We now describe how A' perfectly simulates the t-th step of the execution of A (i.e. the t-th call to COND_D that A makes, and the response of COND_D). We may inductively assume that A' has perfectly simulated the first $t-1$ steps of the execution of A.

For each possible prefix of $t-1$ query-response pairs to COND_D

$$PREFIX = ((S_1, s_1), ..., (S_{t-1}, s_{t-1}))$$

(where each $S_i \subseteq [N]$ and each $s_i \in S_i$), there is some distribution $\mathrm{P}_{A,PREFIX}$ over possible t-th query sets S_t that A would make given that its first $t-1$ query-response pairs were $PREFIX$. So for a set $S_t \subseteq [N]$ and a possible prefix $PREFIX$, the value $P_{A,PREFIX}(S_t)$ is the probability that algorithm A, having had the transcript of its execution thus far be $PREFIX$, generates set S_t as its t-th query set. For any query set $S \subseteq [N]$, let us write S as a disjoint union $S = S_0 \amalg S_1$, where $S_0 = S \cap \left[1, \frac{N}{2}\right]$ and $S_1 = S \cap [\frac{N}{2} + 1, N]$. We may assume that every query S ever used by A has $|S_0|, |S_1| \geq 1$ (for otherwise A could perfectly simulate the response of $\mathsf{COND}_D(S)$ whether D were $\mathcal{U}$ or D_{No} by simply choosing a uniform point from S, so there would be no need to call COND_D on such an S). Thus we may assume that $P_{A,PREFIX}(S)$ is nonzero only for sets S that have $|S_0|, |S_1| \geq 1$.

Consider the bit $b_t \in \{0, 1\}$. As noted above, we inductively have that (whether D is $\mathcal{U}$ or D_{No}) the algorithm A' has perfectly simulated the execution of A for its first $t-1$ query-response pairs; in this simulation some prefix $PREFIX = ((S_1, s_1), \ldots, (S_{t-1}, s_{t-1}))$ of query-response pairs has been constructed. If $b = (b_1, \ldots, b_q)$ is distributed according to D_0 then $PREFIX$ is distributed exactly according to the distribution of A's prefixes of length $t-1$ when A is run with $\mathsf{COND}_{\mathcal{U}}$, and if $b = (b_1, \ldots, b_q)$ is distributed according to $D_{4\epsilon}$ then the distribution of $PREFIX$ is exactly the distribution of A's prefixes of length $t-1$ when A is run with $\mathsf{COND}_{D_{\mathrm{No}}}$.

Algorithm A' simulates the t-th stage of the execution of A as follows:

1. Randomly choose a set $S \subseteq [N]$ according to the distribution $P_{A,PREFIX}$; let $S = S_0 \amalg S_1$ be the set that is selected. Let us write $\alpha(S)$ to denote $|S_1|/|S_0|$ (so $\alpha(S) \in [2/N, N/2]$).

2. If $b_t = 1$ then set the bit $\sigma \in \{0, 1\}$ to be 1 with probability u_t and to be 0 with probability $1 - u_t$. If $b_t = 0$ then set σ to be 1 with probability v_t and to be 0 with probability $1 - v_t$. (We specify the exact values of u_t, v_t below.)

3. Set s to be a uniform random element of S_σ. Output the query-response pair $(S_t, s_t) = (S, s)$.

It is clear that Step 1 above perfectly simulates the t-th query that algorithm A would make (no matter what is the distribution D). To show that the t-th response is simulated perfectly, we must show that

(i) if b_t is uniform random over $\{0,1\}$ then s is distributed exactly as it would be distributed if A were being run on $\mathsf{COND}_{\mathcal{U}}$ and had just proposed S as a query to $\mathsf{COND}_{\mathcal{U}}$; i.e. we must show that s is a uniform random element of S_1 with probability $p(\alpha) \stackrel{\text{def}}{=} \frac{\alpha}{\alpha+1}$ and is a uniform random element of S_0 with probability $1-p(\alpha)$.

(ii) if $b_t \in \{0,1\}$ has $\Pr[b_t = 1] = 1/2 + 4\epsilon$, then s is distributed exactly as it would be distributed if A were being run on $\mathsf{COND}_{D_{\mathrm{No}}}$ and had just proposed S as a query to $\mathsf{COND}_{\mathcal{U}}$; i.e. we must show that s is a uniform random element of S_1 with probability $q(\alpha) \stackrel{\text{def}}{=} \frac{\alpha}{\alpha+(1+2\epsilon)/(1-2\epsilon)}$ and is a uniform random element of S_0 with probability $1-q(\alpha)$.

By (i), we require that

$$\frac{u_t}{2} + \frac{v_t}{2} = p(\alpha) = \frac{\alpha}{\alpha+1}, \tag{22}$$

and by (ii) we require that

$$\left(\frac{1}{2} + 4\epsilon\right) u_t + \left(\frac{1}{2} - 4\epsilon\right) v_t = q(\alpha) = \frac{\alpha}{\alpha + \frac{1+2\epsilon}{1-2\epsilon}} \tag{23}$$

It is straightforward to check that

$$u_t = \frac{\alpha}{\alpha+1}\left(1 - \frac{1}{2((1-2\epsilon)\alpha + 1 + 2\epsilon)}\right), \qquad v_t = \frac{\alpha}{\alpha+1}\left(1 + \frac{1}{2((1-2\epsilon)\alpha + 1 + 2\epsilon)}\right)$$

satisfy the above equations, and that for $0 < \alpha$, $0 < \epsilon \leq 1/8$ we have $0 \leq u_t, v_t \leq 1$. So indeed A' perfectly simulates the execution of A in all stages $t = 1, \ldots, q$. Finally, after simulating the t-th stage algorithm A' outputs whatever is output by its simulation of A, so Equation (21) indeed holds. This concludes the proof of Theorem 6. ■

5 Testing equivalence to a known distribution D^*

5.1 A $\mathrm{poly}(\log n, 1/\epsilon)$-query PCOND_D algorithm

In this subsection we present an algorithm PCOND-Test-Known and prove the following theorem:

Theorem 7 PCOND-Test-Known *is a $\tilde{O}((\log N)^4/\epsilon^4)$-query PCOND_D testing algorithm for testing equivalence to a known distribution D^*. That is, for every pair of distributions D, D^* over $[N]$ (such that D^* is fully specified and there is* PCOND *query access to D) the algorithm outputs* ACCEPT *with probability at least $2/3$ if $D = D^*$ and outputs* REJECT *with probability at least $2/3$ if $d_{\mathrm{TV}}(D, D^*) \geq \epsilon$.*

Intuition. Let D^* be a fully specified distribution, and let D be a distribution that may be accessed via a PCOND_D oracle. The high-level idea of the PCOND-Test-Known algorithm is the following: As in the case of testing uniformity, we shall try to "catch" a pair of points x, y such that $\frac{D(x)}{D(y)}$ differs significantly from $\frac{D^*(x)}{D^*(y)}$ (so that calling $\textsc{Compare}_D$ on $\{x\}, \{y\}$ will reveal this

difference). In the uniformity case, where $D^*(z) = 1/N$ for every z (so that $\frac{D^*(x)}{D^*(x)+D^*(y)} = 1/2$), to get a poly$(1/\epsilon)$-query algorithm it was sufficient to show that sampling $\Theta(1/\epsilon)$ points uniformly (i.e., according to D^*) with high probability yields a point x for which $D(x) < D^*(x) - \Omega(\epsilon/N)$, and that sampling $\Theta(1/\epsilon)$ points from SAMP_D with high probability yields a point y for which $D(x) > D^*(y) + \Omega(\epsilon/N)$. However, for general D^* it is not sufficient to get such a pair because it is possible that $D^*(y)$ could be much larger than $D^*(x)$. If this were the case then it might happen that both $\frac{D^*(x)}{D^*(y)}$ and $\frac{D(x)}{D(y)}$ are very small, so calling $\textsc{Compare}_D$ on $\{x\}, \{y\}$ cannot efficiently demonstrate that $\frac{D^*(x)}{D^*(y)}$ differs from $\frac{D(x)}{D(y)}$.

To address this issue we partition the points into $O(\log N/\epsilon)$ "buckets" so that within each bucket all points have similar probability according to D^*. We show that if D is ϵ-far from D^*, then either the probability weight of one of these buckets according to D differs significantly from what it is according to D^* (which can be observed by sampling from D), or we can get a pair $\{x, y\}$ *that belong to the same bucket* and for which $D(x)$ is sufficiently smaller than $D^*(x)$ and $D(y)$ is sufficiently larger than $D^*(y)$. For such a pair $\textsc{Compare}$ will efficiently give evidence that D differs from D^*.

The algorithm and its analysis. We define some quantities that are used in the algorithm and its analysis. Let $\eta \stackrel{\text{def}}{=} \epsilon/c$ for some sufficiently large constant c that will be determined later. As described above we partition the domain elements $[N]$ into "buckets" according to their probability weight in D^*. Specifically, for $j = 1, \ldots, \lceil \log(N/\eta) + 1 \rceil$, we let

$$B_j \stackrel{\text{def}}{=} \{x \in [N] : 2^{j-1} \cdot \eta/N \leq D^*(x) < 2^j \cdot \eta/N\} \tag{24}$$

and we let $B_0 \stackrel{\text{def}}{=} \{x \in [N] : D^*(x) < \eta/N\}$. Let $b \stackrel{\text{def}}{=} \lceil \log(N/\eta) + 1 \rceil + 1$ denote the number of buckets.

We further define $J^h \stackrel{\text{def}}{=} \{j : D^*(B_j) \geq \eta/b\}$ to denote the set of indices of "heavy" buckets, and let $J^\ell \stackrel{\text{def}}{=} \{j : D^*(B_j) < \eta/b\}$ denote the set of indices of "light" buckets. Note that we have

$$\sum_{j \in J^\ell \cup \{0\}} D^*(B_j) < 2\eta. \tag{25}$$

The query complexity of the algorithm is dominated by the number of PCOND_D queries performed in the executions of $\textsc{Compare}$, which by Lemma 2 is upper bounded by

$$O(s^2 \cdot b^2 \cdot (\log s)/\eta^2) = O\left(\frac{(\log \frac{N}{\epsilon})^4 \cdot \log\left((\log \frac{N}{\epsilon})/\epsilon\right)}{\epsilon^4}\right).$$

We argue completeness and soundness below.

Completeness: Suppose that $D = D^*$. Since the expected value of $\widehat{D}(B_j)$ (defined in Line 3) is precisely $D^*(B_j)$, for any fixed value of $j \in \{0, \ldots, \lceil \log(N/\eta) + 1 \rceil\}$ an additive Chernoff bound implies that $\left|D^*(B_j) - \widehat{D}(B_j)\right| > \eta/b$ with failure probability at most $1/(10b)$. By a union bound over all b values of j, the algorithm outputs REJECT in Line 5 with probability at most $1/10$.

Algorithm 5: PCOND_D-Test-Known

Input: error parameter $\epsilon > 0$; query access to PCOND_D oracle; explicit description $(D^*(1), \ldots, D^*(N))$ of distribution D^*

1: Call the SAMP_D oracle $m = \Theta(b^2(\log b)/\eta^2)$ times to obtain points $h_1, \ldots, h_m$ distributed according to D.
2: **for** $j = 0$ to b **do**
3: Let $\widehat{D}(B_j)$ be the fraction of points $h_1, \ldots, h_m$ that lie in B_j (where the buckets B_j are as defined in Equation (24)).
4: **if** some j has $|D^*(B_j) - \widehat{D}(B_j)| > \eta/b$ **then**
5: output REJECT and exit
6: **end if**
7: **end for**
8: Select $s = \Theta(b/\epsilon)$ points $x_1, \ldots, x_s$ independently from D^*.
9: Call the SAMP_D oracle $s = \Theta(b/\epsilon)$ times to obtain points $y_1, \ldots, y_s$ distributed according to D.
10: **for all** pairs (x_i, y_j) (where $1 \le i, j \le s$) such that $\frac{D^*(x)}{D^*(y)} \in [1/2, 2]$ **do**
11: Call Compare$(\{x\}, \{y\}, \eta/(4b), 2, 1/(10s^2))$
12: **if** Compare returns Low or a value smaller than $(1 - \eta/(2b)) \cdot \frac{D^*(x)}{D^*(y)}$ **then**
13: output REJECT (and exit)
14: **end if**
15: **end for**
16: output ACCEPT

Later in the algorithm, since $D = D^*$, no matter what points x_i, y_j are sampled from D^* and D respectively, the following holds for each pair (x_i, y_j) such that $D^*(x)/D^*(y) \in [1/2, 2]$. By Lemma 2 (and the setting of the parameters in the calls to Compare), the probability that Compare returns Low or a value smaller than $(1 - \delta/(2b)) \cdot (D^*(x)/D^*(y))$, is at most $1/(10s^2)$. A union bound over all (at most s^2) pairs (x_i, y_j) for which $D^*(x)/D^*(y) \in [1/2, 2]$, gives that the probability of outputting REJECT in Line 13 is at most $1/10$. Thus with overall probability at least $8/10$ the algorithm outputs ACCEPT.

Soundness: Now suppose that $d_{\mathrm{TV}}(D, D^*) \ge \epsilon$; our goal is to show that the algorithm rejects with probability at least $2/3$. Since the algorithm rejects if any estimate $\widehat{D}(B_j)$ obtained in Line 3 deviates from $D^*(B_j)$ by more than $\pm\eta/b$, we may assume that all these estimates are indeed $\pm\eta/b$-close to the values $D^*(B_j)$ as required. Moreover, by an additive Chernoff bound (as in the completeness analysis), we have that with overall failure probability at most $1/10$, each j has $|\widehat{D}(B_j) - D(B_j)| \le \eta/b$; we condition on this event going forth. Thus, for every $0 \le j \le b$,

$$D^*(B_j) - 2\eta/b \le D(B_j) \le D^*(B_j) + 2\eta/b\,. \tag{26}$$

Recalling the definition of J^ℓ and Equation (25), we see that

$$\sum_{j \in J^\ell \cup \{0\}} D(B_j) < 4\eta\,. \tag{27}$$

Let

$$d_j \stackrel{\text{def}}{=} \sum_{x \in B_j} |D^*(x) - D(x)| \,, \tag{28}$$

so that $\|D^* - D\|_1 = \sum_j d_j$. By Equations (25) and (27), we have

$$\sum_{j \in J^\ell \cup \{0\}} d_j \leq \sum_{j \in J^\ell \cup \{0\}} (D^*(B_j) + D(B_j)) \leq 6\eta \,. \tag{29}$$

Since we have (by assumption) that $\|D^* - D\|_1 = 2d_{\text{TV}}(D^*, D) \geq 2\epsilon$, we get that

$$\sum_{j \in J^h \setminus \{0\}} d_j > 2\epsilon - 6\eta \,. \tag{30}$$

Let $N_j \stackrel{\text{def}}{=} |B_j|$ and observe that $N_j \leq D^*(B_j)/p_j \leq 1/p_j$, where $p_j \stackrel{\text{def}}{=} 2^{j-1} \cdot \eta/N$ is the lower bound on the probability (under D^*) of all elements in B_j. For each B_j such that $j \in J^h \setminus \{0\}$, let $H_j \stackrel{\text{def}}{=} \{x \in B_j : D(x) > D^*(x)\}$ and $L_j \stackrel{\text{def}}{=} \{x \in B_j : D(x) < D^*(x)\}$. Similarly to the "testing uniformity" analysis, we have that

$$\sum_{x \in L_j} (D^*(x) - D(x)) + \sum_{x \in H_j} (D(x) - D^*(x)) = d_j \,. \tag{31}$$

Equation (26) may be rewritten as

$$\left| \sum_{x \in L_j} (D^*(x) - D(x)) - \sum_{x \in H_j} (D(x) - D^*(x)) \right| \leq 2\eta/b \,, \tag{32}$$

and so we have both

$$\sum_{x \in L_j} (D^*(x) - D(x)) \geq d_j/2 - \eta/b \quad \text{and} \quad \sum_{x \in H_j} (D(x) - D^*(x)) \geq d_j/2 - \eta/b \,. \tag{33}$$

Also similarly to what we had before, let $H'_j \stackrel{\text{def}}{=} \{x \in B_j : D(x) > D^*(x) + \eta/(bN_j)\}$, and $L'_j \stackrel{\text{def}}{=} \{x \in B_j : D(x) < D^*(x) - \eta/(bN_j)\}$ (recall that $N_j = |B_j|$); these are the elements of B_j that are "significantly heavier" (lighter, respectively) under D than under D^*. We have

$$\sum_{x \in L_j \setminus L'_j} (D^*(x) - D(x)) \leq \eta/b \quad \text{and} \quad \sum_{x \in H_j \setminus H'_j} (D(x) - D^*(x)) \leq \eta/b \,. \tag{34}$$

By Equation (30), there exists $j^* \in J^h \setminus \{0\}$ for which $d_{j^*} \geq (2\epsilon - 6\eta)/b$. For this index, applying Equations (33) and (34), we get that

$$\sum_{x \in L'_{j^*}} D^*(x) \geq \sum_{x \in L'_{j^*}} (D^*(x) - D(x)) \geq (\epsilon - 5\eta)/b \,, \tag{35}$$

and similarly,

$$\sum_{x \in H'_{j^*}} D(x) \geq \sum_{x \in H'_{j^*}} (D(x) - D^*(x)) \geq (\epsilon - 5\eta)/b \, . \tag{36}$$

Recalling that $\eta = \epsilon/c$ and setting the constant c to 6, we have that $(\epsilon - 5\eta)/b = \epsilon/6b$. Since $s = \Theta(b/\epsilon)$, with probability at least 9/10 it is the case both that some x_i drawn in Line 8 belongs to L'_{j^*} and that some $y_{i'}$ drawn in Line 9 belongs to H'_{j^*}. By the definitions of L'_{j^*} and H'_{j^*} and the fact for each $j > 0$ it holds that $N_j \leq 1/p_j$ and $p_j \leq D^*(x) < 2p_j$ for each $x_i \in B_j$, we have that

$$D(x_i) < D^*(x_i) - \eta/(bN_{j^*}) \leq D^*(x_i) - (\eta/b)p_{j^*} \leq (1 - \eta/(2b))D^*(x_i) \tag{37}$$

and

$$D(y_{i'}) > D^*(y_{i'}) + \eta/(bN_{j^*}) \geq D^*(y_{i'}) + (\eta/b)p_j \geq (1 + \eta/(2b))D^*(y_{i'}) \, . \tag{38}$$

Therefore,

$$\frac{D(x_i)}{D(y_{i'})} < \frac{1 - \eta/(2b)}{1 + \eta/(2b)} \cdot \frac{D^*(x_i)}{D^*(y_{i'})} < \left(1 - \frac{3\eta}{4b}\right) \cdot \frac{D^*(x_i)}{D^*(y_{i'})} \, . \tag{39}$$

By Lemma 2, with probability at least $1 - 1/(10s^2)$, the output of COMPARE is either Low or is at most $\left(1 - \frac{3\eta}{4b}\right) \cdot \left(1 + \frac{\eta}{4b}\right) < \left(1 - \frac{\eta}{2b}\right)$, causing the algorithm to reject. Thus the overall probability that the algorithm outputs REJECT is at least $8/10 - 1/(10s^2) > 2/3$, and the theorem is proved. ■

5.2 A $(\log N)^{\Omega(1)}$ lower bound for PCOND_D

In this subsection we prove that any PCOND_D algorithm for testing equivalence to a known distribution must have query complexity at least $(\log N)^{\Omega(1)}$:

Theorem 8 *Fix $\epsilon = 1/2$. There is a distribution D^* over $[N]$ (described below), which is such that any PCOND_D algorithm for testing whether $D = D^*$ versus $d_{\mathrm{TV}}(D, D^*) \geq \epsilon$ must make $\Omega\left(\sqrt{\frac{\log N}{\log \log N}}\right)$ queries.*

The distribution D^*. Fix parameters $r = \Theta\left(\frac{\log N}{\log \log N}\right)$ and $K = \Theta(\log N)$. We partition $[N]$ from left (1) to right (N) into $2r$ consecutive intervals $B_1, \ldots, B_{2r}$, which we henceforth refer to as "buckets." The i-th bucket has $|B_i| = K^i$ (we may assume without loss of generality that N is of the form $\sum_{i=1}^{2r} K^i$). The distribution D^* assigns equal probability weight to each bucket, so $D^*(B_i) = 1/(2r)$ for all $1 \leq i \leq 2r$. Moreover D^* is uniform within each bucket, so for all $j \in B_i$ we have $D^*(j) = 1/(2rK^i)$. This completes the specification of D^*.

To prove the lower bound we construct a probability distribution $\mathcal{P}_{\mathrm{No}}$ over possible "No"-distributions. To define the distribution $\mathcal{P}_{\mathrm{No}}$ it will be useful to have the notion of a "bucket-pair." A bucket-pair U_i is $U_i = B_{2i-1} \cup B_{2i}$, i.e. the union of the i-th pair of consecutive buckets.

A distribution D drawn from $\mathcal{P}_{\mathrm{No}}$ is obtained by selecting a string $\pi = (\pi_1, \ldots, \pi_r)$ uniformly at random from $\{\downarrow\uparrow, \uparrow\downarrow\}^r$ and setting D to be D_π, which we now define. The distribution D_π is obtained by perturbing D^* in the following way: for each bucket-pair $U_i = (B_{2i-1}, B_{2i})$,

- If $\pi_i = \uparrow\downarrow$ then the weight of B_{2i-1} is uniformly "scaled up" from $1/(2r)$ to $3/(4r)$ (keeping the distribution uniform within B_{2i-1}) and the weight of B_{2i} is uniformly "scaled down" from $1/(2r)$ to $1/(4r)$ (likewise keeping the distribution uniform within B_{2i}).
- If $\pi_i = \downarrow\uparrow$ then the weight of B_{2i-1} is uniformly "scaled down" from $1/(2r)$ to $1/(4r)$ and the weight of B_{2i} is uniformly "scaled up" from $1/(2r)$ to $3/(4r)$.

Note that for any distribution D in the support of $\mathcal{P}_{\mathrm{No}}$ and any $1 \leq i \leq r$ we have that $D(U_i) = D^*(U_i) = 1/r$.

Every distribution D in the support of $\mathcal{P}_{\mathrm{No}}$ has $d_{\mathrm{TV}}(D^*, D) = 1/2$. Thus Theorem 8 follows immediately from the following:

Theorem 9 *Let A be any (possibly adaptive) algorithm. which makes at most $q \leq \frac{1}{3} \cdot \sqrt{r}$ calls to PCOND_D. Then*

$$\left|\Pr_{D \leftarrow \mathcal{P}_{\mathrm{No}}}\left[A^{\mathsf{PCOND}_D} \text{ outputs } \mathsf{ACCEPT}\right] - \Pr\left[A^{\mathsf{PCOND}_{D^*}} \text{ outputs } \mathsf{ACCEPT}\right]\right| \leq 1/5. \tag{40}$$

Note that in the first probability of Equation (40) the randomness is over the draw of D from $\mathcal{P}_{\mathrm{No}}$, the internal randomness of A in selecting its query sets, and the randomness of the responses to the PCOND_D queries. In the second probability the randomness is just over the internal coin tosses of A and the randomness of the responses to the PCOND_D queries.

Intuition for Theorem 9. A very high-level intuition for the lower bound is that PCOND_D queries are only useful for "comparing" points whose probabilities are within a reasonable multiplicative ratio of each other. But D^* and every distribution D in the support of $\mathcal{P}^{\mathrm{No}}$ are such that every two points either have the same probability mass under all of these distributions (so a PCOND_D query is not informative), or else the ratio of their probabilities is so skewed that a small number of PCOND_D queries is not useful for comparing them.

In more detail, we may suppose without loss of generality that in every possible execution, algorithm A first makes q calls to SAMP_D and then makes q (possibly adaptive) calls to PCOND_D. The more detailed intuition for the lower bound is as follows: First consider the SAMP_D calls. Since every possible D (whether D^* or a distribution drawn from $\mathcal{P}_{\mathrm{No}}$) puts weight $1/r$ on each bucket-pair $U_1, \ldots, U_r$, a birthday paradox argument implies that in both scenarios, with probability at least $9/10$ (over the randomness in the responses to the SAMP_D queries) no two of the $q \leq \frac{1}{3}\sqrt{r}$ calls to SAMP_D return points from the same bucket-pair. Conditioned on this, the distribution of responses to the SAMP_D queries is exactly the same under D^* and under D where D is drawn randomly from $\mathcal{P}_{\mathrm{No}}$.

For the pair queries, the intuition is that in either setting (whether the distribution D is D^* or a randomly chosen distribution from $\mathcal{P}_{\mathrm{No}}$), making q pair queries will with $1 - o(1)$ probability provide no information that the tester could not simulate for itself. This is because any pair query $\mathsf{PCOND}_D(\{x, y\})$ either has x, y in the same bucket B_i or in different buckets $B_i \neq B_j$ with $i < j$. If x, y are both in the same bucket B_i then in either setting $\mathsf{PCOND}_D(\{x, y\})$ is equally likely to return x or y. If they belong to buckets B_i, B_j with $i < j$ then in either setting $\mathsf{PCOND}_D(\{x, y\})$ will return the one that belongs to P_i with probability $1 - 1/\Theta(K^{j-i}) \geq 1 - 1/\Omega(K)$.

Proof of Theorem 9: As described above, we may fix A to be any PCOND_D algorithm that makes exactly q calls to SAMP_D followed by exactly q adaptive calls to PCOND_D.

A *transcript* for A is a full specification of the sequence of interactions that A has with the PCOND_D oracle in a given execution. More precisely, it is a pair (Y, Z) where $Y = (s_1, \ldots, s_q) \in [N]^q$ and $Z = ((\{x_1, y_1\}, p_1), \ldots, (\{x_q, y_q\}, p_q))$, where $p_i \in \{x_i, y_i\}$ and $x_i, y_i \in [N]$. The idea is that Y is a possible sequence of responses that A might receive to the initial q SAMP_D queries, $\{x_i, y_i\}$ is a possible pair that could be the input to an i-th PCOND_D query, and p_i is a possible response that could be received from that query.

We say that a *length-i transcript prefix* is a pair (Y, Z^i) where Y is as above and $Z^i = ((\{x_1, y_1\}, p_1), \ldots, (\{x_i, y_i\}, p_i))$. A PCOND algorithm A may be viewed as a collection of distributions over pairs $\{x, y\}$ in the following way: for each length-i transcript-prefix (Y, Z^i) $(0 \leq i \leq q-1)$, there is a distribution over pairs $\{x_{i+1}, y_{i+1}\}$ that A would use to select the $(i+1)$-st query pair for PCOND_D given that the length-i transcript prefix of A's execution thus far was (Y, Z^i). We write $\mathrm{T}_{(Y,Z^i)}$ to denote this distribution over pairs.

Let P^* denote the distribution over transcripts induced by running A with oracle PCOND_{D^*}. Let P^{No} denote the distribution over transcripts induced by first (i) drawing D from $\mathcal{P}_{\mathrm{No}}$, and then (ii) running A with oracle PCOND_D. To prove Theorem 9 it is sufficient to prove that the distribution over transcripts of A is statistically close whether the oracle is D^* or is a random D drawn from $\mathcal{P}^{\mathrm{No}}$, i.e. it is sufficient to prove that

$$d_{\mathrm{TV}}(\mathrm{P}^*, \mathrm{P}^{\mathrm{No}}) \leq 1/5. \tag{41}$$

For our analysis we will need to consider variants of algorithm A that, rather than making q calls to PCOND_D, instead "fake" the final $q-k$ of these PCOND_D queries as described below. For $0 \leq k \leq q$ we define $A^{(k)}$ to be the algorithm that works as follows:

1. $A^{(k)}$ exactly simulates the execution of A in making an initial q SAMP_D calls and making the first k PCOND_D queries precisely like A. Let (Y, Z^k) be the length-k transcript prefix of A's execution thus obtained.

2. Exactly like A, algorithm $A^{(k)}$ draws a pair $\{x_{k+1}, y_{k+1}\}$ from $\mathrm{T}_{(Y,Z^k)}$. However, instead of calling $\mathsf{PCOND}_D(\{x_{k+1}, y_{k+1}\})$ to obtain p_{k+1}, algorithm $A^{(k)}$ generates p_{k+1} in the following manner:

   (i) If x_{k+1} and y_{k+1} both belong to the same bucket B_ℓ then p_{k+1} is chosen uniformly from $\{x_{k+1}, y_{k+1}\}$.

   (ii) If one of $\{x_{k+1}, y_{k+1}\}$ belongs to B_ℓ and the other belongs to $B_{\ell'}$ for some $\ell < \ell'$, then p_{k+1} is set to be the element of $\{x_{k+1}, y_{k+1}\}$ that belongs to B_ℓ.

   Let (Y, Z^{k+1}) be the length-$(k+1)$ transcript prefix obtained by appending $(\{x_{k+1}, y_{k+1}\}, p_{k+1})$ to Z^k. Algorithm $A^{(k)}$ continues in this way for a total of $q-k$ stages; i.e. it next draws $\{x_{k+2}, y_{k+2}\}$ from $\mathrm{T}_{(Y,Z^{k+1})}$ and generates p_{k+2} as described above; then (Y, Z^{k+2}) is the length-$(k+2)$ transcript prefix obtained by appending $(\{x_{k+2}, y_{k+2}\}, p_{k+2})$ to Z^{k+1}; and so on. At the end of the process a transcript (Y, Z^q) has been constructed.

Let $\mathrm{P}^{*,(k)}$ denote the distribution over final transcripts (Y, Z^q) that are obtained by running $A^{(k)}$ on a PCOND_{D^*} oracle. Let $\mathrm{P}^{\mathrm{No},(k)}$ denote the distribution over final transcripts (Y, Z^q) that are obtained by (i) first drawing D from $\mathcal{P}^{\mathrm{No}}$, and then (ii) running $A^{(k)}$ on a PCOND_D oracle. Note that $\mathrm{P}^{*,(q)}$ is identical to P^* and $\mathrm{P}^{\mathrm{No},(q)}$ is identical to P^{No} (since algorithm $A^{(q)}$, which does not fake any queries, is identical to algorithm A).

Recall that our goal is to prove Equation (41). Since $\mathrm{P}^{*,(q)} = \mathrm{P}^*$ and $\mathrm{P}^{\mathrm{No},(q)} = \mathrm{P}^{\mathrm{No}}$, Equation (41) is an immediate consequence (using the triangle inequality for total variation distance) of the following two lemmas, which we prove below:

Lemma 15 $d_{\mathrm{TV}}(\mathrm{P}^{*,(0)}, \mathrm{P}^{\mathrm{No},(0)}) \leq 1/10$.

Lemma 16 *For all* $0 \leq k < q$, *we have* $d_{\mathrm{TV}}(\mathrm{P}^{*,(k)}, \mathrm{P}^{*,(k+1)}) \leq 1/(20q)$ *and* $d_{\mathrm{TV}}(\mathrm{P}^{\mathrm{No},(k)}, \mathrm{P}^{\mathrm{No},(k+1)}) \leq 1/(20q)$.

Proof of Lemma 15: Define P^*_0 to be the distribution over outcomes of the q calls to SAMP_D (i.e. over length-0 transcript prefixes) when $D = D^*$. Define P^{No} to be the distribution over outcomes of the q calls to SAMP_D when D is drawn from $\mathcal{P}_{\mathrm{No}}$. We begin by noting that by the data processing inequality for total variation distance (Lemma 1), we have $d_{\mathrm{TV}}(\mathrm{P}^{*,(0)}, \mathrm{P}^{\mathrm{No},(0)}) \leq d_{\mathrm{TV}}(\mathrm{P}^*_0, \mathrm{P}^{\mathrm{No}}_0)$ (indeed, after the calls to respectively SAMP_D and SAMP_{D^*}, the same randomized function F – which fakes all remaining oracle calls – is applied to the two resulting distributions over length-0 transcript prefixes P^*_0 and $\mathrm{P}^{\mathrm{No}}_0$). In the rest of the proof we show that $d_{\mathrm{TV}}(\mathrm{P}^*_0, \mathrm{P}^{\mathrm{No}}_0) \leq 1/10$.

Let E denote the event that the q calls to SAMP_D yield points $s_1, \ldots, s_q$ such that no bucket-pair U_i contains more than one of these points. Since $D^*(U_i) = 1/r$ for all i,

$$\mathrm{P}^*_0(E) = \prod_{j=0}^{q-1} \left(1 - \frac{j}{r}\right) \geq 9/10\,, \tag{42}$$

where Equation (42) follows from a standard birthday paradox analysis and the fact that $q \leq \frac{1}{3}\sqrt{r}$. Since for each possible outcome of D drawn from $\mathcal{P}_{\mathrm{No}}$ we have $D(U_i) = 1/r$ for all i, we further have that also

$$\mathrm{P}^{\mathrm{No}}_0(E) = \prod_{j=0}^{q-1} \left(1 - \frac{j}{r}\right). \tag{43}$$

We moreover claim that the two conditional distributions $(\mathrm{P}^*_0|E)$ and $(\mathrm{P}^{\mathrm{No}}_0|E)$ are identical, i.e.

$$(\mathrm{P}^*_0|E) = (\mathrm{P}^{\mathrm{No}}_0|E). \tag{44}$$

To see this, fix any sequence $(\ell_1, \ldots, \ell_q) \in [r]^q$ such that $\ell_i \neq \ell_j$ for all $i \neq j$. Let $(s_1, \ldots, s_q) \in [N]^q$ denote a draw from $(\mathrm{P}^*_0|E)$. The probability that $(s_i \in U_{\ell_i}$ for all $1 \leq i \leq q)$ is precisely $1/r^q$. Now given that $s_i \in U_{\ell_i}$ for all i, it is clear that s_i is equally likely to lie in $B_{2\ell_i - 1}$ and in $B_{2\ell_i}$, and given that it lies in a particular one of the two buckets, it is equally likely to be any element in that bucket. This is true independently for all $1 \leq i \leq q$.

Now let $(s_1, \ldots, s_q) \in [N]^q$ denote a draw from $(\mathrm{P}^{\mathrm{No}}_0|E)$. Since each distribution D in the support of $\mathcal{P}_{\mathrm{No}}$ has $D(U_i) = 1/r$ for all i, we likewise have that the probability that $(s_i \in U_{\ell_i}$ for all $1 \leq i \leq q)$

is precisely $1/r^q$. Now given that $s_i \in U_{\ell_i}$ for all i, we have that s_i is equally likely to lie in $B_{2\ell_i-1}$ and in $B_{2\ell_i}$; this is because π_i (recall that π determines $D = D_\pi$) is equally likely to be $\uparrow\downarrow$ (in which case $D(B_{2\ell_i-1}) = 3/(4r)$ and $D(B_{2\ell_i}) = 1/(4r)$) as it is to be $\downarrow\uparrow$ (in which case $D(B_{2\ell_i-1}) = 1/(4r)$ and $D(B_{2\ell_i}) = 3/(4r)$). Additionally, given that s_i lies in a particular one of the two buckets, it is equally likely to be any element in that bucket. This is true independently for all $1 \le i \le q$ (because conditioning on E ensures that no two elements of $s_1, \ldots, s_q$ lie in the same bucket-pair, so there is "fresh randomness for each i"), and so indeed the two conditional distributions $(\mathrm{P}_0^*|E)$ and $(\mathrm{P}_0^{\mathrm{No}}|E)$ are identical.

Finally, the claimed bound $d_{\mathrm{TV}}(\mathrm{P}_0^*, \mathrm{P}_0^{\mathrm{No}}) \le 1/10$ follows directly from Equations (42), (43) and (44). ■

Proof of Lemma 16: Consider first the claim that $d_{\mathrm{TV}}(\mathrm{P}^{*,(k)}, \mathrm{P}^{*,(k+1)}) \le 1/(20q)$. Fix any $0 \le k < q$. The data processing inequality for total variation distance implies that $d_{\mathrm{TV}}(\mathrm{P}^{*,(k)}, \mathrm{P}^{*,(k+1)})$ is at most the variation distance between random variables X and X', where

- X is the random variable obtained by running A on COND_{D^*} to obtain a length-k transcript prefix (Y, Z^k), then drawing $\{x_{k+1}, y_{k+1}\}$ from $\mathrm{T}_{(Y,Z^k)}$, then setting p_{k+1} to be the output of $\mathsf{PCOND}_{D^*}(\{x_{k+1}, y_{k+1}\})$; and
- X' is the random variable obtained by running A on COND_{D^*} to obtain a length-k transcript prefix (Y, Z^k), then drawing $\{x_{k+1}, y_{k+1}\}$ from $\mathrm{T}_{(Y,Z^k)}$, then setting p_{k+1} according to the aforementioned rules 2(i) and 2(ii).

Consider any fixed outcome of (Y, Z^k) and $\{x_{k+1}, y_{k+1}\}$. If rule 2(i) is applied (x_{k+1} and y_{k+1} are in the same bucket), then there is zero contribution to the variation distance between X and X', because choosing a uniform element of $\{x_{k+1}, y_{k+1}\}$ is a perfect simulation of $\mathsf{PCOND}_D(\{x_{k+1}, y_{k+1}\})$. If rule 2(ii) is applied, then the contribution is upper bounded by $O(1/K) < 1/20q$, because $\mathsf{PCOND}_{D^*}(\{x_{k+1}y_{k+1}\})$ would return a different outcome from rule 2(ii) with probability $1/\Theta(K^{\ell'-\ell}) = O(1/K)$. Averaging over all possible outcomes of (Y, Z^k) and $\{x_{k+1}, y_{k+1}\}$ we get that the variation distance between X and X' is at most $1/20q$ as claimed.

An identical argument shows that similarly $d_{\mathrm{TV}}(\mathrm{P}^{\mathrm{No},(k)}, \mathrm{P}^{\mathrm{No},(k+1)}) \le 1/(20q)$. The key observation is that for any distribution D in the support of $\mathcal{P}^{\mathrm{No}}$, as with D^* it is the case that points in the same bucket have equal probability under D and for a pair of points $\{x, y\}$ such that $x \in B_\ell$ and $y \in B_{\ell'}$ for $\ell' > \ell$, the probability that a call to $\mathsf{PCOND}_D(\{x, y\})$ returns y is only $1/\Theta(K^{\ell'-\ell})$. This concludes the proof of Lemma 16 and of Theorem 8. ■

5.3 A $\mathrm{poly}(1/\epsilon)$-query COND_D algorithm

In this subsection we present an algorithm COND-Test-Known and prove the following theorem:

Theorem 10 COND-Test-Known *is a $\tilde{O}(1/\epsilon^4)$-query COND_D testing algorithm for testing equivalence to a known distribution D^*. That is, for every pair of distributions D, D^* over $[N]$ (such that D^* is fully specified and there is COND query access to D), the algorithm outputs ACCEPT with probability at least $2/3$ if $D = D^*$ and outputs REJECT with probability at least $2/3$ if $d_{\mathrm{TV}}(D, D^*) \ge \epsilon$.*

This constant-query testing algorithm stands in interesting contrast to the $(\log N)^{\Omega(1)}$-query lower bound for PCOND_D algorithms for this problem.

High-level overview of the algorithm and its analysis: First, we note that by reordering elements of $[N]$ we may assume without loss of generality that $D^*(1) \leq \cdots \leq D^*(N)$; this will be convenient for us.

Our $(\log N)^{\Omega(1)}$ query lower bound for PCOND_D algorithms exploited the intuition that comparing two points using the PCOND_D oracle might not provide much information (e.g. if one of the two points was a priori "known" to be much heavier than the other). In contrast, with a general COND_D oracle at our disposal, we can compare a given point $j \in [N]$ with *any subset* of $[N] \setminus \{j\}$. Thus the following definition will be useful:

Definition 4 (comparable points) *Fix $0 < \lambda \leq 1$. A point $j \in \mathrm{supp}(D^*)$ is said to be* λ-comparable *if there exists a set $S \subseteq ([N] \setminus \{j\})$ such that*

$$D^*(j) \in [\lambda D^*(S), D^*(S)/\lambda].$$

Such a set S is then said to be a λ-comparable-witness *for j (according to D^*), which is denoted $S \cong^* j$. We say that a set $T \subseteq [N]$ is λ-comparable if every $i \in T$ is λ-comparable.*

We stress that the notion of being λ-comparable deals only with the known distribution D^*; this will be important later.

Fix $\epsilon_1 = \Theta(\epsilon)$ (we specify ϵ_1 precisely in Equation (47) below). Our analysis and algorithm consider two possible cases for the distribution D^* (where it is not hard to verify, and we provide an explanation subsequently, that one of the two cases must hold):

1. The first case is that for some $i^* \in [N]$ we have

$$D^*(\{1, \ldots, i^*\}) > 2\epsilon_1 \quad \text{but} \quad D^*(\{1, \ldots, i^* - 1\}) \leq \epsilon_1. \tag{45}$$

   In this case $1 - \epsilon_1$ of the total probability mass of D^* must lie on a set of at most $1/\epsilon_1$ elements, and in such a situation it is easy to efficiently test whether $D = D^*$ using $\mathrm{poly}(1/\epsilon)$ queries (see Algorithm COND_D-Test-Known-Heavy and Lemma 19).

2. The second case is that there exists an element $k^* \in [N]$ such that

$$\epsilon_1 < D^*(\{1, \ldots, k^*\}) \leq 2\epsilon_1 < D^*(\{1, \ldots, k^* + 1\}). \tag{46}$$

   This is the more challenging (and typical) case. In this case, it can be shown that every element $j > k^*$ has at least one ϵ_1-comparable-witness within $\{1, \ldots, j\}$. In fact, we show (see Claim 17) that either (a) $\{1, \ldots, j-1\}$ is an ϵ_1-comparable witness for j, or (b) the set $\{1, \ldots, j-1\}$ can be partitioned into disjoint sets[9] $S_1, \ldots, S_t$ such that *each* S_i, $1 \leq i \leq t$, is a $\frac{1}{2}$-comparable-witness for j. Case (a) is relatively easy to handle so we focus on (b) in our informal description below.

[9] In fact the sets are intervals (under the assumption $D^*(1) \leq \cdots \leq D^*(n)$), but that is not really important for our arguments.

The partition $S_1, \ldots, S_t$ is useful to us for the following reason: Suppose that $d_{\mathrm{TV}}(D, D^*) \geq \epsilon$. It is not difficult to show (see Claim 18) that unless $D(\{1, \ldots, k^*\}) > 3\epsilon_1$ (which can be easily detected and provides evidence that the tester should reject), a random sample of $\Theta(1/\epsilon)$ draws from D will with high probability contain a "heavy" point $j > k^*$, that is, a point $j > k^*$ such that $D(j) \geq (1+\epsilon_2)D^*(j)$ (where $\epsilon_2 = \Theta(\epsilon)$). Given such a point j, there are two possibilities:

1. The first possibility is that a significant fraction of the sets $S_1, \ldots, S_t$ have $D(j)/D(S_i)$ "noticeably different" from $D^*(j)/D^*(S_i)$. (Observe that since each set S_i is a $\frac{1}{2}$-comparable witness for j, it is possible to efficiently check whether this is the case.) If this is the case then our tester should reject since this is evidence that $D \neq D^*$.

2. The second possibility is that almost every S_i has $D(j)/D(S_i)$ very close to $D^*(j)/D^*(S_i)$. If this is the case, though, then since $D(j) \geq (1+\epsilon_2)D^*(j)$ and the union of $S_1, \ldots, S_t$ is $\{1, \ldots, j-1\}$, it must be the case that $D(\{1, \ldots, j\})$ is "significantly larger" than $D^*(\{1, \ldots, j\})$. This will be revealed by random sampling from D and thus our testing algorithm can reject in this case as well.

Key quantities and useful claims. We define some quantities that are used in the algorithm and its analysis. Let

$$\epsilon_1 \stackrel{\text{def}}{=} \frac{\epsilon}{10}; \quad \epsilon_2 \stackrel{\text{def}}{=} \frac{\epsilon}{2}; \quad \epsilon_3 \stackrel{\text{def}}{=} \frac{\epsilon}{48}; \quad \epsilon_4 \stackrel{\text{def}}{=} \frac{\epsilon}{6}. \tag{47}$$

Claim 17 *Suppose there exists an element $k^* \in [N]$ that satisfies Equation (46). Fix any $j > k^*$. Then*

1. *If $D^*(j) \geq \epsilon_1$, then $S_1 \stackrel{\text{def}}{=} \{1, \ldots, j-1\}$ is an ϵ_1-comparable witness for j;*

2. *If $D^*(j) < \epsilon_1$ then the set $\{1, \ldots, j-1\}$ can be partitioned into disjoint sets $S_1, \ldots, S_t$ such that each S_i, $1 \leq i \leq t$, is a $\frac{1}{2}$-comparable-witness for j.*

Proof: First consider the case that $D^*(j) \geq \epsilon_1$. In this case $S_1 = \{1, \ldots, j-1\}$ is an ϵ_1-comparable witness for j because $D^*(j) \geq \epsilon_1 \geq \epsilon_1 D^*(\{1, \ldots, j-1\})$ and $D^*(j) \leq 1 \leq \frac{1}{\epsilon_1} D^*(\{1, \ldots, k^*\}) \leq \frac{1}{\epsilon_1} D^*(\{1, \ldots, j-1\})$, where the last inequality holds since $k^* \leq j-1$.

Next, consider the case that $D^*(j) < \epsilon_1$. In this case we build our intervals iteratively from right to left, as follows. Let $j_1 = j-1$ and let j_2 be the minimum index in $\{0, \ldots, j_1 - 1\}$ such that

$$D^*(\{j_2+1, \ldots, j_1\}) \leq D^*(j).$$

(Observe that we must have $j_2 \geq 1$, because $D^*(\{1, \ldots, k^*\}) > \epsilon_1 > D^*(j)$.) Since $D^*(\{j_2, \ldots, j_1\}) > D^*(j)$ and the function $D^*(\cdot)$ is monotonically increasing, it must be the case that

$$\frac{1}{2} D^*(j) \leq D^*(\{j_2+1, \ldots, j_1\}) \leq D^*(j).$$

Thus the interval $S_1 \stackrel{\text{def}}{=} \{j_2+1, \ldots, j_1\}$ is a $\frac{1}{2}$-comparable witness for j as desired.

We continue in this fashion from right to left; i.e. if we have defined $j_2, \ldots, j_t$ as above and there is an index $j' \in \{0, \ldots, j_t - 1\}$ such that $D^*(\{j'+1, \ldots, j_t\}) > D^*(j)$, then we define j_{t+1} to be the minimum index in $\{0, \ldots, j_t - 1\}$ such that

$$D^*(\{j_{t+1}+1, \ldots, j_t\}) \leq D^*(j),$$

and we define S_t to be the interval $\{j_{t+1}+1, \ldots, j_t\}$. The argument of the previous paragraph tells us that

$$\frac{1}{2}D^*(j) \leq D^*(\{j_{t+1}+1, \ldots, j_t\}) \leq D^*(j) \tag{48}$$

and hence S_t is an $\frac{1}{2}$-comparable witness for j.

At some point, after intervals $S_1 = \{j_2+1, \ldots, j_1\}, \ldots, S_t = \{j_{t+1}+1, \ldots, j_t\}$ have been defined in this way, it will be the case that there is no index $j' \in \{0, \ldots, j_t - 1\}$ such that $D^*(\{j'+1, \ldots, j_t\}) > D^*(j)$. At this point there are two possibilities: first, if $j_{t+1}+1 = 1$, then $S_1, \ldots, S_t$ give the desired partition of $\{1, \ldots, j-1\}$. If $j_{t+1}+1 > 1$ then it must be the case that $D^*(\{1, \ldots, j_{t+1}\}) \leq D^*(j)$. In this case we simply add the elements $\{1, \ldots, j_{t+1}\}$ to S_t, i.e. we redefine S_t to be $\{1, \ldots, j_t\}$. By Equation (48) we have that

$$\frac{1}{2}D^*(j) \leq D^*(S_t) \leq 2D^*(j)$$

and thus S_t is an $\frac{1}{2}$-comparable witness for j as desired. This concludes the proof. ■

Definition 5 (Heavy points) *A point $j \in \mathrm{supp}(D^*)$ is said to be η-*heavy *if $D(j) \geq (1+\eta)D^*(j)$.*

Claim 18 *Suppose that $d_{\mathrm{TV}}(D, D^*) \geq \epsilon$ and Equation (46) holds. Suppose moreover that $D(\{1, \ldots, k^*\}) \leq 4\epsilon_1$. Let $i_1, \ldots, i_\ell$ be i.i.d. points drawn from D. Then for $\ell = \Theta(1/\epsilon)$, with probability at least 99/100 (over the i.i.d. draws of $i_1, \ldots, i_\ell \sim D$) there is some point $i_j \in \{i_1, \ldots, i_\ell\}$ such that $i_j > k^*$ and i_j is ϵ_2-heavy.*

Proof: Define H_1 to be the set of all ϵ_2-heavy points and H_2 to be the set of all "slightly lighter" points as follows:

$$H_1 = \{\, i \in [N] \mid D(i) \geq (1+\epsilon_2)D^*(i) \,\}$$
$$H_2 = \{\, i \in [N] \mid (1+\epsilon_2)D^*(i) > D(i) \geq D^*(i) \,\}$$

By definition of the total variation distance, we have

$$\begin{aligned}
\epsilon \leq d_{\mathrm{TV}}(D, D^*) &= \sum_{i: D(i) \geq D^*(i)} (D(i) - D^*(i)) = (D(H_1) - D^*(H_1)) + (D(H_2) - D^*(H_2)) \\
&\leq D(H_1) + ((1+\epsilon_2)D^*(H_2) - D^*(H_2)) \\
&= D(H_1) + \epsilon_2 D^*(H_2) < D(H_1) + \epsilon_2 = D(H_1) + \frac{\epsilon}{2}.
\end{aligned}$$

So it must be the case that $D(H_1) \geq \epsilon/2 = 5\epsilon_1$. Since by assumption we have $D(\{1, \ldots, k^*\}) \leq 4\epsilon_1$, it must be the case that $D(H_1 \setminus \{1, \ldots, k^*\}) \geq \epsilon_1$. The claim follows from the definition of H_1 and the size, ℓ, of the sample. ■

Algorithm 6: COND_D-Test-Known

Input: error parameter $\epsilon > 0$; query access to COND_D oracle; explicit description $(D^*(1), \ldots, D^*(N))$ of distribution D^* satisfying $D^*(1) \leq \cdots \leq D^*(N)$

Let i^* be the minimum index $i \in [N]$ such that $D^*(\{1, \ldots, i\}) > 2\epsilon_1$.
if $D^*(\{1, \ldots, i^* - 1\}) \leq \epsilon_1$ **then**
  Call algorithm COND_D-Test-Known-Heavy$(\epsilon, \mathsf{COND}_D, D^*, i^*)$ (and exit)
else
  Call algorithm COND_D-Test-Known-Main$(\epsilon, \mathsf{COND}_D, D^*, i^* - 1)$ (and exit).
end if

Algorithm 7: COND_D-Test-Known-Heavy

Input: error parameter $\epsilon > 0$; query access to COND_D oracle; explicit description $(D^*(1), \ldots, D^*(N))$ of distribution D^* satisfying $D^*(1) \leq \cdots \leq D^*(N)$; value $i^* \in [N]$ satisfying $D^*(\{1, \ldots, i^* - 1\}) \leq \epsilon_1$, $D^*(\{1, \ldots, i^*\}) > 2\epsilon_1$

1: Call the SAMP_D oracle $m = \Theta((\log(1/\epsilon))/\epsilon^4)$ times. For each $i \in [i^*, N]$ let $\widehat{D}(i)$ be the fraction of the m calls to SAMP_D that returned i. Let $\widehat{D}' = 1 - \sum_{i \in [i^*, N]} \widehat{D}(i)$ be the fraction of the m calls that returned values in $\{1, \ldots, i^* - 1\}$.
2: **if** either (any $i \in [i^*, N]$ has $|\widehat{D}(i) - D^*(i)| > {\epsilon_1}^2$) or $(\widehat{D}' - D^*(\{1, \ldots, i^* - 1\}) > \epsilon_1)$ **then**
3: output REJECT (and exit)
4: **end if**
5: Output ACCEPT

5.3.1 Proof of Theorem 10

It is straightforward to verify that the query complexity of COND_D-Test-Known-Heavy is $\tilde{O}(1/\epsilon^4)$ and the query complexity of COND_D-Test-Known-Main is also $\tilde{O}(1/\epsilon^4)$, so the overall query complexity of COND-Test-Known is as claimed.

By the definition of i^* (in the first line of the algorithm), either Equation (45) holds for this setting of i^*, or Equation (46) holds for $k^* = i^* - 1$. To prove correctness of the algorithm, we first deal with the simpler case, which is that Equation (45) holds:

Lemma 19 *Suppose that D^* is such that $D^*(\{1, \ldots, i^*\}) > 2\epsilon_1$ but $D^*(\{1, \ldots, i^* - 1\}) \leq \epsilon_1$. Then COND_D-Test-Known-Heavy$(\epsilon, \mathsf{COND}_D, D^*, i^*)$ returns ACCEPT with probability at least $2/3$ if $D = D^*$ and returns REJECT with probability at least $2/3$ if $d_{\mathrm{TV}}(D, D^*) \geq \epsilon$.*

Proof: The conditions of Lemma 19, together with the fact that $D^*(\cdot)$ is monotone non-decreasing, imply that each $i \geq i^*$ has $D^*(i) \geq \epsilon_1$. Thus there can be at most $1/\epsilon_1$ many values $i \in \{i^*, \ldots, N\}$, i.e. it must be the case that $i^* \geq N - 1/\epsilon_1 + 1$. Since the expected value of $\widehat{D}(i)$ (defined in Line 1 of COND_D-Test-Known-Heavy) is precisely $D(i)$, for any fixed value of $i \in \{i^*, \ldots, n\}$ an additive Chernoff bound implies that $|D(i) - \widehat{D}(i)| \leq (\epsilon_1)^2$ with failure probability at most $\frac{1}{10\left(1+\frac{1}{\epsilon_1}\right)}$. Similarly $|\widehat{D}' - D(\{1, \ldots, i^* - 1\})| \leq \epsilon_1$ with failure probability at most $\frac{1}{10\left(1+\frac{1}{\epsilon_1}\right)}$. A union bound over all failure events gives that with probability at least $9/10$ each value $i \in \{i^*, \ldots, N\}$

Algorithm 8: COND_D-Test-Known-Main

Input: error parameter $\epsilon > 0$; query access to COND_D oracle; explicit description $(D^*(1), \ldots, D^*(N))$ of distribution D^* satisfying $D^*(1) \leq \cdots \leq D^*(N)$; value $k^* \in [N]$ satisfying $\epsilon_1 < D^*(\{1, \ldots, k^*\}) \leq 2\epsilon_1 < D^*(\{1, \ldots, k^*+1\})$

Call the SAMP_D oracle $\Theta(1/\epsilon^2)$ times and let $\widehat{D}(\{1, \ldots, k^*\})$ denote the fraction of responses that lie in $\{1, \ldots, k^*\}$. If $\widehat{D}(\{1, \ldots, k^*\}) \notin [\frac{\epsilon_1}{2}, \frac{5\epsilon_1}{2}]$ then output REJECT (and exit).

Call the SAMP_D oracle $\ell = \Theta(1/\epsilon)$ times to obtain points $i_1, \ldots, i_\ell$.

for all $j \in \{1, \ldots, \ell\}$ such that $i_j > k^*$ **do**

  Call the SAMP_D oracle $m = \Theta(\log(1/\epsilon)/\epsilon^2)$ times and let $\widehat{D}(\{1, \ldots, i_j\})$ be the fraction of responses that lie in $\{1, \ldots, i_j\}$. If $\widehat{D}(\{1, \ldots, i_j\}) \notin [1-\epsilon_3, 1+\epsilon_3] D^*(\{1, \ldots, i_j\})$ then output REJECT (and exit).

  if $D^*(i_j) \geq \epsilon_1$ **then**

    Run $\textsc{Compare}(\{i_j\}, \{1, \ldots, i_j - 1\}, \frac{\epsilon_2}{16}, \frac{2}{\epsilon_1}, \frac{1}{10\ell})$ and let v denote its output. If $v \notin [1 - \frac{\epsilon_2}{8}, 1 + \frac{\epsilon_2}{8}] \frac{D^*(\{1,\ldots,i_j-1\})}{D^*(\{i_j\})}$ then output REJECT (and exit).

  else

    Let $S_1, \ldots, S_t$ be the partition of $\{1, \ldots, i_j - 1\}$ such that each S_i is an ϵ_1-comparable witness for i_j, which is provided by Claim 17.

    Select a list of $h = \Theta(1/\epsilon)$ elements $S_{a_1}, \ldots, S_{a_h}$ independently and uniformly from $\{S_1, \ldots, S_j\}$.

    For each S_{a_r}, $1 \leq r \leq h$, run $\textsc{Compare}(\{i_j\}, S_{a_r}, \frac{\epsilon_4}{8}, 4, \frac{1}{10\ell h})$ and let v denote its output. If $v \notin [1 - \frac{\epsilon_4}{4}, 1 + \frac{\epsilon_4}{4}] \frac{D^*(S_{a_r})}{D^*(\{i_j\})}$ then output REJECT (and exit).

  end if

end for

Output ACCEPT.

has $|D(i) - \widehat{D}(i)| \leq {\epsilon_1}^2$ and additionally $|\widehat{D}' - D(\{1,\ldots,i^*-1\})| \leq \epsilon_1$; we refer to this compound event as (*).

If $D^* = D$, by (*) the algorithm outputs ACCEPT with probability at least 9/10.

Now suppose that $d_{\mathrm{TV}}(D, D^*) \geq \epsilon$. With probability at least 9/10 we have (*) so we suppose that indeed (*) holds. In this case we have

$$\begin{aligned}
\epsilon \leq d_{\mathrm{TV}}(D, D^*) &= \sum_{i<i^*} |D(i) - D^*(i)| + \sum_{i \geq i^*} |D(i) - D^*(i)| \\
&\leq \sum_{i<i^*} (D(i) + D^*(i)) + \sum_{i \geq i^*} |D(i) - D^*(i)| \\
&\leq D(\{1,\ldots,i^*-1\}) + \epsilon_1 + \sum_{i \geq i^*} \left(|\widehat{D}(i) - D^*(i)| + {\epsilon_1}^2 \right) \\
&\leq \widehat{D}' + \epsilon_1 + 2\epsilon_1 + \sum_{i \geq i^*} \left(|\widehat{D}(i) - D^*(i)| \right)
\end{aligned}$$

where the first inequality is by the triangle inequality, the second is by (*) and the fact that $D^*(\{1,\ldots,i^*-1\}) \leq \epsilon_1$, and the third inequality is by (*) and the fact that there are at most $1/\epsilon_1$ elements in $\{i^*,\ldots,N\}$. Since $\epsilon_1 = \epsilon/10$, the above inequality implies that

$$\frac{7}{10}\epsilon \leq \widehat{D}' + \sum_{i \geq i^*} \left(|\widehat{D}(i) - D^*(i)| \right).$$

If any $i \in \{i^*,\ldots,N\}$ has $|\widehat{D}(i) - D^*(i)| > (\epsilon_1)^2$ then the algorithm outputs REJECT so we may assume that $|\widehat{D}(i) - D^*(i)| \leq {\epsilon_1}^2$ for all i. This implies that

$$6\epsilon_1 = \frac{6}{10}\epsilon \leq \widehat{D}'$$

but since $D^*(\{1,\ldots,i^*-1\}) \leq \epsilon_1$ the algorithm must REJECT. ■

Now we turn to the more difficult (and typical) case, that Equation (46) holds (for $k^* = i^* - 1$), i.e.

$$\epsilon_1 < D^*(\{1,\ldots,k^*\}) \leq 2\epsilon_1 < D^*(\{1,\ldots,k^*+1\}).$$

With the claims we have already established it is straightforward to argue completeness:

Lemma 20 *Suppose that $D = D^*$ and Equation (46) holds. Then with probability at least 2/3 algorithm* COND_D-TEST-KNOWN-MAIN *outputs* ACCEPT.

Proof: We first observe that the expected value of the quantity $\widehat{D}(\{1,\ldots,k^*\})$ defined in Line 1 is precisely $D(\{1,\ldots,k^*\}) = D^*(\{1,\ldots,k^*\})$ and hence lies in $[\epsilon_1, 2\epsilon_1]$ by Equation (46). The additive Chernoff bound implies that the probability the algorithm outputs REJECT in Line 1 is at most 1/10. Thus we may assume the algorithm continues to Line 2.

In any given execution of Line 4, since the expected value of $\widehat{D}(\{1,\ldots,i_j\})$ is precisely $D(\{1,\ldots,i_j\}) = D^*(\{1,\ldots,i_j\}) > \epsilon_1$, a multiplicative Chernoff bound gives that the algorithm

outputs REJECT with probability at most $1/(10\ell)$. Thus the probability that the algorithm outputs REJECT in any execution of Line 4 is at most $1/10$. We henceforth assume that the algorithm never outputs REJECT in this step.

Fix a setting of $j \in \{1,\ldots,\ell\}$ such that $i_j > k^*$. Consider first the case that $D^*(i_j) \geq \epsilon_1$ so the algorithm enters Line 6. By item (1) of Claim 17 and item (1) of Lemma 2, we have that with probability at least $1 - \frac{1}{10\ell}$ COMPARE outputs a value v in the range $[1-\frac{\epsilon_2}{16}, 1+\frac{\epsilon_2}{16}]\frac{D^*(\{1,\ldots,i_j-1\})}{D^*(\{i_j\})}$ (recall that $D = D^*$), so the algorithm does not output REJECT in Line 6. Now suppose that $D^*(i_j) < \epsilon_1$ so the algorithm enters Line 8. Fix a value $1 \leq r \leq h$ in Line 10. By Claim 17 we have that S_{a_r} is a $\frac{1}{2}$-comparable witness for i_j. By item (1) of Lemma 2, we have that with probability at least $1 - \frac{1}{10\ell h}$ COMPARE outputs a value v in the range $[1-\frac{\epsilon_4}{4}, 1+\frac{\epsilon_4}{4}]\frac{D^*(S_{a_r})}{D^*(\{i_j\})}$ (recall that $D = D^*$). A union bound over all h values of r gives that the algorithm outputs REJECT in Line 10 with probability at most $1/(10\ell)$. So in either case, for this setting of j, the algorithm outputs REJECT on that iteration of the outer loop with probability at most $1/(10\ell)$. A union bound over all ℓ iterations of the outer loop gives that the algorithm outputs REJECT at any execution of Line 6 or Line 10 is at most $1/10$.

Thus the overall probability that the algorithm outputs REJECT is at most $3/10$, and the lemma is proved. ■

Next we argue soundness:

Lemma 21 *Suppose that $d_{\mathrm{TV}}(D, D^*) \geq \epsilon$ and Equation (46) holds. Then with probability at least $2/3$ algorithm COND_D-TEST-KNOWN-MAIN outputs* REJECT.

Proof: If $D(\{1,\ldots,k^*\}) \notin [\epsilon_1, 3\epsilon_1]$ then a standard additive Chernoff bound implies that the algorithm outputs REJECT in Line 1 with probability at least $9/10$. Thus we may assume going forward in the argument that $D(\{1,\ldots,k^*\}) \in [\epsilon_1, 3\epsilon_1]$. As a result we may apply Claim 18, and we have that with probability at least $99/100$ there is an element $i_j \in \{i_1,\ldots,i_\ell\}$ such that $i_j > k^*$ and i_j is ϵ_2-heavy, i.e. $D(i_j) \geq (1+\epsilon_2)D^*(i_j)$. We condition on this event going forward (the rest of our analysis will deal with this specific element i_j).

We now consider two cases:

Case 1: Distribution D has $D(\{1,\ldots,i_j\}) \notin [1-3\epsilon_3, 1+3\epsilon_3]D^*(\{1,\ldots,i_j\})$. Since the quantity $\widehat{D}(\{1,\ldots,i_j\})$ obtained in Line 4 has expected value $D(\{1,\ldots,i_j\}) \geq D(\{1,\ldots,k^*\}) \geq \epsilon_1$, applying the multiplicative Chernoff bound implies that $\widehat{D}(\{1,\ldots,i_j\}) \in [1-\epsilon_3, 1+\epsilon_3]D(\{1,\ldots,i_j\})$ except with failure probability at most $\epsilon/10 \leq 1/10$. If this failure event does not occur then since $D(\{1,\ldots,i_j\}) \notin [1-3\epsilon_3, 1+3\epsilon_3]D^*(\{1,\ldots,i_j\})$ it must hold that $\widehat{D}(\{1,\ldots,i_j\}) \notin [1-\epsilon_3, 1+\epsilon_3]D^*(\{1,\ldots,i_j\})$ and consequently the algorithm outputs REJECT. Thus in Case 1 the algorithm outputs REJECT with overall failure probability at least $89/100$.

Case 2: Distribution D has $D(\{1,\ldots,i_j\}) \in [1-3\epsilon_3, 1+3\epsilon_3]D^*(\{1,\ldots,i_j\})$. This case is divided into two sub-cases depending on the value of $D^*(i_j)$.

Case 2(a): $D^*(i_j) \geq \epsilon_1$. In this case the algorithm reaches Line 6. We use the following claim:

Claim 22 *In Case 2(a), suppose that $i_j > k^*$ is such that $D(i_j) \geq (1+\epsilon_2)D^*(i_j)$, and $D(\{1,\ldots,i_j\}) \in [1-3\epsilon_3, 1+3\epsilon_3]D^*(\{1,\ldots,i_j\})$. Then*

$$\frac{D(\{1,\ldots,i_j-1\})}{D(i_j)} \leq \left(1-\frac{\epsilon_2}{4}\right)\cdot\frac{D^*(\{1,\ldots,i_j-1\})}{D^*(i_j)}.$$

Proof: To simplify notation we write

$$a \stackrel{\text{def}}{=} D(i_j); \quad b \stackrel{\text{def}}{=} D^*(i_j); \quad c \stackrel{\text{def}}{=} D(\{1,\ldots,i_j-1\}); \quad d \stackrel{\text{def}}{=} D^*(\{1,\ldots,i_j-1\}).$$

We have that

$$a \geq (1+\epsilon_2)b \quad \text{and} \quad a+c \leq (1+3\epsilon_3)(b+d). \tag{49}$$

This gives

$$c \leq (1+3\epsilon_3)(b+d) - (1+\epsilon_2)b = (1+3\epsilon_3)d + (3\epsilon_3-\epsilon_2)b \ < (1+3\epsilon_3)d\,, \tag{50}$$

where in the last inequality we used $\epsilon_2 > 3\epsilon_3$. Recalling that $a \geq (1+\epsilon_2)b$ and using $\epsilon_3 = \epsilon_2/24$ we get

$$\frac{c}{a} < \frac{(1+3\epsilon_3)d}{(1+\epsilon_2)b} = \frac{d}{b}\cdot\frac{1+\epsilon_2/8}{1+\epsilon_2} < \frac{d}{b}\cdot\left(1-\frac{\epsilon_2}{4}\right)\,. \tag{51}$$

This proves the claim. ■

Applying Claim 22, we get that in Line 6 we have

$$\frac{D(\{1,\ldots,i_j-1\})}{D(i_j)} \leq \left(1-\frac{\epsilon_2}{4}\right)\cdot\frac{D^*(\{1,\ldots,i_j-1\})}{D^*(i_j)}\,. \tag{52}$$

Recalling that by the premise of this case $D^*(i_j) \geq \epsilon_1$, by applying Claim 17 we have that $\{1,\ldots,i_j-1\}$ is an ϵ_1-comparable witness for i_j. Therefore, by Lemma 2, with probability at least $1-\frac{1}{10\ell}$ the call to $\textsc{Compare}(\{i_j\},\{1,\ldots,i_j-1\},\frac{\epsilon_2}{16},\frac{2}{\epsilon_1},\frac{1}{10\ell})$ in Line 6 either outputs an element of $\{\mathsf{High},\mathsf{Low}\}$ or outputs a value $v \leq (1-\frac{\epsilon_2}{4})(1+\frac{\epsilon_2}{16})\frac{D^*(\{1,\ldots,i_j-1\})}{D^*(i_j)} < (1-\frac{\epsilon_2}{8})\frac{D^*(\{1,\ldots,i_j-1\})}{D^*(i_j)}$. In either case the algorithm outputs REJECT in Line 6, so we are done with Case 2(a).

Case 2(b): $D^*(i_j) < \epsilon_1$. In this case the algorithm reaches Line 10, and by item 2 of Claim 17, we have that $S_1,\ldots,S_t$ is a partition of $\{1,\ldots,i_j-1\}$ and each set $S_1,\ldots,S_t$ is a $\frac{1}{2}$-comparable witness for i_j, i.e.,

$$\text{for all } i \in \{1,\ldots,t\}, \quad \frac{1}{2}D^*(i_j) \leq D^*(S_i) \leq 2D^*(i_j). \tag{53}$$

We use the following claim:

Claim 23 *In Case 2(b) suppose $i_j > k^*$ is such that $D(i_j) \geq (1+\epsilon_2)D^*(i_j)$ and $D(\{1,\ldots,i_j\}) \in [1-3\epsilon_3, 1+3\epsilon_3]D^*(\{1,\ldots,i_j\})$. Then at least $(\epsilon_4/8)$-fraction of the sets $S_1,\ldots,S_t$ are such that*

$$D(S_i) \leq (1+\epsilon_4)D^*(S_i).$$

Proof: The proof is by contradiction. Let $\rho = 1 - \epsilon_4/8$ and suppose that there are w sets (without loss of generality we call them $S_1, \ldots, S_w$) that satisfy $D(S_i) > (1+\epsilon_4)D^*(S_i)$, where $\rho' = \frac{w}{t} > \rho$. We first observe that the weight of the w subsets $S_1, \ldots, S_w$ under D^*, as a fraction of $D^*(\{1, \ldots, i_j - 1\})$, is at least

$$\frac{D^*(S_1 \cup \cdots \cup S_w)}{D^*(S_1 \cup \cdots \cup S_w) + (t-w) \cdot 2D^*(i_j)} \geq \frac{w\frac{D^*(i_j)}{2}}{w\frac{D^*(i_j)}{2} + (t-w) \cdot 2D^*(i_j)} = \frac{w}{4t - 3w} = \frac{\rho'}{4 - 3\rho'},$$

where we used the right inequality in Equation (53) on $S_{w+1}, \ldots, S_t$ to obtain the leftmost expression above, and the left inequality in Equation (53) (together with the fact that $\frac{x}{x+c}$ is an increasing function of x for all $c > 0$) to obtain the inequality above. This implies that

$$\begin{aligned} D(\{1, \ldots, i_j - 1\}) &= \sum_{i=1}^{w} D(S_i) + \sum_{i=w+1}^{t} D(S_i) \geq (1+\epsilon_4)\sum_{i=1}^{w} D^*(S_i) + \sum_{i=w+1}^{t} D(S_i) \\ &\geq (1+\epsilon_4)\frac{\rho'}{4-3\rho'}D^*(\{1, \ldots, i_j - 1\}) \\ &\geq (1+\epsilon_4)\frac{\rho}{4-3\rho}D^*(\{1, \ldots, i_j - 1\}) \,. \end{aligned} \tag{54}$$

From Equation (54) we have

$$\begin{aligned} D(\{1, \ldots, i_j\}) &\geq (1+\epsilon_4)\frac{\rho}{4-3\rho}D^*(\{1, \ldots, i_j - 1\}) + (1+\epsilon_2)D^*(i_j) \\ &\geq \left(1 + \frac{3\epsilon_4}{8}\right) D^*(\{1, \ldots, i_j - 1\}) + (1+\epsilon_2)D^*(i_j) \end{aligned}$$

where for the first inequality above we used $D(i_j) \geq (1+\epsilon_2)D^*(i_j)$ and for the second inequality we used $(1+\epsilon_4)\frac{\rho}{4-3\rho} \geq 1 + \frac{3\epsilon_4}{8}$. This implies that

$$D(\{1, \ldots, i_j\}) > \left(1 + \frac{3\epsilon_4}{8}\right) D^*(\{1, \ldots, i_j - 1\}) + \left(1 + \frac{3\epsilon_4}{8}\right) D^*(i_j) = \left(1 + \frac{3\epsilon_4}{8}\right) D^*(\{1, \ldots, i_j\})$$

where the inequality follows from $\epsilon_2 > \frac{3\epsilon_4}{8}$. Since $\frac{3\epsilon_4}{8} = 3\epsilon_3$, though, this is a contradiction and the claim is proved. ■

Applying Claim 23, and recalling that $h = \Theta(1/\epsilon) = \Theta(1/\epsilon_4)$ sets are chosen randomly in Line 9, we have that with probability at least $9/10$ there is some $r \in \{1, \ldots, h\}$ such that $D(S_{a_r}) \leq (1+\epsilon_4)D^*(S_{a_r})$. Combining this with $D(i_j) \geq (1+\epsilon_2)D^*(i_j)$, we get that

$$\frac{D(S_{a_r})}{D(i_j)} \leq \frac{1+\epsilon_4}{1+\epsilon_2} \cdot \frac{D^*(S_{a_r})}{D^*(i_j)} \leq \left(1 - \frac{\epsilon_4}{2}\right) \cdot \frac{D^*(S_{a_r})}{D^*(i_j)}.$$

By Lemma 2, with probability at least $1 - \frac{1}{10\ell h}$ the call to $\textsc{Compare}(\{i_j\}, S_{a_r}, \frac{\epsilon_4}{8}, 4, \frac{1}{10\ell n})$ in Line 10 either outputs an element of $\{\mathsf{High}, \mathsf{Low}\}$ or outputs a value $v \leq (1 - \frac{\epsilon_4}{2})(1 + \frac{\epsilon_4}{8})\frac{D^*(S_{a_r})}{D^*(i_j)} < (1 - \frac{\epsilon_4}{4})\frac{D^*(S_{a_r})}{D^*(i_j)}$. In either case the algorithm outputs REJECT in Line 10, so we are done in Case 2(b). This concludes the proof of soundness and the proof of Theorem 7. ■

6 Testing equality between two unknown distributions

6.1 An approach based on PCOND queries

In this subsection we consider the problem of testing whether two unknown distributions D_1, D_2 are identical versus ϵ-far, given PCOND access to these distributions. Although this is known to require $\Omega(N^{2/3})$ many samples in the standard model [BFR+10, Val11], we are able to give a $\mathrm{poly}(\log N, 1/\epsilon)$-query algorithm using PCOND queries, by taking advantage of comparisons to perform some sort of *clustering* of the domain.

On a high level the algorithm works as follows. First it obtains (with high probability) a small set of points R such that almost every element in $[N]$, except possibly for some negligible subset according to D_1, has probability weight (under D_1) close to some "representative" in R. Next, for each representative r in R it obtains an estimate of the weight, according to D_1, of a set of points $U(r)$ such that $D_1(u)$ is close to $D_1(r)$ for each u in $U(r)$ (i.e., r's "neighborhood under D_1"). This is done using the procedure Estimate-Neighborhood from Subsection 3.2. Note that these neighborhoods can be interpreted roughly as a succinct *cover* of the support of D_1 into (not necessarily disjoint) sets of points, where within each set the points have similar weight (according to D_1). Our algorithm is based on the observation that, if D_1 and D_2 are far from each other, it must be the case that one of these sets, denoted $U(r^*)$, reflects it in one of the following ways: (1) $D_2(U(r^*))$ differs significantly from $D_1(U(r^*))$; (2) $U(r^*)$ contains a subset of points $V(r^*)$ such that $D_2(v)$ differs significantly from $D_2(r^*)$ for each v in $V(r^*)$, and either $D_1(V(r^*))$ is relatively large or $D_2(V(r^*))$ is relatively large. (This structural result is made precise in Lemma 25). We thus take additional samples, both from D_1 and from D_2, and compare the weight (according to both distributions) of each point in these samples to the representatives in R (using the procedure Compare from Subsection 3.1). In this manner we detect (with high probability) that either (1) or (2) holds.

We begin by formalizing the notion of a cover discussed above:

Definition 6 (Weight-Cover) *Given a distribution D on $[N]$ and a parameter $\epsilon_1 > 0$, we say that a point $i \in [N]$ is* ϵ_1-covered *by a set $R = \{r_1, \ldots, r_t\} \subseteq [N]$ if there exists a point $r_j \in R$ such that $D(i) \in [1/(1+\epsilon_1), 1+\epsilon_1]D(r_j)$. Let the set of points in $[N]$ that are ϵ_1-covered by R be denoted by $U^D_{\epsilon_1}(R)$. We say that R is an* (ϵ_1, ϵ_2)-cover for D *if $D([N] \setminus U^D_{\epsilon_1}(R)) \leq \epsilon_2$.*

For a singleton set $R = \{r\}$ we slightly abuse notation and write $U^D_\epsilon(r)$ to denote $U^D_\epsilon(R)$; note that this aligns with the notation established in (11).

The following lemma says that a small sample of points drawn from D gives a cover with high probability:

Lemma 24 *Let D be any distribution over $[N]$. Given any fixed $c > 0$, there exists a constant $c' > 0$ such that with probability at least $99/100$, a sample R of size $m = c' \frac{\log(N/\epsilon)}{\epsilon^2} \cdot \log\left(\frac{\log(N/\epsilon)}{\epsilon}\right)$ drawn according to distribution D is an $(\epsilon/c, \epsilon/c)$-cover for D.*

Proof: Let t denote $\lceil \ln(2cN/\epsilon) \cdot \frac{c}{\epsilon} \rceil$. We define t "buckets" of points with similar weight under D as follows: for $i = 0, 1, \ldots, t-1$, define $B_i \subseteq [N]$ to be

$$B_i \stackrel{\text{def}}{=} \left\{ x \in [N] : \frac{1}{(1+\epsilon/c)^{i+1}} < D(x) \leq \frac{1}{(1+\epsilon/c)^i} \right\}.$$

Let L be the set of points x which are not in any of $B_0, \ldots, B_{t-1}$ (because $D(x)$ is too small); since every point in L has $D(x) < \frac{\epsilon}{2cN}$, one can see that $D(L) \leq \frac{\epsilon}{2c}$.

It is easy to see that if the sample R contains a point from a bucket B_j then every point $y \in B_j$ is $\frac{\epsilon}{c}$-covered by R. We say that bucket B_i is *insignificant* if $D(B_i) \leq \frac{\epsilon}{2ct}$; otherwise bucket B_i is *significant*. It is clear that the total weight under D of all insignificant buckets is at most $\epsilon/2c$. Thus if we can show that for the claimed sample size, with probability at least 99/100 every significant bucket has at least one of its points in R, we will have established the lemma.

This is a simple probabilistic calculation: fix any significant bucket B_j. The probability that m random draws from D all miss B_j is at most $(1 - \frac{\epsilon}{2ct})^m$, which is at most $\frac{1}{100t}$ for a suitable (absolute constant) choice of c'. Thus a union bound over all (at most t) significant buckets gives that with probability at least 99/100, no significant bucket is missed by R. ■

Lemma 25 *Suppose $d_{\mathrm{TV}}(D_1, D_2) \geq \epsilon$, and let $R = \{r_1, \ldots, r_t\}$ be an $(\tilde{\epsilon}, \tilde{\epsilon})$-cover for D_1 where $\tilde{\epsilon} \leq \epsilon/100$. Then, there exists $j \in [t]$ such that at least one of the following conditions holds for every $\alpha \in [\tilde{\epsilon}, 2\tilde{\epsilon}]$:*

1. *$D_1(U_\alpha^{D_1}(r_j)) \geq \frac{\tilde{\epsilon}}{t}$ and $D_2(U_\alpha^{D_1}(r_j)) \notin [1-\tilde{\epsilon}, 1+\tilde{\epsilon}] D_1(U_\alpha^{D_1}(r_j))$, or $D_1(U_\alpha^{D_1}(r_j)) < \frac{\tilde{\epsilon}}{t}$ and $D_2(U_\alpha^{D_1}(r_j)) > \frac{2\tilde{\epsilon}}{t}$;*
2. *$D_1(U_\alpha^{D_1}(r_j)) \geq \frac{\tilde{\epsilon}}{t}$, and at least a $\tilde{\epsilon}$-fraction of the points i in $U_\alpha^{D_1}(r_j)$ satisfy $\frac{D_2(i)}{D_2(r_j)} \notin [1/(1+\alpha+\tilde{\epsilon}), 1+\alpha+\tilde{\epsilon}]$;*
3. *$D_1(U_\alpha^{D_1}(r_j)) \geq \frac{\tilde{\epsilon}}{t}$, and the total weight according to D_2 of the points i in $U_\alpha^{D_1}(r_j)$ for which $\frac{D_2(i)}{D_2(r_j)} \notin [1/(1+\alpha+\tilde{\epsilon}), 1+\alpha+\tilde{\epsilon}]$ is at least $\frac{\tilde{\epsilon}^2}{t}$;*

Proof: Without loss of generality, we can assume that $\epsilon \leq 1/4$. Suppose, contrary to the claim, that for each r_j there exists $\alpha_j \in [\tilde{\epsilon}, 2\tilde{\epsilon}]$ such that if we let $U_j \stackrel{\text{def}}{=} U_{\alpha_j}^{D_1}(r_j)$, then the following holds:

1. If $D_1(U_j) < \frac{\tilde{\epsilon}}{t}$, then $D_2(U_j) \leq \frac{2\tilde{\epsilon}}{t}$;
2. If $D_1(U_j) \geq \frac{\tilde{\epsilon}}{t}$, then:
   (a) $D_2(U_j) \in [1-\tilde{\epsilon}, 1+\tilde{\epsilon}] D_1(U_j)$;
   (b) Less than an $\tilde{\epsilon}$-fraction of the points y in U_j satisfy $\frac{D_2(y)}{D_2(r_j)} \notin [1/(1+\alpha_j+\tilde{\epsilon}), 1+\alpha_j+\tilde{\epsilon}]$;
   (c) The total weight according to D_2 of the points y in U_j for which $\frac{D_2(y)}{D_2(r_j)} \notin [1/(1+\alpha_j+\tilde{\epsilon}), 1+\alpha_j+\tilde{\epsilon}]$ is at most $\frac{\tilde{\epsilon}^2}{t}$;

We show that in such a case $d_{\mathrm{TV}}(D_1, D_2) < \epsilon$, contrary to the premise of the claim.

Consider each point $r_j \in R$ such that $D_1(U_j) \geq \frac{\tilde{\epsilon}}{t}$. By the foregoing discussion (point 2(a)), $D_2(U_j) \in [1-\tilde{\epsilon}, 1+\tilde{\epsilon}]D_1(U_j)$. By the definition of U_j (and since $\alpha_j \leq 2\tilde{\epsilon}$),

$$D_1(r_j) \in [1/(1+2\tilde{\epsilon}), 1+2\tilde{\epsilon}]\,\frac{D_1(U_j)}{|U_j|}\ . \tag{55}$$

Turning to bound $D_2(r_j)$, on one hand (by 2(b))

$$D_2(U_j) = \sum_{y \in U_j} D_2(y) \geq \tilde{\epsilon}|U_j| \cdot 0 + (1-\tilde{\epsilon})|U_j| \cdot \frac{D_2(r_j)}{1+3\tilde{\epsilon}}\ , \tag{56}$$

and so

$$D_2(r_j) \leq \frac{(1+3\tilde{\epsilon})D_2(U_j)}{(1-\tilde{\epsilon})\,|U_j|} \leq (1+6\tilde{\epsilon})\frac{D_1(U_j)}{|U_j|}\ . \tag{57}$$

On the other hand (by 2(c)),

$$D_2(U_j) = \sum_{y \in U_j} D_2(y) \leq \frac{\tilde{\epsilon}^2}{t} + |U_j| \cdot (1+3\tilde{\epsilon})D_2(r_j)\ , \tag{58}$$

and so

$$D_2(r_j) \geq \frac{D_2(U_j) - \tilde{\epsilon}^2/t}{(1+3\tilde{\epsilon})\,|U_j|} \geq \frac{(1-\tilde{\epsilon})D_1(U_j) - \tilde{\epsilon}D_1(U_j)}{(1+3\tilde{\epsilon})\,|U_j|} \geq (1-5\tilde{\epsilon})\frac{D_1(U_j)}{|U_j|}\ . \tag{59}$$

Therefore, for each such r_j we have

$$D_2(r_j) \in [1-8\tilde{\epsilon}, 1+10\tilde{\epsilon}]D_1(r_j)\ . \tag{60}$$

Let $C \stackrel{\text{def}}{=} \bigcup_{j=1}^{t} U_j$. We next partition the points in C so that each point $i \in C$ is assigned to some $r_{j(i)}$ such that $i \in U_{j(i)}$. We define the following "bad" subsets of points in $[N]$:

1. $B_1 \stackrel{\text{def}}{=} [N] \setminus C$, so that $D_1(B_1) \leq \tilde{\epsilon}$ (we later bound $D_2(B_1)$);
2. $B_2 \stackrel{\text{def}}{=} \left\{i \in C : D_1(U_{j(i)}) < \tilde{\epsilon}/t\right\}$, so that $D_1(B_2) \leq \tilde{\epsilon}$ and $D_2(B_2) \leq 2\tilde{\epsilon}$;
3. $B_3 \stackrel{\text{def}}{=} \left\{i \in C \setminus B_2 : D_2(i) \notin [1/(1+3\tilde{\epsilon}), 1+3\tilde{\epsilon}]D_2(r_{j(i)})\right\}$, so that $D_1(B_3) \leq 2\tilde{\epsilon}$ and $D_2(B_3) \leq \tilde{\epsilon}^2$.

Let $B \stackrel{\text{def}}{=} B_1 \cup B_2 \cup B_3$. Observe that for each $i \in [N] \setminus B$ we have that

$$D_2(i) \in [1/(1+3\tilde{\epsilon}), 1+3\tilde{\epsilon}]D_2(r_{j(i)}) \subset [1-15\tilde{\epsilon}, 1+15\tilde{\epsilon}]D_1(r_{j(i)}) \subset [1-23\tilde{\epsilon}, 1+23\tilde{\epsilon}]D_1(i)\ , \tag{61}$$

where the first containment follows from the fact that $i \notin B$, the second follows from Equation (60), and the third from the fact that $i \in U_{j(i)}$. In order to complete the proof we need a bound on $D_2(B_1)$, which we obtain next.

$$\begin{aligned} D_2(B_1) &= 1 - D_2([N] \setminus B_1) \leq 1 - D_2([N] \setminus B) \leq 1 - (1-23\tilde{\epsilon})D_1([N] \setminus B) \\ &\leq 1 - (1-23\tilde{\epsilon})(1-4\tilde{\epsilon}) \leq 27\tilde{\epsilon}\ . \end{aligned} \tag{62}$$

Therefore,

$$\begin{aligned} d_{\mathrm{TV}}(D_1, D_2) &= \frac{1}{2}\sum_{i=1}^{N} |D_1(i) - D_2(i)| \\ &\leq \frac{1}{2}\Big(D_1(B) + D_2(B) + \sum_{i \notin B} 23\tilde{\epsilon} D_1(i)\Big) \quad < \quad \epsilon\,, \end{aligned} \tag{63}$$

and we have reached a contradiction. ■

Algorithm 9: Algorithm PCOND_{D_1,D_2}-Test-Equality-Unknown

Input: PCOND query access to distributions D_1 and D_2 and a parameter ϵ.

1. Set $\tilde{\epsilon} = \epsilon/100$.
2. Draw a sample R of size $t = \tilde{\Theta}\Big(\frac{\log N}{\epsilon^2}\Big)$ from D_1.
3. For each $r_j \in R$:
   (a) Call Estimate-Neighborhood$_{D_1}$ on r_j with $\kappa = \tilde{\epsilon}$, $\eta = \frac{\tilde{\epsilon}}{8}$, $\beta = \frac{\tilde{\epsilon}}{2t}$, $\delta = \frac{1}{100t}$ and let the output be denoted by $(\hat{w}_j^{(1)}, \alpha_j)$.
   (b) Set $\theta = \kappa\eta\beta\delta/64 = \tilde{\Theta}(\epsilon^7/\log^2 N)$.
   (c) Draw a sample S_1 from D_1, of size $s_1 = \Theta\big(\frac{t}{\epsilon^2}\big) = \tilde{\Theta}\Big(\frac{\log N}{\epsilon^4}\Big)$.
   (d) Draw a sample S_2 from D_2, of size $s_2 = \Theta\Big(\frac{t\log t}{\epsilon^3}\Big) = \tilde{\Theta}\Big(\frac{\log N}{\epsilon^5}\Big)$.
   (e) For each point $i \in S_1 \cup S_2$ call Compare$_{D_1}(\{r_j\}, \{i\}, \theta/4, 4, 1/(200t(s_1+s_2)))$ and Compare$_{D_2}(\{r_j\}, \{i\}, \theta/4, 4, 1/(200t(s_1+s_2)))$, and let the outputs be denoted $\rho_{r_j}^{(1)}(i)$ and $\rho_{r_j}^{(2)}(i)$, respectively (where in particular these outputs may be High or Low).
   (f) Let $\hat{w}_j^{(2)}$ be the fraction of occurrences of $i \in S_2$ such that $\rho_{r_j}^{(1)}(i) \in [1/(1+\alpha_j+\theta/2), 1+\alpha_j+\theta/2]$.
   (g) If ($\hat{w}_j^{(1)} \leq \frac{3}{4}\frac{\tilde{\epsilon}}{t}$ and $\hat{w}_j^{(2)} > \frac{3}{2}\frac{\tilde{\epsilon}}{t}$) or ($\hat{w}_j^{(1)} > \frac{3}{4}\frac{\tilde{\epsilon}}{t}$ and $\hat{w}_j^{(2)}/\hat{w}_j^{(1)} \notin [1-\tilde{\epsilon}/2, 1+\tilde{\epsilon}/2]$), then output REJECT.
   (h) If there exists $i \in S_1 \cup S_2$ such that $\rho_{r_j}^{(1)}(i) \in [1/(\alpha_j+\tilde{\epsilon}/2), 1+\alpha_j+\tilde{\epsilon}/2]$ and $\rho_{r_j}^{(2)}(i) \notin [1/(\alpha_j+3\tilde{\epsilon}/2), 1+\alpha_j+3\tilde{\epsilon}/2]$, then output REJECT.
4. Output ACCEPT.

Theorem 11 *If $D_1 = D_2$ then with probability at least 2/3 Algorithm* PCOND-Test-Equality-Unknown *returns* ACCEPT*, and if $d_{\mathrm{TV}}(D_1, D_2) \geq \epsilon$, then with probability at least 2/3 Algorithm* PCOND-Test-Equality-Unknown *returns* REJECT*. The number of* PCOND *queries performed by the algorithm is $\tilde{O}\big(\frac{\log^6 N}{\epsilon^{21}}\big)$.*

Proof: The number of queries performed by the algorithm is the sum of: (1) t times the number of queries performed in each execution of ESTIMATE-NEIGHBORHOOD (in Line 3-a) and (2) $t\cdot(s_1+s_2) = O(t\cdot s_2)$ times the number of queries performed in each execution of COMPARE (in Line 3-e). By Lemma 3 (and the settings of the parameters in the calls to ESTIMATE-NEIGHBORHOOD), the first term is $O\left(t\cdot\frac{\log(1/\delta)\cdot\log(\log(1/\delta)/(\beta\eta^2))}{\kappa^2\eta^4\beta^3\delta^2}\right) = \tilde{O}\left(\frac{\log^6 N}{\epsilon^{19}}\right)$, and by Lemma 2 (and the settings of the parameters in the calls to COMPARE), the second term is $O\left(t\cdot s_2\cdot\frac{\log(t\cdot s_2)}{\theta^2}\right) = \tilde{O}\left(\frac{\log^6 N}{\epsilon^{21}}\right)$, so that we get the bound stated in the theorem.

We now turn to establishing the correctness of the algorithm. We shall use the shorthand U_j for $U^{D_1}_{\alpha_j}(r_j)$, and U'_j for $U^{D_1}_{\alpha_j+\theta}(r_j)$. We consider the following "desirable" events.

1. The event E_1 is that the sample R is a $(\tilde{\epsilon},\tilde{\epsilon})$-weight-cover for D_1 (for $\tilde{\epsilon}=\epsilon/100$). By Lemma 24 (and an appropriate constant in the $\Theta(\cdot)$ notation for the size of R), the probability that E_1 holds is at least 99/100.

2. The event E_2 is that all calls to the procedure ESTIMATE-NEIGHBORHOOD are as specified by Lemma 3. By the setting of the confidence parameter in the calls to the procedure, the event E_2 holds with probability at least 99/100.

3. The event E_3 is that all calls to the procedure COMPARE are as specified by Lemma 2. By the setting of the confidence parameter in the calls to the procedure, the event E_3 holds with probability at least 99/100.

4. The event E_4 is that $D_2(U'_j \setminus U_j) \leq \eta\beta/16 = \tilde{\epsilon}^2/(256t)$ for each j. If $D_2 = D_1$ then this event follows from E_2. Otherwise, it holds with probability at least 99/100 by the setting of θ and the choice of α_j (as shown in the proof of Lemma 3 in the analysis of the event E_1 there.

5. The event E_5 is defined as follows. For each j, if $D_2(U_j) \geq \tilde{\epsilon}/(4t)$, then $|S_2\cap U_j|/|S_2| \in [1-\tilde{\epsilon}/10, 1+\tilde{\epsilon}/10]D_2(U_j)$, and if $D_2(U_j) < \tilde{\epsilon}/(4t)$ then $|S_2\cap U_j|/|S_2| < (1+\tilde{\epsilon}/10)\tilde{\epsilon}/(4t)$. This event holds with probability at least 99/100 by applying a multiplicative Chernoff bound in the first case, and Corollary 2 in the second.

6. The event E_6 is that for each j we have $|S_2\cap(U'_j\setminus U_j)|/|S_2| \leq \tilde{\epsilon}^2/(128t)$. Conditioned on E_4, the event E_6 holds with probability at least 99/100 by applying Corollary 2.

From this point on we assume that events $E_1 - E_6$ all hold. Note that in particular this implies the following:

1. By E_2, for every j:
   - If $D_1(U_j) \geq \beta = \tilde{\epsilon}/(2t)$, then $\hat{w}^{(1)}_j \in [1-\eta, 1+\eta]D_1(U_j) = [1-\tilde{\epsilon}/8, 1+\tilde{\epsilon}/8]D_1(U_j)$.
   - If $D_1(U_j) < \tilde{\epsilon}/(2t)$, then $\hat{w}^{(1)}_j \leq (1+\tilde{\epsilon}/8)(\tilde{\epsilon}/(2t))$.
2. By E_3, for every j and for each point $i \in S_1\cup S_2$:
   - If $i \in U_j$, then $\rho^{(1)}_{r_j}(i) \in [1/(1+\alpha_j+\frac{\theta}{2}), 1+\alpha_j+\frac{\theta}{2}]$.

- If $i \notin U'_j$, then $\rho^{(1)}_{r_j}(i) \notin [1/(1+\alpha_j+\frac{\theta}{2}), 1+\alpha_j+\frac{\theta}{2}]$.

3. By the previous item and E_4–E_6:
   - If $D_2(U_j) \geq \tilde{\epsilon}/(4t)$, then $\hat{w}^{(2)}_j \geq (1-\tilde{\epsilon}/10)D_2(U_j)$ and $\hat{w}^{(2)}_j \leq (1+\tilde{\epsilon}/10)D_2(U_j) + \tilde{\epsilon}^2/(128t) \leq (1+\tilde{\epsilon}/8)D_2(U_j)$.
   - If $D_2(U_j) < \tilde{\epsilon}/(4t)$ then $\hat{w}^{(2)}_j \leq (1+\tilde{\epsilon}/10)\tilde{\epsilon}/(4t) + \tilde{\epsilon}^2/(128t) \leq (1+\tilde{\epsilon}/4)(\tilde{\epsilon}/(4t))$.

Completeness. Assume D_1 and D_2 are the same distribution D. For each j, if $D(U_j) \geq \tilde{\epsilon}/t$, then by the foregoing discussion, $\hat{w}^{(1)}_j \geq (1-\tilde{\epsilon}/8)D(U_j) > 3\tilde{\epsilon}/(4t)$ and $\hat{w}^{(2)}_j/\hat{w}^{(1)}_j \in [(1-\tilde{\epsilon}/8)^2, (1+\tilde{\epsilon}/8)^2] \subset [1-\tilde{\epsilon}/2, 1+\tilde{\epsilon}/2]$, so that the algorithm does not reject in Line 3-g. Otherwise (i.e., $D(U_j) < \tilde{\epsilon}/t$), we consider two subcases. Either $D(U_j) \leq \tilde{\epsilon}/(2t)$, in which case $\hat{w}^{(1)}_j \leq 3\tilde{\epsilon}/(4t)$, or $\tilde{\epsilon}/(2t) < D(U_j) < \tilde{\epsilon}/t$, and then $\hat{w}^{(1)}_j \in [1-\tilde{\epsilon}/8, 1+\tilde{\epsilon}/8]D_1(U_j)$. Since in both cases $\hat{w}^{(2)}_j \leq (1+\tilde{\epsilon}/8)D(U_j) \leq 3\tilde{\epsilon}/(2t)$, the algorithm does not reject in Line 3-g. By E_3, the algorithm does not reject in Line 3-h either. We next turn to establish soundness.

Soundness. Assume $d_{\mathrm{TV}}(D_1, D_2) \geq \epsilon$. By applying Lemma 25 on R (and using E_1), there exists an index j for which one of the items in the lemma holds. We denote this index by j^*, and consider the three items in the lemma.

1. If Item 1 holds, then we consider its two cases:
   (a) In the first case, $D_1(U_{j^*}) \geq \tilde{\epsilon}/t$ and $D_2(U_{j^*}) \notin [1-\tilde{\epsilon}, 1+\tilde{\epsilon}]D_1(U_{j^*})$. Due to the lower bound on $D_1(U_{j^*})$ we have that $\hat{w}^{(1)}_{j^*} \in [1-\tilde{\epsilon}/8, 1+\tilde{\epsilon}/8]D_1(U_{j^*})$, so that in particular $\hat{w}^{(1)}_{j^*} > 3\tilde{\epsilon}/(4t)$. As for $\hat{w}^{(2)}_{j^*}$, either $\hat{w}^{(2)}_{j^*} < (1-\tilde{\epsilon})(1+\tilde{\epsilon}/8)D_1(U_{j^*})$ (this holds both when $D_2(U_{j^*}) \geq \tilde{\epsilon}/(4t)$ and when $D_2(U_{j^*}) < \tilde{\epsilon}/(4t)$) or $\hat{w}^{(2)}_{j^*} > (1+\tilde{\epsilon})(1-\tilde{\epsilon}/10)D_1(U_{j^*})$. In either (sub)case $\hat{w}^{(2)}_{j^*}/\hat{w}^{(1)}_{j^*} \notin [1-\tilde{\epsilon}/2, 1+\tilde{\epsilon}/2]$, causing the algorithm to reject in (the second part of) Line 3-g.
   (b) In the second case, $D_1(U_{j^*}) < \tilde{\epsilon}/t$ and $D_2(U_{j^*}) > 2\tilde{\epsilon}/t$. Due to the lower bound on $D_2(U_{j^*})$ we have that $\hat{w}^{(2)}_{j^*} \geq (1-\tilde{\epsilon}/10)D_2(U_{j^*}) > (1-\tilde{\epsilon}/10)(2\tilde{\epsilon}/t)$, so that in particular $\hat{w}^{(2)}_{j^*} > (3\tilde{\epsilon}/(2t))$. As for $\hat{w}^{(1)}_{j^*}$, if $D_1(U_{j^*}) \leq \tilde{\epsilon}/(2t)$, then $\hat{w}^{(1)}_{j^*} \leq 3\tilde{\epsilon}/(4t)$, causing the algorithm to reject in (the first part of) Line 3-g. If $\tilde{\epsilon}/(2t) < D_1(U_{j^*}) \leq \tilde{\epsilon}/t$, then $\hat{w}^{(1)}_{j^*} \in [1-\tilde{\epsilon}/8, 1+\tilde{\epsilon}/8]D_1(U_{j^*}) \leq (1+\tilde{\epsilon}/8)(\tilde{\epsilon}/t)$, so that $\hat{w}^{(2)}_{j^*}/\hat{w}^{(1)}_{j^*} \geq \frac{(1-\tilde{\epsilon}/10)(2\tilde{\epsilon}/t)}{(1+\tilde{\epsilon}/8)\tilde{\epsilon}/t} > (1+\tilde{\epsilon}/2)$, causing the algorithm to reject in (either the first or second part of) Line 3-g.
2. If Item 2 holds, then, by the choice of the size of S_1, which is $\Theta(t/\tilde{\epsilon}^2)$, and since all points in U_{j^*} have approximately the same weight according to D_1, with probability at least 99/100, the sample S_1 will contain a point i for which $\frac{D_2(i)}{D_2(r_{j^*})} \notin [1/(1+\alpha_{j^*}+\tilde{\epsilon}), 1+\alpha_{j^*}+\tilde{\epsilon}]$, and by E_3 this will be detected in Line 3-h.
3. Similarly, if Item 3 holds, then by the choice of the size of S_2, with probability at least 99/100, the sample S_2 will contain a point i for which $\frac{D_2(i)}{D_2(r_{j^*})} \notin [1/(1+\alpha_{j^*}+\tilde{\epsilon}), 1+\alpha_{j^*}+\tilde{\epsilon}]$, and by E_3 this will be detected in Line 3-h.

The theorem is thus established. ■

6.2 An approach based on simulating EVAL

In this subsection we present an alternate approach for testing whether two unknown distributions D_1, D_2 are identical versus ϵ-far. We prove the following theorem:

Theorem 12 COND-TEST-EQUALITY-UNKNOWN *is a*

$$\tilde{O}\left(\frac{(\log N)^5}{\epsilon^4}\right)$$

-query algorithm with the following properties: given COND_{D_1}, COND_{D_2} *oracles for any two distributions* D_1, D_2 *over* $[N]$*, it outputs* ACCEPT *with probability at least* $2/3$ *if* $D_1 = D_2$ *and outputs* REJECT *with probability at least* $2/3$ *if* $d_{\mathrm{TV}}(D_1, D_2) \geq \epsilon$.

At the heart of this result is our efficient simulation of an "approximate EVAL_D oracle" using a COND_D oracle. (Recall that an EVAL_D oracle is an oracle which, given as input an element $i \in [N]$, outputs the numerical value $D(i)$.) We feel that this efficient simulation of an approximate EVAL oracle using a COND oracle is of independent interest since it sheds light on the relative power of the COND and EVAL models.

In more detail, the starting point of our approach to prove Theorem 12 is a simple algorithm from [RS09] that uses an EVAL_D oracle to test equality between D and a known distribution D^*. We first show (see Theorem 13) that a modified version of the algorithm, which uses a SAMP oracle and an "approximate" EVAL oracle, can be used to efficiently test equality between two unknown distributions D_1 and D_2. As we show (in Subsection 3.3.2) the required "approximate" EVAL oracle can be efficiently implemented using a COND oracle, and so Theorem 12 follows straightforwardly by combining Theorems 13 and 4.

6.2.1 Testing equality between D_1 and D_2 using an approximate EVAL oracle.

We now show how an approximate EVAL_{D_1} oracle, an approximate EVAL_{D_2} oracle, and a SAMP_{D_1} oracle can be used together to test whether $D_1 = D_2$ versus $d_{\mathrm{TV}}(D_1, D_2) \geq \epsilon$. As mentioned earlier, the approach is a simple extension of the EVAL algorithm given in Observation 24 of [RS09].

Theorem 13 *Let* ORACLE_1 *be an* $(\epsilon/100, \epsilon/100)$*-approximate* EVAL_{D_1} *simulator and let* ORACLE_2 *be an* $(\epsilon/100, \epsilon/100)$*-approximate* EVAL_{D_2} *simulator. There is an algorithm* TEST-EQUALITY-UNKNOWN *with the following properties: for any distributions* D_1, D_2 *over* $[N]$*, algorithm* TEST-EQUALITY-UNKNOWN *makes* $O(1/\epsilon)$ *queries to* ORACLE_1, ORACLE_2, SAMP_{D_1}, SAMP_{D_2}*, and it outputs* ACCEPT *with probability at least* $7/10$ *if* $D_1 = D_2$ *and outputs* REJECT *with probability at least* $7/10$ *if* $d_{\mathrm{TV}}(D_1, D_2) \geq \epsilon$.

It is clear that TEST-EQUALITY-UNKNOWN makes $O(1/\epsilon)$ queries as claimed. To prove Theorem 13 we argue completeness and soundness below.

Algorithm 10: TEST-EQUALITY-UNKNOWN

Input: query access to ORACLE_1, to ORACLE_2, and access to SAMP_{D_1} , SAMP_{D_2} oracles

1: Call the SAMP_{D_1} oracle $m = 5/\epsilon$ times to obtain points $h_1, \ldots, h_m$ distributed according to D_1.

2: Call the SAMP_{D_2} oracle $m = 5/\epsilon$ times to obtain points $h_{m+1}, \ldots, h_{2m}$ distributed according to D_2.

3: **for** $j = 1$ to $2m$ **do**

4: Call $\mathsf{ORACLE}_1(h_j)$. If it returns $\mathsf{UNKNOWN}$ then output REJECT, otherwise let $v_{1,i} \in [0,1]$ be the value it outputs.

5: Call $\mathsf{ORACLE}_2(h_j)$. If it returns $\mathsf{UNKNOWN}$ then output REJECT, otherwise let $v_{2,i} \in [0,1]$ be the value it outputs.

6: **if** $v_{1,j} \notin [1-\epsilon/8, 1+\epsilon/8]v_{2,j}$ **then**

7: output REJECT and exit

8: **end if**

9: **end for**

10: output ACCEPT

Completeness: Suppose that $D_1 = D_2$. Since ORACLE_1 is an $(\epsilon/100, \epsilon/100)$-approximate EVAL_{D_1} simulator, the probability that any of the $2m = 10/\epsilon$ points $h_1, \ldots, h_{2m}$ drawn in Lines 1 and 2 lies in $S^{(\epsilon/100, D_1)}$ is at most $1/10$. Going forth, let us assume that all points h_i indeed lie outside $S^{(\epsilon/100, D_1)}$. Then for each execution of Line 4 we have that with probability at least $1 - \epsilon/100$ the call to $\mathsf{ORACLE}(h_i)$ yields a value $v_{1,i}$ satisfying $v_{1,i} \in [1 - \frac{\epsilon}{100}, 1 + \frac{\epsilon}{100}]D_1(i)$. The same holds for each execution of Line 5. Since there are $20/\epsilon$ total executions of Lines 4 and 5, with overall probability at least $7/10$ we have that each $1 \leq j \leq m$ has $v_{1,j}, v_{2,j} \in [1 - \frac{\epsilon}{100}, 1 + \frac{\epsilon}{100}]D_1(i)$. If this is the case then $v_{1,j}, v_{2,j}$ pass the check in Line 6, and thus the algorithm outputs ACCEPT with overall probability at least $7/10$.

Soundness: Now suppose that $d_{\mathrm{TV}}(D_1, D_2) \geq \epsilon$. Let us say that $i \in [N]$ is *good* if $D_1(i) \in [1-\epsilon/5, 1+\epsilon/5]D_2(i)$. Let $\mathrm{BAD} \subseteq [N]$ denote the set of all $i \in [N]$ that are not good. We have

$$2d_{\mathrm{TV}}(D_1, D_2) = \sum_{i \text{ is good}} |D_1(i) - D_2(i)| + \sum_{i \text{ is bad}} |D_1(i) - D_2(i)| \geq 2\epsilon.$$

Since

$$\sum_{i \text{ is good}} |D_1(i) - D_2(i)| \leq \sum_{i \text{ is good}} \frac{\epsilon}{5}|D_2(i)| \leq \frac{\epsilon}{5},$$

we have

$$\sum_{i \text{ is bad}} (|D_1(i)| + |D_2(i)|) \geq \sum_{i \text{ is bad}} |D_1(i) - D_2(i)| \geq \frac{9}{5}\epsilon.$$

Consequently it must be the case that either $D_1(\mathrm{BAD}) \geq \frac{9}{10}\epsilon$ or $D_2(\mathrm{BAD}) \geq \frac{9}{10}\epsilon$. For the rest of the argument we suppose that $D_1(\mathrm{BAD}) \geq \frac{9}{10}\epsilon$ (by the symmetry of the algorithm, an identical argument to the one we give below but with the roles of D_1 and D_2 flipped throughout handles the other case).

Since $D_1(\mathrm{BAD}) \geq \frac{9}{10}\epsilon$, a simple calculation shows that with probability at least $98/100$ at least one of the $5/\epsilon$ points $h_1, \ldots, h_m$ drawn in Line 1 belongs to BAD. For the rest of the argument we

suppose that indeed (at least) one of these points is in BAD; let h_{i^*} be such a point. Now consider the execution of Line 4 when ORACLE_1 is called on h_{i^*}. By Definition 3, whether or not i^* belongs to $S^{(\epsilon/100,D_1)}$, with probability at least $1-\epsilon/100$ the call to ORACLE_1 either causes TEST-EQUALITY-UNKNOWN to REJECT in Line 4 (because ORACLE_1 returns UNKNOWN) or it returns a value $v_{1,i^*} \in [1-\frac{\epsilon}{100}, 1+\frac{\epsilon}{100}]D_1(i^*)$. We may suppose that it returns a value $v_{1,i^*} \in [1-\frac{\epsilon}{100}, 1+\frac{\epsilon}{100}]D_1(i^*)$. Similarly, in the execution of Line 5 when ORACLE_2 is called on h_{i^*}, whether or not i^* belongs to $S^{(\epsilon/100,D_2)}$, with probability at least $1-\epsilon/100$ the call to ORACLE_2 either causes TEST-EQUALITY-UNKNOWN to reject in Line 5 or it returns a value $v_{2,i^*} \in [1-\frac{\epsilon}{100}, 1+\frac{\epsilon}{100}]D_2(i^*)$. We may suppose that it returns a value $v_{2,i^*} \in [1-\frac{\epsilon}{100}, 1+\frac{\epsilon}{100}]D_2(i^*)$. But recalling that $i^* \in$ BAD, an easy calculation shows that the values v_{1,i^*} and v_{2,i^*} must be multiplicatively far enough from each other that the algorithm will output REJECT in Line 7. Thus with overall probability at least 96/100 the algorithm outputs REJECT. ■

7 An algorithm for estimating the distance to uniformity

In this section we describe an algorithm that estimates the distance between a distribution D and the uniform distribution $\mathcal{U}$ by performing $\mathrm{poly}(1/\epsilon)$ PCOND (and SAMP) queries. We start by giving a high level description of the algorithm.

By the definition of the variation distance (and the uniform distribution),

$$d_{\mathrm{TV}}(D,\mathcal{U}) \;=\; \sum_{i:D(i)<1/N} \left(\frac{1}{N} - D(i)\right) \;. \tag{64}$$

We define the following function over $[N]$:

$$\psi^D(i) = (1 - N\cdot D(i)) \;\text{ for }\; D(i) < \frac{1}{N}\;, \;\text{ and } \psi^D(i) = 0 \;\text{ for }\; D(i) \geq \frac{1}{N}\;. \tag{65}$$

Observe that $\psi^D(i) \in [0,1]$ for every $i \in [N]$ and

$$d_{\mathrm{TV}}(D,\mathcal{U}) \;=\; \frac{1}{N}\sum_{i=1}^{N} \psi^D(i)\;. \tag{66}$$

Thus $d_{\mathrm{TV}}(D,\mathcal{U})$ can be viewed as an average value of a function whose range is in $[0,1]$. Since D is fixed throughout this subsection, we shall use the shorthand $\psi(i)$ instead of $\psi^D(i)$. Suppose we were able to compute $\psi(i)$ exactly for any i of our choice. Then we could obtain an estimate $\hat{d}$ of $d_{\mathrm{TV}}(D,\mathcal{U})$ to within an additive error of $\epsilon/2$ by simply selecting $\Theta(1/\epsilon^2)$ points in $[N]$ uniformly at random and setting $\hat{d}$ to be the average value of $\psi(\cdot)$ on the sampled points. By an additive Chernoff bound (for an appropriate constant in the $\Theta(\cdot)$ notation), with high constant probability the estimate $\hat{d}$ would deviate by at most $\epsilon/2$ from $d_{\mathrm{TV}}(D,\mathcal{U})$.

Suppose next that instead of being able to compute $\psi(i)$ exactly, we were able to compute an estimate $\hat{\psi}(i)$ such that $|\hat{\psi}(i) - \psi(i)| \leq \epsilon/2$. By using $\hat{\psi}(i)$ instead of $\psi(i)$ for each of the $\Theta(1/\epsilon^2)$ sampled points we would incur an additional additive error of at most $\epsilon/2$. Observe first that for i such that $D(i) \leq \epsilon/(2N)$ we have that $\psi(i) \geq 1-\epsilon/2$, so the estimate $\hat{\psi}(i) = 1$ meets

our requirements. Similarly, for i such that $D(i) \geq 1/N$, any estimate $\hat{\psi}(i) \in [0, \epsilon/2]$ can be used. Finally, for i such that $D(i) \in [\epsilon/(2N), 1/N]$, if we can obtain an estimate $\widehat{D}(i)$ such that $\widehat{D}(i) \in [1 - \epsilon/2, 1 + \epsilon/2]D(i)$, then we can use $\hat{\psi}(i) = 1 - N \cdot \widehat{D}(i)$.

In order to obtain such estimates $\hat{\psi}(i)$, we shall be interested in finding a *reference point* x. Namely, we shall be interested in finding a pair $(x, \widehat{D}(x))$ such that $\widehat{D}(x) \in [1 - \epsilon/c, 1 + \epsilon/c]D(x)$ for some sufficiently large constant c, and such that $D(x) = \Omega(\epsilon/N)$ and $D(x) = O(1/(\epsilon N))$. In Subsection 7.1 we describe a procedure for finding such a reference point. More precisely, the procedure is required to find such a reference point (with high constant probability) only under a certain condition on D. It is not hard to verify (and we show this subsequently), that if this condition is not met, then $d_{\mathrm{TV}}(D, \mathcal{U})$ is very close to 1. In order to state the lemma we introduce the following notation. For $\gamma \in [0, 1]$, let

$$H_\gamma^D \stackrel{\text{def}}{=} \left\{ i : \; D(i) \geq \frac{1}{\gamma N} \right\} . \tag{67}$$

Lemma 26 *Given an input parameter $\kappa \in (0, 1/4]$ as well as* SAMP *and* PCOND *query access to a distribution D, the procedure* Find-Reference *(Algorithm 12) either returns a pair $(x, \widehat{D}(x))$ where $x \in [N]$ and $\widehat{D}(x) \in [0, 1]$ or returns* No-Pair. *The procedure satisfies the following:*

1. *If $D(H_\kappa^D) \leq 1 - \kappa$, then with probability at least $9/10$, the procedure returns a pair $(x, \widehat{D}(x))$ such that $\widehat{D}(x) \in [1 - 2\kappa, 1 + 3\kappa]D(x)$ and $D(x) \in \left[\frac{\kappa}{8}, \frac{4}{\kappa}\right] \cdot \frac{1}{N}$.*
2. *If $D(H_\kappa^D) > 1 - \kappa$, then with probability at least $9/10$, the procedure either returns* No-Pair *or it returns a pair $(x, \widehat{D}(x))$ such that $\widehat{D}(x) \in [1 - 2\kappa, 1 + 3\kappa]D(x)$ and $D(x) \in \left[\frac{\kappa}{8}, \frac{4}{\kappa}\right] \cdot \frac{1}{N}$.*

The procedure performs $\tilde{O}(1/\kappa^{20})$ PCOND *and* SAMP *queries.*

Once we have a reference point x we can use it to obtain an estimate $\hat{\psi}(i)$ for any i of our choice, using the procedure Compare, whose properties are stated in Lemma 2 (see Subsection 3.1).

Theorem 14 *With probability at least $2/3$, the estimate $\hat{d}$ returned by Algorithm 11 satisfies: $\hat{d} = d_{\mathrm{TV}}(D, \mathcal{U}) \pm O(\epsilon)$. The number of queries performed by the algorithm is $\tilde{O}(1/\epsilon^{20})$.*

Proof: In what follows we shall use the shorthand H_γ instead of H_γ^D. Let E_0 denote the event that the procedure Find-Reference (Algorithm 12) obeys the requirements in Lemma 26, where by Lemma 26 the event E_0 holds with probability at least $9/10$. Conditioned on E_0, the algorithm outputs $\hat{d} = 1$ right after calling the procedure (because the procedure returns No-Pair) only when $D(H_\kappa) > 1 - \kappa = 1 - \epsilon/8$. We claim that in this case $d_{\mathrm{TV}}(D, \mathcal{U}) \geq 1 - 2\epsilon/8 = 1 - \epsilon/4$. To verify this, observe that

$$d_{\mathrm{TV}}(D, \mathcal{U}) = \sum_{i : D(i) > 1/N} \left(D(i) - \frac{1}{N} \right) \geq \sum_{i \in H_\kappa} \left(D(i) - \frac{1}{N} \right) = D(H_\kappa) - \frac{|H_\kappa|}{N} \geq D(H_\kappa) - \kappa \, . \tag{68}$$

Thus, in this case the estimate $\hat{d}$ is as required.

Algorithm 11: Estimating the Distance to Uniformity

Input : PCOND and SAMP query access to a distribution D and a parameter $\epsilon \in [0,1]$.

1. Call the procedure FIND-REFERENCE (Algorithm 12) with κ set to $\epsilon/8$. If it returns No-Pair, then output $\hat{d} = 1$ as the estimate for the distance to uniformity. Otherwise, let $(x, \widehat{D}(x))$ be its output.
2. Select a sample S of $\Theta(1/\epsilon^2)$ points uniformly.
3. Let $K = \max\left\{\frac{2/N}{\widehat{D}(x)}, \frac{\widehat{D}(x)}{\epsilon/(4N)}\right\}$.
4. For each point $y \in S$:
   (a) Call COMPARE$\Big(\{x\}, \{y\}, \kappa, K, \frac{1}{10|S|}\Big)$.
   (b) If COMPARE returns High or it returns a value $\rho(y)$ such that $\rho(y) \cdot \widehat{D}(x) \geq 1/N$, then set $\hat{\psi}(y) = 0$;
   (c) Else, if COMPARE returns Low or it returns a value $\rho(y)$ such that $\rho(y) \cdot \widehat{D}(x) \leq \epsilon/4N$, then set $\hat{\psi}(y) = 1$;
   (d) Else set $\hat{\psi}(y) = 1 - N \cdot \rho(y) \cdot \widehat{D}(x)$.
5. Output $\hat{d} = \frac{1}{|S|}\sum_{y \in S} \hat{\psi}(y)$.

We turn to the case in which the procedure FIND-REFERENCE returns a pair $(x, \widehat{D}(x))$ such that $\widehat{D}(x) \in [1-2\kappa, 1+3\kappa]D(x)$ and $D(x) \in \left[\frac{\kappa}{8}, \frac{4}{\kappa}\right] \cdot \frac{1}{N}$.

We start by defining two more "desirable" events, which hold (simultaneously) with high constant probability, and then show that conditioned on these events holding (as well as E_0), the output of the algorithm is as required. Let E_1 be the event that the sample S satisfies

$$\left| \frac{1}{|S|} \sum_{y \in S} \psi(y) - d_{\mathrm{TV}}(D, \mathcal{U}) \right| \leq \epsilon/2 \,. \tag{69}$$

By an additive Chernoff bound, the event E_1 holds with probability at least $9/10$.

Next, let E_2 be the event that all calls to the procedure COMPARE return answers as specified in Lemma 2. Since COMPARE is called $|S|$ times, and for each call the probability that it does not return an answer as specified in the lemma is at most $1/(10|S|)$, by the union bound the probability that E_2 holds is at least $9/10$.

From this point on assume events E_0, E_1 and E_2 all occur, which holds with probability at least $1 - 3/10 \geq 2/3$. Since E_2 holds, we get the following.

1. When COMPARE returns High for $y \in S$ (so that $\hat{\psi}(y)$ is set to 0) we have that

$$D(y) > K \cdot D(x) \geq \frac{2/N}{\widehat{D}(x)} \cdot D(x) > \frac{1}{N} \,, \tag{70}$$

implying that $\hat{\psi}(y) = \psi(y)$.

2. When Compare returns Low for $y \in S$ (so that $\hat{\psi}(y)$ is set to 1) we have that

$$D(y) < \frac{D(x)}{K} \leq \frac{D(x)}{\widehat{D}(x)/(\epsilon/4N)} \leq \frac{\epsilon}{2N} \,, \tag{71}$$

implying that $\hat{\psi}(y) \leq \psi(y) + \epsilon/2$ (and clearly $\psi(y) \leq \hat{\psi}(y)$).

3. When Compare returns a value $\rho(y)$ it holds that $\rho(y) \in [1-\kappa, 1+\kappa](D(y)/D(x))$, so that $\rho(y) \cdot \widehat{D}(x) \in [(1-\kappa)(1-2\kappa), (1+\kappa)(1+3\kappa)]D(y)$. Since $\kappa = \epsilon/8$, if $\rho(y) \cdot \widehat{D}(x) \geq 1/N$ (so that $\hat{\psi}(y)$ is set to 0), then $\psi(y) < \epsilon/2$, if $\rho(y) \cdot \widehat{D}(x) \leq \epsilon/4N$ (so that $\hat{\psi}(y)$ is set to 1), then $\psi(y) \geq 1 - \epsilon/2$, and otherwise $|\hat{\psi}(y) - \psi(y)| \leq \epsilon/2$.

It follows that

$$\hat{d} = \frac{1}{|S|}\sum_{y \in S} \hat{\psi}(y) \in \left[\frac{1}{|S|}\sum_{y \in S} \psi(y) - \epsilon/2, \frac{1}{|S|}\sum_{y \in S} \psi(y) + \epsilon/2\right] \subseteq [d_{\mathrm{TV}}(D,\mathcal{U}) - \epsilon, d_{\mathrm{TV}}(D,\mathcal{U}) + \epsilon] \tag{72}$$

as required.

The number of queries performed by the algorithm is the number of queries performed by the procedure Find-Reference, which is $\tilde{O}(1/\epsilon^{20})$, plus $\Theta(1/\epsilon^2)$ times the number of queries performed in each call to Compare. The procedure Compare is called with the parameter K, which is bounded by $O(1/\epsilon^2)$, the parameter η, which is $\Omega(\epsilon)$, and δ, which is $\Omega(1/\epsilon^2)$. By Lemma 2, the number of queries performed in each call to Compare is $O(\log(1/\epsilon)/\epsilon^4)$. The total number of queries performed is hence $\tilde{O}(1/\epsilon^{20})$. ■

7.1 Finding a reference point

In this subsection we prove Lemma 26. We start by giving the high-level idea behind the procedure. For a point $x \in [N]$ and $\gamma \in [0,1]$, let $U_\gamma^D(x)$ be as defined in Equation (11). Since D is fixed throughout this subsection, we shall use the shorthand $U_\gamma(x)$ instead of $U_\gamma^D(x)$. Recall that κ is a parameter given to the procedure. Assume we had a point x for which $D(U_\kappa(x)) \geq \kappa^{d_1}$ and $|U_\kappa(x)| \geq \kappa^{d_2}N$ for some constants d_1 and d_2 (so that necessarily $D(x) = \Omega(\kappa^{d_1}/N)$ and $D(x) = O(1/(\kappa^{d_2}N))$. It is not hard to verify (and we show this in detail subsequently), that if $D(H) \leq 1-\kappa$, then a sample of size $\Theta(1/\mathrm{poly}(\kappa))$ distributed according to D will contain such a point x with high constant probability. Now suppose that we could obtain an estimate $\hat{w}$ of $D(U_\kappa(x))$ such that $\hat{w} \in [1-\kappa, 1+\kappa]D(U_\kappa(x))$ and an estimate $\hat{u}$ of $|U_\kappa(x)|$ such that $\hat{u} \in [1-\kappa, 1+\kappa]|U_\kappa(x)|$. By the definition of $U_\kappa(x)$ we have that $(\hat{w}/\hat{u}) \in [1-O(\kappa), 1+O(\kappa)]D(x)$.

Obtaining good estimates of $D(U_\kappa(x))$ and $|U_\kappa(x)|$ (for x such that both $|U_\kappa(x)|$ and $D(U_\kappa(x))$ are sufficiently large) might be infeasible. This is due to the possible existence of many points y for which $D(y)$ is very close to $(1+\kappa)D(x)$ or $D(x)/(1+\kappa)$ which define the boundaries of the set $U_\kappa(x)$. For such points it is not possible to efficiently distinguish between those among them that belong to $U_\kappa(x)$ (so that they are within the borders of the set) and those that do not belong to $U_\kappa(x)$ (so that they are just outside the borders of the set). However, for our purposes it suffices to estimate

the weight and size of *some* set $U_\alpha(x)$ such that $\alpha \geq \kappa$ (so that $U_\kappa(x) \subseteq U_\alpha(x)$) and α is not much larger than κ (e.g., $\alpha \leq 2\kappa$)). To this end we can apply Procedure Estimate-Neighborhood (see Subsection 3.2), which (conditioned on $D(U_\kappa(x))$ being above a certain threshold), returns a pair $(\hat{w}(x), \alpha)$ such that $\hat{w}(x)$ is a good estimate of $D(U_\alpha(x))$. Furthermore, α is such that for α' slightly larger than α, the weight of $U_{\alpha'}(x) \setminus U_\alpha(x)$ is small, allowing us to obtain also a good estimate $\hat{\mu}(x)$ of $|U_\alpha(x)|/N$.

Algorithm 12: Procedure Find-Reference

Input : PCOND and SAMP query access to a distribution D and a parameter $\kappa \in (0, 1/4]$

1. Select a sample X of $\Theta(\log(1/\kappa)/\kappa^2)$ points distributed according to D.
2. For each $x \in X$ do the following:
   (a) Call Estimate-Neighborhood with the parameters κ as in the input to Find-Reference, $\beta = \kappa^2/(40\log(1/\kappa))$, $\eta = \kappa$, and $\delta = 1/(40|X|)$. Let $\theta = \kappa\eta\beta\delta/64 = \Theta(\kappa^6/\log^2(1/\kappa))$ (as in Find-Reference).
   (b) If Estimate-Neighborhood returns a pair $(\hat{w}(x), \alpha(x))$ such that $\hat{w}(x) < \kappa^2/20\log(1/\kappa)$, then go to Line 2 and continue with next $x \in X$.
   (c) Select a sample Y_x of size $\Theta(\log^2(1/\kappa)/\kappa^5)$ distributed uniformly.
   (d) For each $y \in Y_x$ call Compare$(\{x\}, \{y\}, \theta/4, 4, 1/40|X||Y_x|)$, and let the output be denoted $\rho_x(y)$.
   (e) Let $\hat{\mu}(x)$ be the fraction of occurrences of $y \in Y_x$ such that $\rho_x(y) \in [1/(1+\alpha+\theta/2), 1+\alpha+\theta/2]$.
   (f) Set $\widehat{D}(x) = \hat{w}(x)/(\hat{\mu}(x)N)$.
3. If for some point $x \in X$ we have $\hat{w}(x) \geq \kappa^2/20\log(1/\kappa)$, $\hat{\mu}(x) \geq \kappa^3/20\log(1/\kappa)$, and $\kappa/4N \leq \widehat{D}(x) \leq 2/(\kappa N)$, then return $(x, \widehat{D}(x))$. Otherwise return No-Pair.

Proof of Lemma 26: We first introduce the following notation.

$$L \stackrel{\text{def}}{=} \left\{ i : \; D(i) < \frac{\kappa}{2N} \right\}, \; M \stackrel{\text{def}}{=} \left\{ i : \; \frac{\kappa}{2N} \leq D(i) < \frac{1}{\kappa N} \right\} . \tag{73}$$

Let $H = H_\kappa^D$ where H_κ^D is as defined in Equation (67). Observe that $D(L) < \kappa/2$, so that if $D(H) \leq 1-\kappa$, then $D(M) \geq \kappa/2$. Consider further partitioning the set M of "medium weight" points into buckets $M_1, \ldots, M_r$ where $r = \log_{1+\kappa}(2/\kappa^2) = \Theta(\log(1/\kappa)/\kappa)$ and the bucket M_j is defined as follows:

$$M_j \stackrel{\text{def}}{=} \left\{ i : \; (1+\kappa)^{j-1} \cdot \frac{\kappa}{2N} \leq D(i) < (1+\kappa)^j \cdot \frac{\kappa}{2N} \right\} . \tag{74}$$

We consider the following "desirable" events.

1. Let E_1 be the event that conditioned on the existence of a bucket M_j such that $D(M_j) \geq \kappa/2r = \Omega(\kappa^2/\log(1/\kappa))$, there exists a point $x^* \in X$ that belongs to M_j. By the setting of the size of the sample X, the (conditional) event E_1 holds with probability at least $1 - 1/40$.

2. Let E_2 be the event that all calls to ESTIMATE-NEIGHBORHOOD return an output as specified by Lemma 3. By Lemma 3, the setting of the confidence parameter δ in each call and a union bound over all $|X|$ calls, E_2 holds with probability at least $1 - 1/40$.

3. Let E_3 be the event that for each $x \in X$ we have the following.

   (a) If $\frac{|U_{\alpha(x)}(x)|}{N} \geq \frac{\kappa^3}{40\log(1/\kappa)}$, then $\frac{|Y_x \cap U_{\alpha(x)}(x)|}{|Y_x|} \in [1-\eta/2, 1+\eta/2]\frac{|U_{\alpha(x)}(x)|}{N}$;
   If $\frac{|U_{\alpha(x)}(x)|}{N} < \frac{\kappa^3}{40\log(1/\kappa)}$, then $\frac{|Y_x \cap U_{\alpha(x)}(x)|}{|Y_x|} < \frac{\kappa^3}{30\log(1/\kappa)}$;

   (b) Let $\Delta_{\alpha(x),\theta}(x) \stackrel{\text{def}}{=} U_{\alpha(x)+\theta}(x) \setminus U_{\alpha(x)}(x)$ (where θ is as specified by the algorithm).
   If $\frac{|\Delta_{\alpha(x),\theta}(x)|}{N} \geq \frac{\kappa^4}{240\log(1/\kappa)}$, then $\frac{|Y_x \cap \Delta_{\alpha(x),\theta}(x)|}{|Y_x|} \leq 2 \cdot \frac{|\Delta_{\alpha(x),\theta}(x)|}{N}$;
   If $\frac{|\Delta_{\alpha(x),\theta}(x)|}{N} < \frac{\kappa^4}{240\log(1/\kappa)}$, then $\frac{|Y_x \cap \Delta_{\alpha(x),\theta}(x)|}{|Y_x|} < \frac{\kappa^4}{120\log(1/\kappa)}$.

   By the size of each set Y_x and a union bound over all $x \in X$, the event E_3 holds with probability at least $1 - 1/40$.

4. Let E_4 be the event that all calls to COMPARE return an output as specified by Lemma 2. By Lemma 2, the setting of the confidence parameter δ in each call and a union bound over all (at most) $|X| \cdot |Y|$ calls, E_3 holds with probability at least $1 - 1/40$.

Assuming events E_1–E_4 all hold (which occurs with probability at least $9/10$) we have the following.

1. By E_2, for each $x \in X$ such that $\hat{w}(x) \geq \kappa^2/20\log(1/\kappa)$ (so that x may be selected for the output of the procedure) we have that $D(U_{\alpha(x)}(x)) \geq \kappa^2/40\log(1/\kappa)$.

   The event E_2 also implies that for each $x \in X$ we have that $D(\Delta_{\alpha(x),\theta}(x)) \leq \eta\beta/16 \leq (\eta/16) \cdot D(U_{\alpha(x)}(x))$, so that

   $$\frac{|\Delta_{\alpha(x),\theta}(x)|}{N} \leq \frac{\eta(1+\alpha(x))(1+\alpha(x)+\theta)}{16} \cdot \frac{|U_{\alpha(x)}(x)|}{N} \leq \frac{\eta}{6} \cdot \frac{|U_{\alpha(x)}(x)|}{N} . \tag{75}$$

2. Consider any $x \in X$ such that $\hat{w}(x) \geq \kappa^2/20\log(1/\kappa)$. Let $T_x \stackrel{\text{def}}{=} \{y \in Y_x : \rho_x(y) \in [1/(1+\alpha+\theta/2), (1+\alpha+\theta/2]\}$, so that $\hat{\mu}(x) = |T_x|/|Y_x|$. By E_4, for each $y \in Y_x \cap U_{\alpha(x)}(x)$ we have that $\rho_x(y) \leq (1+\alpha)(1+\theta/4) \leq (1+\alpha+\theta/2)$ and $\rho_x(y) \geq (1+\alpha)^{-1}(1-\theta/4) \geq (1+\alpha+\theta/2)^{-1}$, so that $y \in T_x$. On the other hand, for each $y \notin Y_x \cap U_{\alpha(x)+\theta}(x)$ we have that $\rho_x(y) > (1+\alpha+\theta)(1-\theta/4) \geq 1+\alpha+\theta/2$ or $\rho_x(y) < (1+\alpha+\theta)^{-1}(1-\theta/4) < (1+\alpha+\theta/2)^{-1}$, so that $y \notin T_x$. It follows that

   $$Y_x \cap U_{\alpha(x)}(x) \;\subseteq\; T_x \;\subseteq\; Y_x \cap (U_{\alpha(x)}(x) \cup \Delta_{\alpha(x),\theta}(x)) . \tag{76}$$

   By E_3, when $\hat{\mu}(x) = |T_x|/|Y_x| \geq \kappa^3/20\log(1/\kappa)$, then necessarily $\hat{\mu}(x) \in [1-\eta, 1+\eta]|U_{\alpha(x)}(x)|/N$. To verify this consider the following cases.

   (a) If $\frac{|U_{\alpha(x)}(x)|}{N} \geq \frac{\kappa^3}{40\log(1/\kappa)}$, then (by the left-hand-side of Equation (76) and the definition of E_3) we get that $\hat{\mu}(x) \geq (1-\eta/2)\frac{|U_{\alpha(x)}(x)|}{N}$, and (by the right-hand-side of Equation (76), Equation (75) and E_3) we get that $\hat{\mu}(x) \leq (1+\eta/2)\frac{|U_{\alpha(x)}(x)|}{N} + 2(\eta/6)\frac{|U_{\alpha(x)}(x)|}{N} < (1+\eta)\frac{|U_{\alpha(x)}(x)|}{N}$.

(b) If $\frac{|U_{\alpha(x)}(x)|}{N} < \frac{\kappa^3}{40\log(1/\kappa)}$, then (by the right-hand-side of Equation (76), Equation (75) and E_3) we get that $\hat{\mu}(x) < \frac{\kappa^3}{30\log(1/\kappa)} + \frac{\kappa^4}{120\log(1/\kappa)} < \kappa^3/20\log(1/\kappa)$.

3. If $D(H) \leq 1-\kappa$, so that $D(M) \geq \kappa/2$, then there exists at least one bucket M_j such that $D(M_j) \geq \kappa/2r = \Omega(\kappa^2/\log(1/\kappa))$. By E_1, the sample X contains a point $x^* \in M_j$. By the definition of the buckets, for this point x^* we have that $D(U_\kappa(x^*)) \geq \kappa/2r \geq \kappa^2/(10\log(1/\kappa))$ and $|U_\kappa(x^*)| \geq (\kappa^2/2r)N \geq \kappa^3/(10\log(1/\kappa))$.

By the first two items above and the setting $\eta = \kappa$ we have that for each x such that $\hat{w}(x) \geq \kappa^2/20\log(1/\kappa)$ and $\hat{\mu}(x) \geq \kappa^3/20\log(1/\kappa)$,

$$\widehat{D}(x) \in \left[\frac{1-\kappa}{1+\kappa}, \frac{1+\kappa}{1-\kappa}\right] D(x) \subset [1-2\kappa, 1+3\kappa]D(x) \ .$$

Thus, if the algorithm outputs a pair $(x, \widehat{D}(x))$ then it satisfies the condition stated in both items of the lemma. This establishes the second item in the lemma. By combining all three items we get that if $D(H) \geq 1-\kappa$ then the algorithm outputs a pair $(x, \widehat{D}(x))$ (where possibly, but not necessarily, $x = x^*$), and the first item is established as well.

Turning to the query complexity, the total number of PCOND queries performed in the $|X| = O(\log(1/\kappa)/\kappa^2)$ calls to ESTIMATE-NEIGHBORHOOD is $O\left(\frac{|X|\log(1/\delta)^2\log(1/(\beta\eta))}{\kappa^2\eta^4\beta^3\delta^2}\right) = \tilde{O}(1/\kappa^{18})$, and the total number of PCOND queries performed in the calls to COMPARE (for at most all pairs $x \in X$ and $y \in Y_x$) is $\tilde{O}(1/\kappa^{20})$. ■

8 A $\tilde{O}\left(\frac{\log^3 N}{\epsilon^3}\right)$-query ICOND_D algorithm for testing uniformity

In this and the next section we consider ICOND algorithms for testing whether an unknown distribution D over $[N]$ is the uniform distribution versus ϵ-far from uniform. Our results show that ICOND algorithms are not as powerful as PCOND algorithms for this basic testing problem; in this section we give a $\mathrm{poly}(\log N, 1/\epsilon)$-query ICOND_D algorithm, and in the next section we prove that any ICOND_D algorithm must make $\tilde{\Omega}(\log N)$ queries.

In more detail, in this section we describe an algorithm ICOND_D-TEST-UNIFORM and prove the following theorem:

Theorem 15 ICOND_D-TEST-UNIFORM *is a* $\tilde{O}(\frac{\log^3 N}{\epsilon^3})$*-query* ICOND_D *testing algorithm for uniformity, i.e. it outputs* ACCEPT *with probability at least* $2/3$ *if* $D = \mathcal{U}$ *and outputs* REJECT *with probability at least* $2/3$ *if* $d_{\mathrm{TV}}(D,\mathcal{U}) \geq \epsilon$.

Intuition. Recall that, as mentioned in Subsection 4.1, any distribution D which is ϵ-far from uniform must put $\Omega(\epsilon)$ probability mass on "significantly heavy" elements (that is, if we define $H' = \left\{ h \in [N] \;\middle|\; D(h) \geq \frac{1}{N} + \frac{\epsilon}{4N} \right\}$, it must hold that $D(H') \geq \epsilon/4$). Consequently a sample of $O(1/\epsilon)$ points drawn from D will contain such a point with high probability. Thus, a natural

approach to testing whether D is uniform is to devise a procedure that, given an input point y, can distinguish between the case that $y \in H'$ and the case that $D(y) = 1/N$ (as it is when $D = \mathcal{U}$).

We give such a procedure, which uses the ICOND_D oracle to perform a sort of binary search over intervals. The procedure successively "weighs" narrower and narrower intervals until it converges on the single point y. In more detail, we consider the *interval tree* whose root is the whole domain $[N]$, with two children $\{1, \ldots, N/2\}$ and $\{N/2+1, \ldots, N\}$, and so on, with a single point at each of the N leaves. Our algorithm starts at the root of the tree and goes down the path that corresponds to y; at each child node it uses COMPARE to compare the weight of the current node to the weight of its sibling under D. If at any point the estimate deviates significantly from the value it should have if D were uniform (namely the weights should be essentially equal, with slight deviations because of even/odd issues), then the algorithm rejects. Assuming the algorithm does not reject, it provides a $(1 \pm O(\epsilon))$-accurate multiplicative estimate of $D(y)$, and the algorithm checks whether this estimate is sufficiently close to $1/N$ (rejecting if this is not the case). If no point in a sample of $\Theta(1/\epsilon)$ points (drawn according to D) causes rejection, then the algorithm accepts.

Algorithm 13: BINARY-DESCENT

Input: parameter $\epsilon > 0$; integers $1 \le a \le b \le N$; $y \in [a, b]$; query access to ICOND_D oracle

if $a = b$ **then**
  return 1
end if
Let $c = \lfloor \frac{a+b}{2} \rfloor$; $\Delta = (b - a + 1)/2$.
if $y \le c$ **then**
  Define $I_y = \{a, \ldots, c\}$, $I_{\bar{y}} = \{c+1, \ldots, b\}$ and $\rho = \lceil \Delta \rceil / \lfloor \Delta \rfloor$
else
  Define $I_{\bar{y}} = \{a, \ldots, c\}$, $I_y = \{c+1, \ldots, b\}$ and $\rho = \lfloor \Delta \rfloor / \lceil \Delta \rceil$
end if
Call COMPARE on I_y, $I_{\bar{y}}$ with parameters $\eta = \frac{\epsilon}{48 \log N}$, $K = 2$, $\delta = \frac{\epsilon}{100(1+\log N)}$ to get an estimate $\hat{\rho}$ of $D(I_y)/D(I_{\bar{y}})$
if $\hat{\rho} \notin [1 - \frac{\epsilon}{48 \log N}, 1 + \frac{\epsilon}{48 \log N}] \cdot \rho$ (this includes the case that $\hat{\rho}$ is High or Low) **then**
  return REJECT
end if
Call recursively BINARY-DESCENT on input (ϵ, the endpoints of I_y, y);
if BINARY-DESCENT returns a value ν **then**
  return $\frac{\hat{\rho}}{1+\hat{\rho}} \cdot \nu$
else
  return REJECT
end if

The algorithm we use to perform the "binary search" described above is Algorithm 13, BINARY-DESCENT. We begin by proving correctness for it:

Lemma 27 *Suppose the algorithm* BINARY-DESCENT *is run with inputs $\epsilon \in (0, 1]$, $a = 1$, $b = N$, and $y \in [N]$, and is provided* ICOND *oracle access to distribution D over $[N]$. It performs $\tilde{O}(\log^3 N/\epsilon^2)$ queries and either outputs a value $\hat{D}(y)$ or* REJECT*, where the following holds:*

Algorithm 14: ICOND_D-TEST-UNIFORM

Input: error parameter $\epsilon > 0$; query access to ICOND_D oracle

Draw $t = \frac{20}{\epsilon}$ points $y_1, \ldots, y_t$ from SAMP_D.
for $j = 1$ to t **do**
  Call BINARY-DESCENT$(\epsilon, 1, N, y_j)$ and return REJECT if it rejects, otherwise let $\hat{d}_j$ be the value it returns as its estimate of $D(y_j)$
  if $\hat{d}_j \notin [1 - \frac{\epsilon}{12}, 1 + \frac{\epsilon}{12}] \cdot \frac{1}{N}$ **then**
    return REJECT
  end if
end for
return ACCEPT

1. *if* $D(y) \geq \frac{1}{N} + \frac{\epsilon}{4N}$, *then with probability at least* $1 - \frac{\epsilon}{100}$ *the procedure either outputs a value* $\hat{D}(y) \in [1 - \epsilon/12, 1 + \epsilon/12]D(y)$ *or* REJECT*;*

2. *if* $D = \mathcal{U}$, *then with probability at least* $1 - \frac{\epsilon}{100}$ *the procedure outputs a value* $\hat{D}(y) \in [1 - \epsilon/12, 1 + \epsilon/12] \cdot \frac{1}{N}$.

Proof of Lemma 27: The claimed query bound is easily verified, since the recursion depth is at most $1 + \log N$ and the only queries made are during calls to COMPARE, each of which performs $O(\log(1/\delta)/\gamma^2) = \tilde{O}(\log^2 N/\epsilon^2)$ queries.

Let E_0 be the event that all calls to COMPARE satisfy the conditions in Lemma 2; since each of them succeeds with probability at least $1 - \delta = 1 - \frac{\epsilon}{100(1+\log N)}$, a union bound shows that E_0 holds with probability at least $1 - \epsilon/100$. We hereafter condition on E_0.

We first prove the second part of the lemma where $D = \mathcal{U}$. Fix any specific recursive call, say the j-th, during the execution of the procedure. The intervals $I_y^{(j)}, I_{\bar{y}}^{(j)}$ used in that execution of the algorithm are easily seen to satisfy $D(I_y)/D(I_{\bar{y}}) \in [1/K, K]$ (for $K = 2$), so by event E_0 it must be the case that COMPARE returns an estimate $\hat{\rho}_j \in [1 - \frac{\epsilon}{48 \log N}, 1 + \frac{\epsilon}{48 \log N}] \cdot D(I_y^{(j)})/D(I_{\bar{y}}^{(j)})$. Since $D = U$, we have that $D(I_y^{(j)})/D(I_{\bar{y}}^{(j)}) = \rho^{(j)}$, so the overall procedure returns a numerical value rather than REJECT.

Let $M = \lceil \log N \rceil$ be the number of recursive calls (i.e., the number of executions of Line 14). Note that we can write $D(y)$ as a product

$$D(y) = \prod_{j=1}^{M} \frac{D(I_y^{(j)})}{D(I_y^{(j)}) + D(I_{\bar{y}}^{(j)})} = \prod_{j=1}^{M} \frac{D(I_y^{(j)})/D(I_{\bar{y}}^{(j)})}{D(I_y^{(j)})/D(I_{\bar{y}}^{(j)}) + 1} \,. \tag{77}$$

We next observe that for any $0 \leq \epsilon' < 1$ and $\rho, d > 0$, if $\hat{\rho} \in [1 - \epsilon', 1 + \epsilon']d$ then we have $\frac{\hat{\rho}}{\hat{\rho}+1} \in [1 - \frac{\epsilon'}{2}, 1 + \epsilon']\frac{d}{d+1}$ (by straightforward algebra). Applying this M times, we get

$$\begin{aligned}\prod_{j=1}^{M}\frac{\hat{\rho}_j}{\hat{\rho}_j+1} &\in \left[\left(1-\frac{\epsilon}{96\log N}\right)^M,\left(1+\frac{\epsilon}{48\log N}\right)^M\right]\cdot\prod_{j=1}^{M}\frac{D(I_y^{(j)})/D(I_{\bar{y}}^{(j)})}{D(I_y^{(j)})/D(I_{\bar{y}}^{(j)})+1}\\ &\in \left[\left(1-\frac{\epsilon}{96\log N}\right)^M,\left(1+\frac{\epsilon}{48\log N}\right)^M\right]\cdot D(y)\\ &\in \left[1-\frac{\epsilon}{12},1+\frac{\epsilon}{12}\right]D(y).\end{aligned}$$

Since $\prod_{j=1}^{M}\frac{\hat{\rho}_j}{\hat{\rho}_j+1}$ is the value that the procedure outputs, the second part of the lemma is proved.

The proof of the first part of the lemma is virtually identical. The only difference is that now it is possible that COMPARE outputs High or Low at some call (since D is not uniform it need not be the case that $D(I_y^{(j)})/D(I_{\bar{y}}^{(j)})=\rho^{(j)}$), but this is not a problem for (i) since in that case BINARY-DESCENT would output REJECT. ■

See Algorithm 13 for a description of the testing algorithm ICOND_D-TEST-UNIFORM. We now prove Theorem 15:

Proof of Theorem 15: Define E_1 to be the event that all calls to BINARY-DESCENT satisfy the conclusions of Lemma 27. With a union bound over all these $t=20/\epsilon$ calls, we have $\Pr[E_1]\geq 8/10$.

Completeness: Suppose $D=\mathcal{U}$, and condition again on E_1. Since this implies that BINARY-DESCENT will always return a value, the only case ICOND_D-TEST-UNIFORM might reject is by reaching Line 5. However, since it is the case that every value $\hat{d}_j$ returned by the procedure satisfies $\hat{D}(y)\in[1-\epsilon/12,1+\epsilon/12]\cdot\frac{1}{N}$, this can never happen.

Soundness: Suppose $d_{\mathrm{TV}}(D,\mathcal{U})\geq\epsilon$. Let E_2 be the event that at least one of the y_i's drawn in Line 1 belongs to H'. As $D(H')\geq\epsilon/4$, we have $\Pr[E_2]\geq 1-(1-\epsilon/4)^{20/\epsilon}\geq 9/10$. Conditioning on both E_1 and E_2, for such a y_j, one of two cases below holds:

- either the call to BINARY-DESCENT outputs REJECT and ICOND_D-TEST-UNIFORM outputs REJECT;
- or a value $\hat{d}_j$ is returned, for which $\hat{d}_j\geq(1-\frac{\epsilon}{12})(1+\frac{\epsilon}{4})\cdot\frac{1}{N}>(1+\epsilon/12)/N$ (where we used the fact that E_1 holds); and ICOND_D-TEST-UNIFORM reaches Line 5 and rejects.

Since $\Pr[E_1\cup E_2]\geq 7/10$, ICOND_D-TEST-UNIFORM is correct with probability at least 2/3. Finally, the claimed query complexity directly follows from the $t=\Theta(1/\epsilon)$ calls to BINARY-DESCENT, each of which makes $\tilde{O}(\log^3 N/\epsilon^2)$ queries to ICOND_D. ■

9 An $\Omega(\log N/\log\log N)$ lower bound for ICOND_D algorithms that test uniformity

In this section we prove that any ICOND_D algorithm that ϵ-tests uniformity even for constant ϵ must have query complexity $\tilde{\Omega}(\log N)$. This shows that our algorithm in the previous subsection is not too far from optimal, and sheds light on a key difference between ICOND and PCOND oracles.

Theorem 16 *Fix $\epsilon = 1/3$. Any ICOND_D algorithm for testing whether $D = \mathcal{U}$ versus $d_{\mathrm{TV}}(D, D^*) \geq \epsilon$ must make $\Omega\left(\frac{\log N}{\log \log N}\right)$ queries.*

To prove this lower bound we define a probability distribution $\mathcal{P}_{\mathrm{No}}$ over possible "No"-distributions (i.e. distributions that have variation distance at least $1/3$ from $\mathcal{U}$). A distribution drawn from $\mathcal{P}_{\mathrm{No}}$ is constructed as follows: first (assuming without loss of generality that N is a power of 2), we partition $[N]$ into $b = 2^X$ consecutive intervals of the same size $\Delta = \frac{N}{2^X}$, which we refer to as "blocks", where X is a random variable distributed uniformly on the set $\{\frac{1}{3}\log N, \frac{1}{3}\log N + 1, \ldots, \frac{2}{3}\log N\}$. Once the block size Δ is determined, a random offset y is drawn uniformly at random in $[N]$, and all block endpoints are shifted by y modulo $[N]$ (intuitively, this prevents the testing algorithm from "knowing" a priori that specific points are endpoints of blocks). Finally, independently for each block, a fair coin is thrown to determine its *profile*: with probability $1/2$, each point in the first half of the block will have probability weight $\frac{1-2\epsilon}{N}$ and each point in the second half will have probability $\frac{1+2\epsilon}{N}$ (such a block is said to be a *low-high* block, with profile $\downarrow\uparrow$). With probability $1/2$ the reverse is true: each point in the first half has probability $\frac{1+2\epsilon}{N}$ and each point in the second half has probability $\frac{1-2\epsilon}{N}$ (a *high-low* block $\uparrow\downarrow$). It is clear that each distribution D in the support of $\mathcal{P}_{\mathrm{No}}$ defined in this way indeed has $d_{\mathrm{TV}}(D,\mathcal{U}) = \epsilon$.

To summarize, each "No"-distribution D in the support of $\mathcal{P}_{\mathrm{No}}$ is parameterized by $(b+2)$ parameters: its block size Δ, offset y, and profile $\vartheta \in \{\downarrow\uparrow, \uparrow\downarrow\}^b$. Note that regardless of the profile vector, each block always has weight exactly Δ/N.

We note that while there is only one "Yes"-distribution $\mathcal{U}$, it will sometimes be convenient for the analysis to think of $\mathcal{U}$ as resulting from the same initial process of picking a block size and offset, but without the subsequent choice of a profile vector. We sometimes refer to this as the "fake construction" of the uniform distribution $\mathcal{U}$ (the reason for this will be clear later).

The proof of Theorem 16 will be carried out in two steps. First we shall restrict the analysis to *non-adaptive algorithms*, and prove the lower bound for such algorithms. This result will then be extended to the general setting by introducing (similarly to Subsection 5.2) the notion of a *query-faking algorithm*, and reducing the behavior of adaptive algorithms to non-adaptive ones through an appropriate sequence of such query-faking algorithms.

Before proceeding, we define the *transcript* of the interaction between an algorithm and a ICOND_D oracle. Informally, the transcript captures the entire history of interaction between the algorithm and the ICOND_D oracle during the whole sequence of queries.

Definition 7 *Fix any (possibly adaptive) testing algorithm A that queries an ICOND_D oracle. The* transcript *of A is a sequence $\mathcal{T} = (I_\ell, s_\ell)_{\ell \in \mathbb{N}^*}$ of pairs, where I_ℓ is the ℓ-th interval provided by the algorithm as input to ICOND_D, and $s_\ell \in I_\ell$ is the response that ICOND_D provides to this query. Given a transcript $\mathcal{T}$, we shall denote by $\mathcal{T}|_k$ the partial transcript induced by the first k queries, i.e. $\mathcal{T}|_k = (I_\ell, s_\ell)_{1 \leq \ell \leq k}$.*

Equipped with these definitions, we now turn to proving the theorem in the special case of non-adaptive testing algorithms. Observe that there are three different sources of randomness in our arguments: (i) the draw of the "No"-instance from $\mathcal{P}_{\mathrm{No}}$, (ii) the internal randomness of the testing

algorithm; and (iii) the random draws from the oracle. Whenever there could be confusion we shall explicitly state which probability space is under discussion.

9.1 A lower bound against non-adaptive algorithms

Throughout this subsection we assume that A is an arbitrary, fixed, non-adaptive, randomized algorithm that makes exactly $q \leq \tau \cdot \frac{\log N}{\log \log N}$ queries to ICOND_D; here $\tau > 0$ is some absolute constant that will be determined in the course of the analysis. (The assumption that A always makes exactly q queries is without loss of generality since if in some execution the algorithm makes $q' < q$ queries, it can perform additional "dummy" queries). In this setting algorithm A corresponds to a distribution P_A over q-tuples $\bar{I} = (I_1, \ldots, I_q)$ of query intervals. The following theorem will directly imply Theorem 16 in the case of non-adaptive algorithms:

Theorem 17

$$\left| \Pr_{D \sim \mathcal{P}_{\mathrm{No}}}[A^{\mathsf{ICOND}_D} \textit{ outputs } \mathsf{ACCEPT}] - \Pr[A^{\mathsf{ICOND}_{\mathcal{U}}} \textit{ outputs } \mathsf{ACCEPT}] \right| \leq 1/5. \tag{78}$$

Observe that in the first probability of Equation (78) the randomness is taken over the draw of D from $\mathcal{P}_{\mathrm{No}}$, the draw of $\bar{I} \sim P_A$ that A performs to select its sequence of query intervals, and the randomness of the ICOND_D oracle. In the second one the randomness is just over the draw of $\bar{I}$ from P_A and the randomness of the $\mathsf{ICOND}_{\mathcal{U}}$ oracle.

Intuition for Theorem 17. The high-level idea is that the algorithm will not be able to distinguish between the uniform distribution and a "No"-distribution unless it manages to learn something about the "structure" of the blocks in the "No"-case, either by guessing (roughly) the right block size, or by guessing (roughly) the location of a block endpoint and querying a short interval containing such an endpoint.

In more detail, we define the following "bad events" (over the choice of D and the points s_i) for a fixed sequence $\bar{I} = (I_1, \ldots, I_q)$ of queries (the dependence on $\bar{I}$ is omitted in the notation for the sake of readability):

$$\begin{aligned}
B^{\mathrm{N}}_{\mathrm{size}} &= \{\ \exists \ell \in [q] \mid \Delta/\log N \leq |I_\ell| \leq \Delta \cdot (\log N)^2\ \} \\
B^{\mathrm{N}}_{\mathrm{boundary}} &= \{\ \exists \ell \in [q] \mid |I_\ell| < \Delta/\log N \text{ and } I_\ell \text{ intersects two blocks}\ \} \\
B^{\mathrm{N}}_{\mathrm{middle}} &= \{\ \exists \ell \in [q] \mid |I_\ell| < \Delta/\log N \text{ and } I_\ell \text{ intersects both halves of the same block}\ \} \\
B^{\mathrm{N}}_{\ell,\mathrm{outer}} &= \{\Delta \cdot (\log N)^2 < |I_\ell| \text{ and } s_\ell \text{ belongs to a block not contained entirely in } I_\ell\} \qquad \ell \in [q] \\
B^{\mathrm{N}}_{\ell,\mathrm{collide}} &= \{\Delta \cdot (\log N)^2 < |I_\ell| \text{ and } \exists j < \ell,\ s_\ell \text{ and } s_j \text{ belong to the same block}\} \qquad \ell \in [q]
\end{aligned}$$

The first three events depend only on the draw of D from $\mathcal{P}_{\mathrm{No}}$, which determines Δ and y, while the last $2q$ events also depend on the random draws of s_ℓ from the ICOND_D oracle. We define in the same fashion the corresponding bad events for the "Yes"-instance (i.e. the uniform distribution $\mathcal{U}$) $B^{\mathrm{Y}}_{\mathrm{size}}$, $B^{\mathrm{Y}}_{\mathrm{boundary}}$, $B^{\mathrm{Y}}_{\mathrm{middle}}$, $B^{\mathrm{Y}}_{\ell,\mathrm{outer}}$ and $B^{\mathrm{Y}}_{\ell,\mathrm{collide}}$, using the notion of the "fake construction" of $\mathcal{U}$ mentioned above.

Events $B^{\mathrm{N}}_{\mathrm{size}}$ and $B^{\mathrm{Y}}_{\mathrm{size}}$ correspond to the possibility, mentioned above, that algorithm A "guesses" essentially the right block size, and events $B^{\mathrm{N}}_{\mathrm{boundary}}$, $B^{\mathrm{Y}}_{\mathrm{boundary}}$ and $B^{\mathrm{N}}_{\mathrm{middle}}$, $B^{\mathrm{Y}}_{\mathrm{middle}}$ correspond to the possibility that algorithm A "guesses" a short interval containing respectively a block endpoint or a block midpoint. The final bad events correspond to A guessing a "too-large" block size but "getting lucky" with the sample returned by ICOND, either because the sample belongs to one of the (at most two) outer blocks not entirely contained in the query interval, or because A has already received a sample from the same block as the current sample.

We can now describe the *failure events* for both the uniform distribution and for a "No"-distribution as the union of the corresponding bad events:

$$B^{\mathrm{N}}_{(\bar{I})} = B^{\mathrm{N}}_{\mathrm{size}} \cup B^{\mathrm{N}}_{\mathrm{boundary}} \cup B^{\mathrm{N}}_{\mathrm{middle}} \cup \left(\bigcup_{\ell=1}^{q} B^{\mathrm{N}}_{\ell,\mathrm{outer}} \right) \cup \left(\bigcup_{\ell=1}^{q} B^{\mathrm{N}}_{\ell,\mathrm{collide}} \right)$$

$$B^{\mathrm{Y}}_{(\bar{I})} = B^{\mathrm{Y}}_{\mathrm{size}} \cup B^{\mathrm{Y}}_{\mathrm{boundary}} \cup B^{\mathrm{Y}}_{\mathrm{middle}} \cup \left(\bigcup_{\ell=1}^{q} B^{\mathrm{Y}}_{\ell,\mathrm{outer}} \right) \cup \left(\bigcup_{\ell=1}^{q} B^{\mathrm{Y}}_{\ell,\mathrm{collide}} \right)$$

These failure events can be interpreted, from the point of view of the algorithm A, as the "opportunity to potentially learn something;" we shall argue below that if the failure events do not occur then the algorithm gains no information about whether it is interacting with the uniform distribution or with a "No"-distribution.

Structure of the proof of Theorem 17. First, observe that since the transcript is the result of the interaction of the algorithm and the oracle on a randomly chosen distribution, it is itself a random variable; we will be interested in the distribution over this random variable induced by the draws from the oracle and the choice of D. More precisely, for a fixed sequence of query sets $\bar{I}$, let $Z^{\mathrm{N}}_{\bar{I}}$ denote the random variable over "No"-transcripts generated when D is drawn from $\mathcal{P}_{\mathrm{No}}$. Note that this is a random variable over the probability space defined by the random draw of D and the draws of s_i by $\mathsf{ICOND}_D(I_\ell)$. We define $\mathfrak{A}^{\mathrm{N}}_{\bar{I}}$ as the resulting distribution over these "No"-transcripts. Similarly, $Z^{\mathrm{Y}}_{\bar{I}}$ will be the random variable over "Yes"-transcripts, with corresponding distribution $\mathfrak{A}^{\mathrm{Y}}_{\bar{I}}$.

As noted earlier, the nonadaptive algorithm A corresponds to a distribution P_A over q-tuples $\bar{I}$ of query intervals. We define $\mathfrak{A}^{\mathrm{N}}$ as the distribution over transcripts corresponding to first drawing $\bar{I}$ from P_A and then making a draw from $\mathfrak{A}^{\mathrm{N}}_{\bar{I}}$. Similarly, we define $\mathfrak{A}^{\mathrm{Y}}$ as the distribution over transcripts corresponding to first drawing $\bar{I}$ from P_A and then making a draw from $\mathfrak{A}^{\mathrm{Y}}_{\bar{I}}$.

To prove Theorem 17 it is sufficient to show that the two distributions over transcripts described above are statistically close:

Lemma 28 $d_{\mathrm{TV}}\big(\mathfrak{A}^{\mathrm{Y}}, \mathfrak{A}^{\mathrm{N}}\big) \leq 1/5$.

The proof of this lemma is structured as follows: first, for any *fixed* sequence of q queries $\bar{I}$, we bound the probability of the failure events, both for the uniform and the "No"-distributions:

Claim 29 *For each fixed sequence $\bar{I}$ of q query intervals, we have*

$$\Pr\big[B^{\mathrm{Y}}_{(\bar{I})}\big] \leq 1/10 \qquad \textit{and} \qquad \Pr_{D \leftarrow \mathcal{P}_{\mathrm{No}}}\big[B^{\mathrm{N}}_{(\bar{I})}\big] \leq 1/10.$$

(Note that the first probability above is taken over the randomness of the $\mathsf{ICOND}_{\mathcal{U}}$ responses and the choice of offset and size in the "fake construction" of $\mathcal{U}$, while the second is over the random draw of $D \sim \mathcal{P}_{\mathrm{No}}$ and over the ICOND_D responses.)

Next we show that, provided the failure events do not occur, the distribution over transcripts is exactly the same in both cases:

Claim 30 *Fix any sequence $\bar{I} = (I_1, \ldots, I_q)$ of q queries. Then, conditioned on their respective failure events not happening, $Z^{\mathrm{N}}_{\bar{I}}$ and $Z^{\mathrm{Y}}_{\bar{I}}$ are identically distributed:*

$$\text{for every transcript } \mathcal{T} = ((I_1, s_1), \ldots, (I_q, s_q)), \quad \Pr\left[Z^{\mathrm{N}}_{\bar{I}} = \mathcal{T} \;\middle|\; \overline{B^{\mathrm{N}}_{(\bar{I})}}\right] = \Pr\left[Z^{\mathrm{Y}}_{\bar{I}} = \mathcal{T} \;\middle|\; \overline{B^{\mathrm{Y}}_{(\bar{I})}}\right].$$

Finally we combine these two claims to show that the two overall distributions of transcripts are statistically close:

Claim 31 *Fix any sequence of q queries $\bar{I} = (I_1, \ldots, I_q)$. Then $d_{\mathrm{TV}}\big(\mathfrak{A}^{\mathrm{N}}_{\bar{I}}, \mathfrak{A}^{\mathrm{Y}}_{\bar{I}}\big) \le 1/5$.*

Lemma 28 (and thus Theorem 17) directly follows from Claim 31 since, using the notation $\bar{s} = (s_1, \ldots, s_q)$ for a sequence of q answers to a sequence $\bar{I} = (I_1, \ldots, I_q)$ of q queries, which together define a transcript $\mathcal{T}(\bar{I}, \bar{s}) = ((I_1, s_1), \ldots, (I_q, s_q))$,

$$\begin{aligned} d_{\mathrm{TV}}\left(\mathfrak{A}^{\mathrm{Y}}, \mathfrak{A}^{\mathrm{N}}\right) &= \frac{1}{2} \sum_{\bar{I}} \sum_{\bar{s}} \left| P_A(\bar{I}) \cdot \Pr\left[Z^{Y}_{\bar{I}} = \mathcal{T}(\bar{I}, \bar{s})\right] - P_A(\bar{I}) \cdot \Pr\left[Z^{N}_{\bar{I}} = \mathcal{T}(\bar{I}, \bar{s})\right] \right| \\ &= \frac{1}{2} \sum_{\bar{I}} P_A(\bar{I}) \cdot \sum_{\bar{s}} \left| \Pr\left[Z^{Y}_{\bar{I}} = \mathcal{T}(\bar{I}, \bar{s})\right] - \Pr\left[Z^{N}_{\bar{I}} = \mathcal{T}(\bar{I}, \bar{s})\right] \right| \\ &\le \max_{\bar{I}} \left\{ d_{\mathrm{TV}}\left(\mathfrak{A}^{\mathrm{Y}}_{\bar{I}}, \mathfrak{A}^{\mathrm{N}}_{\bar{I}}\right) \right\} \le 1/5 \,. \end{aligned} \tag{79}$$

This concludes the proof of Lemma 28 modulo the proofs of the above claims; we give those proofs in Subsection 9.1.1 below.

9.1.1 Proof of Claims 29 to 31

To prove Claim 29 we bound the probability of each of the bad events separately, starting with the "No"-case.

(i) Defining the event $B^{\mathrm{N}}_{\ell,\mathrm{size}}$ as

$$B^{\mathrm{N}}_{\ell,\mathrm{size}} = \{\Delta/\log N \le |I_\ell| \le \Delta \cdot (\log N)^2\} \,,$$

we can use a union bound to get $\Pr[B^{\mathrm{N}}_{\mathrm{size}}] \le \sum_{\ell=1}^{q} \Pr[B^{\mathrm{N}}_{\ell,\mathrm{size}}]$. For any fixed setting of I_ℓ there are $O(\log\log N)$ values of $\Delta \in \{\frac{N}{2^X} \mid X \in \{\frac{1}{3}\log N, \ldots, \frac{2}{3}\log N\}\}$ for which $\Delta/\log N \le I_\ell \le \Delta \cdot (\log N)^2$. Hence we have $\Pr[B^{\mathrm{N}}_{\ell,\mathrm{size}}] = O((\log\log N)/\log N)$, and consequently $\Pr[B^{\mathrm{N}}_{\mathrm{size}}] = O(q(\log\log N)/\log N)$.

(ii) Similarly, defining the event $B^{\mathrm{N}}_{\ell,\mathrm{boundary}}$ as

$$B^{\mathrm{N}}_{\ell,\mathrm{boundary}} = \{|I_\ell| < \Delta/\log N \text{ and } I_\ell \text{ intersects two blocks}\} \ ,$$

we have $\Pr[B^{\mathrm{N}}_{\mathrm{boundary}}] \leq \sum_{\ell=1}^{q} \Pr[B^{\mathrm{N}}_{\ell,\mathrm{boundary}}]$. For any fixed setting of I_ℓ, recalling the choice of a uniform random offset $y \in [N]$ for the blocks, we have that $\Pr[B^{\mathrm{N}}_{\ell,\mathrm{boundary}}] \leq O(1/\log N)$, and consequently $\Pr[B^{\mathrm{N}}_{\mathrm{boundary}}] = O(q/\log N)$.

(iii) The analysis of $B^{\mathrm{N}}_{\mathrm{middle}}$ is identical (by considering the midpoint of a block instead of its endpoint), yielding directly $\Pr[B^{\mathrm{N}}_{\mathrm{middle}}] = O(q/\log N)$.

(iv) Fix $\ell \in [q]$ and recall that $B^{\mathrm{N}}_{\ell,\mathrm{outer}} = \{\Delta \cdot (\log N)^2 < |I_\ell| \text{ and } s_\ell \text{ is drawn from a block} \subsetneq I_\ell\}$. Fix any outcome for Δ such that $\Delta \cdot (\log N)^2 < |I_\ell|$ and let us consider only the randomness over the draw of s_ℓ from I_ℓ. Since there are $\Omega((\log N)^2)$ blocks contained entirely in I_ℓ, the probability that s_ℓ is drawn from a block not contained entirely in I_ℓ (there are at most two such blocks, one at each end of I_ℓ) is $O(1/(\log N)^2)$. Hence we have $\Pr[B^{\mathrm{N}}_{\ell,\mathrm{outer}}] \leq O(1)/(\log N)^2$.

(v) Finally, recall that

$$B^{\mathrm{N}}_{\ell,\mathrm{collide}} = \{\Delta \cdot (\log N)^2 < |I_\ell| \text{ and } \exists j < \ell \text{ s.t. } s_\ell \text{ and } s_j \text{ belong to the same block } \} \ .$$

Fix $\ell \in [q]$ and a query interval I_ℓ. Let r_ℓ be the number of blocks in I_ℓ within which resides some previously sampled point s_j, $j \in [\ell - 1]$. Since there are $\Omega((\log N)^2)$ blocks in I_ℓ and $r_\ell \leq \ell - 1$, the probability that s_ℓ is drawn from a block containing any s_j, $j < \ell$, is $O(\ell/(\log N)^2)$. Hence we have $\Pr[B^{\mathrm{N}}_{\ell,\mathrm{collide}}] = O(\ell/(\log N)^2)$.

With these probability bounds for bad events in hand, we can prove Claim 29:

Proof of Claim 29: Recall that $q \leq \tau \cdot \frac{\log N}{\log \log N}$. Recalling the definition of $B^{\mathrm{N}}_{(\bar{I})}$, a union bound yields

$$\begin{aligned}
\Pr[B^{\mathrm{N}}_{(\bar{I})}] &\leq \Pr[B^{\mathrm{N}}_{\mathrm{size}}] + \Pr[B^{\mathrm{N}}_{\mathrm{boundary}}] + \Pr[B^{\mathrm{N}}_{\mathrm{middle}}] + \sum_{\ell=1}^{q} \Pr[B^{\mathrm{N}}_{\ell,\mathrm{outer}}] + \sum_{\ell=1}^{q} \Pr[B^{\mathrm{N}}_{\ell,\mathrm{collide}}] \\
&= O\left(\frac{q \cdot \log\log N}{\log N}\right) + O\left(\frac{q}{\log N}\right) + O\left(\frac{q}{\log N}\right) + \sum_{\ell=1}^{q} O\left(\frac{1}{(\log N)^2}\right) + \sum_{\ell=1}^{q} O\left(\frac{\ell}{(\log N)^2}\right) \\
&\leq \frac{1}{10} \ ,
\end{aligned}$$

where the last inequality holds for a sufficiently small choice of the absolute constant τ.

The same analysis applies unchanged for $\Pr[B^{\mathrm{Y}}_{\mathrm{size}}]$, $\Pr[B^{\mathrm{Y}}_{\mathrm{middle}}]$ and $\Pr[B^{\mathrm{Y}}_{\mathrm{boundary}}]$, using the "fake construction" view of $\mathcal{U}$ as described earlier. The arguments for $\Pr[B^{\mathrm{Y}}_{\ell,\mathrm{outer}}]$ and $\Pr[B^{\mathrm{Y}}_{\ell,\mathrm{collide}}]$ go through unchanged as well, and Claim 29 is proved. ■

Proof of Claim 30: Fix any $\bar{I} = (I_1, \ldots, I_q)$ and any transcript $\mathcal{T} = ((I_1, s_1), \ldots, (I_q, s_q))$. Recall that the length-ℓ partial transcript $\mathcal{T}|_\ell$ is defined to be $((I_1, s_1), \ldots, (I_\ell, s_\ell))$. We define the

random variables $Z^{\mathrm{N}}_{\bar{I},\ell}$ and $Z^{\mathrm{Y}}_{\bar{I},\ell}$ to be the length-ℓ prefixes of $Z^{\mathrm{N}}_{\bar{I}}$ and $Z^{\mathrm{Y}}_{\bar{I}}$ respectively. We prove Claim 30 by establishing the following, which we prove by induction on ℓ:

$$\Pr\left[Z^{\mathrm{N}}_{\bar{I},\ell} = \mathcal{T}|_\ell \;\middle|\; \overline{B^{\mathrm{N}}_{(\bar{I})}}\right] = \Pr\left[Z^{\mathrm{Y}}_{\bar{I},\ell} = \mathcal{T}|_\ell \;\middle|\; \overline{B^{\mathrm{Y}}_{(\bar{I})}}\right]. \tag{80}$$

For the base case, it is clear that (80) holds with $\ell = 0$. For the inductive step, suppose (80) holds for all $k \in [\ell - 1]$. When querying I_ℓ at the ℓ-th step, one of the following cases must hold (since we conditioned on the "bad events" not happening):

(1) I_ℓ is contained within a half-block (more precisely, either entirely within the first half of a block or entirely within the second half). In this case the "yes" and "no" distribution oracles behave exactly the same since both generate s_ℓ by sampling uniformly from I_ℓ.

(2) The point s_ℓ belongs to a block, contained entirely in I_ℓ, which is "fresh" in the sense that it contains no s_j, $j < \ell$. In the "No"-case this block may either be high-low or low-high; but since both outcomes have the same probability, there is another transcript with equal probability in which the two profiles are switched. Consequently (over the randomness in the draw of $D \sim \mathcal{P}_{\mathrm{No}}$) the probability of picking s_ℓ in the "No"-distribution case is the same as in the uniform distribution case (i.e., uniform on the fresh blocks contained in I_ℓ).

This concludes the proof of Claim 30. ■

Proof of Claim 31: Given Claims 29 and 30, Claim 31 is an immediate consequence of the following basic fact:

Fact 32 *Let D_1, D_2 be two distributions over the same finite set X. Let E_1, E_2, be two events such that $D_i[E_i] = \alpha_i \le \alpha$ for $i = 1, 2$ and the conditional distributions $(D_i)_{\overline{E_i}}$ are identical, i.e. $d_{\mathrm{TV}}((D_1)_{\overline{E_1}}, (D_2)_{\overline{E_2}}) = 0$. Then $d_{\mathrm{TV}}(D_1, D_2) \le \alpha$.*

Proof: We first observe that since $(D_2)_{\overline{E_2}}(E_2) = 0$ and $(D_1)_{\overline{E_1}}$ is identical to $(D_2)_{\overline{E_2}}$, it must be the case that $(D_1)_{\overline{E_1}}(E_2) = 0$, and likewise $(D_2)_{\overline{E_2}}(E_1) = 0$. This implies that $D_1(E_2 \setminus E_1) = D_2(E_1 \setminus E_2) = 0$. Now let us write

$$\begin{aligned} 2d_{\mathrm{TV}}(D_1, D_2) \;=\; & \sum_{x \in X \setminus (E_1 \cup E_2)} |D_1(x) - D_2(x)| + \sum_{x \in E_1 \cap E_2} |D_1(x) - D_2(x)| + \\ & \sum_{x \in E_1 \setminus E_2} |D_1(x) - D_2(x)| + \sum_{x \in E_2 \setminus E_1} |D_1(x) - D_2(x)|. \end{aligned}$$

We may upper bound $\sum_{x \in E_1 \cap E_2} |D_1(x) - D_2(x)|$ by $\sum_{x \in E_1 \cap E_2} (D_1(x) + D_2(x)) = D_1(E_1 \cap E_2) + D_2(E_1 \cap E_2)$, and the above discussion gives $\sum_{x \in E_1 \setminus E_2} |D_1(x) - D_2(x)| = D_1(E_1 \setminus E_2)$ and $\sum_{x \in E_2 \setminus E_1} |D_1(x) - D_2(x)| = D_2(E_2 \setminus E_1)$. We thus have

$$\begin{aligned} 2d_{\mathrm{TV}}(D_1, D_2) \;&\le\; \sum_{x \in X \setminus (E_1 \cup E_2)} |D_1(x) - D_2(x)| + D_1(E_1) + D_2(E_2) \\ &\le\; \sum_{x \in X \setminus (E_1 \cup E_2)} |D_1(x) - D_2(x)| + \alpha_1 + \alpha_2. \end{aligned}$$

Finally, since $d_{\mathrm{TV}}((D_1)_{\overline{E_1}}, (D_2)_{\overline{E_2}}) = 0$, we have

$$\begin{aligned}\sum_{x \in X \setminus (E_1 \cup E_2)} |D_1(x) - D_2(x)| &= |D_1(X \setminus (E_1 \cup E_2)) - D_2(X \setminus (E_1 \cup E_2))| \\ &= \left|D_1(\overline{E_1}) - D_2(\overline{E_2})\right| = |\alpha_1 - \alpha_2|.\end{aligned}$$

Thus $2d_{\mathrm{TV}}(D_1, D_2) \leq |\alpha_1 - \alpha_2| + \alpha_1 + \alpha_2 = 2\max\{\alpha_1, \alpha_2\} \leq 2\alpha$, and the fact is established. ■

This concludes the proof of Claim 31. ■

9.2 A lower bound against adaptive algorithms: Outline of the proof of Theorem 16

Throughout this subsection A denotes a general adaptive algorithm that makes $q \leq \tau \cdot \frac{\log N}{\log \log N}$ queries, where as before $\tau > 0$ is an absolute constant. Theorem 16 is a consequence of the following theorem, which deals with adaptive algorithms:

Theorem 18

$$\left|\Pr_{D \sim \mathcal{P}_{\mathrm{No}}}[A^{\mathsf{ICOND}_D} \textit{ outputs } \mathsf{ACCEPT}] - \Pr[A^{\mathsf{ICOND}_{\mathcal{U}}} \textit{ outputs } \mathsf{ACCEPT}]\right| \leq 1/5. \tag{81}$$

The idea here is to extend the previous analysis for non-adaptive algorithms, and argue that "adaptiveness does not really help" to distinguish between $D = \mathcal{U}$ and $D \sim \mathcal{P}_{\mathrm{No}}$ given access to ICOND_D.

As in the non-adaptive case, in order to prove Theorem 18, it is sufficient to prove that the transcripts for uniform and "No"-distributions are close in total variation distance; i.e., that

$$d_{\mathrm{TV}}\left(\mathfrak{A}^{\mathrm{Y}}, \mathfrak{A}^{\mathrm{N}}\right) \leq 1/5. \tag{82}$$

The key idea used to prove this will be to introduce a *sequence* $\mathfrak{A}_{\mathrm{otf}}^{(k),\mathrm{N}}$ of distributions over transcripts (where "otf" stands for "on the fly"), for $0 \leq k \leq q$, such that (i) $\mathfrak{A}_{\mathrm{otf}}^{(0),\mathrm{N}} = \mathfrak{A}^{\mathrm{Y}}$ and $\mathfrak{A}_{\mathrm{otf}}^{(q),\mathrm{N}} = \mathfrak{A}^{\mathrm{N}}$, and (ii) the distance $d_{\mathrm{TV}}\left(\mathfrak{A}_{\mathrm{otf}}^{(k),\mathrm{N}}, \mathfrak{A}_{\mathrm{otf}}^{(k+1),\mathrm{N}}\right)$ for each $0 \leq k \leq q-1$ is "small". This will enable us to conclude by the triangle inequality, as

$$d_{\mathrm{TV}}\left(\mathfrak{A}^{\mathrm{N}}, \mathfrak{A}^{\mathrm{Y}}\right) = d_{\mathrm{TV}}\left(\mathfrak{A}_{\mathrm{otf}}^{(0),\mathrm{N}}, \mathfrak{A}_{\mathrm{otf}}^{(q),\mathrm{N}}\right) \leq \sum_{k=0}^{q-1} d_{\mathrm{TV}}\left(\mathfrak{A}_{\mathrm{otf}}^{(k),\mathrm{N}}, \mathfrak{A}_{\mathrm{otf}}^{(k+1),\mathrm{N}}\right). \tag{83}$$

To define this sequence, in the next subsection we will introduce the notion of an *extended transcript*, which in addition to the queries and samples includes additional information about the "local structure" of the distribution at the endpoints of the query intervals and the sample points. Intuitively, this extra information will help us analyze the interaction between the adaptive algorithm and the oracle. We will then describe an alternative process according to which a "faking algorithm" (reminiscent of the similar notion from Subsection 5.2) can interact with an oracle to generate such an extended transcript. More precisely, we shall define a sequence of such faking algorithms,

paramaterized by "how much faking" they perform. For both the original ("non-faking") algorithm A and for the faking algorithms, we will show how extended transcripts can be generated "on the fly". The aforementioned distributions $\mathfrak{A}_{\mathrm{otf}}^{(k),\mathrm{N}}$ over (regular) transcripts are obtained by *truncating* the extended transcripts that are generated on the fly (i.e., discarding the extra information), and we shall argue that they satisfy requirements (i) and (ii) above.

Before turning to the precise definitions and the analysis of extended transcripts and faking algorithms, we provide the following variant of Fact 32, which will come in handy when we bound the right hand side of Equation (83).

Fact 33 *Let D_1, D_2 be two distributions over the same finite set X. Let E be an event such that $D_i[E] = \alpha_i \leq \alpha$ for $i = 1, 2$ and the conditional distributions $(D_1)_{\overline{E}}$ and $(D_2)_{\overline{E}}$ are statistically close, i.e. $d_{\mathrm{TV}}((D_1)_{\overline{E}}, (D_2)_{\overline{E}}) = \beta$. Then $d_{\mathrm{TV}}(D_1, D_2) \leq \alpha + \beta$.*

Proof: As in the proof of Fact 32, let us write

$$2d_{\mathrm{TV}}(D_1, D_2) = \sum_{x \in X \setminus E} |D_1(x) - D_2(x)| + \sum_{x \in E} |D_1(x) - D_2(x)|.$$

We may upper bound $\sum_{x\in E} |D_1(x) - D_2(x)|$ by $\sum_{x\in E}(D_1(x) + D_2(x)) = D_1(E) + D_2(E) = \alpha_1 + \alpha_2$; furthermore,

$$\begin{aligned}
\sum_{x \in \bar{E}} |D_1(x) - D_2(x)| &= \sum_{x \in \bar{E}} \left|(D_1)_{\bar{E}}(x) \cdot D_1(\bar{E}) - (D_2)_{\bar{E}}(x) \cdot D_2(\bar{E})\right| \\
&\leq D_1(\bar{E}) \cdot \sum_{x \in \bar{E}} |(D_1)_{\bar{E}}(x) - (D_2)_{\bar{E}}(x)| + \left|D_1(\bar{E}) - D_2(\bar{E})\right| \cdot (D_2)_{\bar{E}}(\bar{E}) \\
&\leq (1 - \alpha_1) \cdot (2\beta) + |\alpha_2 - \alpha_1| \cdot 1 \leq 2\beta + |\alpha_2 - \alpha_1|
\end{aligned}$$

Thus $2d_{\mathrm{TV}}(D_1, D_2) \leq 2\beta + |\alpha_1 - \alpha_2| + \alpha_1 + \alpha_2 = 2\beta + 2\max\{\alpha_1, \alpha_2\} \leq 2(\alpha + \beta)$, and the fact is established. ■

9.3 Extended transcripts and drawing $D \sim \mathcal{P}_{\mathrm{No}}$ on the fly.

Observe that the testing algorithm, seeing only pairs of queries and answers, does not have direct access to all the underlying information – namely, in the case of a "No"-distribution, whether the profile of the block that the sample point comes from is $\downarrow\uparrow$ or $\uparrow\downarrow$. It will be useful for us to consider an "extended" version of the transcripts, which includes this information along with information about the profile of the "boundary" blocks for each queried interval, even though this information is not directly available to the algorithm.

Definition 8 *With the same notation as in Definition 7, the* extended transcript *of a sequence of queries made by A and the corresponding responses is a sequence $\mathcal{E} = (I_\ell, s_\ell, b_\ell)_{\ell \in [q]}$ of triples, where I_ℓ and s_ℓ are as before, and $b_\ell = (b_\ell^L, b_\ell^{\mathrm{samp}}, b_\ell^R) \in \{\downarrow\uparrow, \uparrow\downarrow\}^3$ is a triple defined as follows: Let $B_{i^L}, \ldots, B_{i^R}$ be the blocks that I_ℓ intersects, going from left to right. Then*

1. b_ℓ^L *is the profile of the block* B_{i^L}*;*

2. b_ℓ^R *is the profile of the block* B_{i^R}*;*

3. b_ℓ^{samp} *is the profile of the block* $B_\ell \in \{B_{i^L}, \ldots, B_{i^R}\}$ *that* s_ℓ *belongs to.*

We define $\mathcal{E}|_k$ *to be the length-*k *prefix of an extended transcript* $\mathcal{E}$*.*

As was briefly discussed prior to the current subsection, we shall be interested in considering algorithms that *fake* some answers to their queries. Specifically, given an adaptive algorithm A, we define $A^{(1)}$ as the algorithm that *fakes* its first query, in the following sense: If the first query made by A to the oracle is some interval I, then the algorithm $A^{(1)}$ does not call ICOND on I but instead chooses a point s uniformly at random from I and then behaves exactly as A would behave if the ICOND oracle had returned s in response to the query I. More generally, we define $A^{(k)}$ for all $0 \le k \le q$ as the algorithm behaving like A but faking its first k queries (note that $A^{(0)} = A$).

In Subsection 9.3.1 we explain how extended transcripts can be generated for $A^{(0)} = A$ in an "on the fly" fashion so that the resulting distribution over extended transcripts is the same as the one that would result from first drawing D from $\mathcal{P}_{\text{No}}$ and then running algorithm A on it. It follows that when we remove the extension to the transcript so as to obtain a regular transcript, we get a distribution over transcripts that is identical to $\mathfrak{A}^{\text{N}}$. In Subsection 9.3.2 we explain how to generate extended transcripts for $A^{(k)}$ where $0 \le k \le q$. We note that for $k \ge 1$ the resulting distribution over extended transcripts is *not* the same as the one that would result from first drawing D from $\mathcal{P}_{\text{No}}$ and then running algorithm $A^{(k)}$ on it. However, this is not necessary for our purposes. For our purposes it is sufficient that the distributions corresponding to pairs of consecutive indices $(k, k+1)$ are similar (including the pair $(0, 1)$), and that for $k = q$ the distribution over regular transcripts obtained by removing the extension to the transcript is identical to $\mathfrak{A}^{\text{Y}}$.

9.3.1 Extended transcripts for $A = A^{(0)}$

Our proof of Equation (82) takes advantage of the fact that one can view the draw of a "No"-distribution from $\mathcal{P}_{\text{No}}$ as being done "on the fly" during the course of algorithm A's execution. First, the size Δ and the offset y are drawn at the very beginning, but we may view the profile vector ϑ as having its components chosen independently, coordinate by coordinate, only as A interacts with ICOND – each time an element s_ℓ is obtained in response to the ℓ-th query I_ℓ, only then are the elements of the profile vector ϑ corresponding to the three coordinates of b_ℓ chosen (if they were not already completely determined by previous calls to ICOND). More precise details follow.

Consider the ℓ-th query I_ℓ that A makes to ICOND_D. Inductively some coordinates of ϑ may have been already set by previous queries. Let $B_{i^L}, \ldots, B_{i^R}$ be the blocks that I_ℓ intersects. First, if the coordinate of ϑ corresponding to block B_{i^L} was not already set by a previous query, a fair coin is tossed to choose a setting from $\{\downarrow\uparrow, \uparrow\downarrow\}$ for this coordinate. Likewise, if the coordinate of ϑ corresponding to block B_{i^R} was not already set (either by a previous query or because $i^R = i^L$), a fair coin is tossed to choose a setting from $\{\downarrow\uparrow, \uparrow\downarrow\}$ for this coordinate.

At this point, the values of b_ℓ^L and b_ℓ^R have been set. A simple but important observation is that these outcomes of b_ℓ^L and b_ℓ^R completely determine the probabilities (call them α^L and α^R

respectively) that the block B_ℓ from which s_ℓ will be chosen is B_{i^L} (is B_{i^R} respectively), as we explain in more detail next. If $i^R = i^L$ then there is no choice to be made, and so assume that $i^R > i^L$. For $K \in \{L, R\}$ let $\rho_1^K \cdot \Delta$ be the size of the intersection of I_ℓ with the first (left) half of B_{i^K} and let $\rho_2^K \cdot \Delta$ be the size of the intersection of I_ℓ with the second (right) half of B_{i^K}. Note that $0 < \rho_1^K + \rho_2^K \leq 1$ and that $\rho_1^L = 0$ when $\rho_2^L \leq 1/2$ and similarly $\rho_2^R = 0$ when $\rho_1^R \leq 1/2$. If $b_\ell^K = \uparrow\downarrow$ then let $w^K = \rho_1^K \cdot (1 + 2\epsilon) + \rho_2^K \cdot (1 - 2\epsilon) = \rho_1^K + \rho_2^K + 2\epsilon(\rho_1^K - \rho_2^K)$, and if $b_\ell^K = \downarrow\uparrow$ then let $w^K = \rho_1^K + \rho_2^K - 2\epsilon(\rho_1^K - \rho_2^K)$. We now set $\alpha^K = \frac{w^K}{w^L + w^R + (i^L - i^R - 1)}$. The block B_{i^L} is selected with probability α^L, the block B_{i^R} is selected with probability α^R, and for $i^R \geq i^L + 2$, each of the other blocks is selected with equal probability, $\frac{1}{w^L + w^R + (i^L - i^R - 1)}$.

Given the selection of the block B_ℓ as described above, the element s_ℓ and the profile b_ℓ^{samp} of the block to which it belongs are selected as follows. If the coordinate of ϑ corresponding to B_ℓ has already been determined, then b_ℓ^{samp} is set to this value and s_ℓ is drawn from B_ℓ as determined by the $\downarrow\uparrow$ or $\uparrow\downarrow$ setting of b_ℓ^{samp}. Otherwise, a fair coin is tossed, b_ℓ^{samp} is set either to $\downarrow\uparrow$ or to $\uparrow\downarrow$ depending on the outcome, and s_ℓ is drawn from B_ℓ as in the previous case (as determined by the setting of b_ℓ^{samp}). Now all of I_ℓ, s_ℓ, and $b_\ell = (b_\ell^L, b_\ell^{\text{samp}}, b_\ell^R)$ have been determined and the triple (I_ℓ, s_ℓ, b_ℓ) is taken as the ℓ-th element of the extended transcript.

We now define $\mathfrak{A}_{\text{otf}}^{(0),\text{N}}$ as follows. A draw from this distribution over (non-extended) transcripts is obtained by first drawing an extended transcript $(I_1, s_1, b_1), \ldots, (I_q, s_q, b_q)$ from the on-the-fly process described above, and then removing the third element of each triple to yield $(I_1, s_1), \ldots, (I_q, s_q)$. This is exactly the distribution over transcripts that is obtained by first drawing D from $\mathcal{P}_{\text{No}}$ and then running A on it.

9.3.2 Extended transcripts for $A^{(k)}$, $k \geq 0$

In this subsection we define the distribution $\mathfrak{A}_{\text{otf}}^{(k),N}$ for $0 \leq k \leq q$ (the definition we give below will coincide with our definition from the previous subsection for $k = 0$). Here too the size Δ and the offset y are drawn at the very beginning, and the coordinates of the profile vector ϑ are chosen on the fly, together with the sample points. For each $\ell > k$, the pair (s_ℓ, b_ℓ) is selected exactly as was described for A, conditioned on the length-k prefix of the extended transcript and the new query I_ℓ (as well as the choice of (Δ, y)). It remains to explain how the selection is made for $1 \leq \ell \leq k$.

Consider a value $1 \leq \ell \leq k$ and the ℓ-th query interval I_ℓ. As in our description of the "on-the-fly" process for A, inductively some coordinates of ϑ may have been already set by previous queries. Let $B_{i^L}, \ldots, B_{i^R}$ be the blocks that I_ℓ intersects. As in the process for A, if the coordinate of ϑ corresponding to block B_{i^L} was not already set by a previous query, a fair coin is tossed to choose a setting from $\{\downarrow\uparrow, \uparrow\downarrow\}$ for this coordinate. Likewise, if the coordinate of ϑ corresponding to block B_{i^R} was not already set (either by a previous query or because $i^L = i^R$), a fair coin is tossed to choose a setting from $\{\downarrow\uparrow, \uparrow\downarrow\}$ for this coordinate. Hence, b_ℓ^L and b_ℓ^R are set exactly the same as described for A.

We now explain how to set the probabilities α^L and α^R of selecting the block B_ℓ (from which s_ℓ is chosen) to be B_{i^L} and B_{i^R}, respectively. Since the "faking" process should choose s_ℓ to be a uniform point from I_ℓ, the probability α^L is simply $|B_{i^L} \cap I_\ell|/|I_\ell|$, and similarly for α^R. (If $i^L = i^R$ we take $\alpha^L = 1$ and $\alpha^R = 0$.) Thus the values of α^L and α^R are completely determined by the

number of blocks j and the relative sizes of the intersection of I_ℓ with B_{i^L} and with B_{i^R}. Now, with probability α^L the block B_ℓ is chosen to be B_{i^L}, with probability α^R it is chosen to be B_{i^R} and with probability $1-\alpha^L-\alpha^R$ it is chosen uniformly among $\{B_{i^L+1},\ldots,B_{i^R-1}\}$.

Given the selection of the block B_ℓ as described above, s_ℓ is chosen to be a uniform random element of $B_\ell \cap I_\ell$. The profile b_ℓ^{samp} of B_ℓ is selected as follows:

1. If the coordinate of ϑ corresponding to B_ℓ has already been determined (either by a previous query or because $B_\ell \in \{B_{i^L}, B_{i^R}\}$), then b_ℓ^{samp} is set accordingly.
2. Otherwise, the profile of B_ℓ was not already set; note that in this case it must hold that $B_\ell \notin \{B_{i^L}, B_{i^R}\}$. We look at the half of B_ℓ that s_ℓ belongs to, and toss a biased coin to set its profile $b_\ell^{\text{samp}} \in \{\downarrow\uparrow, \uparrow\downarrow\}$: If s_ℓ belongs to the first half, then the coin toss's probabilities are $((1-2\epsilon)/2, (1+2\epsilon)/2)$; otherwise, they are $((1+2\epsilon)/2, (1-2\epsilon)/2)$.

Let $\mathfrak{E}_{\text{otf}}^{(k),\text{N}}$ denote the distribution induced by the above process over extended transcripts, and let $\mathfrak{A}_{\text{otf}}^{(k),\text{N}}$ be the corresponding distribution over regular transcripts (that is, when removing the profiles from the transcript). As noted in Subsection 9.3.1, for $k=0$ we have that $\mathfrak{A}_{\text{otf}}^{(0),\text{N}} = \mathfrak{A}^{\text{N}}$. In the other extreme, for $k=q$, since each point s_ℓ is selected uniformly in I_ℓ (with no dependence on the selected profiles) we have that $\mathfrak{A}_{\text{otf}}^{(q),\text{N}} = \mathfrak{A}^{\text{Y}}$. In the next subsection we bound the total variation distance between $\mathfrak{A}_{\text{otf}}^{(k),\text{N}}$ and $\mathfrak{A}_{\text{otf}}^{(k+1),\text{N}}$ for every $0 \le k \le q-1$ by bounding the distance between the corresponding distributions $\mathfrak{E}_{\text{otf}}^{(k),\text{N}}$ and $\mathfrak{E}_{\text{otf}}^{(k+1),\text{N}}$. Roughly speaking, the only difference between the two (for each $0 \le k \le q-1$) is in the distribution over $(s_{k+1}, b_{k+1}^{\text{samp}})$. As we argue in more detail and formally in the next subsection, conditioned on certain events (determined, among other things, by the choice of (Δ, y)), we have that $(s_{k+1}, b_{k+1}^{\text{samp}})$ are distributed the same under $\mathfrak{E}_{\text{otf}}^{(k),\text{N}}$ and $\mathfrak{E}_{\text{otf}}^{(k+1),\text{N}}$.

9.4 Bounding $d_{\text{TV}}\left(\mathfrak{A}_{\text{otf}}^{(k),\text{N}}, \mathfrak{A}_{\text{otf}}^{(k+1),\text{N}}\right)$

As per the foregoing discussion, we can focus on bounding the total variation distance between extended transcripts

$$d_{\text{TV}}\left(\mathfrak{E}_{\text{otf}}^{(k),\text{N}}, \mathfrak{E}_{\text{otf}}^{(k+1),\text{N}}\right)$$

for arbitrary fixed $k \in \{0,\ldots,q-1\}$. Before diving into the proof, we start by defining the probability space we shall be working in, as well as explaining the different sources of randomness that are in play and how they fit into the random processes we end up analyzing.

The probability space. Recall the definition of an extended transcript: for notational convenience, we reserve the notation $\mathcal{E} = (I_\ell, s_\ell, b_\ell)_{\ell\in[q]}$ for extended transcript valued random variables, and will write $E = (\iota_\ell, \sigma_\ell, \pi_\ell)_{\ell\in[q]}$ for a fixed outcome. We denote by Σ the space of all such tuples E, and by Λ the set of all possible outcomes for (Δ, y). The sample space we are considering is now defined as $X \stackrel{\text{def}}{=} \Sigma \times \Lambda$: that is, an extended transcript along with the underlying choice of block size and offset[10]. The two probability measures on X we shall consider will be induced by the execution of

[10] We emphasize the fact that the algorithm, whether faking or not, has access neither to the "extended" part of the transcript nor to the choice of (Δ, y); however, these elements are part of the events we analyze.

$A^{(k)}$ and $A^{(k+1)}$, as per the process detailed below.

A key thing to observe is that, as we focus on two "adjacent" faking algorithms $A^{(k)}$ and $A^{(k+1)}$, it will be sufficient to consider the following equivalent view of the way an extended transcript is generated:

1. up to (and including) stage k, the faking algorithm generates on its own both the queries ι_ℓ and the uniformly distributed samples $\sigma_\ell \in \iota_\ell$; it also chooses its $(k+1)$-st query ι_{k+1};
2. then, at that point only is the choice of (Δ, y) made; and the profiles π_ℓ $(1 \leq \ell \leq k)$ of the *previous* blocks decided upon, as described in Subsection 9.3;
3. after this, the sampling and block profile selection is made exactly according to the previous "on-the-fly process" description.

The reason that we can defer the choice of (Δ, y) and the setting of the profiles in the manner described above is the following: For both $A^{(k)}$ and $A^{(k+1)}$, the choice of each σ_ℓ for $1 \leq \ell \leq k$ depends only on ι_ℓ and the choice of each ι_ℓ for $1 \leq \ell \leq k+1$ depends only on $(\iota_1, \sigma_1), \ldots, (\iota_{\ell-1}, \sigma_{\ell-1})$. That is, there is no dependence on (Δ, y) nor on any $\pi_{\ell'}$ for $\ell' \leq \ell$. By deferring the choice of the pair (Δ, y) we may consider the randomness coming in its draw only at the $(k+1)$-st stage (which is the pivotal stage here). Note that, both for $A^{(k)}$ and $A^{(k+1)}$, the resulting distribution over X induced by the description above exactly matches the one from the "on-the-fly" process. In the next paragraph, we go into more detail, and break down further the randomness and choices happening in this new view.

Sources of randomness. To define the probability measure on this space, we describe the process that, up to stage $k+1$, generates the corresponding part of the extended transcript and the (Δ, y) for $A^{(m)}$ (where $m \in \{k, k+1\}$) (see the previous subsections for precise descriptions of how the following random choices are made):

(R1) $A^{(m)}$ draws $\iota_1, \sigma_1, \ldots, \iota_k, \sigma_k$ and finally ι_{k+1} by itself;

(R2) the outcome of (Δ, y) is chosen: this "retroactively" fixes the partition of the ι_ℓ's $(1 \leq \ell \leq k+1)$ into blocks $B_{i_L}^{(\ell)}, \ldots, B_{i_R}^{(\ell)}$;

(R3) the profiles of $B_{i_L}^{(\ell)}$, $B_{i_R}^{(\ell)}$ and B_ℓ (i.e., the values of the triples π_ℓ, for $1 \leq \ell \leq k$) are drawn;

(R4) the profiles of $B_{i_L}^{(k+1)}$, $B_{i_R}^{(k+1)}$ are chosen;

(R5) the block selection (choice of the block B_{k+1} to which σ_{k+1} will belong to) is made:

   (a) whether it will be one of the two end blocks, or one of the inner ones (for $A^{(k+1)}$ this is based on the respective sizes of the end blocks, and for $A^{(k)}$ this is based on the weights of the end blocks, using the profiles of the end blocks);

   (b) the choice itself:

   - if one of the outer ones, draw it based on either the respective sizes (for $A^{(k+1)}$) or the respective weights (for $A^{(k)}$, using the profiles of the end blocks)

- if one of the inner ones, uniformly at random among all inner blocks;

(R6) the sample σ_{k+1} and the profile π^{samp}_{k+1} are chosen;

(R7) the rest of the transcript, for $k+1, \ldots, q$, is iteratively chosen (in the same way for $A^{(k)}$ and $A^{(k+1)}$) according to the on-the-fly process discussed before.

Note that the only differences between the processes for $A^{(k)}$ and $A^{(k+1)}$ lie in steps (R5a), (R5b) and (R6) of the $(k+1)$-st stage.

Bad events and outline of the argument

Let $G(\iota_{k+1})$ (where 'G' stands for 'Good') denote the settings of (Δ, y) that satisfy the following: Either (i) $|\iota_{k+1}| > \Delta \cdot (\log N)^2$ or (ii) $|\iota_{k+1}| < \Delta / \log N$ and ι_{k+1} is contained entirely within a single half block. We next define three indicator random variables for a given element $\omega = (E, (\Delta, y))$ of the sample space X, where $E = ((\iota_1, \sigma_1, \pi_1), \ldots, (\iota_q, \sigma_q, \pi_q))$. The first, Γ_1, is zero when $(\Delta, y) \notin G(\iota_{k+1})$. Note that the randomness for Γ_1 is over the choice of (Δ, y) and the choice of ι_{k+1}. The second, Γ_2, is zero when ι_{k+1} intersects at least two blocks and the block B_{k+1} is one of the two extreme blocks intersected by ι_{k+1}. The third, Γ_3, is zero when ι_{k+1} is not contained entirely within a single half block and B_{k+1} is a block whose profile had already been set (either because it contains a selected point σ_ℓ for $\ell \leq k$ or because it belongs to one of the two extreme blocks for some queried interval ι_ℓ for $\ell \leq k$). For notational ease we write $\overline{\Gamma}(E)$ to denote the triple $(\Gamma_1, \Gamma_2, \Gamma_3)$. Observe that these indicator variables are well defined, and correspond to events that are indeed subsets of our space X: given any element $\omega \in X$, whether $\Gamma_i(\omega) = 1$ (for $i \in \{1,2,3\}$) is fully determined.

Define D_1, D_2 as the two distributions over X induced by the executions of respectively $A^{(k)}$ and $A^{(k+1)}$ (in particular, by only keeping the first marginal of D_1 we get back $\mathfrak{E}^{(k),\mathrm{N}}$). Applying Fact 33 to D_1 and D_2, we obtain that

$$\begin{aligned} d_{\mathrm{TV}}(D_1, D_2) &\leq \Pr\big[\,\overline{\Gamma} \neq (1,1,1)\,\big] + d_{\mathrm{TV}}\big(D_1 \mid \overline{\Gamma} = (1,1,1), D_2 \mid \overline{\Gamma} = (1,1,1)\big) \\ &\leq \Pr[\,\Gamma_1 = 0\,] + \Pr[\,\Gamma_2 = 0 \mid \Gamma_1 = 1\,] + \Pr[\,\Gamma_3 = 0 \mid \Gamma_1 = \Gamma_2 = 1\,] \\ &\quad + d_{\mathrm{TV}}\big(D_1 \mid \overline{\Gamma} = (1,1,1), D_2 \mid \overline{\Gamma} = (1,1,1)\big)\,. \end{aligned} \tag{84}$$

To conclude, we can now deal which each of these 4 summands separately:

Claim 34 *We have that* $\Pr[\,\Gamma_1 = 0\,] \leq \eta(N)$, *where* $\eta(N) = O\left(\frac{\log\log N}{\log N}\right)$.

Proof: Similarly to the proof of Claim 29, for any fixed setting of ι_{k+1}, there are $O(\log\log N)$ values of $\Delta \in \left\{ \frac{N}{2^j} \;\middle|\; j \in \{\frac{1}{3}\log N, \ldots, \frac{2}{3}\log N\} \right\}$ for which $\Delta/\log N \leq \iota_{k+1} \leq \Delta \cdot (\log N)^2$. Therefore, the probability that one of these ("bad") values of Δ is selected is $O\left(\frac{\log\log N}{\log N}\right)$. If the choice of Δ is such that $|\iota_{k+1}| < \Delta/\log N$, then, by the choice of the random offset y, the probability that ι_{k+1} is not entirely contained within a single half block is $O(1/\log N)$. The claim follows. ■

Claim 35 *We have that* $\Pr[\,\Gamma_2 = 0 \mid \Gamma_1 = 1\,] \leq \eta(N)$.

Proof: If $\Gamma_1 = 1$ because $|\iota_{k+1}| < \Delta/(\log N)^2$ and ι_{k+1} is entirely contained within a single half block, then $\Gamma_2 = 1$ (with probability 1). Otherwise, $|\iota_{k+1}| > \Delta \cdot (\log N)^2$, so that ι_{k+1} intersects at least $(\log N)^2$ blocks. The probability that one of the two extreme blocks is selected is hence $O(1/(\log N)^2)$, and the claim follows. ■

Claim 36 *We have that* $\Pr[\,\Gamma_3 = 0 \mid \Gamma_1 = \Gamma_2 = 1\,] \leq \eta(N)$.

Proof: If $\Gamma_1 = 1$ because $|\iota_{k+1}| < \Delta/(\log N)^2$ and ι_{k+1} is entirely contained within a single half block, then $\Gamma_3 = 1$ (with probability 1). Otherwise, $|\iota_{k+1}| > \Delta \cdot (\log N)^2$, so that ι_{k+1} intersects at least $(\log N)^2$ blocks. Since $\Gamma_2 = 1$, the block B_{k+1} is uniformly selected from $(\log N)^2 - 2$ non-extreme blocks. Among them there are at most $3k = O\Big(\frac{\log N}{\log\log N}\Big)$ blocks whose profiles were already set. The probability that one of them is selected (so that $\Gamma_3 = 1$) is $O\Big(\frac{1}{\log N \log\log N}\Big) = O\Big(\frac{\log\log N}{\log N}\Big)$, and the claim follows. ■

We are left with only the last term, $d_{\mathrm{TV}}\big(D_1 \mid \overline{\Gamma} = (1,1,1), D_2 \mid \overline{\Gamma} = (1,1,1)\big)$. But as we are now ruling out all the "bad events" that would induce a difference between the distributions of the extended transcripts under $A^{(k)}$ and $A^{(k+1)}$, it becomes possible to argue that this distance is actually zero:

Claim 37 $d_{\mathrm{TV}}\big(D_1 \mid \overline{\Gamma} = (1,1,1), D_2 \mid \overline{\Gamma} = (1,1,1)\big) = 0$.

Proof: Unrolling the definition, we can write $d_{\mathrm{TV}}\big(D_1 \mid \overline{\Gamma} = (1,1,1), D_2 \mid \overline{\Gamma} = (1,1,1)\big)$ as

$$\sum_{E,(\Delta,y)} \left| \Pr\Big[\, \mathcal{E}^{(k)} = E, \mathcal{Y}^{(m)} = (\Delta,y) \;\Big|\; \overline{\Gamma} = (1,1,1) \,\Big] - \Pr\Big[\, \mathcal{E}^{(k+1)} = E, \mathcal{Y}^{(m)} = (\Delta,y) \;\Big|\; \overline{\Gamma} = (1,1,1) \,\Big] \right| .$$

where $\mathcal{Y}^{(m)}$ denotes the Λ-valued random variable corresponding to $A^{(m)}$. In order to bound this sum, we will show that each of its terms is zero: i.e., that for any fixed $(E,(\Delta,y)) \in \Sigma \times \Lambda$ we have

$$\Pr\Big[\, \mathcal{E}^{(k)} = E, \mathcal{Y}^{(k)} = (\Delta,y) \;\Big|\; \overline{\Gamma} = (1,1,1) \,\Big] = \Pr\Big[\, \mathcal{E}^{(k+1)} = E, \mathcal{Y}^{(k+1)} = (\Delta,y) \;\Big|\; \overline{\Gamma} = (1,1,1) \,\Big] .$$

We start by observing that, for $m \in \{k, k+1\}$,

$$\begin{aligned} &\Pr\Big[\, \mathcal{E}^{(m)} = E, \mathcal{Y}^{(m)} = (\Delta,y) \;\Big|\; \overline{\Gamma} = (1,1,1) \,\Big] \\ &\qquad = \Pr\Big[\, \mathcal{E}^{(m)} = E \;\Big|\; \overline{\Gamma} = (1,1,1), \mathcal{Y}^{(m)} = (\Delta,y) \,\Big] \Pr\Big[\, \mathcal{Y}^{(m)} = (\Delta,y) \;\Big|\; \overline{\Gamma} = (1,1,1) \,\Big] \end{aligned}$$

and that the term $\Pr\big[\, \mathcal{Y}^{(m)} = (\Delta,y) \;\big|\; \overline{\Gamma} = (1,1,1) \,\big] = \Pr\big[\, \mathcal{Y}^{(m)} = (\Delta,y) \,\big]$ is identical for $m = k$ and $m = k+1$. Therefore, it is sufficient to show that

$$\Pr\Big[\, \mathcal{E}^{(k)} = E \;\Big|\; \overline{\Gamma} = (1,1,1), \mathcal{Y}^{(k)} = (\Delta,y) \,\Big] = \Pr\Big[\, \mathcal{E}^{(k+1)} = E \;\Big|\; \overline{\Gamma} = (1,1,1), \mathcal{Y}^{(k+1)} = (\Delta,y) \,\Big] .$$

Let $\omega = (E,(\Delta,y)) \in X$ be arbitrary, with $E = ((\iota_1,\sigma_1,\pi_1),\ldots,(\iota_q,\sigma_q,\pi_q)) \in \Sigma$, and let $m \in \{k,k+1\}$. We can express $\Phi^{(m)}(\omega) \stackrel{\text{def}}{=} \Pr\big[\, \mathcal{E}^{(m)} = E \;\big|\; \overline{\Gamma} = (1,1,1), \mathcal{Y}^{(m)} = (\Delta,y) \,\big]$ as the product of the following 5 terms:

(T1) $p_k^{(m),\mathrm{int},\mathrm{samp}}(\omega)$, defined as

$$\begin{aligned} p_k^{(m),\mathrm{int},\mathrm{samp}}(\omega) &\stackrel{\mathrm{def}}{=} \Pr\Big[\, \mathcal{E}^{(m),\mathrm{int},\mathrm{samp}}|_k = E^{\mathrm{int},\mathrm{samp}}|_k \;\Big|\; \overline{\Gamma} = (1,1,1), \mathcal{Y}^{(m)} = (\Delta, y) \,\Big] \\ &= \Pr\Big[\, \mathcal{E}^{(m),\mathrm{int},\mathrm{samp}}|_k = E^{\mathrm{int},\mathrm{samp}}|_k \,\Big] \;, \end{aligned}$$

where $E_\ell^{\mathrm{int},\mathrm{samp}}$ denotes $(\iota_\ell, \sigma_\ell)$ and $E^{\mathrm{int},\mathrm{samp}}|_k$ denotes $(E_1^{\mathrm{int},\mathrm{samp}}, \ldots, E_k^{\mathrm{int},\mathrm{samp}})$;

(T2) $p_k^{(m),\mathrm{prof}}(\omega)$, defined as

$$\begin{aligned} p_k^{(m),\mathrm{prof}}(\omega) &\stackrel{\mathrm{def}}{=} \Pr\Big[\, \mathcal{E}^{(m),\mathrm{prof}}|_k = E^{\mathrm{prof}}|_k \;\Big|\; \mathcal{E}^{(m),\mathrm{int},\mathrm{samp}}|_k = E^{\mathrm{int},\mathrm{samp}}|_k, \overline{\Gamma} = (1,1,1), \mathcal{Y}^{(m)} = (\Delta, y) \,\Big] \\ &= \Pr\Big[\, \mathcal{E}^{(m),\mathrm{prof}}|_k = E^{\mathrm{prof}}|_k \;\Big|\; \mathcal{E}^{(m),\mathrm{int},\mathrm{samp}}|_k = E^{\mathrm{int},\mathrm{samp}}|_k, \mathcal{Y}^{(m)} = (\Delta, y) \,\Big] \;, \end{aligned}$$

where $E^{\mathrm{prof}}|_k$ denotes $(\pi_1, \ldots, \pi_k)$;

(T3) $p_{k+1}^{(m),\mathrm{int}}(\omega)$, defined as

$$\begin{aligned} p_{k+1}^{(m),\mathrm{int}}(\omega) &\stackrel{\mathrm{def}}{=} \Pr\Big[\, I_{k+1} = \iota_{k+1} \;\Big|\; \mathcal{E}^{(m),\mathrm{int},\mathrm{samp}}|_k = E^{\mathrm{int},\mathrm{samp}}|_k, \overline{\Gamma} = (1,1,1), \mathcal{Y}^{(m)} = (\Delta, y) \,\Big] \\ &= \Pr\Big[\, I_{k+1} = \iota_{k+1} \;\Big|\; \mathcal{E}^{(m),\mathrm{int},\mathrm{samp}}|_k = E^{\mathrm{int},\mathrm{samp}}|_k \,\Big] \;; \end{aligned}$$

(T4) $p_{k+1}^{(m),\mathrm{samp},\mathrm{prof}}(\omega)$, defined as

$$\begin{aligned} &p_{k+1}^{(m),\mathrm{samp},\mathrm{prof}}(\omega) \\ &\quad\stackrel{\mathrm{def}}{=} \Pr\Big[\, (s_{k+1}, b_{k+1}) = (\sigma_{k+1}, \pi_{k+1}) \;\Big|\; I_{k+1} = \iota_{k+1}, \mathcal{E}|_k^{(m)} = E|_k, \overline{\Gamma} = (1,1,1), \mathcal{Y}^{(m)} = (\Delta, y) \,\Big] \;; \end{aligned}$$

(T5) and the last term $p_{k+2}^{(m)}(\omega)$, defined as

$$p_{k+2}^{(m)}(\omega) \stackrel{\mathrm{def}}{=} \Pr\Big[\, \mathcal{E}^{(m)}|_{k+2,\ldots,q} = E|_{k+2,\ldots,q} \;\Big|\; \mathcal{E}^{(m)}|_{k+1} = E|_{k+1}, \overline{\Gamma} = (1,1,1), \mathcal{Y}^{(m)} = (\Delta, y) \,\Big] \;,$$

where $E|_{k+1} = ((\iota_{k+1}, \sigma_{k+1}, \pi_{k+1}), \ldots, (\iota_q, \sigma_q, \pi_q))$.

Note that we could remove the conditioning on $\bar{\Gamma}$ for the first three terms, as they only depend on the length-k prefix of the (extended) transcript and the choice of ι_{k+1}, that is, on the randomness from (R1). The important observation is that the above probabilities are independent of whether $m = k$ or $m = k+1$. We first verify this for (T1), (T2), (T3) and (T5), and then turn to the slightly less straightforward term (T4). This is true for $p_k^{(m),\mathrm{int},\mathrm{samp}}(E)$ because $A^{(k)}$ and $A^{(k+1)}$ select their interval queries in exactly the same manner, and for $1 \le \ell \le k$, the ℓ-th sample point is uniformly selected in the ℓ-th queried interval. Similarly we get that $p_{k+1}^{(k),\mathrm{int}}(E) = p_{k+1}^{(k+1),\mathrm{int}}(E)$. The probabilities $p_k^{(k),\mathrm{prof}}(E)$ and $p_k^{(k+1),\mathrm{prof}}(E)$ are induced in the same manner by (R2) and (R3), and $p_{k+2}^{(k)}(E) = p_{k+2}^{(k+1)}(E)$ since for both $A^{(k)}$ and $A^{(k+1)}$, the pair (s_ℓ, b_ℓ) is distributed the same for every $\ell \ge k+2$ (conditioned on any length-$(k+1)$ prefix of the (extended) transcript and the choice of (Δ, y)).

Turning to (T4), observe that $\Gamma_1 = \Gamma_2 = \Gamma_3 = 1$ (by conditioning). Consider first the case that $\Gamma_1 = 1$ because $|\iota_{k+1}| < \Delta/\log N$ and ι_{k+1} is contained entirely within a single half block. For this case there are two subcases. In the first subcase, the profile of the block that contains ι_{k+1} was already set. This implies that b_{k+1} is fully determined (in the same manner) for both $m = k$ and $m = k+1$. In the second subcase, the profile of the block that contains ι_{k+1} (which is an extreme block) is set independently and with equal probability to either $\downarrow\uparrow$ or $\uparrow\downarrow$ for both $m = k$ and $m = k+1$. In either subcase, s_{k+1} is uniformly distributed in ι_{k+1} for both $m = k$ and $m = k+1$.

Next, consider the remaining case that $\Gamma_1 = 1$ because $|\iota_{k+1}| > \Delta \cdot (\log N)^2$. In this case, since $\Gamma_2 = 1$, the block B_{k+1} is not an extreme block, and since $\Gamma_3 = 1$, the profile of the block B_{k+1} was not previously set. Given this, it follows from the discussion at the end of Subsection 9.3 that the distribution of (s_{k+1}, b_{k+1}) is identical whether $m = k$ (and $A^{(m)}$ does not fake the $(k+1)$-th query) or $m = k+1$ (and $A^{(m)}$ fakes the $(k+1)$-th query). ■

Assembling the pieces, the 4 claims above together with Equation (84) yield $d_{\mathrm{TV}}\big(\mathfrak{E}^{(k),\mathrm{N}}, \mathfrak{E}^{(k+1),\mathrm{N}}\big) \leq d_{\mathrm{TV}}(D_1, D_2) \leq 3\eta(N)$, and finally

$$\begin{aligned} d_{\mathrm{TV}}\big(\mathfrak{A}^{\mathrm{N}}, \mathfrak{A}^{\mathrm{Y}}\big) = d_{\mathrm{TV}}\Big(\mathfrak{A}_{\mathrm{otf}}^{(0),\mathrm{N}}, \mathfrak{A}_{\mathrm{otf}}^{(q),\mathrm{N}}\Big) &\leq \sum_{k=0}^{q-1} d_{\mathrm{TV}}\Big(\mathfrak{A}_{\mathrm{otf}}^{(k),\mathrm{N}}, \mathfrak{A}_{\mathrm{otf}}^{(k+1),\mathrm{N}}\Big) \\ &\leq \sum_{k=0}^{q-1} d_{\mathrm{TV}}\Big(\mathfrak{E}^{(k),\mathrm{N}}, \mathfrak{E}^{(k+1),\mathrm{N}}\Big) \leq 3q \cdot \eta(N) \\ &\leq 1/5 \end{aligned}$$

for a suitable choice of the absolute constant τ. ■

10 Conclusion

We have introduced a new conditional sampling framework for testing probability distributions and shown that it allows significantly more query-efficient algorithms than the standard framework for a range of problems. This new framework presents many potential directions for future work.

One specific goal is to strengthen the upper and lower bounds for problems studied in this paper. As a concrete question along these lines, we conjecture that COND algorithms for testing equality of two unknown distributions D_1 and D_2 over $[N]$ require $(\log N)^{\Omega(1)}$ queries. A broader goal is to study more properties of distributions beyond those considered in this paper; natural candidates here, which have been well-studied in the standard model, are monotonicity (for which we have preliminary results), independence between marginals of a joint distribution, and entropy.Yet another goal is to study distributions over other structured domains such as the Boolean hypercube $\{0,1\}^n$ – here it would seem natural to consider "subcube" queries, analogous to the ICOND queries we considered when the structured domain is the linearly ordered set $[N]$. A final broad goal is to study distribution *learning* (rather than testing) problems in the conditional sampling framework.

Acknowledgements

We are sincerely grateful to the anonymous referees for their close reading of this paper and for their many helpful suggestions, which significantly improved the exposition of the final version.

References

[AJ06] José A. Adell and Pedro Jodra. Exact Kolmogorov and total variation distances between some familiar discrete distributions. *J. of Ineq. App.*, 2006(1):64307, 2006. 4.2

[BDKR05] Tuğkan Batu, Sanjoy Dasgupta, Ravi Kumar, and Ronitt Rubinfeld. The complexity of approximating the entropy. *SICOMP*, 35(1):132–150, 2005. 1.1

[BFF+01] Tuğkan Batu, Eldar Fischer, Lance Fortnow, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. Testing random variables for independence and identity. In *Proceedings of FOCS*, pages 442–451, 2001. 1.1, 1.3, 4, 7

[BFR+00] Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing that distributions are close. In *Proceedings of FOCS*, pages 189–197, 2000. 1.1, 10

[BFR+10] Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing closeness of discrete distributions. (abs/1009.5397), 2010. This is a long version of [BFR+00]. 1.1, 1, 1.3, 4, 6.1

[BFRV11] Arnab Bhattacharyya, Eldar Fischer, Ronitt Rubinfeld, and Paul Valiant. Testing monotonicity of distributions over general partial orders. In *Proceedings of ITCS*, pages 239–252, 2011. 1.1

[BKR04] Tuğkan Batu, Ravi Kumar, and Ronitt Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *Proceedings of STOC*, pages 381–390, New York, NY, USA, 2004. ACM. 1.1

[CDVV14] Siu-On Chan, Ilias Diakonikolas, Gregory Valiant, and Paul Valiant. Optimal algorithms for testing closeness of discrete distributions. pages 1193–1203. Society for Industrial and Applied Mathematics (SIAM), 2014. 1.3

[CFGM13] 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. (document), 1.2, 1.4

[DDS+13] Constantinos Daskalakis, Ilias Diakonikolas, Rocco A. Servedio, Gregory Valiant, and Paul Valiant. Testing k-modal distributions: Optimal algorithms via reductions. In *Proceedings of SODA*, pages 1833–1852. Society for Industrial and Applied Mathematics (SIAM), 2013. 1.1

[DP09] Devdatt P. Dubhashi and Alessandro Panconesi. *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press, Cambridge, 2009. 2.2

[Fis01] Eldar Fischer. The art of uninformed decisions: A primer to property testing. *BEATCS*, 75:97–126, 2001. 1.1

[GGR98] Oded Goldreich, Shafi Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. *JACM*, 45(4):653–750, 1998. 1.1

[Gol10] Oded Goldreich, editor. *Property Testing: Current Research and Surveys*. Springer, 2010. LNCS 6390. 1.1

[GR00] Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. Technical Report TR00-020, ECCC, 2000. 1.1, 1.3

[ILR12] Piotr Indyk, Reut Levi, and Ronitt Rubinfeld. Approximating and Testing k-Histogram Distributions in Sub-linear Time. In *Proceedings of PODS*, pages 15–22, 2012. 1.1

[Ma81] Shang-Keng Ma. Calculation of entropy from data of motion. *J. Stat. Phys.*, 26(2):221–240, 1981. 1.1

[MR95] Rajeev Motwani and Prabhakar Raghavan. *Randomized Algorithms*. Cambridge University Press, New York, NY, 1995. 2.2

[Ney34] Jerzy Neyman. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4):558–625, 1934. 1.2

[Pan04] Liam Paninski. Estimating entropy on m bins given fewer than m samples. *IEEE-IT*, 50(9):2200–2203, 2004. 1.1

[Pan08] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE-IT*, 54(10):4750–4755, 2008. 1.1, 1.3

[Rey11] Leo Reyzin. Extractors and the leftover hash lemma. http://www.cs.bu.edu/~reyzin/teaching/s11cs937/notes-leo-1.pdf, March 2011. Lecture notes. 2.2

[Ron08] Dana Ron. Property Testing: A Learning Theory Perspective. *FnTML*, 1(3):307–402, 2008. 1.1

[Ron10] Dana Ron. Algorithmic and analysis techniques in property testing. *FnTCS*, 5:73–205, 2010. 1.1

[RRSS09] Sofya Raskhodnikova, Dana Ron, Amir Shpilka, and Adam Smith. Strong lower bonds for approximating distributions support size and the distinct elements problem. *SICOMP*, 39(3):813–842, 2009. 1.1

[RS96] Ronitt Rubinfeld and Madhu Sudan. Robust characterization of polynomials with applications to program testing. *SICOMP*, 25(2):252–271, 1996. 1.1

[RS09] Ronitt Rubinfeld and Rocco A. Servedio. Testing monotone high-dimensional distributions. *RSA*, 34(1):24–44, January 2009. 1.1, 6.2, 6.2.1

[Val11] Paul Valiant. Testing symmetric properties of distributions. *SICOMP*, 40(6):1927–1968, 2011. 1.1, 1.3, 6.1

[VV10a] Gregory Valiant and Paul Valiant. A CLT and tight lower bounds for estimating entropy. Technical Report TR10-179, ECCC, 2010. 1.3, 1.3.1, 10

[VV10b] Gregory Valiant and Paul Valiant. Estimating the unseen: A sublinear-sample canonical estimator of distributions. Technical Report TR10-180, ECCC, 2010. 1.3, 10

[VV11] Gregory Valiant and Paul 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 [VV10a] and [VV10b]. 1.1, 1.3, 1.3.1

[VV14] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. In *Proceedings of FOCS*, 2014. 1.3

[Wik13] Wikipedia contributors. Stratified Sampling. http://en.wikipedia.org/wiki/Stratified_sampling, accessed July 1, 2013. 1.2

ECCC ISSN 1433-8092

http://eccc.hpi-web.de