View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Edinburgh Research Archive

THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Algorithms and Lower Bounds for Testing Properties of Structured Distributions

Vladimir Nikishkin

Doctor of Philosophy

Laboratory for Foundations of Computer Science

School of Informatics

University of Edinburgh

2016

Abstract

In this doctoral thesis we consider various property testing problems for structured distributions. A distribution is said to be structured if it belongs to a certain class which can be simply described in approximation terms. Such distributions often arise in practice, e.g. log-concave distributions, easily approximated by polynomials (see [Bir87a]), often appear in econometric research. For structured distributions, testing a property often requires far less samples than for general unrestricted distributions.

In this thesis we prove that this is indeed the case for several distance-related properties. Namely, we give explicit sub-linear time algorithms for L_1 and L_2 distance testing between two structured distributions for the cases when either one or both of them are available as a "black box".

We also prove that the given algorithms have the best possible asymptotic complexity by proving matching lower bounds in the form of explicit problem instances (albeit constructed using randomized techniques) demanding at least a specified amount of data to be tested successfully.

As the main numerical result, we prove that testing that total variation distance to an explicitly given distribution is at least ε requires $O\left(\frac{\sqrt{k}}{\varepsilon^2}\right)$ samples, where k is an approximation parameter, dependent on the class of distribution being tested and independent of the support size. Testing that the total variation distance between two "black box" distributions is at least ε requires $O\left(\frac{k^{4/5}}{\varepsilon^{6/5}}\right)$. In some cases, when $k \sim n$, this result may be worse than using an unrestricted testing algorithm (which requires $O(\frac{n^{2/3}}{\varepsilon^2})$ samples where n is the domain size). To address this issue, we develop a third algorithm, which requires $O\left(\frac{k^{2/3}}{\varepsilon^{4/3}} \log^{4/3}(\frac{n}{k}) \log\log(\frac{n}{k})\right)$ and serves as a bridge between the cases of small and large domain sizes.

Acknowledgements

First and foremost, I would like to thank my supervisor Dr. Ilias Diakonikolas for his wise guidance and leadership of my doctoral program.

I would also like to thank Dr. Daniel Kane from UCSD, with whom we have published three papers constituting most of the material behind this thesis.

I want to greatly thank all the people who played a vital role in the appearance of this thesis — Dr. Rahul Santhanam, Dr. Kousha Etessami — for their invaluable support and guidance.

I would also like to thank my former supervisor in Information Theory, Dr. Andrei Romashchenko, for bringing me into the subject.

I would also like to thank my former supervisors in Software Development, Dr. Dmitri Nikolaev and Simon Karpenko.

I would also like to thank the members of the faculty and my office mates with whom we have worked and spent a lot of time together Dr. Mary Cryan, D. Hab. Richard Mayr, Dr. Elham Kashefi, Dr. Steven Renals, Dr. Theodoris Kapourniotis, Kristijan Liiva, Alexandru Gheorghiu, Mikhail Basios, Veselin Blagoev, Chrystalla Pavlou.

Finally, I want to thank my mother and father for everything they did for me.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Vladimir Nikishkin*)

Table of Contents

Chapter 1

Introduction

1.1 Distribution Testing and Structured Distributions

1.1.1 Property Testing

The problem of property testing comes from a natural case when a statistical analyst has a large dataset, that comes from a certain distribution, and wishes to find a certain property of this distribution, such as the mean value or the maximal value.

Of course, the most natural and naive approach would be to estimate the distribution itself at every point and compute the property based on the distribution estimate. This, however, would require processing a lot of data, which is not always desirable.

The idea of property testing – that is, applying statistical hypothesis testing theory to combinatorial objects (in our case, distributions), was proposed as a way to reduce the amount of required data and computation if a precise property value is not required, and just a rough estimate would suffice.

More strictly, a distribution testing algorithm that is run on a sample set coming from a distribution returns "true", when a distribution has the property (belongs to a class of objects having the property), or "false" if it is at least ε-far from having the property with respect to some metric (e.g. total variation distance) in the distribution space.

Intuitively, this should greatly reduce the data requirements; indeed, there are many cases illustrating this intuition. The idea of switching from value estimation to a binary response is one of the main pillars of this thesis.

In theoretical computer science community the area of distribution property testing was initiated by the work of Batu *et al.* [BFR+00, BFR+13], and has developed into a very active research area with close connections to information theory, learning and statistics. The paradigmatic algorithmic problem in this area is the following: given sample access to an unknown distribution q over an n-element set, we want to determine whether q has some property or is "far" (in statistical distance or, equivalently, L_1 norm) from any distribution that has the property. The overarching goal is to obtain a computationally efficient algorithm that uses as few samples as possible – certainly asymptotically fewer than the support size n, and ideally much less than that.

One of the first problems studied in this line of work is that of "identity testing against a known distribution": Given samples from an unknown distribution q and an explicitly given distribution p distinguish between the case that $q = p$ versus the case that q is ε-far from p in L_1 norm. The problem of *uniformity testing* – the special case of identity testing when p is the uniform distribution – was first considered by Goldreich and Ron [GR00] who, motivated by a connection to testing expansion in graphs, obtained a uniformity tester using $O(\sqrt{n}/\varepsilon^4)$ samples. Subsequently, Paninski gave the tight bound of $\Theta(\sqrt{n}/\varepsilon^2)$ [Pan08] for this problem. Batu *et al.* [BFF+01] obtained an identity testing algorithm against an arbitrary explicit distribution with sample complexity $O(n\text{polylog}(n)/\varepsilon^4)$. The tight bound of $\Theta(\sqrt{n}/\varepsilon^2)$ for the general identity testing problem was given only recently in [VV13].

We are also interested in the problem of *identity testing against an unknown distribution* from which we can sample. This is another classical problem in statistical hypothesis testing [NP33, LR05] that has received considerable attention by the theoretical computer science community: given sample access to distributions p, q, and a parameter $\varepsilon > 0$, we want to distinguish between the cases that p and q are identical

versus ε-far from each other in L_1 norm (statistical distance). Previous work on this problem focused on characterising the sample size needed to test the identity of two arbitrary distributions of a given support size [BFR+00, CDVV14]. It is now known that the optimal sample complexity (and running time) of this problem for distributions with support of size n is $\Theta(\max\{n^{2/3}/\varepsilon^{4/3}, n^{1/2}/\varepsilon^2\})$.

1.1.2 Structured Distributions

One may, on the other hand, approach the issue of large datasets from a different viewpoint. In reality,it is very seldom that analysts have to work with arbitrary adversarial distributions. Far more often the distributions studied belong to families which are "regular" in one sense or another. It is plausible to assume that having certain reasonable restrictions on the distribution shape should also significantly reduce the requirements to the size of data. In an extreme case of such restriction one would arrive at a "parametric statistics" case, but in this thesis we do not want to restrict ourselves so much.

The VC dimension of the function space is closely related to how well we can approximate our distribution function by primitive functions, such as polynomials, which are easy to work with.

It turns out that the correct way to account for the distribution shape (and thus, the VC-dimension) is to adjust the metric, the way how we compare distributions to each other. In this thesis we heavily use the so called $\mathcal{A}_k$ metric, which we will properly define later. For now we will note that the k is a parameter dependent on how complicated shape we permit out distributions to have.

The algorithms given in this thesis indeed show that, if a reasonable approximation of the distribution family can be found, the amount of data required is greatly reduced, and in many cases can be even made independent of the distribution's domain size.

The area of inference under shape constraints – that is, inference about a probability distribution under the constraint that its probability density function (pdf) sat-

isfies certain qualitative properties – is a classical topic in statistics starting with the pioneering work of Grenander [Gre56] on monotone distributions (see [BBBB72] for an early book on the topic). Various structural restrictions have been studied in the statistics literature, starting from monotonicity, unimodality, and concavity [Gre56, Bru58, Rao69, Weg70, HP76, Gro85, Bir87a, Bir87b, Fou97, CT04, JW09], and more recently focusing on structural restrictions such as log-concavity and k-monotonicity [BW07, DR09, BRW09, GW09, BW10, KM10].

Shape restricted inference is well-motivated in its own right, and has seen a recent surge of research activity in the statistics community, in part due to the ubiquity of structured distributions in the natural sciences. Such structural constraints on the underlying distributions are sometimes direct consequences of the studied application problem (see e.g., Hampel [Ham87], or Wang *et al.* [WWW+05]), or they are a plausible explanation of the model under investigation (see e.g., [Reb05] and references therein for applications to economics and reliability theory). We also point the reader to a recent survey [Wal09] highlighting the importance of log-concavity in statistical inference. The hope is that, under such structural constraints, the quality of the resulting estimators may dramatically improve, both in terms of sample size and in terms of computational efficiency.

The statistics literature on the topic has focused primarily on the problem of *density estimation* or learning an unknown structured distribution. That is, given a sample set, drawn from a distribution q, which is promised to belong to some distribution class C, we would like to output a hypothesis distribution that is a good approximation to q. In recent years, there has been a flurry of results in the theoretical computer science community on learning structured distributions, with a focus on both sample complexity and computational complexity, see [KMR+94, FOS05, BS10, KMV10, MV10, DDS12a, DDS12b, CDSS13, DDO+13, CDSS14] for some representative works.

1.2 Basic Definitions

1.2.1 Distance Between Distributions

We start with some notation that will be used throughout this paper. Our main objects of study are discrete probability distributions over $[n] := \{1, \ldots, n\}$, which are given by probability density functions $p : [n] \to [0,1]$ such that $\sum_{i=1}^{n} p_i = 1$, where p_i is the probability of element i in distribution p. Abusing the notation, we will sometimes use p to denote the distribution with density function p_i. The L_1 (resp. L_2) norm of a distribution is identified with the L_1 (resp. L_2) norm of the corresponding n-vector, i.e., $\|p\|_1 = \sum_{i=1}^{n} |p_i|$ and $\|p\|_2 = \sqrt{\sum_{i=1}^{n} p_i^2}$. The L_1 (resp. L_2) distance between distributions p and q is defined as the L_1 (resp. L_2) norm of the vector of their difference, i.e., $\|p-q\|_1 = \sum_{i=1}^{n} |p_i - q_i|$ and $\|p-q\|_2 = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$. We will denote by U_n the uniform distribution over $[n]$. For $I \subseteq \mathbb{R}$ and density functions $p, q : I \to \mathbb{R}_+$, we have $\|p-q\|_1 = \int_I |p(x) - q(x)| dx$.

In some places the computations will be easier to perform for continuous distributions. In such cases p and q will denote probability density functions for continuous distributions and the summation signs need to be replaced by integrals.

1.2.2 Interval Partitions and $\mathcal{A}_k$-distance

Fix a partition of $[n]$ into disjoint intervals $I := (I_i)_{i=1}^{\ell}$. For such a partition I we will denote its cardinality by $|I|$, i.e., $|I| = \ell$. For an interval $J \subseteq [n]$, we denote by $|J|$ its cardinality or length, i.e., if $J = [a,b]$, with $a \leq b \in [n]$, then $|J| = b - a + 1$. The *reduced distribution* p_r^I corresponding to p and I is the distribution over $[\ell]$ that assigns the ith "point" the mass that p assigns to the interval I_i; i.e., for $i \in [\ell]$, $p_r^I(i) = p(I_i)$.

Let $\mathfrak{J}_k$ be the collection of all partitions of $[n]$ into k intervals, i.e., $I \in \mathfrak{J}_k$ if and only if $I = (I_i)_{i=1}^{k}$ is a partition of $[n]$ into intervals $I_1, \ldots, I_k$. For $p, q : [n] \to [0,1]$ and $k \in \mathbb{Z}_+$, $2 \leq k \leq n$, we define the $\mathcal{A}_k$-distance between p and q by

Definition 1.2.1 ($\mathcal{A}_k$-distance)**.**

$$\|p-q\|_{\mathcal{A}_k} \stackrel{\text{def}}{=} \max_{I=(I_i)_{i=1}^k \in \mathfrak{I}_k} \sum_{i=1}^{k} |p(I_i)-q(I_i)| = \max_{I\in\mathfrak{I}_k} \|p_r^I - q_r^I\|_1.$$

The $\mathcal{A}_k$-distance between distributions [1] is well-studied (see [DL01]) in probability theory and statistics. Note that $\|p-q\|_{\mathcal{A}_k} \leq \|p-q\|_1$, and the two metrics are identical for $k=n$. Also note that $\|p-q\|_{\mathcal{A}_2} = 2d_{\mathrm{K}}(p,q)$, where d_{K} is the Kolmogorov metric (i.e., L_∞ distance between the CDF's).

Also note that this definition makes no distinction between discrete and continuous distributions and is valid for both cases.

1.2.3 Problem Formulation

The well-known *Vapnik-Chervonenkis (VC) inequality* (see e.g., [DL01, p.31]) provides the optimal sample size to *learn* an arbitrary distribution q over $[n]$ in this metric. In particular, it implies that $m = \Omega(k/\varepsilon^2)$ i.i.d. draws from q are sufficient to learn q within $\mathcal{A}_k$-distance ε (with probability at least $9/10$). This fact has recently proved useful in the context of learning structured distributions: By exploiting this fact, Chan *et al.* [CDSS14] recently obtained computationally efficient and near-sample optimal algorithms for learning various classes of structured distributions *with respect to the* L_1 *distance*.

It is thus natural to ask the following question: What is the sample complexity of *testing* properties of distributions with respect to the $\mathcal{A}_k$-distance? Can we use property testing algorithms in this metric to obtain sample-optimal testing algorithms for interesting classes of structured distributions *with respect to the* L_1 *distance*? In this work we answer both questions in the affirmative for the problem of identity testing.

[1] We note that the definition of $\mathcal{A}_k$-distance in this work is slightly different than [DL01, CDSS14], but is easily seen to be essentially equivalent. In particular, [CDSS14] considers the quantity $\max_{S\in\mathcal{S}_k} |p(S)-q(S)|$, where $\mathcal{S}_k$ is the collection of all unions of at most k intervals in $[n]$. It is a simple exercise to verify that $\|p-q\|_{\mathcal{A}_k} \leq 2\cdot\max_{S\in\mathcal{S}_k}|p(S)-q(S)| \leq \|p-q\|_{\mathcal{A}_{2k+1}}$, which implies that the two definitions are equivalent up to constant factors for the purpose of both upper and lower bounds.

1.3 Our Results and Applications

In this section we summarise the results achieved in this thesis. In the next subsection 1.3.1 we discuss some of the most important corollaries following from these results and give tables that sum up the resulting asymptotics.

We solve the problem of identity testing for structured distributions. We solve it optimally for the cases when one of the distributions is known explicitly, and for the case when both distributions are unknown and the domain size is very large. We solve it nearly optimally (up to a double logarithmic factor) when both distributions are unknown and the domain size is small.

We solve this problem in a very general setting. Specifically, for all values of k (the approximation parameter) our result is nearly optimal and if k is sufficiently big it performs with the same efficiency as the algorithm for the unrestricted case.

Our first important result is an optimal algorithm for the identity testing problem under the $\mathcal{A}_k$-distance metric:

Theorem 1.3.1 (**Testing identity to a known distribution**). *Given $\varepsilon > 0$, an integer k with $2 \leq k \leq n$, sample access to a distribution q over $[n]$, and an explicit distribution p over $[n]$, there is a computationally efficient (nearly linear in the sample size) algorithm that uses $O(\sqrt{k}/\varepsilon^2)$ samples from q, and with probability at least $2/3$ distinguishes whether $q = p$ versus $\|p-q\|_{\mathcal{A}_k} \geq \varepsilon$. Additionally there is a lower bound. It claims that $\Omega(\sqrt{k}/\varepsilon^2)$ samples are necessary for any tester to work with probability bounded away from $1/2$.*

The sample lower bound of $\Omega(\sqrt{k}/\varepsilon^2)$ can be easily deduced from the known lower bound of $\Omega(\sqrt{n}/\varepsilon^2)$ for uniformity testing over $[n]$ under the L_1 norm [Pan08]. Indeed, if the underlying distribution q over $[n]$ is piecewise constant with k pieces, and p is the uniform distribution over $[n]$, we have $\|q-p\|_{\mathcal{A}_k} = \|q-p\|_1$. Hence, our $\mathcal{A}_k$-uniformity testing problem in this case is at least as hard as L_1-uniformity testing over support of size k.

This theorem is discussed in detail in Section 3.2. We give two versions of the proof, one for the case when the distribution is precisely k-flat (that is when there is a partition of the domain into k disjoint intervals on each of which the distribution is constant) in Section 3.2.3, and one for the arbitrary case in Section 3.2.4. The first one is a particular case of the second one; this redundancy allows us to illustrate the intuition behind the proof without being distracted by technical difficulties.

If we look and compare the sample complexity for the general case ([Pan08], $\Omega(\sqrt{n}/\varepsilon^2)$) with our result ($\Omega(\sqrt{k}/\varepsilon^2)$), we can see that the formulae have similar structure and coefficients. We therefore could suggest that asymptotic dependency on n probably may be replaced with a similar asymptotic dependence on k for all problems dealing with properties dependent on an "optimal partition" (in other words, using the $\mathcal{A}_k$ distance).

Since the formulae are almost identical, we can perform an immediate reduction $k \to n$ and see that the lower bound for the $\mathcal{A}_k$-distance follows from the lower bound of Paninski [Pan08] since when all values of p_i are different, the $\mathcal{A}_k$ distance is identical to the L_1-distance.

Using this logic, one may think that for the case of testing identity of two *unknown* distributions, the sample complexity should be analogous (replacing n with k) to the one obtained for an unrestricted testing problem with an optimal algorithm given in [CDVV14]. The algorithm there tests identity between two unknown p and q using $O(\max\{n^{2/3}/\varepsilon^2, n^{1/2}/\varepsilon^2\})$ samples. This is, however, not true as our next theorem proves.

Theorem 1.3.2 (**Testing identity to an unknown distribution**). *Given $\varepsilon > 0$, an integer $k \geq 2$, and sample access to two distributions with probability density functions $p, q : [0,1] \to \mathbb{R}_+$, there is a computationally efficient (nearly linear in the sample size) algorithm which uses $O(\max\{k^{4/5}/\varepsilon^{6/5}, k^{1/2}/\varepsilon^2\})$ samples from p, q, and with probability at least $2/3$ distinguishes whether $q = p$ versus $\|q - p\|_{\mathcal{A}_k} \geq \varepsilon$.*

Additionally there is a lower bound, which claims that $\Omega(\max\{k^{4/5}/\varepsilon^{6/5}, k^{1/2}/\varepsilon^2\})$

samples are necessary for any tester to work with probability bounded away from $1/2$.

Note that Theorem 1.3.2 applies to arbitrary univariate distributions (over both continuous and discrete domains). In particular, the sample complexity of the algorithm does not depend on the support size of the underlying distributions.

We discuss the upper bound of the Theorem 1.3.2 in detail in Section 3.3. Here we will note that the ability to appropriately "stretch" the domain size and reduce the identity testing problem to uniformity is crucial. Specifically, if we know that even if both of the distributions are unknown, but we have some a priori knowledge that they are close to being uniform, we can achieve a very strong upper bound (the second part of the max in Theorem 1.3.2) by "stretching" the distribution proportionally to the distances between the samples. Generally, however, the distances between the samples do not tell us anything and we have to rely on their order only; this is where the first part of the max is the optimal upper bound.

We also give a matching lower bound (Section 4.1). This bound (i.e. the hard instance) is achieved by computing mutual information between the sample and a random variable that is predicting the distance (zero or ε) between the distributions (a standard technique). But to use this technique successfully, we have to rigorously formulate and prove the claim, which was informally described in the previous paragraph — that the distances between samples are not relevant to an optimal tester. This lemma is proved using a variation of Ramsey theorem.

The main result of [CDVV14] and the Theorem 1.3.2 above seem to have a gap between them. Indeed, if the domain size n and the structural parameter k are comparable (e.g. $k = cn$), the corresponding sample complexities would be $O(n^{2/3}/\varepsilon^2)$ (given by [CDVV14]) and $O(n^{4/5}/\varepsilon^{6/5})$ (given by Theorem 1.3.2). That is, it looks like it can be more efficient to use an unrestricted testing algorithm to obtain better sample complexity. To address this issue, we need a connecting theorem.

Theorem 1.3.3 (Testing identity to an unknown distribution, finite domain). *Given sample access to distributions* p *and* q *on* $[n]$ *and* $\varepsilon > 0$ *there exists a computationally*

efficient (nearly linear in the sample size) algorithm that takes

$$O\left(\max\left(\min\left(k^{4/5}/\varepsilon^{6/5}, k^{2/3}\log^{4/3}(3+n/k)\log\log(3+n/k)/\varepsilon^{4/3}\right), k^{1/2}/\varepsilon^2\right)\right)$$

samples from each of p and q and distinguishes with $2/3$ probability between the cases that $p = q$ and $\|p-q\|_{\mathcal{A}_k} \geq \varepsilon$.

This theorem matches Theorem 1.3.2 when n is very large, it also matches the result of [CDVV14] when n and k are comparable, but between these two extremes there seems to be a suboptimal logarithmic factor. In fact this factor is unavoidable, and thus our algorithm is (nearly) optimal. Informally, the main achievement of this theorem is that we more carefully established the dependency between how much information a tester can extract from the distances between the samples versus the ordering of the samples. The former information was too small to be relevant for the problem of Theorem 1.3.2, but when k and n are of the same order, this information can no longer be disregarded. We show, however, that by doing an accurate re-balancing of the domain, we can either reduce the problem to either the unrestricted testing problem or to the case of the Theorem 1.3.2 and no more than a logarithmic amount of re-balancings is needed.

We give the corresponding lower bound in Section 4.2, Theorem 4.2.1. It is largely equivalent to the lower bound of Theorem 1.3.2 with few additional randomisations required to "disguise" the samples which are "close" to each other.

1.3.1 Applications

Theorems 1.3.1, 1.3.2 have a wide range of applications to the problem of L_1 identity testing for various classes of natural and well-studied structured distributions. That is, perhaps surprisingly, many broad distribution classes can be well approximated by polynomials of low degree. For example, in the case of log-concave distributions, any log-concave distribution can be arbitrarily well approximated by piecewise-constant

distributions with no more than $O(\text{polylog}\varepsilon^{-\frac{1}{4}})$ pieces (see [Bir87a]). Since an "optimal partition" (the one the $\mathcal{A}_k$-distance would be looking at) will be identical to the intervals of these pieces, on could successfully test L_1-distance, using $\mathcal{A}_k$-distance as a proxy. This intuition is further explained in this section.

Similar efficient partitions exist for various other classes such as Monotone Hazard Rate distributions, used in Mechanism Design (see [NNV07]).

At a high level, the main message of this work is that the $\mathcal{A}_k$ distance can be used to characterise the sample complexity of L_1 identity testing for broad classes of structured distributions. The following simple proposition underlies our approach and proposes a generic method of creating testers if k is given:

Proposition 1.3.4. *For a distribution class C over $[n]$ and $\varepsilon > 0$, let $k = k(C,\varepsilon)$ be the smallest integer such that for any $f_1, f_2 \in C$ it holds that $\|f_1 - f_2\|_1 \leq \|f_1 - f_2\|_{\mathcal{A}_k} + \varepsilon/2$. Then there exists an algorithm that for any $p, q \in C$ tests $p = q$ versus $d_{L_1}(p,q) > \varepsilon$ with the same asymptotic performance as an unrestricted $\mathcal{A}_k$-distance tester.*

Namely, when one of the distributions is known explicitly, there is a tester that requires $O(\frac{\sqrt{k}}{\varepsilon^2})$ samples, and when both distributions are unknown there is a tester that requires $O(\frac{k^{4/5}}{\varepsilon^{6/5}})$ samples.

In this proposition we use the definition of the $\mathcal{A}_k$-distance from Definition 1.2.1.

Informally, this proposition means that we can use the $\mathcal{A}_k$-distance tester to test L_1 identity if (for some k) the values of the two distances are close to each other.

The proof of the proposition is straightforward: Given sample access to $q \in C$ and a $p \in C$ (which can either be given explicitly, if we want to use the tester from Theorem 1.3.1, or be given as samples, if we want to use the tester from Theorem 1.3.2), we apply the $\mathcal{A}_k$-identity testing algorithm of Theorems 1.3.1 or 1.3.2 for the value of k in the statement of the proposition, and error $\varepsilon' = \varepsilon/2$. If $q = p$, the algorithm will output "YES" with probability at least $2/3$. If $\|q - p\|_1 \geq \varepsilon$, then by the condition of Proposition 1.3.4 we have that $\|q - p\|_{\mathcal{A}_k} \geq \varepsilon'$, and the algorithm will output "NO" with probability at least $2/3$. Hence, as long as the underlying distribution satisfies

the condition of Proposition 1.3.4 for appropriate values of $k = o(n)$, Theorems 1.3.1 and 1.3.2 yield asymptotic improvements over the sample complexities of $\Theta(\sqrt{n}/\varepsilon^2)$ and $\Theta(n^{2/3}/\varepsilon^2)$.

We remark that the value of k in the proposition is a natural complexity measure for the difference between two probability density functions in class C. It follows from the definition of the $\mathcal{A}_k$ distance that this value corresponds to the number of "essential" crossings between f_1 and f_2 – i.e., the number of crossings between the functions f_1 and f_2 that significantly affect their L_1 distance. Intuitively, the number of essential crossings – as opposed to the domain size – is, in some sense, the "right" parameter to characterise the sample complexity of L_1 identity testing for C.

More specifically, our framework can be applied to all structured distribution classes C that can be well-approximated in L_1 distance by *piecewise low-degree polynomials*. We say that a distribution p over $[n]$ is *t-piecewise degree-d* if there exists a partition of $[n]$ into t intervals such that p is a (discrete) degree-d polynomial within each interval. Let $\mathcal{P}_{t,d}$ denote the class of all t-piecewise degree-d distributions over $[n]$. We say that a distribution class C is *ε-close in L_1* to $\mathcal{P}_{t,d}$ if for any $f \in C$ there exists $p \in \mathcal{P}_{t,d}$ such that $\|f - p\|_1 \leq \varepsilon$. It is easy to see that any pair of distributions $p, q \in \mathcal{P}_{t,d}$ have at most $2t(d+1)$ crossings, which implies that $\|p - q\|_{\mathcal{A}_k} = \|p - q\|_1$, for $k = 2t(d+1)$ (see e.g., Proposition 6 in [CDSS14]). We therefore obtain the following:

Corollary 1.3.5. *Let C be a distribution class over $[n]$ and $\varepsilon > 0$. Consider parameters $t = t(C, \varepsilon)$ and $d = d(C, \varepsilon)$ such that C is $\varepsilon/4$-close in L_1 to $\mathcal{P}_{t,d}$. Then there exists an L_1 identity testing algorithm for C using $O(\sqrt{t(d+1)}/\varepsilon^2)$ samples if the target distribution is given explicitly or $O((t(d+1))^{4/5}/\varepsilon^{6/5})$ if the distribution is given sample access.*

Note that any pair of values (t, d) satisfying the condition above suffices for the conclusion of the corollary. Since our goal is to minimise the sample complexity, for a given class C, we would like to apply the corollary for values t and d satisfying the above condition and are such that the product $t(d+1)$ is minimised. The appropriate choice of these values is crucial, and is based on properties of the underlying distribu-

tion family.

The concrete testing results of Tables 1.1, 1.2 and 1.3 are obtained from Corollary 1.3.5 by using known existential approximation theorems [Bir87a, CDSS13, CDSS14] for the corresponding structured distribution classes. In particular, we obtain efficient identity testers, in most cases with optimal sample complexity, for all the structured distribution classes studied in [CDSS13, CDSS14] in the context of learning.

Tables 1.2 and 1.3 correspond to different regimes (with respect to k) of the same problem, when both distributions are unknown. When the problem arises in practice, one should choose whichever result is more evvicient in their case.

Perhaps surprisingly, our upper bounds are tight not only for the class of piecewise polynomials, but also for the specific shape restricted classes of Tables 1.1, 1.2 and 1.3. The corresponding lower bounds for specific classes are either known from previous work (as e.g., in the case of t-modal distributions [DDS$^+$13]) or can be obtained using standard constructions.

Finally, we remark that although Theorems 1.3.1, 1.3.2 are formulated for discrete and continuous distributions respectively, both can be appropriately generalised to the the other setting of testing the identity of continuous distributions over the real line. It is easy to see that both Proposition 1.3.4 and Corollary 1.3.5 hold.

Distribution Family	**k**	**Our upper bound**	**Previous work**
t-piecewise constant	t	$O(\sqrt{t}/\varepsilon^2)$	$O(t/\varepsilon^2)$ [CDSS14]
t-piecewise degree-d	$t(d+1)$	$O\left(\sqrt{t(d+1)}/\varepsilon^2\right)$	$O\left(t(d+1)/\varepsilon^2\right)$ [CDSS14]
log-concave	$\varepsilon^{-1/2}$	$\widetilde{O}(1/\varepsilon^{9/4})$	$\widetilde{O}(1/\varepsilon^{5/2})$ [CDSS14]
s-mixture of log-concave	$s\varepsilon^{-1/2}$	$\sqrt{s}\cdot\widetilde{O}(1/\varepsilon^{9/4})$	$\widetilde{O}(s/\varepsilon^{5/2})$ [CDSS14]
t-modal	$\frac{t\log n}{\varepsilon}$	$O(\sqrt{t\log(n)}/\varepsilon^{5/2})$	$O\left(\frac{\sqrt{t\log(n)}}{\varepsilon^3}+\frac{t^2}{\varepsilon^4}\right)$ [DDS+13]
s-mixture of t-modal	$\frac{st\log n}{\varepsilon}$	$O(\sqrt{st\log(n)}/\varepsilon^{5/2})$	$O\left(\frac{\sqrt{st\log(n)}}{\varepsilon^3}+\frac{s^2t^2}{\varepsilon^4}\right)$ [DDS+13]
monotone hazard rate (MHR)	$\frac{\log\frac{n}{\varepsilon}}{\varepsilon}$	$O(\sqrt{\log(n/\varepsilon)}/\varepsilon^{5/2})$	$O(\log(n/\varepsilon)/\varepsilon^3)$ [CDSS14]
s-mixture of MHR	$\frac{s\log\frac{n}{\varepsilon}}{\varepsilon}$	$O(\sqrt{s\log(n/\varepsilon)}/\varepsilon^{5/2})$	$O(s\log(n/\varepsilon)/\varepsilon^3)$ [CDSS14]

Table 1.1: Algorithmic results for identity testing of various classes of probability distributions. The third column indicates the sample complexity of our general algorithm applied to the class under consideration. The fourth column indicates the sample complexity of the best previously known algorithm for the same problem, in most cases obtained by learning.

Distribution Family	**k**	**Our upper bound**	**Previous work**
t-piecewise constant	t	$O\big(\max\big\{\frac{t^{4/5}}{\varepsilon^{6/5}}, \frac{t^{1/2}}{\varepsilon^2}\big\}\big)$	$O\big(\frac{t}{\varepsilon^2}\big)$ [CDSS14]
t-piecewise degree-d	$t(d+1)$	$O\big(\max\big\{\frac{(t(d+1))^{4/5}}{\varepsilon^{6/5}}, \frac{(t(d+1))^{1/2}}{\varepsilon^2}\big\}\big)$	$O\big(\frac{t(d+1)}{\varepsilon^2}\big)$ [CDSS14]
log-concave	$\varepsilon^{-1/2}$	$O\big(\frac{1}{\varepsilon^{9/4}}\big)$	$O\big(\frac{1}{\varepsilon^{5/2}}\big)$ [CDSS14]
s-mixture of log-concave	$s\varepsilon^{-1/2}$	$O\big(\max\big\{\frac{s^{4/5}}{\varepsilon^{8/5}}, \frac{s^{1/2}}{\varepsilon^{9/4}}\big\}\big)$	$O\big(\frac{s}{\varepsilon^{5/2}}\big)$ [CDSS14]
t-modal over $[n]$	$\frac{t\log n}{\varepsilon}$	$O\big(\max\big\{\frac{(t\log n)^{4/5}}{\varepsilon^2}, \frac{(t\log n)^{1/2}}{\varepsilon^{5/2}}\big\}\big)$	$O\big(\frac{(t\log n)^{2/3}}{\varepsilon^{8/3}} + \frac{t^2}{\varepsilon^4}\big)$ [DDS+13]
MHR over $[n]$	$\frac{\log\frac{n}{\varepsilon}}{\varepsilon}$	$O\big(\max\big\{\frac{\log(n/\varepsilon)^{4/5}}{\varepsilon^2}, \frac{\log(n/\varepsilon)^{1/2}}{\varepsilon^{5/2}}\big\}\big)$	$O\big(\frac{\log(n/\varepsilon)}{\varepsilon^3}\big)$ [CDSS14]

Table 1.2: Algorithmic results for closeness testing of selected families of structured probability distributions. The third column indicates the sample complexity of our general algorithm applied to the class under consideration. The fourth column indicates the sample complexity of the best previously known algorithm for the same problem, in most cases obtained by learning. This table corresponds to the case when n is much larger than k and the algorithm given in the proof of Theorem 1.3.2 is optimal.

Distribution Family	**k**	**Our upper bound**	**Previous work**
t-piecewise constant	t	$O\big(\frac{1}{\varepsilon^{4/3}} t^{2/3} \cdot \cdot \log^{4/3}(\frac{n}{t}) \cdot \cdot \log\log(\frac{n}{t})\big)$	$O\big(\frac{t}{\varepsilon^2}\big)$ [CDSS14]
t-piecewise degree-d	$t(d+1)$	$O\big(\frac{1}{\varepsilon^{4/3}} (t(d+1))^{2/3} \cdot \cdot \log^{4/3}(\frac{n}{t(d+1)}) \cdot \cdot \log\log(\frac{n}{t(d+1)})\big)$	$O\big(\frac{t(d+1)}{\varepsilon^2}\big)$ [CDSS14]
log-concave	$\varepsilon^{-1/2}$	$O\big(\frac{1}{\varepsilon^{5/3}} \cdot \cdot \log^{4/3}(n\sqrt{\varepsilon}) \cdot \cdot \log\log(n\sqrt{\varepsilon})\big)$	$O\big(\frac{1}{\varepsilon^{5/2}}\big)$ [CDSS14]
t-modal over $[n]$	$\frac{t\log n}{\varepsilon}$	$O\big(\frac{1}{\varepsilon^2} (t\log n)^{2/3} \cdot \cdot \log^{4/3}(\frac{n\varepsilon}{t\log n}) \cdot \cdot \log\log(\frac{n\varepsilon}{t\log n})\big)$	$O\big(\frac{(t\log n)^{2/3}}{\varepsilon^{8/3}} + \frac{t^2}{\varepsilon^4}\big)$ [DDS+13]
MHR over $[n]$	$\frac{\log\frac{n}{\varepsilon}}{\varepsilon}$	$O\big(\frac{1}{\varepsilon^{8/3}} (\log n)^{2/3} \cdot \cdot \log^{4/3}(\frac{\varepsilon^2 n}{\log n}) \cdot \cdot \log\log(\frac{\varepsilon^2 n}{\log n})\big)$	$O\big(\frac{\log(n/\varepsilon)}{\varepsilon^3}\big)$ [CDSS14]

Table 1.3: Algorithmic results for closeness testing of selected families of structured probability distributions. The third column only shows the part which is different from table 1.2. The fourth column indicates the sample complexity of the best previously known algorithm for the same problem. For simplicity, we consider all $\log x = c\,\forall x < 1$. This table corresponds to the case when n of the same order as k and the algorithm given in the proof of Theorem 1.3.3 is optimal.

Chapter 2

Prior Work

In computer science the problem of testing structured distributions arises naturally from two predecessors. The first one is the problem of learning structured distributions, which was initially proposed by Kearns et al. [KMR+94]. The second is the general testing framework, initially formulated by Rubinfeld and Sudan in 1996 ([RS96]).

2.1 Learning Distributions and Shape Restricted Estimation

The problem of learning probability distributions began at least a century ago within the statistics field. The first articles from computer science perspective started to appear in the nineties, the most notable being the work by Kearns et al. ([KMR+94]) For a book-length introduction, refer to [DG85], [DL01], [Sco92], [Sil86].

There followed a series of works on learning distributions in high dimensional spaces. (See the following papers: [Das99], [FM99], [AK01].)

2.1.1 Learning Mixtures

The most recent advancement on the topic comes from Chan et al. [CDSS13], where one of the most generic structures imposed on the distributions is assumed. Chan et al.

treat the problem of learning a mixture of certain types of "simple" probability distributions. In their terms "simple" means either log-concave, flat, unimodal or monotone hazard rate distributions. A "mixture" of k distributions is the following:

If $p_1 \dots p_k$ are probability distributions, and $\mu_1 \dots \mu_k$ are weights which sum up to 1, their mixture p is defined by the following expression:

$$p \stackrel{\text{def}}{=} \sum_{i=1}^{k} \mu_i p_i$$

The fact that the weights sum up to 1 represents an algorithmic (dual) side of the mixture idea. Instead of having a single "mixture" distribution one may select one of the distributions uniformly at random and then draw samples from the selected distribution. As a result, the equivalent distribution will be a mixture.

The main idea consists of constructing a decomposition of the domain into intervals, on each of which the function behaves "well", for example being flat. So essentially all of the algorithms provided in the article follow the same approach - they use the first set of samples to construct a decomposition of the domain into intervals, on each of which the function behaves well. Then they use a general empirical distribution function to learn the underlying distribution and apply some heuristics to fit it to the assumed structure.

The algorithms given allow learning k-mixtures of various structured distributions using from $O(k)$ to $O(k\log(n))$ samples.

There is a simple argument which proves that if there is an algorithm which learns distributions from some class, then learning a mixture of distributions from that class cannot require many more samples.

Theorem 2.1.1 (**Existential theorem from [CDSS13].**). *Let C be a class of distributions over $[n]$. Let A be an algorithm which learns an unknown distribution p in C using $m(n,\varepsilon)\cdot\log 1/\delta$ samples. (i.e., with success probability $1-\delta$ outputs a distribution h such that $d_{TV}(p,h)\leqslant\varepsilon$) Then there is an algorithm A' which learns any mixture of k distributions from C using $O(k/\varepsilon^3)m(n,\varepsilon/20)\log^2(1/\delta)$ samples.*

This theorem, despite giving an optimal sample complexity, is not useful in practice, as exponential calculations are required. Indeed, it iterates over all possible partitions of the sample set into k subsets.

However, the paper [CDSS13] also gives an efficient (polynomial) algorithm for all mixtures of distributions satisfying a mild restriction. The distributions have to be (ε,t)-flat.

Definition 2.1.2. A distribution p over $[n]$ is (ε,t)-flat if there exists a t-piecewide constant distribution q such that the L_1 distance between p and q, $d(p,q)$ is less than ε.

On distributions satisfying the flatness case, the following theorem is true:

Theorem 2.1.3. ***Improved Theorem 2.1.1.*** ***([CDSS13])*** *There is an algorithm that learns any k-mixture of (ε,t)-flat distributions over $[n]$ to accuracy $O(\varepsilon)$ using $O(kt/\varepsilon^3)$ samples and running in $O(kt\log(n)/\varepsilon^3)$.*

The high level idea of the algorithm is that if the decomposition of the domain into intervals on which the distribution function is flat is known, then the learnt distribution is just an empirical distribution function, flattened over those intervals.

Chan et al. give non-trivial methods of constructing those decompositions for various types of (ε,t)-flat distributions.

2.1.2 Estimating the Unseen

The lower bound for density estimation of a distribution over a discrete domain with respect to L_1 metric is linear in the size of the domain:

$$C = \Omega(\frac{n}{\varepsilon^2}) \tag{2.1}$$

However, choosing a different metric allows to reduce the number of samples (see paper by Valiant and Valiant [VV11a]). The resulting learnt distribution can be used instead of the original one to calculate many parameters, such as entropy.

Instead of using a traditional total variation distance, Valiant and Valiant propose a metric, which is more relaxed, but still captures some important notion.

Definition 2.1.4. The histogram of a distribution The histogram of a distribution p is a mapping $h : (0,1] \to \mathbb{Z}$, where $h(x) = |\{i : p(i) = x\}|$.

A "histogram of a distrbution", proposed in [VV11a], captures the idea of "symmetric" properties. Informally, a symmetric property is a property that does not change with permutation of support elements. Entropy and distance to uniformity are examples of such properties. It is easy to show that any symmetric property only depends on a histogram of a distribution.

Definition 2.1.5. The relative-earthmover distance For two histograms h_1, h_2 the relative earth-mover distance $R(h_1, h_2)$ between them is defined as the minimum over all schemes of moving the probability mass of the first histogram to yield the second histogram of the cost of moving that mass. The per-unit cost of moving mass from probability x to y is $|\log(x/y)|$.

Valiant and Valiant ([VV11a]) give an algorithm which allows the learning of a distribution from an oracle so that the relative earthmover distance between their histograms is small enough. Afterwards one computes the value of any symmetric property of the learnt distribution. The property of the relative earthmover distance guarantees that with high probability, the value will be close to the value of this property on the original distribution.

We state the theorem more formally:

Theorem 2.1.6. *(**Learning up to an earthmover distance**) For sufficiently large n and any constant $c > 1$ given $c\frac{\log c}{\sqrt{c}}$ independent samples from a distribution D, with probability at least $1 - e^{-n^{0.03}}$ the algorithm returns a distribution D', representable as an $O(c\frac{n}{\log n})$-length vector, such that the relative earthmover distance between D and D' satisfies*

$$R(D,D') \leqslant O(\frac{logc}{\sqrt{c}})$$

The algorithm runs in time $O(c\frac{n}{\log n})$.

The algorithm works by simply learning the heavier part of the distribution (constructing an empirical distribution function) and constructing a sophisticated linear program to "estimate an unseen" light part of the distribution so that it agrees with the observed samples. For more detail please refer to the paper [VV11a].

The important corollaries from this theorem are algorithms for estimation **entropy**, **support size** and **distance to any known distribution** aka tolerant testing.

2.1.3 Learning via Testing

While learning and testing may seem distinct (although connected) areas, sometimes the results of one area have direct applications in the other. A notable example being the "testing via learning" approach, where the distribution is learned up to some distance for some metric (often the metric itself may be quite impractical), and then the value of the property directly computed on the learned distribution is close enough to a real one. (See paper [VV11a]).

In the case of the paper [DDS+13] the situation is completely dual. If at least some pieces of the distribution are "good" in some sense, then it is possible to partition the domain into intervals supporting these "good" pieces, using the testing algorithm to "test" the intervals. After a good partitioning is obtained, some other algorithm, efficient on "good" pieces, may be applied.

Now let us formulate the theorem:

Theorem 2.1.7. ***(Learning k-modal distributions)*** *There is an algorithm, which learns a k-modal distribution up to* L_1*-accuracy* ε *with confidence* $9/10$ *using:*

$$O(\frac{k\log(n/k)}{\varepsilon^3} + \frac{k^3}{\varepsilon^3}\log(k/\varepsilon)\log\log(k/\varepsilon))$$

samples and performing $(poly)(k, \log n, 1/\varepsilon)$ *bit operation.*

2.2 Testing Distributions.

State of the art results in distributions testing were mostly obtained in the same time as the learning results.

The most recent results in the area are given by the works of Valiant et al. ([VV14]) for identity testing, Chan et al. ([CDVV14]) for 2-unknown testing, and Valiant ([VV11a]) for tolerant testing.

They solve the unstructured testing problems optimally, and in what follows, I will explain how.

2.2.1 Graph Uniformity

The first problem I am going to review actually comes from a different domain called "graph testing". Goldreigh and Ron in their 2000 paper ([GR00]) weren't initially concerned with testing distribution properties at all. They wanted to check if the graph is an expander. Without digging seriously into the graph field, recall that any random walk of sufficient length on an expander will produce a random vertex uniformly. In other words, the distribution on the vertices will be approximately uniform.

This immediately gives a sufficient condition for a graph being an expander together with an algorithm for checking expansion. Just take enough random walks to sample from this distribution and use these samples to test uniformity. If the distribution is uniform, the graph can be an expander. The converse is not necessarily true, i.e. non-expanders can also produce uniform distributions.

This algorithm requires a uniformity test, and indeed it is also given in the paper as a tool.

Denoting the count of samples falling on each element by $\hat{p}$, we can write the statistic calculated by the algorithm in the following way:

$$F = \sum_{i=1}^{n} \binom{\hat{p}_i - 1}{2}$$

For $m = O(\sqrt{N})$ samples, F must not exceed $O(\binom{m}{2} \cdot N^{-1})$, otherwise the uniformity hypothesis is rejected.

The proof is based on the observation that the amount of collisions for the uniform distribution is minimal among all the variety of distributions.

2.2.2 Testing Closeness

One of the interesting results giving both necessary and sufficient conditions for the distributions to be identical was given in the paper [BFR+13]. Though not exactly tight, the bound given by this algorithm is good and the method is interesting.

The algorithms assume nothing about the distributions, i.e. both are given oracle access.

The algorithm is based on two observations. The first is that if two distributions have few elements with high probability mass, it is possible to efficiently test their closeness by naively measuring the L_1 distance between their empirical distributions. The second is that if conversely there are no prominent elements, there is a collision-based L_2 test requiring few samples, which can be easily extended to the L_1 distance using the bound $|v|_1 \leqslant \sqrt{n}|v|_2$. This algorithm is similar to the one used for graph expansion testing in [GR00]. It is based on the fact that when two distributions are equal, the self collision probability will be approximately equal to the mutual collision probability (up to a factor of two).

The edge between these two cases was found by trial and error and is the following: the element is considered "large" if its probability mass is greater than $(\frac{\varepsilon}{n})^{2/3}$.

2.2.3 Optimal Algorithms for Testing Closeness

Imagine a situation when there are two random variables distributed according to some probability distributions p and q. There is a natural question if these distributions are close or far from each other in statistical distance (equivalent to an L_1 norm). One might also think about the L_2 norm. Although the $l2$ norm itself is quite strange - if L_1 norm is ε, the L_2 norm may be as small as $\varepsilon/\sqrt{n}$, it is sometimes very helpful tool to use while proving theorems.

We already saw the problem of L_1 identity testing considered in the paper by Valiant and Valiant ([VV11a]).

In the paper [CDVV14], Chan et al. provide two theorems for the L_1 and L_2 norms respectively and give lower bounds matching their upper bound, which are similar to the bounds given in [VV11a], thus proving that the lower bounds in [VV11a] were optimal.

Theorem 2.2.1. *(L_1 **two-unknown tester**) Given $\varepsilon > 0$ and sample access to distributions p and q over $[n]$, there is an algorithm which uses $O(\max n^{2/3}\varepsilon^{3/4}, n^{1/2}/\varepsilon^2)$ samples, runs in time linear in its sample size and with probability at least $2/3$ distinguishes whether $p = q$ versus $L_1(p,1) \geqslant \varepsilon$. The requirement in the number of samples cannot be improved.*

Theorem 2.2.2. *(L_2 **two-unknown tolerant tester**) For two distributions p, q over $[n]$ with $b \geqslant ||p||_2^2, ||q||_2^2$, there is an algorithm which distinguishes the case that $||p-q||_2 \leqslant \varepsilon$ from the case that $||p-q|| \geqslant 2\varepsilon$ when given $O(\sqrt{b}/\varepsilon^2)$ samples form p and q with probability at least $2/3$. The requirement in the number of samples cannot be improved.*

Both algorithms work in a similar manner: the tester function Z is computed and if the value exceeds the threshold, the functions are considered far, otherwise - close.

For the L_1 tester the tester function is the following:

$$Z = \sum_i \frac{(X_i - Y_i)^2 - X_i - Y_i}{X_i + Y_i}$$

And threshold is $C \cdot \sqrt{n}$

Where X_i and Y_i denote the amount of samples falling on the i-th domain element and C is an absolute constant.

For the L_2 tester the tester function is the following:

$$Z = \sum_i (X_i - Y_i)^2 - X_i - Y_i$$

with the threshold $\frac{\sqrt{Z}}{m}$, where m denotes the number of samples.

The proofs are done by applying Chebyshev's inequality to the tester functions. For the details, please refer to the paper [CDVV14].

These algorithms are expected to be very important for our case, as the problem of testing structured distributions, the one we are working on, is very similar to the problem solved.

2.2.4 Instance-optimal Identity Testing

The paper of Chan et al. ([CDVV14]) dealt with the case when the two distributions are both available as black boxes. Their identity can be tested using $O(n^{2/3})$ samples. Can we do better in the case when one of the distributions is given explicitly? Certainly we can do as good. We would just sample from the explicit distribution as if it was a black box.

It turns out (see the paper [VV13] by Valiant et al.) that it is indeed possible to "save" on the knowledge of one of the distributions. Valiant and Valiant give an algorithm to test identity to a known distribution using $O(\frac{\sqrt{n}}{\varepsilon^2})$. The algorithm has computational complexity linear in the number of samples.

The idea of the algorithm is similar to the idea of the algorithms from the paper by Chan ([CDVV14]), but more complex. Instead of having just one tester function and

a threshold, the algorithm has two. If any of those functions exceed the threshold, the distributions are far from each other with high probability. This trick allows Valiant and Valiant to give a tight bound on sample size not just with respect to the domain size, but even with respect to the structure of the explicitly known function.

The tester functions, along with the thresholds are the following:

$$Z = \sum_{i=s}^{n-1} \frac{(X_i - mp_i)^2 - X_i}{p_i^{2/3}} \geqslant 4m\|p^{n-1}_{\geqslant s}\|^{1/3}_{2/3}$$

$$\sum_{i=1}^{s-1} X_i \geqslant \frac{3}{16}\varepsilon m$$

Here s denotes the smallest index of the domain element such that $\sum_{i \leqslant s} p_i \leqslant \frac{\varepsilon}{8}$. p_i and X_i denote the probability if i-th domain element with respect to p and the number of samples falling on this element respectively.

The proof correctness is analogous to the proof of the theorem in paper by Chan ([CDVV14]) and is also based on the applying Chebyshev's inequality to the tester functions.

To prove the optimality of the tester, the matching lower bounds are given as well.

Valiant and Valiant released an improvement analysis of their algorithm of [VV13]. Their 2014 paper ([VV14]) has a succinct characterisation of a generalised Cauchy-Schwartz inequality, which is used in the proof of the boundedness of variation for the tester function.

2.2.5 An Optimal Uniformity Tester

While the paper [VV14] provides an optimal upper and lower bound for the identity testing problem, there is an earlier result ([Pan08]) which provided an asymptotically equivalent algorithm, which only managed to test identity to a **uniform** distribution. (Or, in other words, a uniformity testing algorithm.) The interesting thing is that while it may seem that the problem of identity testing should be harder than uniformity test-

ing, this is not the case. In some sense, testing uniformity is "the hardest" of all identity testing problems.

The interesting thing is that the uniformity tester at first glance seems completely unrelated to the optimal (as we already know, but it was unknown in 2008) identity testing algorithm.

The algorithm is based on a so-called birthday paradox.

Input: sample access to a distribution q over $[n]$, $\varepsilon > 0$.

Output: "YES" if $q = U_n$; "NO" if $\|q - U_n\|_{L_1} > \varepsilon$.

1. Draw m samples from the tested distribution q supported on $[1...n]$.
2. Define by K_1 the amount of domain elements on which there falls just one sample.
3. If $m(\frac{n-1}{n})^{m-1} - K_1 > \frac{m^2\varepsilon^2}{2n}$ return "FAR". Otherwise return "UNIFORM".

The correctness of the algorithm is proved by applying the Chebyshev's inequality. The author computes the expectation and the variance and fits an appropriate threshold.

Three things are particularly interesting about this algorithm. The first one is that it only looks at the samples which occur once, completely ignoring the rest and thus losing some information. The second thing is that the author does not use the poissonisation trick in the proof. The third and most important one is the fact that the tester seems unrelated to the optimal tester from the paper [VV14], and contrary to that tester is linear. That is - the identity tester calculates the quadratic function of the samples in a bin (domain element), but the uniformity one only calculates a linear function. That raises the question whether there exists a linear tester for identity.

2.2.6 k-Modal Distributions

We have seen optimal identity testing algorithms for the "one unknown"([VV14]) and "two unknown"([CDVV14]) settings. These algorithms give optimal (and quite effi-

cient) sublinear algorithms for the general case, when there are no other assumptions about the distributions.

From another point of view, we have seen that if we know something about the distributions available (their **structure**), then we can save on the sample size for various algorithms. (See [CDSS13])

A reasonable question is whether we can save on testing algorithms as well? Daskalakis et al. ([DDS+13]) give the first step into optimising testing algorithms for distributions structure.

They give algorithms for testing identity and estimating L_1 distance in cases when either one of the distributions, or both are unknown, when both distributions are either monotone or k-modal.

Definition 2.2.3. *k-modal distribution*

A distribution p is called k-modal, if it has at most k min-intervals and max-intervals.

An interval $I = [j \dots l]$ is called a max-interval if $p(j) = p(j+1) = \dots = p(l)$ and $p(j-1) < p(j)$, $p(l+1) < p(l)$. Min-interval is defined analogously.

The paper gives upper bounds for testing identity of k-modal distributions. We must note that this thesis is a direct continuation of their work and gives better upper bounds for same problems.

Also, upper bounds are given for the corresponding cases for monotone distributions.

Such wide range of bounds is a result of a very general method used to obtain them. The key thing in all of them is a domain decomposition algorithm, inspired by the work of Birge ([Bir87b]). Having the domain decomposed properly allows to reduce the problem to a case with a much smaller domain. Then the general testing algorithms can be applied as a black box, or with minor modifications.

Note that in the case of monotone distributions, the decomposition mimics the Birge's construction and is completely oblivious, that is, independent of samples. In

the k-modal case the decomposition is more complicated.

2.2.7 The Power of Linear Estimators

In the previous sections, various algorithms for estimating distributions properties were presented. Some were as easy as presenting the empirical distributions function, others were more sophisticated and involved stronger techniques, such as linear programming.

The paper [VV11b] asks what is the highest necessary power of those algorithms. The essential claim is that for a broad class of properties, namely the so called "symmetric", any estimators (defined similarly to the used in paragraph 2.1.2), even the most tight, have an especially simple form.

The estimators are called "linear" in the sense that the value they return is a dot product of a "fingerprint" of samples with some precomputed vector.

Definition 2.2.4. A fingerprint of a sample set is a vector F, such that $F(i)$ equals the amount of domain elements, encountered in the sample set more than i times.

Any symmetric linear property can be estimated nearly optimally with a fingerprint-based linear estimator.

Mathematically, therefore, the linearity of estimators can be expressed in the following way:

Definition 2.2.5. A symmetric property π is linear if there exists some function $f_\pi : [0,1] \to R$ which, such that for any distribution A with histogram h_A,

$$\pi(A) = \sum_{x:h_A(x)\neq 0} h(x) f_\pi(x)$$

The interesting thing is that the linear upper bounds are connected with corresponding lower bounds via a linear programming duality.

The constructive aspect of the paper is the following:

The paper ([VV11b]) has explicit linear estimators for the following properties: entropy, distance to uniformity and distance between two unknown distributions.

The sample complexity of those properties is

$$O(\frac{n}{\log n})$$

Chapter 3

Upper Bounds

In this chapter we will prove upper bounds for Theorems 1.3.1, 1.3.2 and 1.3.3 to establish the algorithmic part of the thesis.

3.1 A Key Tool: L_2 testing

Before we provide the algorithms and their analysis, we will give two auxiliary testing algorithms to be used as subroutines. Although these algorithms are not the main point of this thesis, they might be interesting in their own right.

Both of these algorithms are used to test identity with respect to the L_2 distance metric. Although our main point is the L_1 distance, there is a strong connection between the two distances, which is one of the key foundations of all our algorithms.

3.1.1 Testing L_2 Uniformity

Theorem 3.1.1. *Given $0 < \varepsilon, \delta < 1$ and sample access to a distribution q over $[n]$, there is an algorithm* Test-Uniformity-$L_2(q, n, \varepsilon, \delta)$ *which uses $m = O\left((\sqrt{n}/\varepsilon^2) \cdot \log(1/\delta)\right)$ samples from q, runs in time linear in its sample size, and with probability at least $1 - \delta$ distinguishes whether $q = U_n$ versus $\|p - q\|_2 \geq \varepsilon/\sqrt{n}$.*

To prove Theorem 3.1.1 we show that a variant of Pearson's chi-squared test [Pea00] – which can be viewed as a special case of the recent "chi-square type" testers in [CDVV14, VV14] – has the desired L_2 guarantee. While this tester has been (implicitly) studied in [CDVV14, VV14], and it is known to be sample optimal with respect to the L_1 norm, it has not been previously analysed for the L_2 norm. The novelty of Theorem 3.1.1 lies in the tight analysis of the algorithm under the L_2 distance.

In this section, we give an algorithm for uniformity testing with respect to the L_2 distance, thereby establishing Theorem 3.1.1. The algorithm Test-Uniformity-$L_2(q,n,\varepsilon)$ described below draws $O(\sqrt{n}/\varepsilon^2)$ samples from a distribution q over $[n]$ and distinguishes between the cases that $q = U_n$ versus $\|q - U_n\|_2 > \varepsilon/\sqrt{n}$ with probability at least $2/3$. Repeating the algorithm $O(\log(1/\delta))$ times and taking the majority answer results in a confidence probability of $1-\delta$, giving the desired algorithm Test-Uniformity-$L_2(q,n,\varepsilon,\delta)$ of Theorem 3.1.1.

Our estimator is a variant of Pearson's chi-squared test [Pea00], and can be viewed as a special case of the recent "chi-square type" testers in [CDVV14, VV14]. We remark that, as follows from the Cauchy-Schwarz inequality, the same estimator distinguishes the uniform distribution from any distribution q such that $\|q-U_n\|_1 > \varepsilon$, i.e., algorithm Test-Uniformity-$L_2(q,n,\varepsilon)$ is an optimal uniformity tester for the L_1 norm. The L_2 guarantee we prove here is new, is strictly stronger than the aforementioned L_1 guarantee, and is crucial for our purposes in Section 3.2.2.

For $\lambda \geq 0$, we denote by $\mathrm{Poi}(\lambda)$ the Poisson distribution with parameter λ. In our algorithm below, we employ the standard "Poissonisation" approach: namely, we assume that, rather than drawing m $(= O\left((\sqrt{n}/\varepsilon^2)\cdot\log(1/\delta)\right))$ independent samples from a distribution, we first select m' from $\mathrm{Poi}(m)$, and then draw m' samples. This Poissonisation makes the number of times different elements occur in the sample independent, with the distribution of the number of occurrences of the i-th domain element distributed as $\mathrm{Poi}(mq_i)$, simplifying the analysis. As $\mathrm{Poi}(m)$ is tightly concentrated about m, we can carry out this Poissonisation trick without loss of generality at the

expense of only sub-constant factors in the sample complexity.

Algorithm Test-Uniformity-$L_2(q,n,\varepsilon)$

Input: sample access to a distribution q over $[n]$, and $\varepsilon > 0$.

Output: "YES" if $q = U_n$; "NO" if $\|q - U_n\|_2 \geq \varepsilon/\sqrt{n}$.

1. Draw $m' \sim \mathrm{Poi}(m)$ i.i.d. samples from q. (Where $m = O\left((\sqrt{n}/\varepsilon^2)\cdot\log(1/\delta)\right)$.)
2. Let X_i be the number of occurrences of the ith domain elements in the sample from q
3. Define $Z = \sum_{i=1}^{n}(X_i - m/n)^2 - X_i$.
4. If $Z \geq 4m/\sqrt{n}$ return "NO"; otherwise, return "YES".

The following theorem characterises the performance of the above estimator:

Theorem 3.1.2. *For any distribution q over $[n]$ the above algorithm distinguishes the case that $q = U_n$ from the case that $||q - U_n||_2 \geq \varepsilon/\sqrt{n}$ when given $O(\sqrt{n}/\varepsilon^2)$ samples from q with probability at least $2/3$.*

Proof. Define $Z_i = (X_i - m/n)^2 - X_i$. Since X_i is distributed as $\mathrm{Poi}(mq_i)$, $\mathbb{E}[Z_i] = m^2\Delta_i^2$, where $\Delta_i := 1/n - q_i$. By linearity of expectation we can write $\mathbb{E}[Z] = \sum_{i=1}^{n}\mathbb{E}[Z_i] = m^2\cdot\sum_{i=1}^{n}\Delta_i^2$. Similarly we can calculate

$$\mathrm{Var}[Z_i] = 2m^2(\Delta_i - 1/n)^2 + 4m^3(1/n - \Delta_i)\Delta_i^2.$$

Since the X_i's (and hence the Z_i's) are independent, it follows that $\mathrm{Var}[Z] = \sum_{i=1}^{n}\mathrm{Var}[Z_i]$.

We start by establishing completeness. Suppose $q = U_n$. We will show that $\Pr[Z \geq 4m/\sqrt{n}] \leq 1/3$. Note that in this case $\Delta_i = 0$ for all $i \in [n]$, hence $\mathbb{E}[Z] = 0$ and $\mathrm{Var}[Z] = 2m^2/n$. Chebyshev's inequality implies that

$$\Pr[Z \geq 4m/\sqrt{n}] = \Pr\left[Z \geq (2\sqrt{2})\sqrt{\mathrm{Var}[Z]}\right] \leq (1/8) < 2/3$$

as desired.

We now proceed to prove the soundness of the tester. Suppose that $\|q-U_n\|_2 \geq \frac{\varepsilon}{\sqrt{n}}$. In this case we will show that $\Pr[Z \leq 4m/\sqrt{n}] \leq 1/3$. Note that Chebyshev's inequality implies that

$$\Pr\left[Z \leq \mathbb{E}[Z] - 2\sqrt{\mathrm{Var}[Z]}\right] \leq 1/4.$$

It thus suffices to show that $\mathbb{E}[Z] \geq 8m/\sqrt{n}$ and $\mathbb{E}[Z]^2 \geq 16\,\mathrm{Var}[Z]$. Establishing the former inequality is easy. Indeed,

$$\mathbb{E}[Z] = m^2 \cdot \|q-U_n\|_2^2 \geq m^2 \cdot (\varepsilon^2/n) \geq 8m/\sqrt{n}$$

for $m \geq 8\sqrt{n}/\varepsilon^2$.

Proving the latter inequality requires a more detailed analysis. We will show that for a sufficiently large constant $C > 0$, if $m \geq C\sqrt{n}/\varepsilon^2$ we will have

$$\mathrm{Var}[Z] \ll \mathbb{E}[Z]^2.$$

Ignoring multiplicative constant factors, we equivalently need to show that

$$m^2 \cdot \left(\sum_{i=1}^{n} \left(\Delta_i^2 - 2\Delta_i/n\right) + 1/n\right) + m^3 \sum_{i=1}^{n} \left(\Delta_i^2/n + \Delta_i^3\right) \ll m^4 \left(\sum_{i=1}^{n} \Delta_i^2\right)^2.$$

To prove the desired inequality, it suffices to bound from above the absolute value of each of the five terms of the LHS separately. For the first term we need to show that $m^2 \cdot \sum_{i=1}^n \Delta_i^2 \ll m^4 \cdot \left(\sum_{i=1}^n \Delta_i^2\right)^2$ or equivalently

$$m \gg 1/\|q-U_n\|_2. \tag{3.1}$$

Since $\|q-U_n\|_2 \geq \varepsilon/\sqrt{n}$, the RHS of (3.1) is bounded from above by $\sqrt{n}/\varepsilon$, hence (3.1) holds true for our choice of m.

For the second term we want to show that $\sum_{i=1}^n |\Delta_i| \ll m^2 n \cdot \left(\sum_{i=1}^n \Delta_i^2\right)^2$. Recalling that $\sum_{i=1}^n |\Delta_i| \leq \sqrt{n} \cdot \sqrt{\sum_{i=1}^n \Delta_i^2}$, as follows from the Cauchy-Schwarz inequality, it suffices to show that $m^2 \gg (1/\sqrt{n}) \cdot 1/(\sum_{i=1}^n \Delta_i^2)^{3/2}$ or equivalently

$$m \gg \frac{1}{n^{1/4}} \cdot \frac{1}{\|q-U_n\|_2^{3/2}}. \tag{3.2}$$

Since $\|q-U_n\|_2 \geq \varepsilon/\sqrt{n}$, the RHS of (3.2) is bounded from above by $\sqrt{n}/\varepsilon^{3/2}$, hence (3.2) is also satisfied.

For the third term we want to argue that $m^2/n \ll m^4 \cdot \left(\sum_{i=1}^n \Delta_i^2\right)^2$ or

$$m \gg \frac{1}{n^{1/2}} \cdot \frac{1}{\|q-U_n\|_2^2}, \tag{3.3}$$

which holds for our choice of m, since the RHS is bounded from above by $\sqrt{n}/\varepsilon^2$.

Bounding the fourth term amounts to showing that $(m^3/n)\sum_{i=1}^n \Delta_i^2 \ll m^4 \left(\sum_{i=1}^n \Delta_i^2\right)^2$ which can be rewritten as

$$m \gg \frac{1}{n} \cdot \frac{1}{\|q-U_n\|_2^2}, \tag{3.4}$$

and is satisfied since the RHS is at most $1/\varepsilon^2$.

Finally, for the fifth term we want to prove that $m^3 \cdot \sum_{i=1}^n |\Delta_i|^3 \ll m^4 \cdot \left(\sum_{i=1}^n \Delta_i^2\right)^2$ or that $\sum_{i=1}^n |\Delta_i|^3 \ll m \cdot \left(\sum_{i=1}^n \Delta_i^2\right)^2$. From Jensen's inequality it follows that $\sum_{i=1}^n |\Delta_i|^3 \leq \left(\sum_{i=1}^n \Delta_i|^2\right)^{3/2}$; hence, it is sufficient to show that $\left(\sum_{i=1}^n \Delta_i|^2\right)^{3/2} \ll m \cdot \left(\sum_{i=1}^n \Delta_i^2\right)^2$ or

$$m \gg \frac{1}{\|q-U_n\|_2}. \tag{3.5}$$

Since $\|q-U_n\|_2 \geq \varepsilon/\sqrt{n}$ the above RHS is at most $\sqrt{n}/\varepsilon$ and (3.5) is satisfied. This completes the soundness proof and the proof of Theorem 3.1.2. □

3.1.2 Testing $\mathcal{A}_k$ Identity to Unknown Distribution

The next theorem is in some sense analogous to Theorem 3.1.1. It appeared as a result of an attempt to create an identity tester for two unknown distributions using a "Chi-squared-like" tester. It turned out, however, that such an attempt only works if the distributions are really close to each other. Nevertheless, it is used as a subroutine in a general algorithm of Section 3.3 in Theorem 3.3.11.

Theorem 3.1.3. *Let p,q be discrete distributions over $[n]$ satisfying $\|p\|_2, \|q\|_2 = O(1/\sqrt{n})$. There exists a testing algorithm with the following properties: On input $k \in \mathbb{Z}_+$, $2 \leq k \leq n$, and $\delta, \varepsilon > 0$, the algorithm draws $O\left((\sqrt{k}/\varepsilon^2) \cdot \log(1/\delta)\right)$ samples from p and q and with probability at least $1-\delta$ distinguishes between the cases $p = q$ and $\|p-q\|_{\mathcal{A}_k} > \varepsilon$.*

The above proposition says that the identity testing problem under the $\mathcal{A}_k$ distance can be solved with $O(\sqrt{k}/\varepsilon^2)$ samples when both distributions p and q are promised to be "nearly" uniform (in the sense that their L_2 norm is $O(1)$ times that of the uniform distribution). To prove Proposition 3.1.3 we follow a similar approach as in [DKN15b]: Starting from the L_2 identity tester of [CDVV14], we consider several oblivious interval decompositions of the domain into intervals of approximately the same mass, and apply a "reduced" identity tester for each such decomposition.

We note that it suffices to attain confidence probability $2/3$ with $O(\sqrt{k}/\varepsilon^2)$ samples, as we can then run $O(\log(1/\delta))$ independent iterations to boost the confidence to $1-\delta$. Our starting point is the following Theorem from [CDVV14]:

Theorem 3.1.4 ([CDVV14], Proposition 3.1)**.** *For any distributions p and q over $[n]$ such that $\|p\|_2 \leqslant \frac{O(1)}{\sqrt{n}}$ and $\|q\|_2 \leqslant \frac{O(1)}{\sqrt{n}}$ there is a testing algorithm that distinguishes with probability at least $2/3$ the case that $q = p$ from the case that $||q-p||_2 \geqslant \varepsilon/\sqrt{n}$ when given $O(\sqrt{n}/\varepsilon^2)$ samples from q and p.*

Our $\mathcal{A}_k$ testing algorithm for this regime is the following:

Algorithm Test-Identity-Flat-$\mathcal{A}_k(p,q,n,\varepsilon)$

Input: sample access to distributions p and q over $[n]$ with $\|p\|_2, \|q\|_2 = O(1/\sqrt{n})$, $k \in \mathbb{Z}_+$ with $2 \le k \le n$, and $\varepsilon > 0$.

Output: "YES" if $q = p$; "NO" if $\|q-p\|_{\mathcal{A}_k} \ge \varepsilon$.

1. Draw samples S_1, S_2 of size $m = O(\sqrt{k}/\varepsilon^2)$ from q and p.
2. By artificially increasing the support if necessary, we can guarantee that $n = k \cdot 2^{j_0}$, where $j_0 \stackrel{\text{def}}{=} \lceil \log_2(1/\varepsilon) \rceil + O(1)$.
3. Consider the collection $\{I^{(j)}\}_{j=0}^{j_0-1}$ of j_0 partitions of $[n]$ into intervals; the partition $I^{(j)} = (I_i^{(j)})_{i=1}^{\ell_j}$ consists of $\ell_j = k \cdot 2^j$ many intervals with $I_i^{(j_0)}$ of length $n/\ell_j + O(1)$, and $I_i^{(j)}$ the union of two adjacent intervals of $I_i^{(j+1)}$.
4. For $j = 0, 1, \ldots, j_0 - 1$:
   (a) Consider the reduced distributions $q_r^{I^{(j)}}$ and $p_r^{I^{(j)}}$. Use the samples S_1, S_2 to simulate samples to $q_r^{I^{(j)}}$ and $p_r^{I^{(j)}}$.
   (b) Run Test-Identity-$L_2(q_r^{I^{(j)}}, p_r^{I^{(j)}}, \ell_j, \varepsilon_j, \delta_j)$ for $\varepsilon_j = C \cdot \varepsilon \cdot 2^{3j/8}$ for $C > 0$ a sufficiently small constant and $\delta_j = 2^{-j}/6$, i.e., test whether $q_r^{I^{(j)}} = p_r^{I^{(j)}}$ versus $\|q_r^{I^{(j)}} - p_r^{I^{(j)}}\|_2 > \gamma_j \stackrel{\text{def}}{=} \varepsilon_j/\sqrt{\ell_j}$.
5. If all the testers in Step 3(b) output "YES", then output "YES"; otherwise output "NO".

Note in the above that when $\varepsilon_j > 1$, that the appropriate tester requires no samples.

The following proposition characterises the performance of the above algorithm.

Proposition 3.1.5. *The algorithm* Test-Identity-Flat-$\mathcal{A}_k(p,q,n,\varepsilon)$*, on input a sample of size* $m = O(\sqrt{k}/\varepsilon^2)$ *drawn from distributions* q *and* p *over* $[n]$ *with* $\|p\|_2, \|q\|_2 = O(1/\sqrt{n})$*,* $\varepsilon > 0$*, and an integer* k *with* $2 \le k \le n$*, correctly distinguishes the case that* $q = p$ *from the case that* $\|q-p\|_{\mathcal{A}_k} \ge \varepsilon$*, with probability at least* $2/3$*.*

Proof. First, it is straightforward to verify the claimed sample complexity, as the algo-

rithm only draws samples in Step 1.

Note that the algorithm uses the same set of samples S_1, S_2 for all testers in Step 4(b). Note that it is easy to see that $\|p_r^{I^{(j)}}\|_2, \|q_r^{I^{(j)}}\|_2 = O(1/\sqrt{\ell_j})$, and therefore, by Theorem 3.1.4, the tester Test-Identity-$L_2(q_r^{I^{(j)}}, p_r^{I^{(j)}}, \ell_j, \varepsilon_j, \delta_j)$, on input a set of $m_j = O((\sqrt{\ell_j}/\varepsilon_j^2)\cdot\log(1/\delta_j))$ samples from $q_r^{I^{(j)}}$ and $p_r^{I^{(j)}}$ distinguishes the case that $q_r^{I^{(j)}} = p_r^{I^{(j)}}$ from the case that $\|q_r^{I^{(j)}} - p_r^{I^{(j)}}\|_2 \geq \gamma_j \stackrel{\text{def}}{=} \varepsilon_j/\sqrt{\ell_j}$ with probability at least $1-\delta_j$. From our choice of parameters it can be verified that $\max_j m_j \leq m = O(\sqrt{k}/\varepsilon^2)$, hence we can use the same sample S_1, S_2 as input to these testers for all $0 \leq j \leq j_0 - 1$. In fact, it is easy to see that $\sum_{j=0}^{j_0-1} m_j = O(m)$, which implies that the overall algorithm runs in sample-linear time. Since each tester in Step 3(b) has error probability δ_j, by a union bound over all $j \in \{0, \ldots, j_0 - 1\}$, the total error probability is at most $\sum_{j=0}^{j_0-1} \delta_j \leq (1/6)\cdot\sum_{j=0}^{\infty} 2^{-j} = 1/3$. Therefore, with probability at least $2/3$ all the testers in Step 4(b) succeed. We will henceforth condition on this "good" event, and establish the completeness and soundness properties of the overall algorithm under this conditioning.

We start by establishing completeness. If $q = p$, then for any partition $I^{(j)}$, $0 \leq j \leq j_0 - 1$, we have that $q_r^{I^{(j)}} = p_r^{I^{(j)}}$. By our aforementioned conditioning, all testers in Step 3(b) will output "YES", hence the overall algorithm will also output "YES", as desired.

We now proceed to establish the soundness of our algorithm. Assuming that $\|q - p\|_{\mathcal{A}_k} \geq \varepsilon$, we want to show that the algorithm Test-Identity-$\mathcal{A}_k(q, n, \varepsilon)$ outputs "NO" with probability at least $2/3$. Toward this end, we prove the following structural lemma:

Lemma 3.1.6. *For $C > 0$ a sufficiently small constant, if $\|q - p\|_{\mathcal{A}_k} \geq \varepsilon$, there exists $j \in \mathbb{Z}_+$ with $0 \leq j \leq j_0 - 1$ such that $\|q_r^{I^{(j)}} - p_r^{I^{(j)}}\|_2^2 \geq \gamma_j^2$.*

Given the lemma, the soundness property of our algorithm follows easily. Indeed, since all testers Test-Identity-$L_2(q_r^{I^{(j)}}, \ell_j, \varepsilon_j, \delta_j)$ of Step 4(b) are successful by our conditioning, Lemma 3.1.6 implies that at least one of them outputs "NO", hence the

overall algorithm will output "NO". □

The proof of Lemma 3.1.6 is very similar to the analogous lemma in [DKN15b]. For the same of completeness, it is given in the following subsection.

3.1.2.1 Proof of Lemma 3.1.6

We claim that it is sufficient to take $C \leqslant 5 \cdot 10^{-6}$. Thus, we are in the case where $n = 2^{j_0-1} \cdot k$ and have argued that it suffices to show that our algorithm works to distinguish $\mathcal{A}_k$-distance in this setting with $\varepsilon_j = 10^{-5} \cdot \varepsilon \cdot 2^{3j/8}$.

We make use of the following definition:

Definition 3.1.7. For p and q arbitrary distributions over $[n]$, we define the *scale-sensitive-L_2* distance between q and p to be

$$\|q-p\|_{[k]}^2 \stackrel{\text{def}}{=} \max_{I=(I_1,\ldots,I_r)\in \mathbf{W}_{1/k}} \sum_{i=1}^{r} \frac{\mathbf{Discr}^2(I_i)}{\mathbf{width}^{1/8}(I_i)}$$

where $\mathbf{W}_{1/k}$ is the collection of all interval partitions of $[n]$ into intervals of width at most $1/k$, $\mathbf{Discr}(I) = |p(I) - q(I)|$, and $\mathbf{width}(I)$ is the number of bins in I divided by n.

The first thing we need to show is that if q and p have large $\mathcal{A}_k$ distance then they also have large scale-sensitive-L_2 distance. Indeed, we have the following lemma:

Lemma 3.1.8. *For p and q an arbitrary distributions over $[n]$, we have that*

$$\|q-p\|_{[k]}^2 \geqslant \frac{\|q-p\|_{\mathcal{A}_k}^2}{(2k)^{7/8}}.$$

Proof. Let $\varepsilon = \|q-p\|_{\mathcal{A}_k}^2$. Consider the optimal I^* in the definition of the $\mathcal{A}_k$ distance. By further subdividing intervals of width more than $1/k$ into smaller ones, we can obtain a new partition, $I' = (I'_i)_{i=1}^s$, of cardinality $s \leq 2k$ all of whose parts have width at most $1/k$. Furthermore, we have that $\sum_i \mathbf{Discr}(I'_i) \geqslant \varepsilon$. Using this partition to bound

from below $\|q-p\|_{[k]}^2$, by Cauchy-Schwarz we obtain that

$$\begin{aligned}\|q-p\|_{[k]}^2 &\geqslant \sum_i \frac{\mathbf{Discr}^2(I_i')}{\mathbf{width}(I_i')^{1/8}} \\ &\geqslant \frac{\left(\sum_i \mathbf{Discr}(I_i')\right)^2}{\sum_i \mathbf{width}(I_i')^{1/8}} \\ &\geqslant \frac{\varepsilon^2}{2k(1/(2k))^{1/8}} \\ &= \frac{\varepsilon^2}{(2k)^{7/8}}.\end{aligned}$$

□

The second important fact about the scale-sensitive-L_2 distance is that if it is large then one of the partitions considered in our algorithm will produce a large L_2 error.

Proposition 3.1.9. *Let p and q be distributions over $[n]$. Then we have that*

$$\|q-p\|_{[k]}^2 \leqslant 10^8 \sum_{j=0}^{j_0-1} \sum_{i=1}^{2^j\cdot k} \frac{\mathbf{Discr}^2(I_i^{(j)})}{\mathbf{width}^{1/8}(I_i^{(j)})}. \tag{3.6}$$

Proof. Let $\mathcal{J} \in \mathbf{W}_{1/k}$ be the optimal partition used when computing the scale-sensitive-L_2 distance $\|q-p\|_{[k]}$. In particular, it is a partition into intervals of width at most $1/k$ so that $\sum_i \frac{\mathbf{Discr}^2(J_i)}{\mathbf{width}(J_i)^{1/8}}$ is as large as possible. To prove (3.6), we prove a notably stronger claim. In particular, we will prove that for each interval $J_\ell \in \mathcal{J}$

$$\frac{\mathbf{Discr}^2(J_\ell)}{\mathbf{width}^{1/8}(J_\ell)} \leqslant 10^8 \sum_{j=0}^{j_0-1} \sum_{i:I_i^{(j)}\subset J_\ell} \frac{\mathbf{Discr}^2(I_i^{(j)})}{\mathbf{width}^{1/8}(I_i^{(j)})}. \tag{3.7}$$

Summing over ℓ would then yield $\|q-p\|_{[k]}^2$ on the left hand side and a strict subset of the terms from (3.6) on the right hand side. From here on, we will consider only a single interval J_ℓ. For notational convenience, we will drop the subscript and merely call it J.

First, note that if $|J| \leqslant 10^8$, then this follows easily from considering just the sum over $j = j_0 - 1$. Then, if $t = |J|$, J is divided into t intervals of size 1. The sum of the discrepancies of these intervals equals the discrepancy of J, and thus, the sum of the squares of the discrepancies is at least $\mathbf{Discr}^2(J)/t$. Furthermore, the widths of these

subintervals are all smaller than the width of J by a factor of t. Thus, in this case the sum of the right hand side of (3.7) is at least $1/t^{7/8} \geqslant \frac{1}{10^7}$ of the left hand side.

Otherwise, if $|J| > 10^8$, we can find a j so that $\mathbf{width}(J)/10^8 < 1/(2^j \cdot k) \leqslant 2 \cdot \mathbf{width}(J)/10^8$. We claim that in this case Equation (3.7) holds even if we restrict the sum on the right hand side to this value of j. Note that J contains at most 10^8 intervals of $I^{(j)}$, and that it is covered by these intervals plus two narrower intervals on the ends. Call these end-intervals R_1 and R_2. We claim that $\mathbf{Discr}(R_i) \leqslant \mathbf{Discr}(J)/3$. This is because otherwise it would be the case that

$$\frac{\mathbf{Discr}^2(R_i)}{\mathbf{width}^{1/8}(R_i)} > \frac{\mathbf{Discr}^2(J)}{\mathbf{width}^{1/8}(J)}.$$

(This is because $(1/3)^2 \cdot (2/10^8)^{-1/8} > 1$.) This is a contradiction, since it would mean that partitioning J into R_i and its complement would improve the sum defining $\|q-p\|_{[k]}$, which was assumed to be maximum. This means that the sum of the discrepancies of the $I_i^{(j)}$ contained in J must be at least $\mathbf{Discr}(J)/3$, so the sum of their squares is at least $\mathbf{Discr}^2(J)/(9 \cdot 10^8)$. On the other hand, each of these intervals is narrower than J by a factor of at least $10^8/2$, thus the appropriate sum of $\frac{\mathbf{Discr}^2(I_i^{(j)})}{\mathbf{width}^{1/8}(I_i^{(j)})}$ is at least $\frac{\mathbf{Discr}^2(J)}{10^8 \mathbf{width}^{1/8}(J)}$. This completes the proof. □

We are now ready to prove Lemma 3.1.6.

Proof. If $\|q-p\|_{\mathcal{A}_k} \geqslant \varepsilon$ we have by Lemma 3.1.8 that

$$\|q-p\|_{[k]}^2 \geqslant \frac{\varepsilon^2}{(2k)^{7/8}}.$$

By Proposition 3.1.9, this implies that

$$\begin{aligned}\frac{\varepsilon^2}{(2k)^{7/8}} &\leqslant 10^8 \sum_{j=0}^{j_0-1} \sum_{i=1}^{2^j \cdot k} \frac{\mathbf{Discr}^2(I_i^{(j)})}{\mathbf{width}^{1/8}(I_i^{(j)})} \\ &= 10^8 \sum_{j=0}^{j_0-1} (2^j k)^{1/8} \|q^{I^{(j)}} - p^{I^{(j)}}\|_2^2.\end{aligned}$$

Therefore,

$$\sum_{j=0}^{j_0-1} 2^{j/8} \|q^{I^{(j)}} - p^{I^{(j)}}\|_2^2 \geqslant 5 \cdot 10^{-9} \varepsilon^2/k. \tag{3.8}$$

On the other hand, if $\|q^{I^{(j)}} - p^{I^{(j)}}\|_2^2$ were at most $10^{-10}2^{-j/4}\varepsilon^2/k$ for each j, then the sum above would be at most

$$10^{-10}\varepsilon^2/k\sum_j 2^{-j/8} < 5\cdot 10^{-9}\varepsilon^2/k.$$

This would contradict Equation (3.8), thus proving that $\|q^{I^{(j)}} - U_{\ell_j}\|_2^2 \geqslant 10^{-10}2^{-j/4}\varepsilon^2/k$ for at least one j, proving Lemma 3.1.6. □

3.2 Testing Identity to a Known Distribution

Theorem 3.2.1 (Identity to known (Theorem 1.3.1))**.** *Given $\varepsilon > 0$, an integer k with $2 \leq k \leq n$, sample access to a distribution q over $[n]$, and an explicit distribution p over $[n]$, there is a computationally efficient algorithm which uses $O(\sqrt{k}/\varepsilon^2)$ samples from q, and with probability at least $2/3$ distinguishes whether $q = p$ versus $\|p-q\|_{\mathcal{A}_k} \geq \varepsilon$.*

The proof of Theorem 3.2.1 proceeds in two stages: In the first stage, we reduce the $\mathcal{A}_k$ identity testing problem to $\mathcal{A}_k$ uniformity testing without incurring any loss in the sample complexity. In the second stage, we use an optimal L_2 *uniformity* tester as a black-box to obtain an $O(\sqrt{k}/\varepsilon^2)$ sample algorithm for $\mathcal{A}_k$ uniformity testing. We remark that the L_2 uniformity tester is not applied to the distribution q directly, but to a sequence of reduced distributions q_r^I, for an appropriate collection of interval partitions I.

We remark that an application of Theorem 3.2.1 for $k = n$, yields a sample optimal L_1 identity tester (for an arbitrary distribution q), giving a new algorithm matching the recent tight upper bound in [VV14]. Our new L_1 identity tester is arguably simpler and more intuitive, as it only uses an L_2 uniformity tester in a black-box manner.

3.2.1 The Intuitive Explanation

We now provide a detailed intuitive explanation of the ideas that lead to our main result, Theorem 3.2.1. Given sample access to a distribution q and an explicit distribution

p, we want to test whether $q = p$ versus $\|q-p\|_{\mathcal{A}_k} \geq \varepsilon$. By definition we have that $\|q-p\|_{\mathcal{A}_k} = \max_I \|q_r^I - p_r^I\|_1$. So, if the "optimal" partition $\mathcal{J}^* = (J_i^*)_{j=1}^k$ maximising this expression was known a priori, the problem would be easy: Our algorithm could then consider the reduced distributions $q_r^{\mathcal{J}^*}$ and $p_r^{\mathcal{J}^*}$, which are supported on sets of size k, and call a standard L_1-identity tester to decide whether $q_r^{\mathcal{J}^*} = p_r^{\mathcal{J}^*}$ versus $\|q_r^{\mathcal{J}^*} - p_r^{\mathcal{J}^*}\|_1 \geq \varepsilon$. (Note that for any given partition I of $[n]$ into intervals and any distribution q, given sample access to q one can simulate sample access to the reduced distribution q_r^I.) The difficulty, of course, is that the optimal k-partition is *not fixed*, as it depends on the unknown distribution q, thus it is not available to the algorithm. Hence, a more refined approach is necessary.

Our starting point is a new, simple reduction of the general problem of identity testing to its special case of uniformity testing. The main idea of the reduction is to appropriately "stretch" the domain size, using the explicit distribution p, in order to transform the identity testing problem between q and p into a uniformity testing problem for a (different) distribution q' (that depends on q and p). To show correctness of this reduction we need to show that it preserves the $\mathcal{A}_k$ distance, and that we can sample from q' given samples from q.

We now proceed with the details. Since p is given explicitly in the input, we assume for simplicity that each p_i is a rational number, hence there exists some (potentially large) $N \in \mathbb{Z}_+$ such that $p_i = \alpha_i/N$, where $\alpha_i \in \mathbb{Z}_+$ and $\sum_{i=1}^n \alpha_i = N$.[1] Given sample access to q and an explicit p over $[n]$, we construct an instance of the uniformity testing problem as follows: Let p' be the uniform distribution over $[N]$ and let q' be the distribution over $[N]$ obtained from q by subdividing the probability mass of q_i, $i \in [n]$ equally among α_i new consecutive points. It is clear that this reduction preserves the $\mathcal{A}_k$ distance, i.e., $\|q-p\|_{\mathcal{A}_k} = \|q'-p'\|_{\mathcal{A}_k}$. The only remaining task is to show how

[1] We remark that this assumption is not necessary: For the case of irrational p_i's we can approximate them by rational numbers $\widetilde{p}_i$ up to sufficient accuracy and proceed with the approximate distribution $\widetilde{p}$. This approximation step does not preserve perfect completeness; however, we point out that our testers have some mild robustness (of at least a global constant with respect to a distance) in the completeness case, which suffices for all the arguments to go through.

to simulate sample access to q', given samples from q. Given a sample i from q, our sample for q' is selected uniformly at random from the corresponding set of α_i many new points. Hence, we have reduced the problem of identity testing between q and p in $\mathcal{A}_k$ distance, to the problem of uniformity testing of q' in $\mathcal{A}_k$ distance. Note that this reduction is also computationally efficient, as it only requires $O(n)$ pre-computation to specify the new intervals.

For the rest of this section, we focus on the problem of $\mathcal{A}_k$ uniformity testing. For notational convenience, we will use q to denote the unknown distribution and p to denote the uniform distribution over $[n]$. The rough idea is to consider an appropriate collection of interval partitions of $[n]$ and call a standard L_1-uniformity tester for each of these partitions. To make such an approach work and give us a *sample optimal* algorithm for our $\mathcal{A}_k$-uniformity testing problem we need to use a subtle and strong property of uniformity testing, namely its performance guarantee under the L_2 norm. We elaborate on this point below.

For any partition I of $[n]$ into k intervals by definition we have that $\|q_r^I - p_r^I\|_1 \leq \|q-p\|_{\mathcal{A}_k}$. Therefore, if $q = p$, we will also have $q_r^I = p_r^I$. The issue is that $\|q_r^I - p_r^I\|_1$ can be much smaller than $\|q-p\|_{\mathcal{A}_k}$; in fact, it is not difficult to construct examples where $\|q-p\|_{\mathcal{A}_k} = \Omega(1)$ and $\|q_r^I - p_r^I\|_1 = 0$. In particular, it is possible for the points where q is larger than p, and where it is smaller than p to cancel each other out within each interval in the partition, thus making the partition useless for distinguishing q from p. In other words, if the partition I is not "good", we may not be able to detect any existing discrepancy. A simple, but suboptimal, way to circumvent this issue is to consider a partition I' of $[n]$ into $k' = \Theta(k/\varepsilon)$ intervals of the same length. Note that each such interval will have probability mass $1/k' = \Theta(\varepsilon/k)$ under the uniform distribution p. If the constant in the big-Θ is appropriately selected, say $k' = 10k/\varepsilon$, it is not hard to show that $\|q_r^{I'} - p_r^{I'}\|_1 \geq \|q-p\|_{\mathcal{A}_k} - \varepsilon/2$; hence, we will necessarily detect a large discrepancy for the reduced distribution. By applying the optimal L_1 uniformity tester this approach will require $\Omega(\sqrt{k'}/\varepsilon^2) = \Omega(\sqrt{k}/\varepsilon^{2.5})$ samples.

A key tool that is essential in our analysis is a strong property of uniformity testing. An optimal L_1 uniformity tester for q can distinguish between the uniform distribution and the case that $\|q-p\|_1 \geq \varepsilon$ using $O(\sqrt{n}/\varepsilon^2)$ samples. However, a stronger guarantee is possible: With the *same* sample size, we can distinguish the uniform distribution from the case that $\|q-p\|_2 \geq \varepsilon/\sqrt{n}$. We emphasise that such a strong L_2 guarantee is specific to uniformity testing, and is provably not possible for the general problem of identity testing. In previous work, Goldreich and Ron [GR00] gave such an L_2 guarantee for uniformity testing, but their algorithm uses $O(\sqrt{n}/\varepsilon^4)$ samples. Paninski's $O(\sqrt{n}/\varepsilon^2)$ uniformity tester works for the L_1 norm, but it is not known whether it achieves the desired L_2 property. As one of our main tools we use the following L_2 tester, which is optimal as a function of n and ε:

Please, refer to the Section 3.1.1 for a complete correctness proof.

Theorem 3.2.2. *Given $0 < \varepsilon, \delta < 1$ and sample access to a distribution q over $[n]$, there is an algorithm* Test-Uniformity-$L_2(q,n,\varepsilon,\delta)$ *which uses $m = O\left((\sqrt{n}/\varepsilon^2)\cdot\log(1/\delta)\right)$ samples from q, runs in time linear in its sample size, and with probability at least $1-\delta$ distinguishes whether $q = U_n$ versus $\|p-q\|_2 \geq \varepsilon/\sqrt{n}$.*

Armed with Theorem 3.2.2 we proceed as follows: We consider a set of $j_0 = O(\log(1/\varepsilon))$ different partitions of the domain $[n]$ into intervals. For $0 \leq j < j_0$ the partition $I^{(j)}$ consists of $\ell_j \stackrel{\text{def}}{=} |I^{(j)}| = k \cdot 2^j$ many intervals $I_i^{(j)}$, $i \in [\ell_j]$, i.e., $I^{(j)} = (I_i^{(j)})_{i=1}^{\ell_j}$. For a fixed value of j, all intervals in $I^{(j)}$ have the same length, or equivalently, the same probability mass *under the uniform distribution*. Then, for any fixed $j \in [j_0]$, we have $p(I_i^{(j)}) = 1/(k \cdot 2^j)$ for all $i \in [\ell_j]$. (Observe that, by our aforementioned reduction to the uniform case, we may assume that the domain size n is a multiple of $k2^{j_0}$, and therefore that it is possible to evenly divide into such intervals of the same length).

Note that if $q = p$, then for all $0 \leq j < j_0$, it holds $q_r^{I^{(j)}} = p_r^{I^{(j)}}$. Recalling that all intervals in $I^{(j)}$ have the same probability mass *under* p, it follows that $p_r^{I^{(j)}} = U_{\ell_j}$, i.e., $p_r^{I^{(j)}}$ is the uniform distribution over its support. So, if $q = p$, for any partition we have

$q_r^{I^{(j)}} = U_{\ell_j}$. Our main structural result (Lemma 3.2.4) is a robust inverse lemma: *If q is far from uniform in $\mathcal{A}_k$ distance then, for at least one of the partitions $I^{(j)}$, the reduced distribution $q_r^{I^{(j)}}$ will be far from uniform in L_2 distance.* The quantitative version of this statement is quite subtle. In particular, we start from the assumption of being ε-far in $\mathcal{A}_k$ distance and can only deduce "far" in L_2 distance. This is absolutely critical for us to be able to obtain the optimal sample complexity.

The key insight for the analysis comes from noting that the optimal partition separating q from p in $\mathcal{A}_k$ distance cannot have too many parts. Specifically, if the "highs" and "lows" cancel out over some small intervals, they must be very large in order to compensate for the fact that they are relatively narrow. Therefore, when p and q differ on a smaller scale, their L_2 discrepancy will be greater, and this compensates for the fact that the partition detecting this discrepancy will need to have more intervals in it.

In Section 3.2.2 we present our sample optimal uniformity tester under the $\mathcal{A}_k$ distance, thereby establishing Theorem 3.2.1.

3.2.2 Testing Uniformity under the $\mathcal{A}_k$-norm

Algorithm Test-Uniformity-$\mathcal{A}_k(q,n,\varepsilon)$

Input: sample access to a distribution q over $[n]$, $k \in \mathbb{Z}_+$ with $2 \le k \le n$, and $\varepsilon > 0$.

Output: "YES" if $q = U_n$; "NO" if $\|q - U_n\|_{\mathcal{A}_k} \ge \varepsilon$.

1. Draw a sample S of size $m = O(\sqrt{k}/\varepsilon^2)$ from q.
2. Fix $j_0 \in \mathbb{Z}_+$ such that $j_0 \stackrel{\text{def}}{=} \lceil \log_2(1/\varepsilon) \rceil + O(1)$. Consider the collection $\{I^{(j)}\}_{j=0}^{j_0-1}$ of j_0 partitions of $[n]$ into intervals; the partition $I^{(j)} = (I_i^{(j)})_{i=1}^{\ell_j}$ consists of $\ell_j = k \cdot 2^j$ many intervals with $p(I_i^{(j)}) = 1/(k \cdot 2^j)$, where $p \stackrel{\text{def}}{=} U_n$.
3. For $j = 0, 1, \ldots, j_0 - 1$:
   (a) Consider the reduced distributions $q_r^{I^{(j)}}$ and $p_r^{I^{(j)}} \equiv U_{\ell_j}$. Use the sample S to simulate samples to $q_r^{I^{(j)}}$.
   (b) Run Test-Uniformity-$L_2(q_r^{I^{(j)}}, \ell_j, \varepsilon_j, \delta_j)$ for $\varepsilon_j = C \cdot \varepsilon \cdot 2^{3j/8}$ for $C > 0$ a sufficiently small constant and $\delta_j = 2^{-j}/6$, i.e., test whether $q_r^{I^{(j)}} = U_{\ell_j}$ versus $\|q_r^{I^{(j)}} - U_{\ell_j}\|_2 > \gamma_j \stackrel{\text{def}}{=} \varepsilon_j/\sqrt{\ell_j}$.
4. If all the testers in Step 3(b) output "YES", then output "YES"; otherwise output "NO".

Proposition 3.2.3. *The algorithm* Test-Uniformity-$\mathcal{A}_k(q,n,\varepsilon)$*, on input a sample of size* $m = O(\sqrt{k}/\varepsilon^2)$ *drawn from a distribution* q *over* $[n]$*,* $\varepsilon > 0$ *and an integer* k *with* $2 \le k \le n$*, correctly distinguishes the case that* $q = U_n$ *from the case that* $\|q - U_n\|_{\mathcal{A}_k} \ge \varepsilon$*, with probability at least* $2/3$*.*

Proof. First, it is straightforward to verify the claimed sample complexity, as the algorithm only draws samples in Step 1. Note that the algorithm uses the same set of samples S for all testers in Step 3(b).

By Theorem 3.2.2, the tester Test-Uniformity-$L_2(q_r^{I^{(j)}}, \ell_j, \varepsilon_j, \delta_j)$, on input a set of $m_j = O((\sqrt{\ell_j}/\varepsilon_j^2) \cdot \log(1/\delta_j))$ samples from $q_r^{I^{(j)}}$ distinguishes the case that $q_r^{I^{(j)}} = U_{\ell_j}$ from the case that $\|q_r^{I^{(j)}} - U_{\ell_j}\|_2 \geq \gamma_j \stackrel{\text{def}}{=} \varepsilon_j/\sqrt{\ell_j}$ with probability at least $1 - \delta_j$. From our choice of parameters it can be verified that $\max_j m_j \leq m = O(\sqrt{k}/\varepsilon^2)$, hence we can use the same sample S as input to these testers for all $0 \leq j \leq j_0 - 1$. In fact, it is easy to see that $\sum_{j=0}^{j_0-1} m_j = O(m)$, which implies that the overall algorithm runs in sample-linear time. Since each tester in Step 3(b) has error probability δ_j, by a union bound over all $j \in \{0, \ldots, j_0 - 1\}$, the total error probability is at most $\sum_{j=0}^{j_0-1} \delta_j \leq (1/6) \cdot \sum_{j=0}^{\infty} 2^{-j} = 1/3$. Therefore, with probability at least $2/3$ all the testers in Step 3(b) succeed. We will henceforth condition on this "good" event, and establish the completeness and soundness properties of the overall algorithm under this conditioning.

We start by establishing completeness. If $q = p = U_n$, then for any partition $I^{(j)}$, $0 \leq j \leq j_0 - 1$, we have that $q_r^{I^{(j)}} = p_r^{I^{(j)}} = U_{\ell_j}$. By our aforementioned conditioning, all testers in Step 3(b) will output "YES", hence the overall algorithm will also output "YES", as desired.

We now proceed to establish the soundness of our algorithm. Assuming that $\|q - p\|_{\mathcal{A}_k} \geq \varepsilon$, we want to show that the algorithm Test-Uniformity-$\mathcal{A}_k(q, n, \varepsilon)$ outputs "NO" with probability at least $2/3$. Toward this end, we prove the following structural lemma:

Lemma 3.2.4. *There exists a constant $C > 0$ such that the following holds: If $\|q - p\|_{\mathcal{A}_k} \geq \varepsilon$, there exists $j \in \mathbb{Z}_+$ with $0 \leq j \leq j_0 - 1$ such that*

$$\|q_r^{I^{(j)}} - U_{\ell_j}\|_2^2 \geq \gamma_j^2 \stackrel{\text{def}}{=} \varepsilon_j^2/\ell_j = C^2 \cdot (\varepsilon^2/k) \cdot 2^{-j/4}.$$

Given the lemma, the soundness property of our algorithm follows easily. Indeed, since all testers Test-Uniformity-$L_2(q_r^{I^{(j)}}, \ell_j, \varepsilon_j, \delta_j)$ of Step 3(b) are successful by our conditioning, Lemma 3.2.4 implies that at least one of them outputs "NO", hence the overall algorithm will output "NO". □

The proof of Lemma 3.2.4 in its full generality is quite technical. For the sake of the intuition, in the following subsection (Section 3.2.3) we provide a proof of the lemma for the important special case that the unknown distribution q is promised to be *k-flat*, i.e., piecewise constant with k pieces. This setting captures many of the core ideas and, at the same time, avoids some of the necessary technical difficulties of the general case. Finally, in Section 3.2.4 we present our proof for the general case.

3.2.3 Proof of Structural Lemma: k-flat Case

For this special case we will prove the lemma for $C = 1/80$. Since q is k-flat there exists a partition $I^* = (I_j^*)_{j=1}^k$ of $[n]$ into k intervals so that q is constant within each such interval. This in particular implies that $\|q-p\|_{\mathcal{A}_k} = \|q-p\|_1$, where $p \stackrel{\text{def}}{=} U_n$. For $J \in I^*$ let us denote by q_J the value of q within interval J, that is, for all $j \in [k]$ and $i \in I_j^*$ we have $q_i = q_{I_j^*}$. For notational convenience, we sometimes use p_J to denote the value of $p = U_n$ within interval J. By assumption we have that $\|q-p\|_1 = \sum_{j=1}^k |I_j^*| \cdot |q_{I_j^*} - 1/n| \geq \varepsilon$.

Throughout the proof, we work with intervals $I_j^* \in I^*$ such that $q_{I_j^*} < 1/n$. We will henceforth refer to such intervals as *troughs* and will denote by $\mathcal{T} \subseteq [k]$ the corresponding set of indices, i.e., $\mathcal{T} = \{j \in [k] \mid q_{I_j^*} < 1/n\}$. For each trough $J \in \{I_j^*\}_{j\in\mathcal{T}}$ we define its *depth* as $\mathbf{depth}(J) = (p_J - q_J)/p_J = n \cdot (1/n - q_J)$ and its *width* as $\mathbf{width}(J) = p(J) = (1/n) \cdot |J|$. Note that the width of J is identified with the probability mass that the uniform distribution assigns to it. The *discrepancy* of a trough J is defined by $\mathbf{Discr}(J) = \mathbf{depth}(J) \cdot \mathbf{width}(J) = |J| \cdot (1/n - q_J)$ and corresponds to the contribution of J to the L_1 distance between q and p.

It follows from Scheffe's identity that half of the contribution to $\|q-p\|_1$ comes from troughs, namely $\|q-p\|_1^{\mathcal{T}} \stackrel{\text{def}}{=} \sum_{j\in\mathcal{T}} \mathbf{Discr}(I_j^*) = (1/2) \cdot \|q-p\|_1 \geq \varepsilon/2$. An important observation is that we may assume that all troughs have *width* at most $1/k$ at the cost of potentially doubling the total number of intervals. Indeed, it is easy to see that we can artificially subdivide "wider" troughs so that each new trough has width

at most $1/k$. This process comes at the expense of at most doubling the number of troughs. Let us denote by $\{\widetilde{I}_j\}_{j\in\mathcal{T}'}$ this set of (new) troughs, where $|\mathcal{T}'| \le 2k$ and each $\widetilde{I}_j$ is a subset of some I_i^*, $i \in \mathcal{T}$. We will henceforth deal with the set of troughs $\{\widetilde{I}_j\}_{j\in\mathcal{T}'}$ each of width at most $1/k$. By construction, it is clear that

$$\|q-p\|_1^{\mathcal{T}'} \stackrel{\text{def}}{=} \sum_{j\in\mathcal{T}'} \mathbf{Discr}(\widetilde{I}_j) = \|q-p\|_1^{\mathcal{T}} \ge \varepsilon/2. \tag{3.9}$$

At this point we note that we can essentially ignore troughs $J \in \{\widetilde{I}_j\}_{j\in\mathcal{T}'}$ with small discrepancy. Indeed, the total contribution of intervals $J \in \{\widetilde{I}_j\}_{j\in\mathcal{T}'}$ with $\mathbf{Discr}(J) \le \varepsilon/20k$ to the LHS of (3.9) is at most $|\mathcal{T}'| \cdot (\varepsilon/20k) \le 2k \cdot (\varepsilon/20k) = \varepsilon/10$. Let $\mathcal{T}^*$ be the subset of $\mathcal{T}'$ corresponding to troughs with discrepancy at least $\varepsilon/20k$, i.e., $j \in \mathcal{T}^*$ if and only if $j \in \mathcal{T}'$ and $\mathbf{Discr}(\widetilde{I}_j) \ge \varepsilon/20k$. Then, we have that

$$\|q-p\|_1^{\mathcal{T}^*} \stackrel{\text{def}}{=} \sum_{j\in\mathcal{T}^*} \mathbf{Discr}(\widetilde{I}_j) \ge 2\varepsilon/5. \tag{3.10}$$

Observe that for any interval J it holds $\mathbf{Discr}(J) \le \mathbf{width}(J)$. Note that this part of the argument depends critically on considering only troughs. Hence, for $j \in \mathcal{T}^*$ we have that

$$\varepsilon/(20k) \le \mathbf{width}(\widetilde{I}_j) \le 1/k. \tag{3.11}$$

Thus far we have argued that a constant fraction of the contribution to $\|q-p\|_1$ comes from troughs whose width satisfies (3.11). Our next crucial claim is that each such trough must have a "large" overlap with one of the intervals $I_i^{(j)}$ considered by our algorithm Test-Uniformity-$\mathcal{A}_k$. In particular, consider a trough $J \in \{\widetilde{I}_j\}_{j\in\mathcal{T}^*}$. We claim that there exists $j \in \{0,\ldots,j_0-1\}$ and $i \in [\ell_j]$ such that $|I_i^{(j)}| \ge |J|/4$ and so that $I_i^{(j)} \subseteq J$. To see this we first pick a j so that $\mathbf{width}(J)/2 > 2^{-j}/k \ge \mathbf{width}(J)/4$. Since the $I_i^{(j)}$ have width less than half that of J, J must intersect at least three of these intervals. Thus, any but the two outermost such intervals will be entirely contained within J, and furthermore has width $2^j/k \ge \mathbf{width}(J)/4$.

Since the interval $L \in I^{(j+1)}$ is a "domain point" for the reduced distribution $q_r^{I^{(j+1)}}$, the L_1 error between $q_r^{I^{(j+1)}}$ and $U_{\ell_{j+1}}$ incurred by this element is at least $\frac{1}{4}\cdot\mathbf{Discr}(J)$,

and the corresponding L_2^2 error is at least $\frac{1}{16} \cdot (\mathbf{Discr}(J))^2 \geq \frac{\varepsilon}{320k} \cdot \mathbf{Discr}(J)$, where the inequality follows from the fact that $\mathbf{Discr}(J) \geq \varepsilon/(20k)$. Hence, we have that

$$\|q_r^{I^{(j+1)}} - U_{\ell_{j+1}}\|_2^2 \geq \varepsilon/(320k) \cdot \mathbf{Discr}(J). \tag{3.12}$$

As shown above, for every trough $J \in \{\widetilde{I}_j\}_{j \in \mathcal{T}^*}$ there exists a level $j \in \{0, \ldots, j_0 - 1\}$ such that (3.12) holds. Hence, summing (3.12) over all levels we obtain

$$\sum_{j=0}^{j_0-1} \|q_r^{I^{(j+1)}} - U_{\ell_{j+1}}\|_2^2 \geq \varepsilon/(320k) \cdot \sum_{j \in \mathcal{T}^*} \mathbf{Discr}(\widetilde{I}_j) \geq \varepsilon^2/(800k), \tag{3.13}$$

where the second inequality follows from (3.10). Note that

$$\sum_{j=0}^{j_0-1} \gamma_j^2 \leq \sum_{j=0}^{j_0-1} \frac{\varepsilon^2 \cdot 2^{3j/4}}{80^2 \cdot k 2^j} = \frac{\varepsilon^2}{6400k} \sum_{j=0}^{j_0-1} 2^{-j/4} < \varepsilon^2/(800k).$$

Therefore, by the above, we must have that $\|q_r^{I^{(j+1)}} - U_{\ell_{j+1}}\|_2^2 > \gamma_j^2$ for some $0 \leq j \leq j_0 - 1$. This completes the proof of Lemma 3.2.4 for the special case of q being k-flat.

3.2.4 Proof of Structural Lemma: General Case

To prove the general version of our structural result for the $\mathcal{A}_k$ distance, we will need to choose an appropriate value for the universal constant C. We show that it is sufficient to take $C \leq 5 \cdot 10^{-6}$. (While we have not attempted to optimise constant factors, we believe that a more careful analysis will lead to substantially better constants.)

A useful observation is that our Test-Uniformity-$\mathcal{A}_k$ algorithm only distinguishes which of the intervals of $I^{(j_0-1)}$ each of our samples lies in, and can therefore equivalently be thought of as a uniformity tester for the reduced distribution $q_r^{I^{(j_0-1)}}$. In order to show that it suffices to consider only this restricted sample set, we claim that

$$\|q_r^{I^{(j_0-1)}} - U_{\ell_{j_0-1}}\|_{\mathcal{A}_k} \geq \|p - q\|_{\mathcal{A}_k} - \varepsilon/2.$$

In particular, these $\mathcal{A}_k$ distances would be equal if the dividers of the optimal partition for q were all on boundaries between intervals of $I^{(j_0-1)}$. If this was not the case though, we could round the endpoint of each trough inward to the nearest such

boundary (note that we can assume that the optimal partition has no two adjacent troughs). This increases the discrepancy of each trough by at most $2k \cdot 2^{-j_0}$, and thus for $j_0 - \log_2(1/\varepsilon)$ a sufficiently large universal constant, the total discrepancy decreases by at most $\varepsilon/2$.

Thus, we have reduced ourselves to the case where $n = 2^{j_0-1} \cdot k$ and have argued that it suffices to show that our algorithm works to distinguish $\mathcal{A}_k$-distance in this setting with $\varepsilon_j = 10^{-5} \cdot \varepsilon \cdot 2^{3j/8}$.

The analysis of the completeness and the soundness of the tester is identical to Proposition 3.2.3. The only missing piece is the proof of Lemma 3.2.4, which we now restate for the sake of convenience:

Lemma 3.2.5. *If $\|q-p\|_{\mathcal{A}_k} \geq \varepsilon$, there exists some $j \in \mathbb{Z}_+$ with $0 \leq j \leq j_0 - 1$ such that*

$$\|q_r^{I^{(j)}} - U_{\ell_j}\|_2^2 \geq \gamma_j^2 := \varepsilon_j^2/\ell_j = 10^{-10} 2^{-j/4} \varepsilon^2/k.$$

The analysis of the general case here is somewhat more complicated than the special case for q being k-flat case that was presented in the previous section. This is because it is possible for one of the intervals J in the optimal partition (i.e., the interval partition $I^* \in \mathfrak{I}_k$ maximizing $\|q_r^I - q_r^I\|_1$ in the definition of the $\mathcal{A}_k$ distance) to have large overlap with an interval I that our algorithm considers – that is, $I \in \cup_{j=0}^{j_0-1} I^{(j)}$ – without having $q(I)$ and $p(I)$ differ substantially. Note that the unknown distribution q is not guaranteed to be constant within such an interval J, and in particular the difference $q-p$ does not necessarily preserve its sign within J.

To deal with this issue, we note that there are two possibilities for an interval J in the optimal partition: Either one of the intervals $I_i^{(j)}$ (considered by our algorithm) of size at least $|J|/2$ has discrepancy comparable to J, or the distribution q differs from p even more substantially on one of the intervals separating the endpoints of $I_i^{(j)}$ from the endpoints of J. Therefore, either an interval contained within this will detect a large L_2 error, or we will need to again pass to a subinterval. To make this intuition rigorous, we will need a mechanism for detecting where this recursion will terminate. To handle

this formally, we introduce the following definition:

Definition 3.2.6. For $p = U_n$ and q an arbitrary distribution over $[n]$, we define the *scale-sensitive-L_2 distance* between q and p to be

$$\|q-p\|_{[k]}^2 \stackrel{\text{def}}{=} \max_{I=(I_1,\ldots,I_r)\in \mathbf{W}_{1/k}} \sum_{i=1}^{r} \frac{\mathbf{Discr}^2(I_i)}{\mathbf{width}^{1/8}(I_i)}$$

where $\mathbf{W}_{1/k}$ is the collection of all interval partitions of $[n]$ into intervals of width at most $1/k$.

The notion of the scale-sensitive-L_2 distance will be a useful intermediate tool in our analysis. The rough idea of the definition is that the optimal partition will be able to detect the correctly sized intervals for our tester to notice. (It will act as an analogue of the partition into the intervals where q is constant for the k-flat case.)

The first thing we need to show is that if q and p have large $\mathcal{A}_k$ distance then they also have large scale-sensitive-L_2 distance. Indeed, we have the following lemma:

Lemma 3.2.7. *For $p = U_n$ and q an arbitrary distribution over $[n]$, we have that*

$$\|q-p\|_{[k]}^2 \geq \frac{\|q-p\|_{\mathcal{A}_k}^2}{(2k)^{7/8}}.$$

Proof. Let $\varepsilon = \|q-p\|_{\mathcal{A}_k}^2$. Consider the optimal I^* in the definition of the $\mathcal{A}_k$ distance. As in our analysis for the k-flat case, by further subdividing intervals of width more than $1/k$ into smaller ones, we can obtain a new partition, $I' = (I'_i)_{i=1}^{s}$, of cardinality $s \leq 2k$ all of whose parts have width at most $1/k$. Furthermore, we have that $\sum_i \mathbf{Discr}(I'_i) \geq \varepsilon$. Using this partition to bound from below $\|q-p\|_{[k]}^2$, by Cauchy-Schwarz we obtain that

$$\begin{aligned}
\|q-p\|_{[k]}^2 &\geq \sum_i \frac{\mathbf{Discr}^2(I'_i)}{\mathbf{width}(I'_i)^{1/8}} \\
&\geq \frac{\left(\sum_i \mathbf{Discr}(I'_i)\right)^2}{\sum_i \mathbf{width}(I'_i)^{1/8}} \\
&\geq \frac{\varepsilon^2}{2k(1/(2k))^{1/8}} \\
&= \frac{\varepsilon^2}{(2k)^{7/8}}.
\end{aligned}$$

□

The second important fact about the scale-sensitive-L_2 distance is that if it is large then one of the partitions considered in our algorithm will produce a large L_2 error.

Proposition 3.2.8. *Let $p = U_n$ be the uniform distribution and q a distribution over $[n]$. Then we have that*

$$\|q-p\|_{[k]}^2 \leq 10^8 \sum_{j=0}^{j_0-1} \sum_{i=1}^{2^j \cdot k} \frac{\mathbf{Discr}^2(I_i^{(j)})}{\mathbf{width}^{1/8}(I_i^{(j)})}. \tag{3.14}$$

Proof. Let $\mathcal{J} \in \mathbf{W}_{1/k}$ be the optimal partition used when computing the scale-sensitive-L_2 distance $\|q-p\|_{[k]}$. In particular, it is a partition into intervals of width at most $1/k$ so that $\sum_i \frac{\mathbf{Discr}^2(J_i)}{\mathbf{width}(J_i)^{1/8}}$ is maximized. To prove Equation (3.14), we prove a notably stronger claim. In particular, we will prove that for each interval $J_\ell \in \mathcal{J}$

$$\frac{\mathbf{Discr}^2(J_\ell)}{\mathbf{width}^{1/8}(J_\ell)} \leq 10^8 \sum_{j=0}^{j_0-1} \sum_{i:I_i^{(j)} \subset J_\ell} \frac{\mathbf{Discr}^2(I_i^{(j)})}{\mathbf{width}^{1/8}(I_i^{(j)})}. \tag{3.15}$$

Summing over ℓ would then yield $\|q-p\|_{[k]}^2$ on the left hand side and a strict subset of the terms from Equation (3.14) on the right hand side. From here on, we will consider only a single interval J_ℓ. For notational convenience, we will drop the subscript and merely call it J.

First, note that if $|J| \leq 10^8$, then this follows easily from considering just the sum over $j = j_0 - 1$. Then, if $t = |J|$, J is divided into t intervals of size one. The sum of the discrepancies of these intervals equals the discrepancy of J, and thus, the sum of the squares of the discrepancies is at least $\mathbf{Discr}^2(J)/t$. Furthermore, the widths of these subintervals are all smaller than the width of J by a factor of t. Thus, in this case the sum of the right hand side of Equation (3.15) is at least $1/t^{7/8} \geq \frac{1}{10^7}$ of the left hand side.

Otherwise, if $|J| > 10^8$, we can find a j so that $\mathbf{width}(J)/10^8 < 1/(2^j \cdot k) \leq 2 \cdot \mathbf{width}(J)/10^8$. We claim that in this case Equation (3.15) holds even if we restrict the sum on the right hand side to this value of j. Note that J contains at most 10^8 intervals of $I^{(j)}$, and that it is covered by these intervals plus two narrower intervals on the ends.

Call these end-intervals R_1 and R_2. We claim that $\mathbf{Discr}(R_i) \leq \mathbf{Discr}(J)/3$. This is because otherwise it would be the case that

$$\frac{\mathbf{Discr}^2(R_i)}{\mathbf{width}^{1/8}(R_i)} > \frac{\mathbf{Discr}^2(J)}{\mathbf{width}^{1/8}(J)}.$$

(This is because $(1/3)^2 \cdot (2/10^8)^{-1/8} > 1$.) This is a contradiction, since it would mean that partitioning J into R_i and its complement would improve the sum defining $\|q-p\|_{[k]}$, which was assumed to be maximum. This in turn implies that the sum of the discrepancies of the $I_i^{(j)}$ contained in J must be at least $\mathbf{Discr}(J)/3$, so the sum of their squares is at least $\mathbf{Discr}^2(J)/(9\cdot 10^8)$. On the other hand, each of these intervals is narrower than J by a factor of at least $10^8/2$, thus the appropriate sum of $\frac{\mathbf{Discr}^2(I_i^{(j)})}{\mathbf{width}^{1/8}(I_i^{(j)})}$ is at least $\frac{\mathbf{Discr}^2(J)}{10^8\mathbf{width}^{1/8}(J)}$. This completes the proof. □

We are now ready to prove Lemma 3.2.5.

Proof. If $\|q-p\|_{\mathcal{A}_k} \geq \varepsilon$ we have by Lemma 3.2.7 that

$$\|q-p\|_{[k]}^2 \geq \frac{\varepsilon^2}{(2k)^{7/8}}.$$

By Proposition 3.2.8, this implies that

$$\begin{aligned}\frac{\varepsilon^2}{(2k)^{7/8}} &\leq 10^8 \sum_{j=0}^{j_0-1}\sum_{i=1}^{2^j\cdot k} \frac{\mathbf{Discr}^2(I_i^{(j)})}{\mathbf{width}^{1/8}(I_i^{(j)})} \\ &= 10^8 \sum_{j=0}^{j_0-1} (2^j k)^{1/8} \|q^{I^{(j)}} - U_{\ell_j}\|_2^2.\end{aligned}$$

Therefore,

$$\sum_{j=0}^{j_0-1} 2^{j/8}\|q^{I^{(j)}} - U_{\ell_j}\|_2^2 \geq 5\cdot 10^{-9}\varepsilon^2/k. \tag{3.16}$$

On the other hand, if $\|q^{I^{(j)}} - U_{\ell_j}\|_2^2$ were at most $10^{-10}2^{-j/4}\varepsilon^2/k$ for each j, then the sum above would be at most

$$10^{-10}\varepsilon^2/k \sum_j 2^{-j/8} < 5\cdot 10^{-9}\varepsilon^2/k.$$

This would contradict Equation (3.16), thus proving that $\|q^{I^{(j)}} - U_{\ell_j}\|_2^2 \geq 10^{-10}2^{-j/4}\varepsilon^2/k$ for at least one j, proving Lemma 3.2.5. □

3.3 Testing Identity to an Unknown Distribution

In this section we give an optimal algorithm for testing identity of two unknown distributions under the $\mathcal{A}_k$ distance metric, establishing the upper bound of Theorem 1.3.2.

Theorem 3.3.1 (Identity to unknown (Theorem 1.3.2))**.** *Given $\varepsilon > 0$, an integer $k \geq 2$, and sample access to two distributions with probability density functions $p, q : [0,1] \to \mathbb{R}_+$, there is a computationally efficient algorithm which uses $O(\max\{k^{4/5}/\varepsilon^{6/5}, k^{1/2}/\varepsilon^2\})$ samples from p, q, and with probability at least $2/3$ distinguishes whether $q = p$ versus $\|q - p\|_{\mathcal{A}_k} \geq \varepsilon$.*

In this subsection, we provide a high-level overview of our techniques in tandem with a comparison to prior work.

Our upper bound is achieved by an explicit, sample near-linear-time algorithm. A good starting point for considering this problem would be the testing algorithm of [DKN15b] (Section 3.2), which deals with the case where p is an *explicitly known* distribution. The basic idea of the testing algorithm in this case [DKN15b] is to partition the domain into intervals in several different ways, and run a known L_2 tester on the reduced distributions (with respect to the intervals in the partition) as a black-box. At a high-level, these interval partitions can be constructed by exploiting our knowledge of p, in order to divide our domain into several equal mass intervals under p. It can be shown that if p and q have a large $\mathcal{A}_k$ distance between them, one of these partitions will be able to detect the difference.

Generalising this algorithm to the case where p is *unknown* turns out to be challenging, because there seems to be no way to find the appropriate interval partitions with $o(k)$ samples. If we allowed ourselves to take $\Omega(k/\varepsilon)$ samples from p, we would be able to approximate an appropriate interval partition, and make the aforementioned approach go through. Alas, this would not lead to an $o(k)$ sample algorithm. If we can only draw m samples from our distributions, the best that we could hope to do would be to use our samples in order to partition the domain into $m+1$ interval re-

gions. This, of course, is not going to be sufficient to allow an analysis along the lines of the above approach to work. In particular, if we partition our domain *deterministically* into $m = o(k)$ intervals, it may well be the case that the reduced distributions over those intervals are identical, despite the fact that the original distributions have large $\mathcal{A}_k$ distance. In essence, the differences between p and q may well cancel each other out on the chosen intervals.

However, it is important to note that our interval boundaries are *not* deterministic. This suggests that unless we get unlucky, the discrepancy between p and q will not actually cancel out in our partition. As a slight modification of this idea, instead of partitioning the domain into intervals (which we expect to have only $O(1)$ samples each) and comparing the number of samples from p versus q in each, we sort our samples and test how many of them came from the same distribution as their neighbours (with respect to natural ordering on the real line).

We intuitively expect that, if $p = q$, the number of pairs of ordered samples drawn from the same distribution versus a different one will be the same. Indeed, this can be formalised and the completeness of this tester is simple to establish. The soundness analysis, however, is somewhat involved. We need to show that the expected value of the statistic that we compute is larger than its standard deviation. Whereas the variance is easy to bound from above, bounding the expectation is quite challenging. To do so, we define a function, $f(t)$, that encodes how likely it is that the samples near point t come from one distribution or the other. It turns out that f satisfies a relatively nice differential equation, and relates in a clean way to the expectation of our statistic. From this, we can show that any discrepancy between p and q, taking place on a scale too short to be detected by the above partitioning approach, will yield a notable contribution to our expectation.

3.3.1 An $O(k^{4/5}/\varepsilon^{6/5})$-sample tester

In this subsection we give a tester with sample complexity $O(k^{4/5}/\varepsilon^{6/5})$ that applies for $\varepsilon = \Omega(k^{-1/6})$. For simplicity, we focus on the case that we take samples from two unknown distributions with probability density functions $p, q : [0,1] \to \mathbb{R}_+$. Our results are easily seen to extend to discrete probability distributions.

Algorithm Simple-Test-Identity-$\mathcal{A}_k(p,q,\varepsilon)$

Input: sample access to pdf's $p, q : [0,1] \to \mathbb{R}_+$, $k \in \mathbb{Z}_+$, and $\varepsilon > 0$.

Output: "YES" if $q = p$; "NO" if $\|q-p\|_{\mathcal{A}_k} \geq \varepsilon$.

1. Let $m = C \cdot (k^{4/5}/\varepsilon^{6/5})$, for a sufficiently large constant C. Draw two sets of samples S_p, S_q each of size $\mathrm{Poi}(m)$ from p and from q respectively.

2. Merge S_p and S_q, while remembering from which distribution each sample comes from. Let S be the union of S_p and S_q sorted in increasing order (breaking ties randomly).

3. Compute the statistic Z defined as follows:

   $Z \stackrel{\text{def}}{=}$

   # (pairs of successive samples in S coming from the same distribution) $-$

   # (pairs of successive samples in S coming from different distributions)

4. If $Z > 3 \cdot (\sqrt{m})$ return "NO". Otherwise return "YES".

Proposition 3.3.2. *The algorithm* Simple-Test-Identity-$\mathcal{A}_k(p,q,\varepsilon)$*, on input two samples each of size* $O(k^{4/5}/\varepsilon^{6/5})$ *drawn from two distributions with densities* $p, q : [0,1] \to \mathbb{R}_+$*, an integer* $k > 2$*, and* $\varepsilon = \Omega(k^{-1/6})$*, correctly distinguishes the case that* $q = p$ *from the case* $\|p-q\|_{\mathcal{A}_k} \geq \varepsilon$*, with probability at least* $2/3$.

Proof. First, it is straightforward to verify the claimed sample complexity, since the algorithm only draws samples in Step 1. To simplify the analysis we make essential

use of the following simple claim:

Claim 3.3.3. *We can assume without loss of generality that the pdf's $p,q : [0,1] \to \mathbb{R}_+$ are continuous functions bounded from above by 2.*

Proof. We start by showing we can assume that p,q are at most 2. Let $p,q : [0,1] \to \mathbb{R}_+$ be arbitrary pdf's. We consider the cumulative distribution function (CDF) Φ of the mixture $(p+q)/2$. Let $X \sim p$, $Y \sim q$, $W \sim (p+q)/2$ be random variables. Since Φ is non-decreasing, replacing X and Y by $\Phi(X)$ and $\Phi(Y)$ does not affect the algorithm (as the ordering on the samples remains the same). We claim that, after making this replacement, $\Phi(X)$ and $\Phi(Y)$ are continuous distributions with probability density functions bounded by 2. In fact, we will show that the sum of their probability density functions is exactly 2. This is because for any $0 \leqslant a \leqslant b \leqslant 1$,

$$\Pr[\Phi(X) \in [a,b]] + \Pr[\Phi(Y) \in [a,b]] = 2\Pr[\Phi(W) \in [a,b]] = 2(b-a)\,,$$

where the second equality is by the definition of a CDF. Thus, we can assume that p and q are bounded from above by 2.

To show that we can assume continuity, note that p and q can be approximated by continuous density functions p' and q' so that the L_1 errors $\|p-p'\|_1, \|q-q'\|_1$ are each at most $1/(10m)$. If our algorithm succeeds with the continuous densities p' and q', it must also succeed for p and q. Indeed, since the L_1 distance between p and p' and q and q' is at most $1/(10m)$, a set of m samples taken from p or q are statistically indistinguishable to m samples taken from p' or q'. This proves that there is no loss of generality to assume that p and q are continuous. □

Note that the algorithm makes use of the well-known "Poissonisation" approach. Namely, instead of drawing $m = O(k^{4/5}/\varepsilon^{6/5})$ samples from p and from q, we draw $m' = \mathrm{Poi}(m)$ samples from p and $m'' = \mathrm{Poi}(m)$ sample from q. The crucial properties of the Poisson distribution are that it is sharply concentrated around its mean, and it makes the number of times different elements occur in the sample independent.

We now establish completeness. Note that our algorithm draws $\mathrm{Poi}(2m)$ samples from p or q. If $p = q$, then our process equivalently selects $\mathrm{Poi}(2m)$ values from p and then randomly and independently with equal probability decides whether or not each sample came from p or from q. Making these decisions one at a time in increasing order of points, we note that each adjacent pair of elements in S randomly and independently contributes either a $+1$ or a -1 to Z. Therefore, the distribution of Z is exactly that of a sum of $\mathrm{Poi}(2m) - 1$ independent $\{\pm 1\}$ random variables. Therefore, Z has mean 0 and variance $2m - 1$. By Chebyshev's inequality it follows that $|Z| \leqslant 3\sqrt{m}$ with probability at least $7/9$. This proves completeness.

We now proceed to prove the soundness of our algorithm. Assuming that $\|p - q\|_{\mathcal{A}_k} > \varepsilon$, we want to show that the value of Z is at most $3 \cdot \sqrt{m}$ with probability at most $1/3$. To prove this statement, we will again use Chebyshev's inequality. In this case it suffices to show that $\mathbb{E}[Z] \gg \sqrt{\mathrm{Var}[Z]} + \sqrt{m}$ for the inequality to be applicable. We begin with an important definition.

Definition 3.3.4. Let $f : [0,1] \to [-1,1]$ equal

$$\begin{aligned} f(t) \stackrel{\text{def}}{=} & \Pr[\text{largest sample in } S \text{ that is at most } t \text{ was drawn from } p] \\ & - \Pr[\text{largest sample in } S \text{ that is at most } t \text{ was drawn from } q] \ . \end{aligned}$$

The importance of this function is demonstrated by the following lemma.

Lemma 3.3.5. *We have that:* $\mathbb{E}[Z] = m \int_0^1 f(t)(p(t) - q(t))dt$.

Proof. Given an interval I, we let Z_I be the contribution to Z coming from pairs of consecutive points of S the larger of which is drawn from I. We wish to approximate the expectation of Z_I. We let $\tau(I) = m(p(I) + q(I))$ be the expected total number of points drawn from I. We note that the contribution coming from cases where more than one point is drawn from I is $O(\tau(I)^2)$. We next consider the contribution under the condition that only one sample is drawn from I. For this, we let EP_I and EQ_I be the events that the largest element of S preceding I comes from p or q respectively.

We have that the expected contribution to Z_I coming from events where exactly one element of S is drawn from I is:

$$(\Pr[\mathrm{EP}_I] - \Pr[\mathrm{QP}_I]) \Pr(\text{The only element drawn from } I \text{ is from } p)$$
$$-(\Pr[\mathrm{EP}_I] - \Pr[\mathrm{QP}_I]) \Pr(\text{The only element drawn from } I \text{ is from } q).$$

Letting x_I be the left endpoint of I, this is

$$f(x_I)(mp(I) - mq(I)) + O(\tau(I)^2).$$

Therefore,

$$\mathbb{E}[Z_I] = f(x_I)(mp(I) - mq(I)) + O(\tau(I)^2).$$

Letting I be a partition of our domain into intervals, we find that

$$\begin{aligned}\mathbb{E}[Z] &= \sum_{I \in I} \mathbb{E}[Z_I] \\ &= \sum_{I \in I} f(x_I)(mp(I) - mq(I)) + O(\tau(I)^2) \\ &= O(m \max_{I \in I} \tau(I)) + \sum_{I \in I} f(x_I)(mp(I) - mq(I)).\end{aligned}$$

As the partition I becomes iteratively more refined, these sums approach Riemann sums for the integral of

$$mf(x)(p(x) - q(x))dx.$$

Therefore, taking a limit over partitions I, we have that

$$\mathbb{E}[Z] = m \int f(x)(p(x) - q(x))dx.$$

□

We will also make essential use of the following technical lemma:

Lemma 3.3.6. *The function f is differentiable with derivative*

$$f'(t) = m\left(p(t) - q(t) - (p(t) + q(t))f(t)\right).$$

Proof. Consider the difference between $f(t)$ and $f(t+h)$ for some small $h > 0$. We note that $f(t) = \mathbb{E}[F_t]$ where F_t is 1 if the sample of S preceding t came from p, -1 if the sample came from q, and 0 if no sample came before t. Note that

$$F_{t+h} = \begin{cases} F_t & \text{if no samples from } p \text{ nor } q \text{ are drawn from } [t,t+h] \\ 1 & \text{if one sample from } p \text{ and none from } q \text{ are drawn from } [t,t+h] \\ -1 & \text{if one sample from } q \text{ and none from } p \text{ are drawn from } [t,t+h] \\ \pm 1 & \text{if at least two samples from } p \text{ or } q \text{ are drawn from } [t,t+h]. \end{cases}$$

Since p and q are continuous at $t \in [0,1]$, these four events happen with probabilities $1 - mh(p(t)+q(t)) + o(h)$, $mhp(t) + o(h)$, $mhq(t) + o(h)$, $o(h)$, respectively. Therefore, taking an expectation we find that $f(t+h) = f(t)(1 - mh(p(t)+q(t))) + mh(p(t)-q(t)) + o(h)$. This, and a similar relation relating $f(t)$ to $f(t-h)$, proves that f is differentiable with the desired derivative. □

To analyse the desired expectation, $\mathbb{E}[Z]$, we consider the quantity $\int_0^1 f'(t)f(t)dt = (1/2)\left(f^2(1) - f^2(0)\right)$. Substituting f' from Lemma 3.3.6 above gives

$$\int_0^1 f'(t)f(t)dt = m\int_0^1 f(t)(p(t)-q(t))dt - m\int_0^1 f^2(t)(p(t)+q(t))dt.$$

Combining this with Lemma 3.3.5, we get

$$\mathbb{E}[Z] = m\int_0^1 f^2(t)(p(t)+q(t))dt + f^2(1)/2\,. \tag{3.17}$$

The second term in (3.17) above is $O(1)$, so we focus our attention to bound the first term from below. To do this, we consider intervals $I \subset [0,1]$ over which $|p(I) - q(I)|$ is "large" and show that they must produce some noticeable contribution to the first term. Fix such an interval I. We want to show that f^2 is large somewhere in I. Intuitively, we attempt to prove that on at least one of the endpoints of the interval, the value of f is big. Since f does not vary too rapidly, f^2 will be large on some large fraction of I. Formally, we have the following lemma:

Lemma 3.3.7. *For $\delta > 0$, let $I \subset [0,1]$ be an interval with $|p(I) - q(I)| = \delta$ and $p(I) + q(I) < 1/m$. Then, there exists an $x \in I$ such that $|f(x)| \geqslant \frac{m\delta}{3}$.*

Proof. Suppose for the sake of contradiction that $|f(x)| < m\delta/3$ for all $x \in I = [X,Y]$. Then, we have that

$$\begin{aligned}
2m\delta/3 &> |f(X) - f(Y)| \\
&= \left| \int_X^Y f'(t)dt \right| \\
&= \left| \int_X^Y (m(p(t) - q(t)) - mf(t)(p(t) + q(t)))\, dt \right| \\
&= \left| m(p(I) - q(I)) - m \int_X^Y f(t)(p(t) + q(t))dt \right| \\
&\geqslant m|p(I) - q(I)| - m \left| \int_X^Y f(t)(p(t) + q(t))dt \right| \\
&> m\delta - m \int_X^Y (m\delta/3)\,(p(t) + q(t))dt \\
&= m\delta\,(1 - m(p(I) + q(I))/3) > 2m\delta/3\,,
\end{aligned}$$

which yields the desired contradiction. □

We are now able to show that the contribution to $\mathbb{E}[Z]$ coming from such an interval is large.

Lemma 3.3.8. *Let I be an interval satisfying the hypotheses of Lemma 3.3.7. Then*

$$\int_I f^2(t)(p(t) + q(t))dt = \Omega(m^2\delta^3)\,.$$

Proof. By Lemma 3.3.7, f is large at some point x of the interval $I = [X,Y]$. Without loss of generality, we assume that $p([X,x]) + q([X,x]) \leqslant (p(I) + q(I))/2$. Let $I' = [x,Y']$ be the interval so that $p(I') + q(I') = \delta/9$. Note that $I' \subset I$ (since by assumption $|p(I) - q(I)| > \delta$ and thus $p(I) + q(I) > \delta$). Furthermore, note that since with probability at least $1 - m\delta/9$, no samples from S lie in I', we have that for all z in I' it holds $|f(x) - f(z)| \leqslant 2m\delta/9$, so $|f(z)| \geqslant m\delta/9$. Therefore,

$$\begin{aligned}
\int_I f^2(t)(p(t) + q(t))dt &\geqslant \int_{I'} f^2(t)(p(t) + q(t))dt \geqslant \int_{I'} \left(\frac{m\delta}{9}\right)^2 (p(t) + q(t))dt \\
&= \frac{m^2\delta^2}{81}(p(I') + q(I')) = \frac{m^2\delta^3}{729}\,.
\end{aligned}$$

□

Since $\|p-q\|_{\mathcal{A}_k} > \varepsilon$, there is a partition I of $[0,1]$ into k intervals so that $\|p_r^I - q_r^I\|_1 > \varepsilon$. By subdividing intervals further if necessary, we can guarantee that I has at most $3k$ intervals, $\|p_r^I - q_r^I\| > \varepsilon$, and for each subinterval $I \in I$ it holds $p(I), q(I) \leqslant 1/k$. For each such interval $I \in I$, let $\delta_I = |p(I) - q(I)|$. Note that $\sum_{I \in I} \delta_I \geqslant \varepsilon$.

By (3.17) we have that

$$\begin{aligned}
\mathbb{E}[Z] &= m \sum_{I \in I} \int_I f^2(t)(p(t)+q(t))dt + O(1) \\
&= \Omega\left(m \sum_{I \in I} m^2 \delta_I^3\right) = \Omega\left(m^3 (\sum_{I \in I} \delta_I)^3/(3k)^2\right) \\
&= \Omega\left(m^3 \varepsilon^3/k^2\right) = \Omega(C^{5/2}\sqrt{m}) \,.
\end{aligned}$$

We note that the second to last line above follows by Hölder's inequality. It remains to bound from above the variance of Z.

Lemma 3.3.9. *We have that* $\mathrm{Var}[Z] = O(m)$.

Proof. We divide the domain $[0,1]$ into m intervals I_i, $i = 1, \ldots, m$, each of total mass $2/m$ under the sum-distribution $p+q$. Consider the random variable X_i denoting the contribution to Z coming from pairs of adjacent samples in S such that the right sample is drawn from I_i. Clearly, $Z = \sum_{i=1}^m X_i$ and $\mathrm{Var}[Z] = \sum_{i=1}^m \mathrm{Var}[X_i] + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j)$.

To bound the first sum, note that the number of pairs of S in an interval I_i is no more than the number of samples drawn from I_i, and the variance of X_i is less than the expectation of the square of the number of samples from I_i. Since the number of samples from I_i is a Poisson random variables with parameter 2, we have $\mathrm{Var}[X_i] = O(1)$. This shows that $\sum_{i=1}^m \mathrm{Var}[X_i] = O(m)$.

To bound the sum of covariance, consider X_i and X_j conditioned on the samples drawn from intervals other than I_i and I_j. Note that if any sample is drawn from an intermediate interval, X_i and X_j are uncorrelated, and otherwise their covariance is at most $\sqrt{\mathrm{Var}(X_i)\,\mathrm{Var}(X_j)} = O(1)$. Since the probability that no sample is drawn from

any intervening interval decreases exponentially with their separation, it follows that $\mathrm{Cov}(X_i, X_j) = O(1) \cdot e^{-\Omega(|j-i|)}$. This completes the proof. □

An application of Chebyshev's inequality completes the analysis of the soundness and the proof of Proposition 3.3.2. □

3.3.2 The General Tester

In this section, we present a tester whose sample complexity is optimal (up to constant factors) for all values of ε and k, thereby proving Theorem 3.3.1 and establishing the upper bound part of Theorem 1.3.2.

Our general tester (Algorithm Test-Identity-$\mathcal{A}_k$) builds on the tester presented in the previous subsection (Algorithm Simple-Test-Identity-$\mathcal{A}_k$). It is not difficult to see that the latter algorithm can fail once ε becomes sufficiently small, if the discrepancy between p and q is concentrated on intervals of mass larger than $1/m$. In this scenario, the tester Simple-Test-Identity-$\mathcal{A}_k$ will not take sufficient advantage of these intervals. To obtain our enhanced tester Test-Identity-$\mathcal{A}_k$, we will need to combine Simple-Test-Identity-$\mathcal{A}_k$ with an alternative tester when this is the case. Note that we can easily bin the distributions p and q into intervals of total mass approximately $1/m$ by taking m random samples. Once we do this, we can use an identity tester similar to that in our previous work [DKN15b] to detect the discrepancy in these intervals.

Proposition 3.3.10. *Let p, q be discrete distributions over $[n]$ satisfying $\|p\|_2, \|q\|_2 = O(1/\sqrt{n})$. There exists a testing algorithm with the following properties: On input $k \in \mathbb{Z}_+$, $2 \le k \le n$, and $\delta, \varepsilon > 0$, the algorithm draws $O\left((\sqrt{k}/\varepsilon^2) \cdot \log(1/\delta)\right)$ samples from p and q and with probability at least $1 - \delta$ distinguishes between the cases $p = q$ and $\|p - q\|_{\mathcal{A}_k} > \varepsilon$.*

This tester has been analysed in detail in the Section 3.1.2 and will be used in the present algorithm as a black box.

We are now ready to present our general testing algorithm:

Algorithm Test-Identity-$\mathcal{A}_k(p,q,\varepsilon)$

Input: sample access to distributions $p,q:[0,1]\to\mathbb{R}_+$, $k\in\mathbb{Z}_+$, and $\varepsilon>0$.

Output: "YES" if $q=p$; "NO" if $\|q-p\|_{\mathcal{A}_k}\geq\varepsilon$.

1. Let $m=Ck^{4/5}/\varepsilon^{6/5}$, for a sufficiently large constant C. Draw two sets of samples S_p, S_q each of size $\mathrm{Poi}(m)$ from p and from q respectively.

2. Merge S_p and S_q while remembering from which distribution each sample comes from. Let S be the union of S_p and S_q sorted in increasing order (breaking ties randomly).

3. Compute the statistic Z defined as follows:

   $Z\stackrel{\text{def}}{=}$

   # (pairs of successive samples in S coming from the same distribution) −

   # (pairs of successive samples in S coming from different distributions)

4. If $Z>5\sqrt{m}$ return "NO".

5. Repeat the following steps $O(C)$ times:

   (a) Draw $\mathrm{Poi}(m)$ samples from $(p+q)/2$.

   (b) Split the domain into intervals with the interval endpoints given by the above samples. Let p' and q' be the reduced distributions with respect to these intervals.

   (c) Run the tester of Proposition 3.3.10 on p' and q' with error probability $1/C^2$ to determine if $\|p'-q'\|_{\mathcal{A}_{2k+1}}>\varepsilon/C$. If the output of this tester is "NO", output "NO".

6. Output "YES".

Our main result for this section is the following:

Theorem 3.3.11. *Algorithm Test-Identity-$\mathcal{A}_k$ draws $O(\max\{k^{4/5}/\varepsilon^{6/5}, k^{1/2}/\varepsilon^2\})$ samples from p, q and with probability at least $2/3$ returns "YES" if $p = q$ and "NO" if $\|p-q\|_{\mathcal{A}_k} > \varepsilon$.*

Proof. First, it is easy to see that the sample complexity of the algorithm is $O(m + k^{1/2}/\varepsilon^2)$. Recall that we can assume that p, q are continuous pdf's bounded from above by 2.

We start by establishing completeness. If $p = q$, it is once again the case that $\mathbb{E}[Z] = 0$ and $\mathrm{Var}[Z] < 2m$, so by Chebyshev' s inequality, Step 4 will fail with probability at most $1/9$. Next, when taking our samples in Step 5(a), note that the expected samples size is $O(m)$, and that the expected squared L_2 norms of the reduced distributions p' and q' are $O(1/m)$. Therefore, with probability at least $1 - 1/C^2$, p' and q' satisfy the hypothesis of Proposition 3.3.10. Hence, this holds for all C iterations with probability at least $8/9$.

Conditioning on this event, since $p' = q'$, the tester in Step 5(c) will return "YES" with probability at least $1 - 1/C^2$ on each iteration. Therefore, it returns "YES" on all iterations with probability at least $8/9$. By a union bound, it follows that if $p = q$, our algorithm returns "YES" with probability at least $2/3$.

We now proceed to establish soundness. Suppose that $\|p-q\|_{\mathcal{A}_k} \geqslant \varepsilon$. Then there exists a partition I of the domain into k intervals such that $\|p_r^I - q_r^I\| \geqslant \varepsilon$. For an interval $I \in I$, let $\delta(I) = |p(I) - q(I)|$. We will call an $I \in I$ small if there is a subinterval $J \subseteq I$ so that $p(J) + q(J) < 1/m$ and $|p(J) - q(J)| \geqslant \delta(I)/3$, otherwise we will call I large. Note that $\sum_{I \in I, I \text{ small}} \delta(I) + \sum_{I \in I, I \text{ large}} \delta(I) = \sum_{I \in I} \delta(I) \geqslant \varepsilon$. Therefore either $\sum_{I \in I, I \text{ small}} \delta(I) \geqslant \varepsilon/2$, or $\sum_{I \in I, I \text{ large}} \delta(I) \geqslant \varepsilon/2$. We analyse soundness separately in each of these cases.

Consider first the case that $\sum_{I \in I, I \text{ small}} \delta(I) \geqslant \varepsilon/2$. The analysis in this case is very similar to the soundness proof of Proposition 3.3.2, which we describe for the sake of completeness.

By definition, for each small interval I, there exists a subinterval J so that $p(J) +$

$q(J) < 1/m$ and $|p(J) - q(J)| > \delta(I)/2$. By Lemma 3.3.8, for such J we have that $\int_J f^2(t)(p(t)+q(t))dt = \Omega(m^2\delta^3(I))$, and therefore, that $\int_I f^2(t)(p(t)+q(t))dt = \Omega(m^2\delta^3(I))$. Hence, we have that

$$\begin{aligned}
\mathbb{E}[Z] &\geqslant m\int_0^1 f^2(t)(p(t)+q(t))dt \\
&\geqslant \sum_{I\in\mathcal{I},I \text{ small}} m\int_I f^2(t)(p(t)+q(t))dt \\
&\geqslant \sum_{I\in\mathcal{I},I \text{ small}} \Omega(m^3\delta^3(I)) \\
&\geqslant \Omega(m^3)\left(\sum_{I\in\mathcal{I},I \text{ small}} \delta(I)\right)^3/k^2 \\
&= \Omega(m^3\varepsilon^3/k^2) \\
&= \Omega(C^{5/2}\sqrt{m}).
\end{aligned}$$

On the other hand, Lemma 3.3.9 gives that $\mathrm{Var}[Z] = O(m)$, so for C sufficiently large, Chebyshev's inequality implies that with probability at least $2/3$ it holds $Z > 5\sqrt{m}$. That is, our algorithm outputs "NO" with probability at least $2/3$.

Now consider that case that $\sum_{I\in\mathcal{I},I \text{ large}} \delta(I) \geqslant \varepsilon/2$. We claim that the second part of our tester will detect the discrepancy between p and q with high constant probability. Once again, we can say that with probability at least $8/9$ the squared L_2 norms of the reduced distributions p' and q' are both $O(1/m)$ and that the size of the reduced domain is $O(m)$. Thus, the conditions of Proposition 3.3.10 are satisfied on all iterations with probability at least $8/9$. To complete the proof, we will show that with constant probability we have $\|p'-q'\|_{\mathcal{A}_{2k+1}} > \varepsilon/C$. To do this, we construct an explicit partition I' of our reduced domain into at most $2k+1$ intervals so that with constant probability $\|p_r^{I'} - q_r^{I'}\|_1 > \varepsilon/C$. This will imply that with probability at least $8/9$ that on at least one of our C trials that $\|p_r^{I'} - q_r^{I'}\|_1 > \varepsilon/C$.

More specifically, for each interval $I \in \mathcal{I}$ we place interval boundaries at the smallest and largest sample points taken from I in Step 5(a) (ignoring them if fewer than two samples landed in I). Since we have selected at most $2k$ points, this process defines a

partition I' of the domain into at most $2k+1$ intervals. We will show that the reduced distributions $p'' = p_r^{I'}$ and $q'' = q_r^{I'}$ have large expected L_1 error.

In particular, for each interval $I \in I$ let I' be the interval between the first and last sample points of I. Note that I' is an interval in the partition I'. We claim that if I is large, then with constant probability

$$|p(I') - q(I')| = \Omega(\delta(I)).$$

Let $I = [X,Y]$ and $I' = [x,y]$ (so x and y are the smallest and largest samples taken from I, respectively). We note that if $p([X,x]) + q([X,x]) < 1/m$ and $p([y,Y]) + q([y,Y]) < 1/m$ then

$$|p(I') - q(I')| \geqslant$$

$$|p(I) - q(I)| - |p([X,x]) - q([X,x])| - |p([y,Y]) - q([y,Y])| \geqslant \delta(I) - \delta(I)/3 - \delta(I)/3 = \delta(I)/3,$$

where the second inequality uses the fact that I is large. On the other hand, we note that $p([X,x]) + q([X,x])$ and $p([y,Y]) + q([y,Y])$ are exponential distributions with mean $1/m$, and thus, this event happens with constant probability. Let N_I be the indicator random variable for the event that $|p(I') - q(I')| \geqslant \delta(I)/3$. We have that

$$\|p'' - q''\|_1 \geqslant \sum_I N_I \delta(I)/3 \geqslant \sum_{I \in I, I \text{ large}} N_I \delta(I)/3.$$

Thus, we have that

$$\|p'' - q''\|_1 \geqslant \sum_{I \in I, I \text{ large}} \delta(I)/3 - \sum_{I \in I, I \text{ large}} (1 - N_I)\delta(I)/3.$$

Therefore, since

$$\mathbb{E}\left[\sum_{I \in I, I \text{ large}} (1 - N_I)\delta(I)/3\right] < \left(\sum_{I \in I, I \text{ large}} \delta(I)/3\right)(1-c)$$

for some fixed $c > 0$, we have that with constant probability that

$$\|p'' - q''\|_1 \geqslant c \sum_{I \in I, I \text{ large}} \delta(I)/3 \geqslant c\varepsilon/6 \geqslant \varepsilon/C.$$

This means that with probability at least $8/9$ for at least one iteration we will have that $\|p' - q'\|_{\mathcal{A}_{2k+1}} > \varepsilon/C$, and therefore, with probability at least $2/3$, our algorithm outputs "NO".

□

3.4 Testing Identity: Small Domain Size

Section 3.3 gives an upper bound for the case when the domain size tends to infinity. Quite unexpectedly, it disagrees with the result of [CDVV14] for the case when $n = k$. The next theorem establishes a bound between the unstructured case and the case of Theorem 1.3.2.

Theorem 3.4.1. *Given sample access to distributions p and q on $[n]$ and $\varepsilon > 0$ there exists an algorithm that takes*

$$O\left(\max\left(\min\left(k^{4/5}/\varepsilon^{6/5}, k^{2/3}\log^{4/3}(3+n/k)\log\log(3+n/k)/\varepsilon^{4/3}\right), k^{1/2}/\varepsilon^2\right)\right)$$

samples from each of p and q and distinguishes with $2/3$ probability between the cases that $p = q$ and $\|p-q\|_{\mathcal{A}_k} \geq \varepsilon$.

We use a technique of iteratively reducing the number of bins (domain elements). In particular, we show that if we merge bins together in consecutive pairs, this does not significantly affect the $\mathcal{A}_k$ distance between the distributions, unless a large fraction of the discrepancy between our distributions is supported on $O(k)$ bins near the boundaries in the optimal partition. In order to take advantage of this, we provide an identity tester that requires few samples to distinguish between the cases where $p = q$ and the case where p and q have a large ℓ_1 distance supported on only k of the bins. We are able to take advantage of the small support essentially because having a discrepancy supported on few bins implies that the ℓ_2 distance between the distributions must be reasonably large.

The basic idea of our algorithm is the following. From the distributions p and q construct new distributions p' and q' by merging pairs of consecutive buckets. Note that p' and q' each have much smaller domains (of size about $n/2$). Furthermore, note that the $\mathcal{A}_k$ distance between p and q is $\sum_{I\in I}|p(I)-q(I)|$ for some partition I into k intervals. By using essentially the same partition, we can show that $\|p'-q'\|_{\mathcal{A}_k}$ should be almost as large as $\|p-q\|_{\mathcal{A}_k}$. This will in fact hold unless much of the error between p and q is supported at points near the endpoints of intervals in I. If this is the case, it turns out there is an easy algorithm to detect this discrepancy. We require the following definitions:

Definition 3.4.2. For a discrete distribution p on [n], the merged distribution obtained from p is the distribution p' on $\lceil n/2\rceil$, so that $p'(i)\stackrel{\text{def}}{=}p(2i)+p(2i+1)$. For a partition I of $[n]$, define the *merged partition* I' of domain $\lceil n/2\rceil$, so that $I'_i\in I'$ has the points obtained by point-wise gluing together odd points and even points.

Definition 3.4.3. Let p and q be distributions on $[n]$. For integers $k\geq 1$, let $\|p-q\|_{1,k}$ be the sum of the largest k values of $|p(i)-q(i)|$ over $i\in[n]$.

We begin by showing that either $\|p'-q'\|_{\mathcal{A}_k}$ is close to $\|p-q\|_{\mathcal{A}_k}$ or $|p-q|_{1,k}$ is large.

Lemma 3.4.4. *For any two distributions p and q on $[n]$, let $p'=\lceil p/2\rceil$ and $q'=\lceil q/2\rceil$ be the merged distributions. Then,*

$$\|p-q\|_{\mathcal{A}_k}\leq\|p'-q'\|_{\mathcal{A}_k}+2\|p-q\|_{1,k}\,.$$

Proof. Let I be the partition of $[n]$ into k intervals so that $\|p-q\|_{\mathcal{A}_k}=\sum_{I\in I}|p(I)-q(I)|$. Let I' be obtained from I by rounding each upper endpoint of each interval except for the last down to the nearest even integer, and rounding the lower endpoint of each interval up to the nearest odd integer. Note that

$$\sum_{I\in I'}|p(I)-q(I)|=\sum_{I\in I'}|p'(I/2)-q'(I/2)|\leq\|p'-q'\|_{\mathcal{A}_k}\,.$$

The partition I' is obtained from I by taking at most k points and moving them from one interval to another. Therefore, the difference

$$\left| \sum_{I \in I} |p(I) - q(I)| - \sum_{I \in I'} |p(I) - q(I)| \right| ,$$

is at most twice the sum of $|p(i) - q(i)|$ over these k points, and therefore at most $2\|p-q\|_{1,k}$. Combing this with the above gives our result. □

Next we need to show that if two distributions have $\|p-q\|_{1,k}$ large that this can be detected easily.

Lemma 3.4.5. *Let p and q be distributions on $[n]$. Let $k > 0$ be a positive integer, and $\varepsilon > k^{-1/4}$. There exists an algorithm which takes $O(k^{2/3}/\varepsilon^{4/3})$ samples from each of p and q and, with probability at least $2/3$, distinguishes between the cases that $p = q$ and $\|p-q\|_{1,k} > \varepsilon$.*

We start by introducing some important terminology from [DK16]. We begin with the definition of a split distribution:

Definition 3.4.6. Given a distribution p on $[n]$ and a multiset S of elements of $[n]$, define the *split distribution* p_S on $[n+|S|]$ as follows: For $1 \le i \le n$, let a_i denote 1 plus the number of elements of S that are equal to i. Thus, $\sum_{i=1}^{n} a_i = n + |S|$. We can therefore associate the elements of $[n+|S|]$ to elements of the set $B = \{(i,j) : i \in [n], 1 \le j \le a_i\}$. We now define a distribution p_S with support B, by letting a random sample from p_S be given by (i,j), where i is drawn randomly from p and j is drawn randomly from $[a_i]$.

A split distribution has the following nice property:

Lemma 3.4.7. *Let p and q be probability distributions on $[n]$, and S a given multiset of $[n]$. Then: (i) We can simulate a sample from p_S or q_S by taking a single sample from p or q, respectively. (ii) It holds $\|p_S - q_S\|_1 = \|p-q\|_1$.*

Lemma 3.4.8 ([DK16]). *Let p be a distribution on $[n]$. Then: (i) For any multisets $S \subseteq S'$ of $[n]$, $\|p_{S'}\|_2 \leq \|p_S\|_2$, and (ii) If S is obtained by taking $\mathrm{Poi}(m)$ samples from p, then $\mathbb{E}[\|p_S\|_2^2] \leq 1/m$.*

The essential meaning of this lemma is the following: no matter how pathological our distribution is, we can transform it into a relatively regular one (having a small L_2 norm), so that many algorithms requiring "near flateness" assumption can be used.

Proof of Lemma 3.4.5: We begin by presenting the algorithm:

Algorithm `Small-Support-Discrepancy-Tester`

Input: sample access to pdf's $p, q : [n] \to \mathbb{R}_+$, $k \in \mathbb{Z}_+$, and $\varepsilon > 0$.

Output: "YES" if $q = p$; "NO" if $\|q - p\|_{1,k} \geq \varepsilon$.

1. Let $m = k^{2/3}/\varepsilon^{4/3}$.
2. Let S be the multiset obtained by taking m independent samples from p.
3. Use the ℓ_2 tester to distinguish between the cases $p_S = q_S$ and $\|p_S - q_S\|_2^2 \geq k^{-1}\varepsilon^2/2$, and return the result.

The analysis is simple. We note that with 90% probability it holds $\|p_S\|_2 = O(1/m)$, and therefore the number of samples needed is $O(m + km^{-1/2}/\varepsilon^{-2}) = O(k^{2/3}/\varepsilon^{4/3})$. If $p = q$, then $p_S = q_S$ and the algorithm will return "YES" with appropriate probability. If $\|q - p\|_{1,k} \geq \varepsilon$, then $\|p_S - q_S\|_{1,k+m} \geq \varepsilon$. Since $k+m$ elements contribute to total L^1 error at least ε, by Cauchy-Schwartz, we have that $\|p_S - q_S\|_2^2 \geq \varepsilon^2/(k+m) \geq k^{-1}\varepsilon^2/2$. Therefore, in this case, the algorithm returns "NO" with appropriate probability. □

Algorithm `Small-Domain-`$\mathcal{A}_k$`-tester`

Input: sample access to pdf's $p, q : [n] \to \mathbb{R}_+$, $k \in \mathbb{Z}_+$, and $\varepsilon > 0$.

Output: "YES" if $q = p$; "NO" if $\|q - p\|_{\mathcal{A}_k} \geq \varepsilon$.

1. For $i := 0$ to $t \stackrel{\text{def}}{=} \lceil \log_2(n/k) \rceil$, let $p^{(i)}, q^{(i)}$ be distributions on $[\lceil 2^{-i} n \rceil]$ defined by $p^{(i)} = \lceil 2^{-i} p \rceil$ and $q^{(i)} = \lceil 2^{-i} q \rceil$.

2. Take $Ck^{2/3} \log^{4/3}(3 + n/k) \log\log(3 + n/k)/\varepsilon^{4/3}$ samples, for C sufficiently large, and use these samples to distinguish between the cases $p^{(i)} = q^{(i)}$ and $\|p^{(i)} - q^{(i)}\|_{1,k} > \varepsilon/(4\log_2(3 + n/k))$ with probability of error at most $1/(10\log_2(3 + n/k))$ for each i from 0 to t, using the same samples for each test.

3. If any test yields that $p^{(i)} \neq q^{(i)}$, return "NO". Otherwise, return "YES".

Proof of Theorem 3.4.1: Given the algorithms from [DK16], we only need an algorithm that distinguishes in $O(k^{2/3} \log^{4/3}(n/k) \log\log(n/k)/\varepsilon^{4/3})$ samples when $\varepsilon > k^{-1/4}$.

We now show correctness. In terms of sample complexity, we note that by taking a majority over $O(\log\log(3 + n/k))$ independent runs of the tester from Lemma 3.4.5, we can run this algorithm in an appropriate sample complexity. Taking a union bound, we can also assume that all tests performed in step 2 returned the correct answer. If $p = q$ then $p^{(i)} = q^{(i)}$ for all i and thus, our algorithm returns "YES". Otherwise, we have that $\|p - q\|_{\mathcal{A}_k} \geq \varepsilon$. By repeated application of Lemma 3.4.4, we have that

$$\|p - q\|_{\mathcal{A}_k} \leq \sum_{i=0}^{t-1} 2\|p^{(i)} - q^{(i)}\|_{1,k} + \|p^{(t)} - q^{(t)}\|_{\mathcal{A}_k} \leq 2\sum_{i=0}^{t} 2\|p^{(i)} - q^{(i)}\|_{1,k} ,$$

where the last step was because $p^{(t)}$ and $q^{(t)}$ have a support of size at most k and so $\|p^{(t)} - q^{(t)}\|_{\mathcal{A}_k} = \|p^{(t)} - q^{(t)}\|_1 = \|p^{(t)} - q^{(t)}\|_{1,k}$. Therefore, if this is at least ε, it must be the case that $\|p^{(i)} - q^{(i)}\|_{1,k} > \varepsilon/(4\log_2(3 + n/k))$ for some $0 \leq i \leq t$, and thus our algorithm returns "NO".

This completes our proof. □

Chapter 4

Lower Bounds

4.1 A Lower Bound for Testing Identity when Both Distributions are Unknown

Our upper bound from Section 3.3 seems potentially suboptimal. Instead of obtaining an upper bound of $O(\max\{k^{2/3}/\varepsilon^{4/3}, k^{1/2}/\varepsilon^2\})$, which would be analogical to the unstructured testing result of [CDVV14], we obtain a very different bound of $O(\max\{k^{4/5}/\varepsilon^{6/5}, k^{1/2}/\varepsilon^2\})$. In this section we show, surprisingly, that our upper bound is optimal for continuous distributions, or discrete distributions with support size n that is sufficiently large as a function of k.

Intuitively, our lower bound proof consists of two steps. In the first step, we show it is no loss of generality to assume that an optimal algorithm only considers the ordering of the samples, and ignores all other information. In the second step, we construct a pair of distributions which is hard to distinguish given the condition that the tester is only allowed to look at the ordering of the samples and nothing more.

Our first step is described in the following theorem. We note that unlike the arguments in the upper bound proofs, this part of our lower bound technique will work best for random variables of discrete support.

Theorem 4.1.1. *For all $n, k, m \in \mathbb{Z}_+$ there exists $N \in \mathbb{Z}_+$ such that the following holds:*

If there exists an algorithm A that for every pair of distributions p and q, supported over $[N]$, distinguishes the case $p = q$ from the case $\|p-q\|_{\mathcal{A}_k} \geqslant \varepsilon$ drawing m samples, then there exists an algorithm A' that for every pair of distributions p' and q' supported on $[n]$ distinguishes the case $p' = q'$ versus $\|p'-q'\|_{\mathcal{A}_k} \geqslant \varepsilon$ using the same number samples m. Moreover, A' only considers the ordering of the samples and ignores all other information.

Proof. As a preliminary simplification, we assume that our algorithm, instead of taking m samples from any combination of p or q of its choosing, takes exactly m samples from p and m samples from q, as such algorithms are strictly more powerful. This also allows us to assume that the algorithm merely takes these random samples and applies some processing to determine its output.

As a critical tool of our proof, we will use the classical Ramsey theorem for hypergraphs. For completeness, we restate it here in a slightly adapted form.

Lemma 4.1.2 (Ramsey theorem for hypergraphs, [CFS10])**.** *Given a set S and an integer t let $\binom{S}{t}$ denote the set of subsets of S of cardinality t. For all positive integers, a, b and c, there exists a positive integer N so that for any function $f : \binom{[N]}{a} \to [b]$, there exists an $S \subset [N]$ with $|S| = c$ so that f is constant on $\binom{S}{a}$.*

In words, this means that if we colour all subsets of size a of a size N set with at most b different colours, then for large enough N we will find a (bigger) subset T such that all its subsets are coloured with the same colour. Note that in our setting c from the theorem equals n.

The idea of our proof is as follows. Given an algorithm A, we will use it to implement the algorithm A'. Given A, we produce some monotonic function $f : [n] \to [N]$, and run A on the distributions $f(p)$ and $f(q)$. Since f is order preserving, $\|f(p) - f(q)\|_{\mathcal{A}_k} = \|p-q\|_{\mathcal{A}_k}$, so our algorithm is guaranteed to work. The tricky part will be to guarantee that the output of this new algorithm A' depends only on the ordering of the samples that it takes. Since we may assume that A is deterministic, once we pick

which $2m$ samples are taken from $[N]$ the output will be some function of the ordering of these samples (and in particular which are from p and which are from q). For the algorithm A, this function may depend upon the values that the samples happened to have. Thus, for A' to depend only on order, we need it to be the case that A behaves the same way on any subset of $\mathrm{Im}(f)$ of size $2m$. Fortunately, we can find such a set using Lemma 4.1.2.

Since our sample set has size at most $2m$, it is clear that the total number of possible sample sets is at most N^{2m}. We colour each of these subsets of $[N]$ of size $a = 2m$ one of a finite number of colours. The colour associates to the sample set the function that A uses to obtain an output given $2m$ samples given by this set coming in a particular order (some of which are potentially equal). The total number of such functions is at most $b = 2^{2^{4m}}$. We let n be the proposed support size for p' and q'. By Lemma 4.1.2, for N sufficiently large, there are sets of size n such that the function has the same value in samples from these sets. Letting f be the unique monotonic function from $[n]$ to $[N]$ with this set as its image, causes the output A' to depend only on the ordering of the samples.

The above reduction works as long as the samples given to our algorithm A' are distinct. To deal with the case where samples are potentially non-distinct, we show that it is possible to reduce to the case where all $2m$ samples are distinct with $9/10$ probability. To do this, we divide each of our original bins into $200m^2$ sub-bins, and upon drawing a sample from a given bin, we assign it instead to a uniformly random sub-bin. This procedure maintains the $\mathcal{A}_k$ distance between our distributions, and guarantees that the probability of a collision is small. Now, our algorithm A' will depend only on the order of the samples so long as there is no collision. As this happens with probability $9/10$, we can also ensure that this is the case when collisions do occur without sacrificing correctness. This completes our proof. □

We will now give the "hard" instance of the testing problem for algorithms that only consider the ordering of the samples. We will first describe a construction that works

for $\varepsilon = \Omega(k^{-1/6})$. We define a mini-bucket to be a segment I, which can be divided into three subsegments I_1, I_2, I_3 in that order so that $p(I_1) = p(I_3) = \varepsilon/(2k)$, $p(I_2) = 0$, and $q(I_1) = q(I_3) = 0, q(I_2) = \varepsilon/k$. We define a bucket to be an interval consisting of a mini-bucket followed by an interval on which $p = q$ and on which both p, q have total mass $(1-\varepsilon)/k$. Our distributions for p and q will consist of k consecutive buckets. See Figure 1 for an illustration.

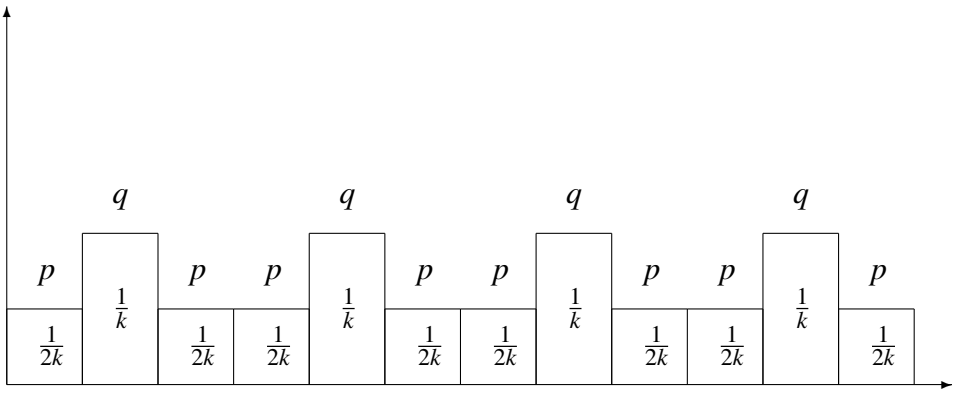

Figure 4.1: $\phi = \frac{1}{2}p + \frac{1}{2}q$ when $\varepsilon = 1$

Next consider partitioning the domain into macro-buckets each of which is a union of buckets of total mass $\Theta(1/m)$. Note that these distributions have $\mathcal{A}_{2k+1}$ distance of 2ε. An important fact to note is the following:

Observation 4.1.3. *If zero, one or two draws are made randomly and independently from $(p+q)/2$ on a mini-bucket, then the distribution of which of p or q the samples came from and their relative ordering is indistinguishable from the case where $p = q$.*

To prove the lower bound for the algorithm A', which is only allowed to look at the ordering of samples. We let X be a random variable that is taken to be 0 or 1 each with probability $1/2$. When $X = 0$ we define p and q as above with mini-buckets, macro-buckets and regular buckets as described. When $X = 1$, we let $p = q$ and define mini-buckets to have total mass ε/k for each of p and q, buckets to have total mass $1/k$ each, and we combine buckets into macro-buckets as in the $X = 0$ case.

Let Y be the distribution on the (ordered) sequences, obtained by drawing $m' = \mathrm{Poi}(m)$ samples from p and $m'' = \mathrm{Poi}(m)$ samples from q, with p and q given by X. We are interested in bounding the mutual information between X and Y, since it must be $\Omega(1)$ if the algorithm is going to succeed with probability bounded away from $1/2$. We show the following:

Theorem 4.1.4. *We have that* $I(X:Y) = O(m^5\varepsilon^6/k^4)$.

Proof. We begin with a couple of definitions. Let Y' denote (Y,α), where α is the information about which draws come from which macro-bucket. Y' consists of Y'_i, the sequence of samples coming from the i-th macro-bucket. Note that

$$I(X:Y) \leqslant I(X:Y') \leqslant \sum_{i=1}^{O(m)} I(X:Y'_i)\ .$$

We will now estimate $I(X:Y'_i)$. We claim that it is $O(\frac{m^4\varepsilon^6}{k^4})$ for each i. This would cause the sum to be small enough and give our theorem. We have that,

$$I(X:Y'_i) = \mathbb{E}_y\left[O\left(1-\frac{\Pr(Y'_i=y|X=0)}{\Pr(Y'_i=y|X=1)}\right)^2\right].$$

We then have that

$$I(X:Y'_i) = \sum_{\ell=0}^{\infty}\sum_{y:|y|=\ell}\frac{O(1)^\ell}{\ell!}O\left(1-\frac{\Pr(Y'_i=y|X=0,|y|=\ell)}{\Pr(Y'_i=y|X=1,|y|=\ell)}\right)^2.$$

We note that if $X=1,|y|=\ell$ that any of the 2^ℓ possible orderings are equally likely. On the other hand, if $X=0$, this also holds in an approximate sense. To show this, first consider picking which mini-buckets our ℓ draws are from. If no three land in the same mini-bucket, then Observation 4.1.3 implies that all orderings are equally likely. Therefore, the statistical distance between $Y'_i|X=0,|y|=\ell$ and $Y'_i|X=1,|y|=\ell$ is at most the probability that some three draws come from the same mini-bucket. This is in turn at most the expected number of triples that land in the same mini-bucket, which is equal to $\binom{\ell}{3}$ times the probability that a particular triple does. The probability of landing in a particular mini-bucket is $O(m\varepsilon/k)^3$. By definition, there are $O(m/k)$

mini-buckets in a macro-bucket, so this probability is $O(\ell^3\varepsilon^3(m/k)^2)$. Therefore, we have that

$$\begin{aligned}
I(X:Y_i') &= \sum_\ell \frac{O(1)^\ell}{\ell!} \sum_{y:|y|=\ell} O(4^\ell)\left(\Pr(Y_i'=y|X=0,|y|=\ell)-\Pr(Y_i'=y|X=1,|y|=\ell)\right)^2 \\
&\leqslant \sum_\ell \frac{O(1)^\ell}{\ell!}\left(\sum_{y:|y|=\ell}\left|\Pr(Y_i'=y|X=0,|y|=\ell)-\Pr(Y_i'=y|X=1,|y|=\ell)\right|\right)^2 \\
&= \sum_\ell \frac{O(1)^\ell}{\ell!} O\left(\ell^6\varepsilon^6 m^4/k^4\right) \\
&= \frac{m^4}{k^4}\sum_\ell \frac{O(1)^\ell \ell^6\varepsilon^6}{\ell!} \\
&= O\left(\frac{m^4\varepsilon^6}{k^4}\right).
\end{aligned}$$

This completes our proof. □

The above construction only works when $k \geqslant m$, or equivalently, when $\varepsilon = \Omega(k^{-1/6})$. When ε is small, we need a slightly different construction. We will similarly split our domain into mini-buckets and macro-buckets and argue based on shared information. Once again we define two distributions p and q, though this time the distributions themselves will need to be randomised. Given k and ε, we begin by splitting the domain into k macro-buckets. Each macro-bucket will have mass $1/k$ under both p and q.

First pick a global variable X to be either 0 or 1 with equal probability. If $X = 1$ then we will have $p = q$ and if $X = 0$, $\|p-q\|_{\mathcal{A}_{2k+1}} = \varepsilon$. For each macro-bucket, pick an x uniformly in $[0,(1-\varepsilon)/k]$. The macro-bucket will consist of an interval on which $p = q$ with mass x (for each of p,q), followed by a mini-bucket, followed by an interval of mass $(1-\varepsilon)/k - x$ on which $p = q$. The mini-bucket is an interval of mass ε/k under either p or q. If $X = 1$, we have $p = q$ on the mini-bucket. If $X = 0$, the mini-bucket consists of an interval of mass $\varepsilon/(2k)$ under q and 0 under p, an interval of mass ε/k under p and 0 under q, and then another interval of mass $\varepsilon/(2k)$ under q and 0 under p.

We let Y be the random variable associated with the ordering of elements from a set of $\mathrm{Poi}(m)$ draws from each of p and q. We show:

Theorem 4.1.5. *If $m\varepsilon = O(k)$, $\log(mk/\varepsilon) = O(\varepsilon^{-1})$, and $k = O(m)$, with implied constants sufficiently small, then $I(X:Y) = O(m^5\varepsilon^6/k^4)$.*

Note that the above statement differs from Theorem 4.1.4 in that X and Y are defined differently.

Proof. Once again, we let Y' be Y along with the information of which draws came from which macro-bucket, and let Y_i' be the information of the draws from the i-th macro-bucket along with their ordering. It suffices for us to show that $I(X:Y_i') = O(m^5\varepsilon^6 k^{-5})$ for each i (as now there are only k macro-buckets rather than m).

Let s be a string of ℓ ordered draws from p and q. In particular, we may consider s to be a string $s_1 s_2 \dots s_\ell$, where $s_i \in \{p,q\}$. We wish to consider the probability that $Y_i' = s$ under the conditions that $X = 0$ or that $X = 1$. In order to do this, we further condition on which elements of s were drawn from the mini-bucket. For $1 \leqslant a \leqslant b \leqslant \ell$ we consider the probability that not only did we obtain sequence s, but that the draws $s_a, \dots, s_b$ were exactly the ones coming from the mini-bucket within this macro-bucket. Let h denote the ordered string coming from elements drawn from the mini-bucket and M the ordered sequence of strings coming from elements not drawn from the mini-bucket. The probability of the event in question is then

$$\Pr(h = s_a \dots s_b)\cdot$$
$$\cdot \Pr(M = s_1 \dots s_{a-1} s_{b+1} \dots s_\ell)\cdot$$
$$\cdot \Pr(\text{the mini-bucket is placed between } s_{a-1} \text{ and } s_{b+1}).$$

Note that the mini-bucket can be thought of as being randomly and uniformly inserted within an interval of length $(1-\varepsilon)/k$ and that this is equally likely to be inserted between any pair of elements of M. Thus, the probability of the third term in the product is exactly $1/(\ell + a - b)$. The second probability is the probability that $\ell + a -$

$b-1$ elements are drawn from the complement of the mini-bucket times $2^{-(\ell+a-b+1)}$, as draws from p and q are equally likely. Thus, letting $t = b-a+1$ (i.e., the number of elements in the mini-bucket), we have that

$$\Pr(Y_i' = s) =$$
$$e^{-m/k} \sum_{t=0}^{\ell} \left(\frac{\left(\frac{m(1-\varepsilon)}{2k}\right)^{\ell-t}}{(\ell-t)!} \right) \left(\frac{\left(\frac{m\varepsilon}{k}\right)^{t}}{t!} \right) \left(\frac{1}{\ell-t} \right) \sum_{a} \Pr(h = s_a \dots s_{a+t-1} : |h| = t).$$

Note that this equality holds even after conditioning upon X. We next simplify this expression further by grouping together terms in the last sum based upon the value of the substring $s_a \dots s_{a+t-1}$, which we call r. We get that

$$\Pr(Y_i' = s) = e^{-m/k} \sum_{t=0}^{\ell} \left(\frac{\left(\frac{m(1-\varepsilon)}{2k}\right)^{\ell-t}}{(\ell-t)!} \right) \left(\frac{\left(\frac{m\varepsilon}{k}\right)^{t}}{t!} \right) \left(\frac{1}{\ell-t} \right) \sum_{|r|=t} \Pr(h = r : |h| = t) N_{r,s},$$

where $N_{r,s}$ is the number of occurrences of r as a substring of s.

Next, we wish to bound

$$\sum_{|s|=\ell} |\Pr(Y_i' = s : X = 0) - \Pr(Y_i' = s : X = 1)|^2. \tag{4.1}$$

By the above formula this is at most

$$e^{-2m/k} \sum_{|s|=\ell} \sum_{t=0}^{\ell} \left(\frac{\left(\frac{m(1-\varepsilon)}{2k}\right)^{\ell-t}}{(\ell-t)!} \right) \left(\frac{\left(\frac{m\varepsilon}{k}\right)^{t}}{t!} \right) \left(\frac{1}{\ell-t} \right) \cdot$$
$$\left| \sum_{|r|=t} N_{r,s} \left(\Pr(h = r : |h| = t, X = 0) - \Pr(h = r : |h| = t, X = 1) \right) \right|^2.$$

For fixed values of t we consider the sum

$$\sum_{|s|=\ell} \left| \sum_{|r|=t} N_{r,s} (\Pr(h = r : |h| = t, X = 0) - \Pr(h = r : |h| = t, X = 1)) \right|^2.$$

Note that if $t \leqslant 2$ then $\Pr(h = r : |h| = t, X = 0) = \Pr(h = r : |h| = t, X = 1)$, and so the above sum is 0. Otherwise, it is at most

$$\sum_{|s|=\ell} \sum_{|r|=t} |N_{r,s} - (\ell+1-t)/2^t|^2$$

because $\sum_r \Pr(h = r : |h| = t, X = 0) = \sum_r \Pr(h = r : |h| = t, X = 1) = 1$. Note on the other hand that the expectation over random strings s of length ℓ of $N_{r,s} - (\ell + 1 - t)/2^t$ is 0. Furthermore, the variance of $N_{r,s}$ is easily bounded by $t\ell 2^{-t}$ as whether or not two disjoint substrings of s are equal to r are independent events. Therefore, the above sum is at most

$$2^\ell 2^t t\ell 2^{-t} = 2^\ell t\ell.$$

Hence, by Cauchy-Schwartz, we have that

$$\sum_{|s|=\ell} \left| \sum_{|r|=t} N_{r,s} - (\ell + 1 - t)/2^t \right|^2 \leqslant 2^\ell 2^t t\ell.$$

Therefore, the expression in (4.1) is at most

$$e^{-2m/k} \left(\sum_{t=3}^{\ell} \left(\frac{\left(\frac{m(1-\varepsilon)}{2k}\right)^{\ell-t}}{(\ell-t)!} \right) \left(\frac{O\left(\frac{m\varepsilon}{k}\right)^t}{t!} \right) \left(\frac{1}{\ell-t} \right) \left(2^\ell 2^t \ell t \right)^{1/2} \right)^2.$$

Assuming that $\ell\varepsilon$ is sufficiently small, these terms are decreasing exponentially with t, and thus this is

$$O\left(e^{-2m/k} \left(\frac{(m^2/(2k^2))^\ell}{(\ell!)^2} \right) \varepsilon^6 \ell^5 \right).$$

Now we have that for N a sufficiently small constant times ε^{-1},

$$\begin{aligned}
I(X : Y_i') &= \sum_s \Pr(Y_i' = s : X = 1) O\left(1 - \frac{\Pr(Y_i' = s : X = 0)}{\Pr(Y_i' = s : X = 1)}\right)^2 \\
&= \sum_\ell \sum_{s:|s|=\ell} e^{m/k} \left(\frac{(m/(2k))^\ell}{\ell!} \right)^{-1} O(\Pr(Y_i' = s : X = 1) - \Pr(Y_i' = s : X = 0))^2 \\
&\leqslant \sum_\ell e^{m/k} \left(\frac{(m/(2k))^\ell}{\ell!} \right)^{-1} O\left(\sum_{s:|s|=\ell} |\Pr(Y_i' = s : X = 1) - \Pr(Y_i' = s : X = 0)|^2 \right) \\
&\leqslant \sum_{\ell > N} O\left(\frac{(2m/k)^\ell}{\ell!} \right) + \sum_{\ell < N} e^{m/k} \left(\frac{(m/(2k))^\ell}{\ell!} \right)^{-1} O\left(e^{-2m/k} \left(\frac{(m^2/(2k^2))^\ell}{(\ell!)^2} \right) \varepsilon^6 \ell^5 \right) \\
&\leqslant \sum_{\ell > N} O\left(\frac{m}{kN} \right)^\ell + \sum_\ell O\left(e^{-m/k} \frac{(m/k)^\ell}{\ell!} \varepsilon^6 \ell^5 \right).
\end{aligned}$$

Since $\frac{m}{kN} \leq \frac{m\varepsilon}{k}$ is sufficiently small, the first term is at most $(1/2)^N$ which is polynomially small in mk/ε, and thus negligible. The second term is the expectation

of $\varepsilon^6\ell^5$ for ℓ a Poisson random variable with mean m/k. Thus, it is easily seen to be $O((m/k)^5\varepsilon^6)$. Therefore, we have that $I(X : Y_i') = O(m^5\varepsilon^6 k^{-5})$, and therefore, $I(X : Y) = O(m^5\varepsilon^6 k^{-4})$, as desired. □

We are now ready to complete the proof of our general lower bound.

Theorem 4.1.6. *For any $k > 2$, there exists an N so that any algorithm that is given sample access to two distributions, p and q over $[N]$, and can distinguish between the cases $p = q$ and $\|p - q\|_{\mathcal{A}_k}$ with probability at least $2/3$, requires at least*

$$\Omega\left(\max\left\{k^{4/5}/\varepsilon^{6/5}, k^{1/2}/\varepsilon^2\right\}\right)$$

samples.

Proof. The lower bound of $k^{1/2}/\varepsilon^2$ follows from the known lower bound [Pan08] even in the case where q is known and p and q have support of size k. It now suffices to consider the case that $\varepsilon > k^{-1/2}$ and m a sufficiently small constant times $k^{4/5}\varepsilon^{-6/5}$.

Note that by Theorem 4.1.1, we may assume that the algorithm in question takes m samples from each of p and q and determines its output based only on the ordering of the samples. We need to show that this is impossible for N sufficiently large.

We note that if we allow p and q to be continuous distributions instead of discrete ones we are already done. If $m < k$, we use our first counter-example construction, and if $m \geqslant k$ use the second one. If we let X be randomly 0 or 1, and set $p = q$ for $X = 1$ and p, q as described above when $X = 0$, then by Theorems 4.1.4 and 4.1.5, the shared information between X and the output of our algorithm is at most $O(m^5\varepsilon^6 k^{-4}) = o(1)$, and thus our algorithm cannot correctly determine X with constant probability.

In order to prove our Theorem, we will need to make this work for distributions p and q with finite support size as follows: By splitting our domain into m^3 intervals each of equal mass under $p + q$, we note that the $\mathcal{A}_k$ distance between the distributions is only negligibly affected. Furthermore, with high probability, m samples will have no pair chosen from the same bin. Thus, the distribution on orderings of samples from

these discrete distributions are nearly identical to the continuous case, and thus our algorithm would behave nearly identically. This completes the proof. □

4.2 A Lower Bound for Testing Identity to an Unknown Distribution with Small Domain Size

The Theorem 3.4.1 provides an upper bound which nearly (up to double logarithmic factors) matches the upper bound of Theorem 3.4.1 (Theorem 1.3.3)

Theorem 4.2.1. *Let p and q be distributions on $[n]$ and let $\varepsilon > 0$ be sufficiently small. Any tester that distinguishes between $p = q$ and $\|p-q\|_{\mathcal{A}_k}$ for some $k \leq n$ must use $\Omega(m)$ samples for $m = \min(k^{2/3}\log^{1/3}(3+n/k)/\varepsilon^{4/3}, k^{4/5}/\varepsilon^{6/5})$.*

In fact, this lower bound holds even if p and q are both guaranteed to be piecewise constant distributions on $O(k+m)$ pieces.

Our new lower bounds are somewhat more complicated. We prove them by exhibiting explicit families of pairs of distributions, where in one case $p = q$ and in the other p and q have large $\mathcal{A}_k$ distance, but so that it is impossible to distinguish between these two families with a small number of samples. In both cases, p and q are explicit piecewise constant distributions with a small number of pieces. In both cases, our domain is partitioned into a small number of bins and the restrictions of the distributions to different bins are independent, making our analysis easier. In some bins we will have $p = q$ each with mass about $1/m$ (where m is the number of samples). These bins will serve the purpose of adding "noise" making harder to read the "signal" from the other bins. In the remaining bins, we will have either that $p = q$ being supported on some interval, or p and q will be supported on consecutive, non-overlapping intervals. If three samples are obtained from any one of these intervals, the order of the samples and the distributions that they come from will provide us with information about which family we came from. Unfortunately, since triple collisions are relatively uncommon,

this will not be useful unless $m \gg \max(k^{4/5}/\varepsilon^{6/5}, k^{1/2}/\varepsilon^2)$. Bins from which we have one or zero samples will tell us nothing, but bins from which we have exactly two samples may provide information.

For these bins, it can be seen that we learn nothing from the ordering of the samples, but we may learn something from their spacing. In particular, in the case where p and q are supported on disjoint intervals, we would suspect that two samples very close to each other are far more likely to be taken from the same distribution rather than from opposite distributions. On the other hand, in order to properly interpret this information, we will need to know something about the scale of the distributions involved in order to know when two points should be considered to be "close". To overcome this difficulty, we will stretch each of our distributions by a random exponential amount. This will effectively conceal any information about the scales involved so long as the total support size of our distributions is exponentially large.

Here we prove a nearly matching sample lower bound. We begin by proving a bound for continuous distributions that are piecewise constant with few pieces. Our bound on discrete distributions will follow from taking the adversarial distribution from this example and rounding its values to the nearest integer. In order for this to work, we will need ensure to that our adversarial distribution does not have its $\mathcal{A}_k$-distance decrease by too much when we apply this operation. To satisfy this requirement, we will guarantee that our distributions will be piecewise constant with all the pieces of length at least 1.

Proposition 4.2.2. *Let $k \in \mathbb{Z}_+$, $\varepsilon > 0$ sufficiently small, and $W > 2$.*
Let $m = \min(k^{2/3}\log^{1/3}(W)/\varepsilon^{4/3}, k^{4/5}/\varepsilon^{6/5})$. There exist distributions $\mathcal{D}, \mathcal{D}'$ over pairs of distributions p and q on $[0, 2(m+k)W]$ so that p and q are $O(m+k)$-flat with pieces of length at least 1, so that when drawn from $\mathcal{D}$, $p = q$ deterministically, when drawn from $\mathcal{D}'$ $\|p-q\|_{\mathcal{A}_k} > \varepsilon$ with 90% probability, and so that $o(m)$ samples are insufficient to distinguish whether or not the pair is drawn from $\mathcal{D}$ or $\mathcal{D}'$ with better than $2/3$ probability.

The basic idea of the proof will be as follows. At a high-level, our construction will mimic the lower bound construction of [DKN15a]. We will divide our domain into $m+k$ bins so that no information about which distributions had samples drawn from a given bin or the ordering of these samples will help to distinguish between the cases of $p = q$ and otherwise, unless at least three samples are taken from the bin in question. Approximately k of these bins will each have mass ε/k and might convey this information if at least three samples are taken from the bin. However, the other m bins will each have mass approximately $1/m$, and will be used to add noise. In all, if we take s samples, we expect to see approximately $s^3\varepsilon^3/k^2$ of the lighter bins with at least three samples. However, we will see approximately s^3/m^2 of our heavy bins with three samples. In order for the signal to overwhelm the noise we will need to ensure that we have $(s^3\varepsilon^3/k^2)^2 > s^3/m^2$.

The above analysis assumes that we cannot obtain information from the bins in which only two samples are drawn. This naively should not be the case. If $p = q$ then the distance between two samples drawn from that bin will be independent of whether or not they are drawn from the same distribution. However, if p and q are supported on disjoint intervals, one would expect that points that are close to each other should be far more likely to be drawn from the same distribution than from different distributions. In order to disguise this, we will scale the length of the intervals by a random, exponential amount, essentially making it impossible to determine what is meant by two points being close to each other. In effect, this will imply that two points drawn from the same bin will only tell us $O(1/\log(W))$ bits of information about whether $p = q$ or not. Thus, in order for this information to be sufficient, we will need that $(s^2\varepsilon^2/k)^2/\log(W) > (s^2/m)$.

Proof of Proposition 4.2.2: We use ideas from [DK16] to obtain this lower bound using an argument from information theory.

We may assume that $\varepsilon > k^{1/2}$, because otherwise we may employ the standard lower bound that $\Omega(\sqrt{k}/\varepsilon^2)$ samples are required to distinguish two distributions on a

support of size k.

First, we note that it is sufficient to take $\mathcal{D}$ and $\mathcal{D}'$ distributions over pairs of non-negative, piecewise constant distributions with total mass $\Theta(1)$ with 90% probability so that running a Poisson process with parameter $o(m)$ is insufficient to distinguish a pair from $\mathcal{D}$ from a pair from $\mathcal{D}'$ [DK16].

We construct this distribution as follows: We divide the domain into $m+k$ bins of length $2W$. For each bin i, we independently generate a random ℓ_i, so that $\log(\ell_i/2)$ is uniformly distributed over $[0, 2\log(W)/3]$. We then produce an interval I_i within bin i of total length ℓ_i and with random offset. In all cases, we will have p and q supported on the union of the I_i's.

For each i with probability $m/(m+k)$, we have the restrictions of p and q to I_i both uniform with $p(I_i) = q(I_i) = 1/m$. The other $k/(m+k)$ of the time we have $p(I_i) = q(I_i) = \varepsilon/k$. In this latter case, if p and q are being drawn from $\mathcal{D}$, p and q are each constant on this interval. If they are being drawn from $\mathcal{D}'$, then $p+q$ will be constant on the interval, with all of that mass coming from p on a random half and coming from q on the other half.

Note that in all cases p and q are piecewise constant with $O(m+k)$ pieces of length at least 1. It is easy to show that with high probability the total mass of each of p and q is $\Theta(1)$, and that if drawn from $\mathcal{D}'$ that $\|p-q\|_{\mathcal{A}_k} \gg \varepsilon$ with at least 90% probability.

Now we will show that if one is given m samples from each of p and q, taken randomly from either $\mathcal{D}$ or $\mathcal{D}'$, that the shared information between the samples and the source family will be small. This implies that one is unable to consistently guess whether our pair was taken from $\mathcal{D}$ or $\mathcal{D}'$.

Let X be a random variable that is uniformly randomly either 0 or 1. Let A be obtained by applying a Poisson process with parameter $s = o(m)$ on the pair of distributions p, q drawn from $\mathcal{D}$ if $X = 0$ or from $\mathcal{D}'$ if $X = 1$. We note that it suffices to show that the shared information $I(X : A) = o(1)$. In particular, by Fano's inequality, we have:

Lemma 4.2.3. *If X is a uniform random bit and A is a correlated random variable, then if f is any function so that $f(A) = X$ with at least 51% probability, then $I(X:A) \geq 2 \cdot 10^{-4}$.*

Let A_i be the samples of A taken from the i^{th} bin. Note that the A_i are conditionally independent on X. Therefore, we have that $I(X:A) \leq \sum_i I(X:A_i) = (m+k)I(X:A_1)$. We will proceed to bound $I(X:A_1)$.

We note that $I(X:A_1)$ is at most the integral over pairs of multisets a (representing a set of samples from q and a set of samples from p), of

$$O\left(\frac{(\Pr(A_1 = a|X=0) - \Pr(A_1 = a|X=1))^2}{\Pr(A_1 = a)}\right).$$

Thus,

$$I(X:A_1) = \sum_{h=0}^{\infty} \int_{|a|=h} O\left(\frac{(\Pr(A_1 = a|X=0) - \Pr(A_1 = a|X=1))^2}{\Pr(A_1 = a)}\right).$$

We will split this sum up based on the h.

For $h = 0$, we note that the distributions for $p+q$ are the same for $X = 0$ and $X = 1$. Therefore, the probability of selecting no samples is the same. Therefore, this contributes 0.

For $h = 1$, we note that the distributions for $p+q$ are the same in both cases, and conditioning on I_1 and $(p+q)(I_1)$ that $\mathbb{E}[p]$ and $\mathbb{E}[q]$ are the same in each of the cases $X = 0$ and $X = 1$. Therefore, again in this case, we have no contribution.

For $h \geq 3$, we note that $I(X:A_1) \leq I(X:A_1,I_1) \leq I(X:A_1|I_1)$, since I_1 is independent of X. We note that $\Pr(A_1 = a|X=0, p(I_1) = 1/m) = \Pr(A_1 = a|X=1, p(I_1) = 1/m)$. Therefore, we have that

$$\Pr(A_1 = a|X=0) - \Pr(A_1 = a|X=1) =$$

$$\Pr(A_1 = a|X=0, p(I_1) = \varepsilon/k) - \Pr(A_1 = a|X=1, p(I_1) =$$

$$\varepsilon/k).$$

If $p(I_1) = \varepsilon/k$, the probability that exactly h elements are selected in this bin is at most $k/(m+k)(2s\varepsilon/k)^h/h!$, and if they are selected, they are uniformly distributed in I_1

(although which of the sets p and q they are taken from is non-uniform). However, the probability that h elements are taken from I_1 is at least $\Omega(m/(m+k)(sm)^{-h}/h!)$ from the case where $p(I_1) = 1/m$, and in this case the elements are uniformly distributed in I_1 and uniformly from each of p and q. Therefore, we have that this contribution to our shared information is at most $k^2/(m(m+k))O(s\varepsilon^2 m/k^2)^h/h!$. We note that $\varepsilon^2 m/k^2 < 1$. Therefore, the sum of this over all $h \geq 3$ is $k^2/(m(m+k))O(s\varepsilon^2 m/k^2)^3$. Summing over all $m+k$ bins, this is $k^{-4}\varepsilon^6 s^3 m^2 = o(1)$. It remains to analyse the case where $h = 2$. Once again, we have that ignoring which of p and q elements of A_1 came from that A_1 is identically distributed conditioned on $p(I_1) = 1/m$ and $|A_1| = 2$ as it is conditioned on $p(I_1) = \varepsilon/k$ and $|A_1| = 2$. Since once again, the distributions $\mathcal{D}$ and $\mathcal{D}'$ are indistinguishable in the former case, we have that the contribution of the $h = 2$ terms to the shared information is at most

$$O\left(\frac{(k/(k+m)(\varepsilon s/k)^2)^2}{m/(k+m)(s/m)^2}\right)\cdot$$
$$\cdot d_{\mathrm{TV}}((A_1|X=0, p(I_1)\varepsilon/k, |A_1|=2), (A_1|X=1, p(I_1)=\varepsilon/k, |A_1|=2))$$

or

$$O\left(s^2 m k^{-2}\varepsilon^4/(k+m)\right)\cdot$$
$$\cdot d_{\mathrm{TV}}((A_1|X=0, p(I_1)=\varepsilon/k, |A_1|=2), (A_1|X=1, p(I_1)=\varepsilon/k, |A_1|=2)) .$$

It will suffice to show that conditioned upon $p(I_1) = \varepsilon/k$ and $|A_1| = 2$ that $d_{\mathrm{TV}}((A_1|X = 0), (A_1|X = 1)) = O(1/\log(W))$.

Let f be the order preserving linear function from $[0,2]$ to I_1. Notice that conditional on $|A_1| = 2$ and $p(I_1) = \varepsilon/k$ that we may sample from A_1 as follows:

- Pick two points $x > y$ uniformly at random from $[0,2]$.
- Assign the points to p and q as follows:
  - If $X = 0$ uniformly randomly assign these points to either distribution p or q.

  - If $X = 1$ randomly do either:
    - Assign points in $[0,1]$ to q and other points to p.
    - Assign points in $[0,1]$ to p and other points to q.

- Randomly pick I_1 and apply f to x and y to get outputs $z = f(x), w = f(y)$.

Notice that the four cases: (i) both points coming from p, (ii) both points coming from q, (iii) a point from p preceding a point from q, (iv) a point from q preceding a point from p, are all equally likely conditioned on either $X = 0$ or $X = 1$. However, we will note that this ordering is no longer independent of the choice of x and y.

We note therefore that we can sample from A_1 subject to $X = 0$ and from A_1 subject to $X = 1$ in such a way that this ordering is the same deterministically. We consider running the above sampling algorithm to select (x,y) while sampling from $X = 0$ and (x',y') when sampling from $X = 1$ so that we are in the same one of the above four cases. We note that

$$d_{\mathrm{TV}}((A_1|X=0),(A_1|X=1)) \leq \mathbb{E}_{x,y,x',y'}[d_{\mathrm{TV}}((f(x),f(y)),(f(x'),f(y')))] ,$$

where variational distance is over the random choices of f.

To show that this is small, we note that $|f(x) - f(y)|$ is distributed like $\ell_1(x - y)$. This means that $\log(|f(x) - f(y)|)$ is uniform over $[\log(f(x) - f(y)), \log(f(x) - f(y)) + 2\log(W)/3]$. Similarly, $\log(|f'(x') - f'(y')|)$ is uniform over $[\log(f(x') - f(y')), \log(f(x') - f(y')) + 2\log(W)/3]$. These differ in total variation distance by

$$O\left(\frac{|\log(f(x) - f(y))| + |\log(f(x') - f(y'))|}{\log(W)}\right) .$$

Taking the expectation over x, y, x', y' we get $O(1/\log(W))$. Therefore, we may further correlate the choices made in selecting our two samples, so that the $z - w = z' - w'$ except with probability $O(1/\log(W))$. We note that after conditioning on this z and z' are both uniformly distributed over subintervals of $[0, 2W]$ of length at least $2(W - W^{2/3})$. Therefore, the distributions on z and z' differ by at most $O(W^{-1/3})$. Hence, the

total variation distance between A_1 conditioned on $|A_1| = 2, p(I_1) = \varepsilon/k, X = 0$ and conditioned on $|A_1| = 2, p(I_1) = \varepsilon/k, X = 1$ is at most $O(1/\log(W)) + O(W^{-1/3}) = O(1/\log(W))$. This completes our proof. □

We can now turn this into a lower bound for testing $\mathcal{A}_k$ distance on discrete domains.

Proof of Theorem 4.2.1: Assume for sake of contradiction that this is not the case, and that there exists a tester taking $o(m)$ samples. We use this tester to come up with a continuous tester that violates Proposition 4.2.2.

We begin by proving a few technical bounds on the parameters involved. Firstly, note that we already have a lower bound of $\Omega(k^{1/2}/\varepsilon^2)$, so we may assume that this is much less than m. We now claim that $m = O(\min(k^{2/3}\log^{1/3}(3 + n/(m + k))/\varepsilon^{4/3}, k^{4/5}/\varepsilon^{6/5})$. If $m \leq k$, there is nothing to prove. Otherwise,

$$k^{2/3}\log^{1/3}(3+n/(m+k))/\varepsilon^{4/3} \geq m(m/k)^{-1/3}\log(3+n/(m+k))^{1/3}.$$

Thus, there is nothing more to prove unless $\log(3+n/(m+k)) \gg m/k$. But, in this case, $\log(3+n/(m+k)) \gg \log(m/k)$ and thus $\log(3+n/(m+k)) = \Theta(\log(3+n/k))$, and we are done.

We now let $W = n/(6(m+k))$, and let $\mathcal{D}$ and $\mathcal{D}'$ be as specified in Proposition 4.2.2. We claim that we have a tester to distinguish a p, q from $\mathcal{D}$ from ones taken from $\mathcal{D}'$ in $o(m)$ samples. We do this as follows: By rounding p and q down to the nearest third of an integer, we obtain p',q' supported on set of size n. Since p and q were piecewise constant on pieces of size at least 1, it is not hard to see that $\|p'-q'\|_{\mathcal{A}_k} \geq \|p-q\|_{\mathcal{A}_k}/3$. Therefore, a tester to distinguish $p' = q'$ from $\|p'-q'\|_{\mathcal{A}_k} \geq \varepsilon$ can be used to distinguish $p = q$ from $\|p-q\|_{\mathcal{A}_k} \geq 3\varepsilon$. This is a contradiction and proves our lower bound. □

Chapter 5

Conclusion and Future Work

This thesis had quite an ambitious aim: to improve our understanding of property testing beyond the "worst case" adversarial scenarios. For two major, probably the most typical problems: testing identity for the cases of one and two unknown distributions, this aim was fullfilled. Moreover, we have proposed a method which not only gave as corollaries optimal testers for broad structured classes, but also created a foundation on which other efficient testers can be built if new approximation bounds are discovered.

In this chaper I will briefly summarise the results and give some topics which could be possibly worth exploring in the future.

This thesis explores the idea of testing under shape restrictions of distributions, that is, the distributions for which their probability density functions or their probability mass functions can be well approximated by polynomials. While this thesis almost closes two, perhaps, most representative problems in hypothesis testing, the idea of shape-restricted testing goes beyond.

The generic method of creating testers for the L_1 distance introduced in Proposition 1.3.4 intuitively suggests that almost all other properties of arbitrary distributions may require much fewer samples to be tested, given that some a priori knowledge about the underlying distribution is known.

This thesis deals with the problem of testing "exact" identity, that is, in the com-

pleteness case p must be identical to q almost everywhere. While this is probably the most known and characteristic problem, several others are often explored in the literature and thus could potentially benefit from the methods developed throughout this thesis. For instance, one may want to relax this requirement and allow p and q to be different by some function of small L_1 norm. This problem is called "tolerant testing" and is explored in the unrestricted case by Valiant and Valiant ([VV11b]).

Definition 5.0.4 (Tolerant testing)**.** Given probability distributions p and q, return "TRUE" if $\underset{\cdot}{(p,q)} < \varepsilon$ and "FALSE" if $\underset{\cdot}{(p,q)} > 2\varepsilon$.

As with the problems considered in this thesis, p and q may be given either explicitly or as sample sets. It is known that this problem is equivalent to the problem of L_1 distance estimation between distributions. Therefore one may want to explore if it is possible to test the $\mathcal{A}_k$-distance, and find better structured testing results as corollaries.

The same paper ([VV11b]) deals with two other popular properties: Another two famous properties to consider are *entropy* and *support size*. The reader may refer to, e.g. [VV11a].

All together, these problems can be described by the term "symmetric" — that is, they do not change their value if the order of elements in the domain is changed. It is known that for these distributions, all relevant information can be condensed to a "fingerprint". The question whether shape restrictions on the distribution imply similar restrictions on the fingerprint, and whether this can be exploited to create better testers is unknown.

Another point to build on top of this thesis is the Corollary 1.3.5. It gives optimal results for various distribution families, many other families arising in practice still are not well described in terms of approximation. Apart from trying to establish the precise approximation constant k values for various classes of distributions, one could also think of finding other general methods of approximating distributions.

Recalling that the theoretical computer science community initially adopted property testing as a tool for establishing properties of graphs, we can also ask ourselves if

the shape restriction of the distribution on the vertices of a graph, generated by a random walk, can tell us anything about the properties of the graph. (And vice-versa — we know that if the graph is an expander then the distribution will be uniform ([GR00]), but many other properties can probably be uncovered as well.)

Another area where structured testing may be useful is the external memory model. That is, if the algorithm is only allowed to take samples from the distribution in blocks of size B similarly to the model described in [AIOR09]:

> typically, massive data sets are not stored in main memory, where each element can be accessed at a unit cost. Instead, the data is stored on external storage devices, such as a hard disk. There, the data is stored in blocks of certain size (say, B), and each disk access returns a block of data, as opposed to an individual element. In such models, it is often possible to solve problems using roughly T/B disk accesses, where T is the time needed to solve the problem in main memory. The 1/B factor is often crucial to the efficiency of the algorithms, given that (a) the block size B tends to be large, on the order of thousands and (b) each block access is many orders of magnitude slower than a main memory lookup.

Their result for testing identity (requiring $O(\sqrt{\frac{m}{B}}\log B\frac{1}{\varepsilon})$ samples) works when restrictions are enforced *on the way the algorithm is allowed to sample. Can we save queries if the underlying distribution comes from a certain family? What if there are some other restrictions of the method of sampling?*

One may also try and expand the result to high-dimensional distributions. The previous results by Batu et al. ([BKR04]) give a sublinear result for a two-dimensional unstructured case, but structured results may happen to be much better. The problem is going to be much harder in this case, since the solution of the equation

$$f(x,y)-g(x,y)=0$$

is not necessarily a set of points, but rather is a union of functions $y_i(x)$ and thus some generalisation on the $\mathcal{A}_k$-distance will be needed.

In conclusion we may speculate that the idea of structured testing seems to be of both theoretical and practical importance, and express hope that many interesting results are still to come in the future.

Bibliography

[AIOR09] *A. Andoni, P. Indyk, K. Onak, and R. Rubinfeld. External sampling. In* ICALP (1)*, pages 83–94, 2009.*

[AK01] *S. Arora and R. Kannan. Learning mixtures of arbitrary Gaussians. In* Proceedings of the 33rd Symposium on Theory of Computing*, pages 247–257, 2001.*

[BBBB72] *R.E. Barlow, D.J. Bartholomew, J.M. Bremner, and H.D. Brunk.* Statistical Inference under Order Restrictions*. Wiley, New York, 1972.*

[BFF+01] *T. Batu, E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White. Testing random variables for independence and identity. In* Proc. 42nd IEEE Symposium on Foundations of Computer Science*, pages 442–451, 2001.*

[BFR+00] *T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing that distributions are close. In* IEEE Symposium on Foundations of Computer Science*, pages 259–269, 2000.*

[BFR+13] *T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing closeness of discrete distributions.* J. ACM*, 60(1):4, 2013.*

[Bir87a] *L. Birgé. Estimating a density under order restrictions: Nonasymptotic minimax risk.* Annals of Statistics*, 15(3):995–1012, 1987.*

[Bir87b] *L. Birgé. On the risk of histograms for estimating decreasing densities.* Annals of Statistics*, 15(3):1013–1022, 1987.*

[BKR04] *T. Batu, R. Kumar, and R. Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In* ACM Symposium on Theory of Computing*, pages 381–390, 2004.*

[Bru58] *H. D. Brunk. On the estimation of parameters restricted by inequalities.* The Annals of Mathematical Statistics*, 29(2):pp. 437–454, 1958.*

[BRW09] *F. Balabdaoui, K. Rufibach, and J. A. Wellner. Limit distribution theory for maximum likelihood estimation of a log-concave density.* The Annals of Statistics*, 37(3):pp. 1299–1331, 2009.*

[BS10] *M. Belkin and K. Sinha. Polynomial learning of distribution families. In* FOCS*, pages 103–112, 2010.*

[BW07] *F. Balabdaoui and J. A. Wellner. Estimation of a k-monotone density: Limit distribution theory and the spline connection.* The Annals of Statistics*, 35(6):pp. 2536–2564, 2007.*

[BW10] *F. Balabdaoui and J. A. Wellner. Estimation of a k-monotone density: characterizations, consistency and minimax lower bounds.* Statistica Neerlandica*, 64(1):45–70, 2010.*

[CDSS13] *S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Learning mixtures of structured distributions over discrete domains. In* SODA*, pages 1380–1394, 2013.*

[CDSS14] *S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Efficient density estimation via piecewise polynomial approximation. In* STOC*, pages 604–613, 2014.*

[CDVV14] S. Chan, I. Diakonikolas, P. Valiant, and G. Valiant. Optimal algorithms for testing closeness of discrete distributions. In SODA*, pages 1193–1203, 2014.*

[CFS10] D. Conlon, J. Fox, and B. Sudakov. Hypergraph ramsey numbers. Journal of the American Mathematical Society*, 23(1):247–266, 2010.*

[CT04] K.S. Chan and H. Tong. Testing for multimodality with dependent data. Biometrika*, 91(1):113–123, 2004.*

[Das99] Sanjoy Dasgupta. Learning mixtures of gaussians. In Foundations of Computer Science, 1999. 40th Annual Symposium on*, pages 634–644. IEEE, 1999.*

[DDO+13] C. Daskalakis, I. Diakonikolas, R. O'Donnell, R.A. Servedio, and L. Tan. Learning Sums of Independent Integer Random Variables. In FOCS*, pages 217–226, 2013.*

[DDS12a] C. Daskalakis, I. Diakonikolas, and R.A. Servedio. Learning k-modal distributions via testing. In SODA*, pages 1371–1385, 2012.*

[DDS12b] C. Daskalakis, I. Diakonikolas, and R.A. Servedio. Learning Poisson Binomial Distributions. In STOC*, pages 709–728, 2012.*

[DDS+13] C. Daskalakis, I. Diakonikolas, R. Servedio, G. Valiant, and P. Valiant. Testing k-modal distributions: Optimal algorithms via reductions. In SODA*, pages 1833–1852, 2013.*

[DG85] Luc Devroye and László Györfi. Nonparametric density estimation: The L B1 S view. *Wiley, 1985.*

[DK16] I. Diakonikolas and D. M. Kane. A new approach for testing properties of discrete distributions. CoRR*, abs/1601.05557, 2016.*

[DKN15a] *I. Diakonikolas, D. M. Kane, and V. Nikishkin. Optimal algorithms and lower bounds for testing closeness of structured distributions. In* 56th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2015*, 2015.*

[DKN15b] *I. Diakonikolas, D. M. Kane, and V. Nikishkin. Testing Identity of Structured Distributions. In* Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015, San Diego, CA, USA, January 4-6, 2015*, 2015.*

[DL01] *L. Devroye and G. Lugosi.* Combinatorial methods in density estimation*. Springer Series in Statistics, Springer, 2001.*

[DR09] *L. D umbgen and K. Rufibach. Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency.* Bernoulli*, 15(1):40–68, 2009.*

[FM99] *Y. Freund and Y. Mansour. Estimating a mixture of two product distributions. In* Proceedings of the twelfth annual conference on Computational learning theory*, pages 53–62. ACM, 1999.*

[FOS05] *J. Feldman, R. O'Donnell, and R. Servedio. Learning mixtures of product distributions over discrete domains. In* Proc. 46th Symposium on Foundations of Computer Science (FOCS)*, pages 501–510, 2005.*

[Fou97] *A.-L. Fougères. Estimation de densités unimodales.* Canadian Journal of Statistics*, 25:375–387, 1997.*

[GR00] *O. Goldreich and D. Ron. On testing expansion in bounded-degree graphs. Technical Report TR00-020, Electronic Colloquium on Computational Complexity, 2000.*

[Gre56] *U. Grenander. On the theory of mortality measurement.* Skand. Aktuarietidskr.*, 39:125–153, 1956.*

[Gro85] *P. Groeneboom. Estimating a monotone density. In* Proc. of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, pages 539–555, 1985.*

[GW09] *F. Gao and J. A. Wellner. On the rate of convergence of the maximum likelihood estimator of a k-monotone density.* Science in China Series A: Mathematics*, 52:1525–1538, 2009.*

[Ham87] *F. R. Hampel. Design, data & analysis. In Colin L. Mallows, editor,* Design, data & analysis*, chapter Design, modelling, and analysis of some biological data sets, pages 93–128. John Wiley & Sons, Inc., New York, NY, USA, 1987.*

[HP76] *D. L. Hanson and G. Pledger. Consistency in concave regression.* The Annals of Statistics*, 4(6):pp. 1038–1050, 1976.*

[JW09] *H. K. Jankowski and J. A. Wellner. Estimation of a discrete monotone density.* Electronic Journal of Statistics*, 3:1567–1605, 2009.*

[KM10] *R. Koenker and I. Mizera. Quasi-concave density estimation.* Ann. Statist.*, 38(5):2998–3027, 2010.*

[KMR+94] *M. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R. Schapire, and L. Sellie. On the learnability of discrete distributions. In* Proceedings of the 26th Symposium on Theory of Computing*, pages 273–282, 1994.*

[KMV10] *A. T. Kalai, A. Moitra, and G. Valiant. Efficiently learning mixtures of two Gaussians. In* STOC*, pages 553–562, 2010.*

[LR05] *E. L. Lehmann and J. P. Romano.* Testing statistical hypotheses*. Springer Texts in Statistics. Springer, 2005.*

[MV10] A. Moitra and G. Valiant. Settling the polynomial learnability of mixtures of Gaussians. In FOCS, pages 93–102, 2010.

[NNV07] E. Tardos N. Nisan, T. Roughgarden, , and V. V. Vazirani. Algorithmic game theory, volume 1. Cambridge University Press Cambridge, 2007.

[NP33] J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 231(694-706):289–337, 1933.

[Pan08] L. Paninski. A coincidence-based test for uniformity given very sparsely-sampled discrete data. IEEE Transactions on Information Theory, 54:4750–4755, 2008.

[Pea00] K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine Series 5, 50(302):157–175, 1900.

[Rao69] B.L.S. Prakasa Rao. Estimation of a unimodal density. Sankhya Ser. A, 31:23–36, 1969.

[Reb05] L. Reboul. Estimation of a function under shape restrictions. Applications to reliability. Ann. Statist., 33(3):1330–1356, 2005.

[RS96] R. Rubinfeld and M. Sudan. Robust characterizations of polynomials with applications to program testing. SIAM J. on Comput., 25:252–271, 1996.

[Sco92] D.W. Scott. Multivariate Density Estimation: Theory, Practice and Visualization. Wiley, New York, 1992.

[Sil86] B. W. Silverman. Density Estimation. Chapman and Hall, London, 1986.

[VV11a] *G. Valiant and P. Valiant. Estimating the unseen: an $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs. In* STOC*, pages 685–694, 2011.*

[VV11b] *G. Valiant and P. Valiant. The power of linear estimators. In* FOCS*, 2011.*

[VV13] *G. Valiant and P. Valiant. Instance-by-instance optimal identity testing.* ECCC*, http://eccc.hpi-web.de/report/2013/111/, 2013.*

[VV14] *G. Valiant and P. Valiant. An automatic inequality prover and instance optimal identity testing. In* FOCS*, 2014.*

[Wal09] *G. Walther. Inference and modeling with log-concave distributions.* Statistical Science*, 24(3):319–327, 2009.*

[Weg70] *E.J. Wegman. Maximum likelihood estimation of a unimodal density. I. and II.* Ann. Math. Statist.*, 41:457–471, 2169–2174, 1970.*

[WWW+05] *X. Wang, M. Woodroofe, M. Walker, M. Mateo, and E. Olszewski. Estimating dark matter distributions.* The Astrophysical Journal*, 626:145–158, 2005.*

Appendix A

Published papers

A.1 Testing Identity of Structured Distributions

Testing Identity of Structured Distributions

Ilias Diakonikolas*
University of Edinburgh
ilias.d@ed.ac.uk.

Daniel M. Kane†
University of California, San Diego
dakane@cs.ucsd.edu.

Vladimir Nikishkin‡
University of Edinburgh
v.nikishkin@sms.ed.ac.uk.

Abstract

We study the question of identity testing for structured distributions. More precisely, given samples from a *structured* distribution q over $[n]$ and an explicit distribution p over $[n]$, we wish to distinguish whether $q = p$ versus q is at least ε-far from p, in L_1 distance. In this work, we present a unified approach that yields new, simple testers, with sample complexity that is information-theoretically optimal, for broad classes of structured distributions, including t-flat distributions, t-modal distributions, log-concave distributions, monotone hazard rate (MHR) distributions, and mixtures thereof.

1 Introduction

How many samples do we need to verify the identity of a distribution? This is arguably *the* single most fundamental question in statistical hypothesis testing [NP33], with Pearson's chi-squared test [Pea00] (and variants thereof) still being the method of choice used in practice. This question has also been extensively studied by the TCS community in the framework of *property testing* [RS96, GGR98]: Given sample access to an unknown distribution q over a finite domain $[n] := \{1, \ldots, n\}$, an explicit distribution p over $[n]$, and a parameter $\varepsilon > 0$, we want to distinguish between the cases that q and p are identical versus ε-far from each other in L_1 norm (statistical distance). Previous work on this problem focused on characterizing the sample size needed to test the identity of an arbitrary distribution of a given support size. After more than a decade of study, this "worst-case" regime is well-understood: there exists a computationally efficient estimator with sample complexity $O(\sqrt{n}/\varepsilon^2)$ [VV14] and a matching information-theoretic lower bound [Pan08].

While it is certainly a significant improvement over naive approaches and is tight in general, the bound of $\Theta(\sqrt{n})$ is still impractical, if the support size n is very large. We emphasize that the aforementioned sample complexity characterizes worst-case instances, and one might hope that drastically better results can be obtained for most natural settings. In contrast to this setting, in which we assume nothing about the structure of the unknown distribution q, in many cases we know a priori that the distribution q in question has some "nice structure". For example, we may have some qualitative information about the density q, e.g., it may be a mixture of a small number of log-concave distributions, or a multimodal distribution with a bounded number of modes. The following question naturally arises: *Can we exploit the underlying structure in order to perform the desired statistical estimation task more efficiently?*

One would optimistically hope for the answer to the above question to be "YES." While this has been confirmed in several cases for the problem of *learning* (see e.g., [DDS12a, DDS12b, DDO+13, CDSS14]), relatively little work has been done for *testing* properties of structured distributions. In this paper, we show that this is indeed the case for the aforementioned problem of identity testing for a broad spectrum of natural and well-studied distribution classes. To describe our results in more detail, we will need some terminology.

Let C be a class of distributions over $[n]$. The problem of *identity testing for* C is the following: Given sample access to an unknown distribution $q \in C$, and an explicit distribution $p \in C$[1], we want to distinguish between the case that $q = p$ versus $\|q - p\|_1 \geq \varepsilon$. We emphasize that the sample complexity of this testing problem depends on the underlying class C, and we believe it is of fundamental interest to obtain efficient algorithms that are *sample optimal* for C. One approach to solve this problem is to learn q up to L_1 distance $\varepsilon/2$ and check that the hypothesis is $\varepsilon/2$-close to p. Thus, the sample complexity of identity testing for C is bounded from above by the sample complexity of *learning* (an

*Supported by EPSRC grant EP/L021749/1, a Marie Curie Career Integration Grant, and a SICSA grant.

†Supported in part by an NSF Postdoctoral Fellowship.

‡Supported by a University of Edinburgh PCD Scholarship.

[1] It is no loss of generality to assume that $p \in C$; otherwise the tester can output "NO" without drawing samples.

arbitrary distribution in) C. It is natural to ask whether a better sample size bound could be achieved for the identity testing problem, since this task is, in some sense, less demanding than the task of learning.

In this work, we provide a comprehensive picture of the sample and computational complexities of identity testing for a broad class of structured distributions. More specifically, we propose a unified framework that yields new, simple, and *provably optimal* identity testers for various structured classes C; see Table 1 for an indicative list of distribution classes to which our framework applies. *Our approach relies on a single unified algorithm that we design, which yields highly efficient identity testers for many shape restricted classes of distributions.*

As an interesting byproduct, we establish that, for various structured classes C, identity testing for C is provably easier than learning. In particular, the sample bounds in the third column of Table 1 from [CDSS14] also apply for *learning* the corresponding class C, and are known to be information-theoretically optimal for the learning problem.

Our main result (see Theorem 2.1 and Proposition 2.1 in Section 2) can be phrased, roughly, as follows: *Let C be a class of univariate distributions such that any pair of distributions $p, q \in C$ have "essentially" at most k crossings, that is, points of the domain where $q - p$ changes its sign. Then, the identity problem for C can be solved with $O(\sqrt{k}/\varepsilon^2)$ samples. Moreover, this bound is information-theoretically optimal.*

By the term "essentially" we mean that a constant fraction of the contribution to $\|q - p\|_1$ is due to a set of k crossings – the actual number of crossings can be arbitrary. For example, if C is the class of t-piecewise constant distributions, it is clear that any two distributions in C have $O(t)$ crossings, which gives us the first line of Table 1. As a more interesting example, consider the class C of log-concave distributions over $[n]$. While the number of crossings between $p, q \in C$ can be $\Omega(n)$, it can be shown (see Lemma 17 in [CDSS14]) that the essential number of crossings is $k = \widetilde{O}(1/\sqrt{\varepsilon})$, which gives us the third line of the table. More generally, we obtain asymptotic improvements over the standard $O(\sqrt{n}/\varepsilon^2)$ bound for any class C such that the essential number of crossings is $k = o(n)$. This condition applies for any class C that can be well-approximated in L_1 distance by piecewise low-degree polynomials (see Corollary 2.1 for a precise statement).

1.1 Related and Prior Work

In this subsection we review the related literature and compare our results with previous work.

Distribution Property Testing The area of distribution property testing, initiated in the TCS community by the work of Batu *et al.* [BFR$^+$00, BFR$^+$13], has developed into a very active research area with intimate connections to information theory, learning and statistics. The paradigmatic algorithmic problem in this area is the following: given sample access to an unknown distribution q over an n-element set, we want to determine whether q has some property or is "far" (in statistical distance or, equivalently, L_1 norm) from any distribution having the property. The overarching goal is to obtain a computationally efficient algorithm that uses as few samples as possible – certainly asymptotically fewer than the support size n, and ideally much less than that. See [GR00, BFR$^+$00, BFF$^+$01, Bat01, BDKR02, BKR04, Pan08, Val11, VV11, DDS$^+$13, ADJ$^+$11, LRR11, ILR12] for a sample of works and [Rub12] for a survey.

One of the first problems studied in this line of work is that of "identity testing against a known distribution": Given samples from an unknown distribution q and an explicitly given distribution p distinguish between the case that $q = p$ versus the case that q is ε-far from p in L_1 norm. The problem of *uniformity testing* – the special case of identity testing when p is the uniform distribution – was first considered by Goldreich and Ron [GR00] who, motivated by a connection to testing expansion in graphs, obtained a uniformity tester using $O(\sqrt{n}/\varepsilon^4)$ samples. Subsequently, Paninski gave the tight bound of $\Theta(\sqrt{n}/\varepsilon^2)$ [Pan08] for this problem. Batu *et al.* [BFF$^+$01] obtained an identity testing algorithm against an arbitrary explicit distribution with sample complexity $\tilde{O}(\sqrt{n}/\varepsilon^4)$. The tight bound of $\Theta(\sqrt{n}/\varepsilon^2)$ for the general identity testing problem was given only recently in [VV14].

Shape Restricted Statistical Estimation The area of inference under shape constraints – that is, inference about a probability distribution under the constraint that its probability density function (pdf) satisfies certain qualitative properties – is a classical topic in statistics starting with the pioneering work of Grenander [Gre56] on monotone distributions (see [BBBB72] for an early book on the topic). Various structural restrictions have been studied in the statistics literature, starting from monotonicity, unimodality, and concavity [Gre56, Bru58, Rao69, Weg70, HP76, Gro85, Bir87a, Bir87b, Fou97, CT04, JW09], and more recently focusing on structural restrictions such as log-concavity and k-monotonicity [BW07, DR09, BRW09, GW09, BW10, KM10].

Shape restricted inference is well-motivated in its own right, and has seen a recent surge of research activity in the statistics community, in part due to the ubiquity of structured distributions in the natural sciences. Such structural constraints on the underlying distributions are sometimes direct consequences of the studied application problem (see e.g., Hampel [Ham87], or Wang *et al.* [WWW$^+$05]), or they are a plausible explanation of the model under investigation (see e.g., [Reb05] and ref-

Class of Distributions over $[n]$	Our upper bound	Previous work
t-piecewise constant	$O(\sqrt{t}/\varepsilon^2)$	$O(t/\varepsilon^2)$ [CDSS14]
t-piecewise degree-d polynomial	$O\left(\sqrt{t(d+1)}/\varepsilon^2\right)$	$O\left(t(d+1)/\varepsilon^2\right)$ [CDSS14]
log-concave	$\widetilde{O}(1/\varepsilon^{9/4})$	$\widetilde{O}(1/\varepsilon^{5/2})$ [CDSS14]
k-mixture of log-concave	$\sqrt{k}\cdot\widetilde{O}(1/\varepsilon^{9/4})$	$\widetilde{O}(k/\varepsilon^{5/2})$ [CDSS14]
t-modal	$O(\sqrt{t\log(n)}/\varepsilon^{5/2})$	$O\left(\sqrt{t\log(n)}/\varepsilon^3 + t^2/\varepsilon^4\right)$ [DDS$^+$13]
k-mixture of t-modal	$O(\sqrt{kt\log(n)}/\varepsilon^{5/2})$	$O\left(\sqrt{kt\log(n)}/\varepsilon^3 + k^2t^2/\varepsilon^4\right)$ [DDS$^+$13]
monotone hazard rate (MHR)	$O(\sqrt{\log(n/\varepsilon)}/\varepsilon^{5/2})$	$O(\log(n/\varepsilon)/\varepsilon^3)$ [CDSS14]
k-mixture of MHR	$O(\sqrt{k\log(n/\varepsilon)}/\varepsilon^{5/2})$	$O(k\log(n/\varepsilon)/\varepsilon^3)$ [CDSS14]

Table 1: Algorithmic results for identity testing of various classes of probability distributions. The second column indicates the sample complexity of our general algorithm applied to the class under consideration. The third column indicates the sample complexity of the best previously known algorithm for the same problem.

erences therein for applications to economics and reliability theory). We also point the reader to the recent survey [Wal09] highlighting the importance of log-concavity in statistical inference. The hope is that, under such structural constraints, the quality of the resulting estimators may dramatically improve, both in terms of sample size and in terms of computational efficiency.

We remark that the statistics literature on the topic has focused primarily on the problem of *density estimation* or learning an unknown structured distribution. That is, given samples from a distribution q promised to belong to some distribution class C, we would like to output a hypothesis distribution that is a good approximation to q. In recent years, there has been a flurry of results in the TCS community on learning structured distributions, with a focus on both sample complexity and computational complexity, see [KMR$^+$94, FOS05, BS10, KMV10, MV10, DDS12a, DDS12b, CDSS13, DDO$^+$13, CDSS14] for some representative works.

Comparison with Prior Work In recent work, Chan, Diakonikolas, Servedio, and Sun [CDSS14] proposed a general approach to *learn* univariate probability distributions that are well approximated by piecewise polynomials. [CDSS14] obtained a computationally efficient and sample near-optimal algorithm to agnostically learn piecewise polynomial distributions, thus obtaining efficient estimators for various classes of structured distributions. For many of the classes C considered in Table 1 the best previously known sample complexity for the identity testing problem for C is identified with the sample complexity of the corresponding learning problem from [CDSS14]. We remark that the results of this paper apply to all classes C considered in [CDSS14], and are in fact more general as our condition (any $p, q \in C$ have a bounded number of "essential" crossings) subsumes the piecewise polynomial condition (see discussion before Corollary 2.1 in Section 2). At the technical level, in contrast to the learning algorithm of [CDSS14], which relies on a combination of linear programming and dynamic programming, our identity tester is simple and combinatorial.

In the context of property testing, Batu, Kumar, and Rubinfeld [BKR04] gave algorithms for the problem of identity testing of unimodal distributions with sample complexity $O(\log^3 n)$. More recently, Daskalakis, Diakonikolas, Servedio, Valiant, and Valiant [DDS$^+$13] generalized this result to t-modal distributions obtaining an identity tester with sample complexity $O(\sqrt{t\log(n)}/\varepsilon^3 + t^2/\varepsilon^4)$. We remark that for the class of t-modal distributions our approach yields an identity tester with sample complexity $O(\sqrt{t\log(n)}/\varepsilon^{5/2})$, matching the lower bound of [DDS$^+$13]. Moreover, our work yields sample optimal identity testing algorithms not only for t-modal distributions, but for a broad spectrum of structured distributions via a unified approach.

It should be emphasized that the main ideas underlying this paper are very different from those of [DDS$^+$13]. The algorithm of [DDS$^+$13] is based on the fact from [Bir87a] that any t-modal distribution is ε-close in L_1 norm to a piecewise constant distribution with $k = O(t \cdot \log(n)/\varepsilon)$ intervals. Hence, if the location and the width of these k "flat" intervals were known in advance, the problem would be easy: The algorithm could just test identity between the "reduced" distributions supported on these k intervals, thus obtaining the optimal sample complexity of $O(\sqrt{k}/\varepsilon^2) = O(\sqrt{t\log(n)}/\varepsilon^{5/2})$. To circumvent the problem that this decomposition is not known a priori, [DDS$^+$13] start by drawing samples from the unknown distribution q to construct such a decomposition. There are two caveats with this strategy: First, the number of samples used to achieve this is $\Omega(t^2)$ and the number of intervals of the constructed decomposition is significantly larger than k, namely $k' = \Omega(k/\varepsilon)$. As a consequence, the sample complexity of identity testing for the reduced distributions on support k' is $\Omega(\sqrt{k'}/\varepsilon^2) =$

$\Omega(\sqrt{t\log(n)}/\varepsilon^3)$.

In conclusion, the approach of [DDS+13] involves constructing an *adaptive* interval decomposition of the domain followed by a single application of an identity tester to the reduced distributions over those intervals. At a high-level our novel approach works as follows: We consider *several oblivious* interval decompositions of the domain (i.e., without drawing any samples from q) and apply a "reduced" identity tester for *each* such decomposition. While it may seem surprising that such an approach can be optimal, our algorithm and its analysis exploit a certain strong property of uniformity testers, namely their performance guarantee with respect to the L_2 *norm*. See Section 2 for a detailed explanation of our techniques.

Finally, we comment on the relation of this work to the recent paper [VV14]. In [VV14], Valiant and Valiant study the sample complexity of the identity testing problem as a function of the explicit distribution. In particular, [VV14] makes no assumptions about the structure of the unknown distribution q, and characterizes the sample complexity of the identity testing problem as a function of the known distribution p. The current work provides a unified framework to exploit structural properties of the unknown distribution q, and yields sample optimal identity testers for various shape restrictions. Hence, the results of this paper are orthogonal to the results of [VV14].

2 Our Results and Techniques

2.1 Basic Definitions We start with some notation that will be used throughout this paper. We consider discrete probability distributions over $[n] := \{1, \ldots, n\}$, which are given by probability density functions $p : [n] \to [0, 1]$ such that $\sum_{i=1}^n p_i = 1$, where p_i is the probability of element i in distribution p. By abuse of notation, we will sometimes use p to denote the distribution with density function p_i. We emphasize that we view the domain $[n]$ as an ordered set. Throughout this paper we will be interested in structured distribution families that respect this ordering.

The L_1 (resp. L_2) norm of a distribution is identified with the L_1 (resp. L_2) norm of the corresponding density function, i.e., $\|p\|_1 = \sum_{i=1}^n |p_i|$ and $\|p\|_2 = \sqrt{\sum_{i=1}^n p_i^2}$. The L_1 (resp. L_2) distance between distributions p and q is defined as the L_1 (resp. L_2) norm of the vector of their difference, i.e., $\|p - q\|_1 = \sum_{i=1}^n |p_i - q_i|$ and $\|p - q\|_2 = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$. We will denote by U_n the uniform distribution over $[n]$.

Interval partitions and $\mathcal{A}_k$-distance Fix a partition of $[n]$ into disjoint intervals $\mathcal{I} := (I_i)_{i=1}^\ell$. For such a collection $\mathcal{I}$ we will denote its cardinality by $|\mathcal{I}|$, i.e., $|\mathcal{I}| = \ell$. For an interval $J \subseteq [n]$, we denote by $|J|$ its cardinality or length, i.e., if $J = [a, b]$, with $a \leq b \in [n]$, then $|J| = b - a + 1$. The *reduced distribution* $p_r^{\mathcal{I}}$ corresponding to p and $\mathcal{I}$ is the distribution over $[\ell]$ that assigns the ith "point" the mass that p assigns to the interval I_i; i.e., for $i \in [\ell]$, $p_r^{\mathcal{I}}(i) = p(I_i)$.

We now define a distance metric between distributions that will be crucial for this paper. Let $\mathfrak{J}_k$ be the collection of all partitions of $[n]$ into k intervals, i.e., $\mathcal{I} \in \mathfrak{J}_k$ if and only if $\mathcal{I} = (I_i)_{i=1}^k$ is a partition of $[n]$ into intervals $I_1, \ldots, I_k$. For $p, q : [n] \to [0, 1]$ and $k \in \mathbb{Z}_+$, $2 \leq k \leq n$, we define the $\mathcal{A}_k$-distance between p and q by

$$\begin{aligned} \|p - q\|_{\mathcal{A}_k} &\stackrel{\text{def}}{=} \max_{\mathcal{I}=(I_i)_{i=1}^k \in \mathfrak{J}_k} \sum_{i=1}^k |p(I_i) - q(I_i)| \\ &= \max_{\mathcal{I} \in \mathfrak{J}_k} \|p_r^{\mathcal{I}} - q_r^{\mathcal{I}}\|_1. \end{aligned}$$

We remark that the $\mathcal{A}_k$-distance between distributions[2] is well-studied in probability theory and statistics. Note that for any pair of distributions $p, q : [n] \to [0, 1]$, and any $k \in \mathbb{Z}_+$ with $2 \leq k \leq n$, we have that $\|p - q\|_{\mathcal{A}_k} \leq \|p - q\|_1$, and the two metrics are identical for $k = n$. Also note that $\|p - q\|_{\mathcal{A}_2} = 2 \cdot d_{\mathrm{K}}(p, q)$, where d_{K} is the Kolmogorov metric (i.e., the L_∞ distance between the CDF's).

Discussion The well-known *Vapnik-Chervonenkis (VC) inequality* (see e.g., [DL01, p.31]) provides the information-theoretically optimal sample size to *learn* an arbitrary distribution q over $[n]$ in this metric. In particular, it implies that $m = \Omega(k/\varepsilon^2)$ iid draws from q suffice in order to learn q within $\mathcal{A}_k$-distance ε (with probability at least $9/10$). This fact has recently proved useful in the context of learning structured distributions: By exploiting this fact, Chan, Diakonikolas, Servedio, and Sun [CDSS14] recently obtained computationally efficient and near-sample optimal algorithms for learning various classes of structured distributions *with respect to the L_1 distance*.

It is thus natural to ask the following question: What is the sample complexity of *testing* properties of distributions with respect to the $\mathcal{A}_k$-distance? Can we use property testing algorithms in this metric to obtain sample-optimal testing algorithms for interesting classes of structured distributions *with respect to the L_1 distance*? In this work we answer both questions in the affirmative for the problem of identity testing.

[2] We note that the definition of $\mathcal{A}_k$-distance in this work is slightly different than [DL01, CDSS14], but is easily seen to be essentially equivalent. In particular, [CDSS14] considers the quantity $\max_{S \in \mathcal{S}_k} |p(S) - q(S)|$, where $\mathcal{S}_k$ is the collection of all unions of at most k intervals in $[n]$. It is a simple exercise to verify that $\|p - q\|_{\mathcal{A}_k} \leq 2 \cdot \max_{S \in \mathcal{S}_k} |p(S) - q(S)| = \|p - q\|_{\mathcal{A}_{2k+1}}$, which implies that the two definitions are equivalent up to constant factors for the purpose of both upper and lower bounds.

2.2 Our Results Our main result is an optimal algorithm for the identity testing problem under the $\mathcal{A}_k$-distance metric:

THEOREM 2.1. (MAIN) *Given $\varepsilon > 0$, an integer k with $2 \leq k \leq n$, sample access to a distribution q over $[n]$, and an explicit distribution p over $[n]$, there is a computationally efficient algorithm which uses $O(\sqrt{k}/\varepsilon^2)$ samples from q, and with probability at least $2/3$ distinguishes whether $q = p$ versus $\|q - p\|_{\mathcal{A}_k} \geq \varepsilon$. Additionally, $\Omega(\sqrt{k}/\varepsilon^2)$ samples are information-theoretically necessary.*

The information-theoretic sample lower bound of $\Omega(\sqrt{k}/\varepsilon^2)$ can be easily deduced from the known lower bound of $\Omega(\sqrt{n}/\varepsilon^2)$ for uniformity testing over $[n]$ under the L_1 norm [Pan08]. Indeed, if the underlying distribution q over $[n]$ is piecewise constant with k pieces, and p is the uniform distribution over $[n]$, we have $\|q-p\|_{\mathcal{A}_k} = \|q - p\|_1$. Hence, our $\mathcal{A}_k$-uniformity testing problem in this case is at least as hard as L_1-uniformity testing over support of size k.

The proof of Theorem 2.1 proceeds in two stages: In the first stage, we reduce the $\mathcal{A}_k$ identity testing problem to $\mathcal{A}_k$ uniformity testing without incurring any loss in the sample complexity. In the second stage, we use an optimal L_2 *uniformity* tester as a black-box to obtain an $O(\sqrt{k}/\varepsilon^2)$ sample algorithm for $\mathcal{A}_k$ uniformity testing. We remark that the L_2 uniformity tester is not applied to the distribution q directly, but to a sequence of reduced distributions $q_r^{\mathcal{I}}$, for an appropriate collection of interval partitions $\mathcal{I}$. See Section 2.3 for a detailed intuitive explanation of the proof.

We remark that an application of Theorem 2.1 for $k = n$, yields a sample optimal L_1 identity tester (for an arbitrary distribution q), giving a new algorithm matching the recent tight upper bound in [VV14]. Our new L_1 identity tester is arguable simpler and more intuitive, as it only uses an L_2 uniformity tester in a black-box manner.

We show that Theorem 2.1 has a wide range of applications to the problem of L_1 identity testing for various classes of natural and well-studied structured distributions. At a high level, the main message of this work is that the $\mathcal{A}_k$ distance can be used to characterize the sample complexity of L_1 identity testing for broad classes of structured distributions. The following simple proposition underlies our approach:

PROPOSITION 2.1. *For a distribution class C over $[n]$ and $\varepsilon > 0$, let $k = k(C, \varepsilon)$ be the smallest integer such that for any $f_1, f_2 \in C$ it holds that $\|f_1 - f_2\|_1 \leq \|f_1 - f_2\|_{\mathcal{A}_k} + \varepsilon/2$. Then there exists an L_1 identity testing algorithm for C using $O(\sqrt{k}/\varepsilon^2)$ samples.*

The proof of the proposition is straightforward: Given sample access to $q \in C$ and an explicit description of $p \in C$, we apply the $\mathcal{A}_k$-identity testing algorithm of Theorem 2.1 for the value of k in the statement of the proposition, and error $\varepsilon' = \varepsilon/2$. If $q = p$, the algorithm will output "YES" with probability at least $2/3$. If $\|q-p\|_1 \geq \varepsilon$, then by the condition of Proposition 2.1 we have that $\|q - p\|_{\mathcal{A}_k} \geq \varepsilon'$, and the algorithm will output "NO" with probability at least $2/3$. Hence, as long as the underlying distribution satisfies the condition of Proposition 2.1 for a value of $k = o(n)$, Theorem 2.1 yields an asymptotic improvement over the sample complexity of $\Theta(\sqrt{n}/\varepsilon^2)$.

We remark that the value of k in the proposition is a natural complexity measure for the difference between two probability density functions in the class C. It follows from the definition of the $\mathcal{A}_k$ distance that this value corresponds to the number of "essential" crossings between f_1 and f_2 – i.e., the number of crossings between the functions f_1 and f_2 that significantly affect their L_1 distance. Intuitively, the number of essential crossings – as opposed to the domain size – is, in some sense, the "right" parameter to characterize the sample complexity of L_1 identity testing for C. As we explain below, the upper bound implied by the above proposition is information-theoretically optimal for a wide range of structured distribution classes C.

More specifically, our framework can be applied to all structured distribution classes C that can be well-approximated in L_1 distance by *piecewise low-degree polynomials*. We say that a distribution p over $[n]$ is *t-piecewise degree-d* if there exists a partition of $[n]$ into t intervals such that p is a (discrete) degree-d polynomial within each interval. Let $\mathcal{P}_{t,d}$ denote the class of all t-piecewise degree-d distributions over $[n]$. We say that a distribution class C is *ε-close in L_1* to $\mathcal{P}_{t,d}$ if for any $f \in C$ there exists $p \in \mathcal{P}_{t,d}$ such that $\|f - p\|_1 \leq \varepsilon$. It is easy to see that any pair of distributions $p, q \in \mathcal{P}_{t,d}$ have at most $2t(d + 1)$ crossings, which implies that $\|p - q\|_{A_k} = \|p - q\|_1$, for $k = 2t(d + 1)$ (see e.g., Proposition 6 in [CDSS14]). We therefore obtain the following:

COROLLARY 2.1. *Let C be a distribution class over $[n]$ and $\varepsilon > 0$. Consider parameters $t = t(C, \varepsilon)$ and $d = d(C, \varepsilon)$ such that C is $\varepsilon/4$-close in L_1 to $\mathcal{P}_{t,d}$. Then there exists an L_1 identity testing algorithm for C using $O(\sqrt{t(d + 1)}/\varepsilon^2)$ samples.*

Note that any pair of values (t, d) satisfying the condition above suffices for the conclusion of the corollary. Since our goal is to minimize the sample complexity, for a given class C, we would like to apply the corollary for values t and d satisfying the above condition and are such that the product $t(d + 1)$ is minimized. The appropriate choice of these values is crucial, and is based on properties of the underlying distribution family. Observe that the sample

bound of $O(\sqrt{t(d+1)}/\varepsilon^2)$ is tight in general, as follows by selecting $C = \mathcal{P}_{t,d}$. This can be deduced from the general lower bound of $\Omega(\sqrt{n}/\varepsilon^2)$ for uniformity testing, and the fact that for $n = t(d+1)$, any distribution over support $[n]$ can be expressed as a t-piecewise degree-d distribution.

The concrete testing results of Table 1 are obtained from Corollary 2.1 by using known existential approximation theorems [Bir87a, CDSS13, CDSS14] for the corresponding structured distribution classes. In particular, we obtain efficient identity testers, in most cases with provably optimal sample complexity, for all the structured distribution classes studied in [CDSS13, CDSS14] in the context of learning. Perhaps surprisingly, our upper bounds are tight not only for the class of piecewise polynomials, but also for the specific shape restricted classes of Table 1. The corresponding lower bounds for specific classes are either known from previous work (as e.g., in the case of t-modal distributions [DDS$^+$13]) or can be obtained using standard constructions.

Finally, we remark that the results of this paper can be appropriately generalized to the setting of testing the identity of continuous distributions over the real line. More specifically, Theorem 2.1 also holds for probability distributions over $\mathbb{R}$. (The only additional assumption required is that the explicitly given continuous pdf p can be efficiently integrated up to any additive accuracy.) In fact, the proof for the discrete setting extends almost verbatim to the continuous setting with minor modifications. It is easy to see that both Proposition 2.1 and Corollary 2.1 hold for the continuous setting as well.

2.3 Our Techniques We now provide a detailed intuitive explanation of the ideas that lead to our main result, Theorem 2.1. Given sample access to a distribution q and an explicit distribution p, we want to test whether $q = p$ versus $\|q - p\|_{\mathcal{A}_k} \geq \varepsilon$. By definition we have that $\|q - p\|_{\mathcal{A}_k} = \max_{\mathcal{I}} \|q_r^{\mathcal{I}} - p_r^{\mathcal{I}}\|_1$. So, if the "optimal" partition $\mathcal{J}^* = (J_i^*)_{j=1}^k$ maximizing this expression was known a priori, the problem would be easy: Our algorithm could then consider the reduced distributions $q_r^{\mathcal{J}^*}$ and $p_r^{\mathcal{J}^*}$, which are supported on sets of size k, and call a standard L_1-identity tester to decide whether $q_r^{\mathcal{J}^*} = p_r^{\mathcal{J}^*}$ versus $\|q_r^{\mathcal{J}^*} - p_r^{\mathcal{J}^*}\|_1 \geq \varepsilon$. (Note that for any given partition $\mathcal{I}$ of $[n]$ into intervals and any distribution q, given sample access to q one can simulate sample access to the reduced distribution $q_r^{\mathcal{I}}$.) The difficulty, of course, is that the optimal k-partition is *not fixed*, as it depends on the unknown distribution q, thus it is not available to the algorithm. Hence, a more refined approach is necessary.

Our starting point is a new, simple reduction of the general problem of identity testing to its special case of uniformity testing. The main idea of the reduction is to appropriately "stretch" the domain size, using the explicit distribution p, in order to transform the identity testing problem between q and p into a uniformity testing problem for a (different) distribution q' (that depends on q and p). To show correctness of this reduction we need to show that it preserves the $\mathcal{A}_k$ distance, and that we can sample from q' given samples from q.

We now proceed with the details. Since p is given explicitly in the input, we assume for simplicity that each p_i is a rational number, hence there exists some (potentially large) $N \in \mathbb{Z}_+$ such that $p_i = \alpha_i/N$, where $\alpha_i \in \mathbb{Z}_+$ and $\sum_{i=1}^n \alpha_i = N$.[3] Given sample access to q and an explicit p over $[n]$, we construct an instance of the uniformity testing problem as follows: Let p' be the uniform distribution over $[N]$ and let q' be the distribution over $[N]$ obtained from q by subdividing the probability mass of q_i, $i \in [n]$, equally among α_i new consecutive points. It is clear that this reduction preserves the $\mathcal{A}_k$ distance, i.e., $\|q - p\|_{\mathcal{A}_k} = \|q' - p'\|_{\mathcal{A}_k}$. The only remaining task is to show how to simulate sample access to q', given samples from q. Given a sample i from q, our sample for q' is selected uniformly at random from the corresponding set of α_i many new points. Hence, we have reduced the problem of identity testing between q and p in $\mathcal{A}_k$ distance, to the problem of uniformity testing of q' in $\mathcal{A}_k$ distance. Note that this reduction is also computationally efficient, as it only requires $O(n)$ precomputation to specify the new intervals.

For the rest of this section, we focus on the problem of $\mathcal{A}_k$ uniformity testing. For notational convenience, we will use q to denote the unknown distribution and p to denote the uniform distribution over $[n]$. The rough idea is to consider an appropriate collection of interval partitions of $[n]$ and call a standard L_1-uniformity tester for each of these partitions. To make such an approach work and give us a *sample optimal* algorithm for our $\mathcal{A}_k$-uniformity testing problem we need to use a subtle and strong property of uniformity testing, namely its performance guarantee under the L_2 norm. We elaborate on this point below.

For any partition $\mathcal{I}$ of $[n]$ into k intervals by definition we have that $\|q_r^{\mathcal{I}} - p_r^{\mathcal{I}}\|_1 \leq \|q - p\|_{\mathcal{A}_k}$. Therefore, if $q = p$, we will also have $q_r^{\mathcal{I}} = p_r^{\mathcal{I}}$. The issue is that $\|q_r^{\mathcal{I}} - p_r^{\mathcal{I}}\|_1$ can be much smaller than $\|q - p\|_{\mathcal{A}_k}$; in fact, it is not difficult to construct examples where $\|q-p\|_{\mathcal{A}_k} = \Omega(1)$ and $\|q_r^{\mathcal{I}} - p_r^{\mathcal{I}}\|_1 = 0$. In particular, it is possible for the points where q is larger than p, and where it is smaller than p to cancel each other out within each interval in the partition, thus making the partition useless

[3] We remark that this assumption is not necessary: For the case of irrational p_i's we can approximate them by rational numbers $\widetilde{p}_i$ up to sufficient accuracy and proceed with the approximate distribution $\widetilde{p}$. This approximation step does not preserve perfect completeness; however, we point out that our testers have some mild robustness in the completeness case, which suffices for all the arguments to go through.

for distinguishing q from p. In other words, if the partition $\mathcal{I}$ is not "good", we may not be able to detect any existing discrepancy. A simple, but suboptimal, way to circumvent this issue is to consider a partition $\mathcal{I}'$ of $[n]$ into $k' = \Theta(k/\varepsilon)$ intervals of the same length. Note that each such interval will have probability mass $1/k' = \Theta(\varepsilon/k)$ under the uniform distribution p. If the constant in the big-Θ is appropriately selected, say $k' = 10k/\varepsilon$, it is not hard to show that $\|q_r^{\mathcal{I}'} - p_r^{\mathcal{I}'}\|_1 \geq \|q - p\|_{\mathcal{A}_k} - \varepsilon/2$; hence, we will necessarily detect a large discrepancy for the reduced distribution. By applying the optimal L_1 uniformity tester this approach will require $\Omega(\sqrt{k'}/\varepsilon^2) = \Omega(\sqrt{k}/\varepsilon^{2.5})$ samples.

A key tool that is essential in our analysis is a strong property of uniformity testing. An optimal L_1 uniformity tester for q can distinguish between the uniform distribution and the case that $\|q - p\|_1 \geq \varepsilon$ using $O(\sqrt{n}/\varepsilon^2)$ samples. However, a stronger guarantee is possible: With the *same* sample size, we can distinguish the uniform distribution from the case that $\|q-p\|_2 \geq \varepsilon/\sqrt{n}$. We emphasize that such a strong L_2 guarantee is specific to uniformity testing, and is provably not possible for the general problem of identity testing. In previous work, Goldreich and Ron [GR00] gave such an L_2 guarantee for uniformity testing, but their algorithm uses $O(\sqrt{n}/\varepsilon^4)$ samples. Paninski's $O(\sqrt{n}/\varepsilon^2)$ uniformity tester works for the L_1 norm, and it is not known whether it achieves the desired L_2 property. As one of our main tools we show the following L_2 guarantee, which is optimal as a function of n and ε:

THEOREM 2.2. *Given $0 < \varepsilon, \delta < 1$ and sample access to a distribution q over $[n]$, there is an algorithm* Test-Uniformity-$L_2(q, n, \varepsilon, \delta)$ *which uses $m = O\left((\sqrt{n}/\varepsilon^2) \cdot \log(1/\delta)\right)$ samples from q, runs in time linear in its sample size, and with probability at least $1 - \delta$ distinguishes whether $q = U_n$ versus $\|p - q\|_2 \geq \varepsilon/\sqrt{n}$.*

To prove Theorem 2.2 we show that a variant of Pearson's chi-squared test [Pea00] – which can be viewed as a special case of the recent "chi-square type" testers in [CDVV14, VV14] – has the desired L_2 guarantee. While this tester has been (implicitly) studied in [CDVV14, VV14], and it is known to be sample optimal with respect to the L_1 norm, it has not been previously analyzed for the L_2 norm. The novelty of Theorem 2.2 lies in the tight analysis of the algorithm under the L_2 distance, and is presented in Appendix A.

Armed with Theorem 2.2 we proceed as follows: We consider a set of $j_0 = O(\log(1/\varepsilon))$ different partitions of the domain $[n]$ into intervals. For $0 \leq j < j_0$ the partition $\mathcal{I}^{(j)}$ consists of $\ell_j \stackrel{\text{def}}{=} |\mathcal{I}^{(j)}| = k \cdot 2^j$ many intervals $I_i^{(j)}$, $i \in [\ell_j]$, i.e., $\mathcal{I}^{(j)} = (I_i^{(j)})_{i=1}^{\ell_j}$. For a fixed value of j, all intervals in $\mathcal{I}^{(j)}$ have the same length, or equivalently, the same probability mass *under the uniform distribution*. Then, for any fixed $j \in [j_0]$, we have $p(I_i^{(j)}) = 1/(k \cdot 2^j)$ for all $i \in [\ell_j]$. (Observe that, by our aforementioned reduction to the uniform case, we may assume that the domain size n is a multiple of $k2^{j_0}$, and thus that it is possible to evenly divide into such intervals of the same length).

Note that if $q = p$, then for all $0 \leq j < j_0$, it holds $q_r^{\mathcal{I}^{(j)}} = p_r^{\mathcal{I}^{(j)}}$. Recalling that all intervals in $\mathcal{I}^{(j)}$ have the same probability mass *under* p, it follows that $p_r^{\mathcal{I}^{(j)}} = U_{\ell_j}$, i.e., $p_r^{\mathcal{I}^{(j)}}$ is the uniform distribution over its support. So, if $q = p$, for any partition we have $q_r^{\mathcal{I}^{(j)}} = U_{\ell_j}$. Our main structural result (Lemma 3.1) is a robust inverse lemma: *If q is far from uniform in $\mathcal{A}_k$ distance then, for at least one of the partitions $\mathcal{I}^{(j)}$, the reduced distribution $q_r^{\mathcal{I}^{(j)}}$ will be far from uniform in L_2 distance.* The quantitative version of this statement is quite subtle. In particular, we start from the assumption of being ε-far in $\mathcal{A}_k$ distance and can only deduce "far" in L_2 distance. This is absolutely critical for us to be able to obtain the optimal sample complexity.

The key insight for the analysis comes from noting that the optimal partition separating q from p in $\mathcal{A}_k$ distance cannot have too many parts. Thus, if the "highs" and "lows" cancel out over some small intervals, they must be very large in order to compensate for the fact that they are relatively narrow. Therefore, when p and q differ on a smaller scale, their L_2 discrepancy will be greater, and this compensates for the fact that the partition detecting this discrepancy will need to have more intervals in it.

In Section 3 we present our sample optimal uniformity tester under the $\mathcal{A}_k$ distance, thereby establishing Theorem 2.1.

3 Testing Uniformity under the $\mathcal{A}_k$-norm

Algorithm Test-Uniformity-$\mathcal{A}_k(q, n, \varepsilon)$
Input: sample access to a distribution q over $[n]$, $k \in \mathbb{Z}_+$ with $2 \leq k \leq n$, and $\varepsilon > 0$.
Output: "YES" if $q = U_n$; "NO" if $\|q - U_n\|_{\mathcal{A}_k} \geq \varepsilon$.

1. Draw a sample S of size $m = O(\sqrt{k}/\varepsilon^2)$ from q.
2. Fix $j_0 \in \mathbb{Z}_+$ such that $j_0 \stackrel{\text{def}}{=} \lceil \log_2(1/\varepsilon) \rceil + O(1)$. Consider the collection $\{\mathcal{I}^{(j)}\}_{j=0}^{j_0-1}$ of j_0 partitions of $[n]$ into intervals; the partition $\mathcal{I}^{(j)} = (I_i^{(j)})_{i=1}^{\ell_j}$ consists of $\ell_j = k \cdot 2^j$ many intervals with $p(I_i^{(j)}) = 1/(k \cdot 2^j)$, where $p \stackrel{\text{def}}{=} U_n$.
3. For $j = 0, 1, \ldots, j_0 - 1$:
   (a) Consider the reduced distributions $q_r^{\mathcal{I}^{(j)}}$ and $p_r^{\mathcal{I}^{(j)}} \equiv U_{\ell_j}$. Use the sample S to simulate samples to $q_r^{\mathcal{I}^{(j)}}$.
   (b) Run Test-Uniformity-$L_2(q_r^{\mathcal{I}^{(j)}}, \ell_j, \varepsilon_j, \delta_j)$ for $\varepsilon_j = C \cdot \varepsilon \cdot 2^{3j/8}$ for $C > 0$ a sufficiently small constant and $\delta_j = 2^{-j}/6$, i.e., test whether $q_r^{\mathcal{I}^{(j)}} = U_{\ell_j}$ versus $\|q_r^{\mathcal{I}^{(j)}} - U_{\ell_j}\|_2 > \gamma_j \stackrel{\text{def}}{=} \varepsilon_j/\sqrt{\ell_j}$.
4. If all the testers in Step 3(b) output "YES", then output "YES"; otherwise output "NO".

PROPOSITION 3.1. *The algorithm* Test-Uniformity-$\mathcal{A}_k(q, n, \varepsilon)$*, on input a sample of size* $m = O(\sqrt{k}/\varepsilon^2)$ *drawn from a distribution* q *over* $[n]$*,* $\varepsilon > 0$ *and an integer* k *with* $2 \leq k \leq n$*, correctly distinguishes the case that* $q = U_n$ *from the case that* $\|q - U_n\|_{\mathcal{A}_k} \geq \varepsilon$*, with probability at least* $2/3$*.*

Proof. First, it is straightforward to verify the claimed sample complexity, as the algorithm only draws samples in Step 1. Note that the algorithm uses the same set of samples S for all testers in Step 3(b). By Theorem 2.2, the tester Test-Uniformity-$L_2(q_r^{\mathcal{I}^{(j)}}, \ell_j, \varepsilon_j, \delta_j)$, on input a set of $m_j = O((\sqrt{\ell_j}/\varepsilon_j^2) \cdot \log(1/\delta_j))$ samples from $q_r^{\mathcal{I}^{(j)}}$ distinguishes the case that $q_r^{\mathcal{I}^{(j)}} = U_{\ell_j}$ from the case that $\|q_r^{\mathcal{I}^{(j)}} - U_{\ell_j}\|_2 \geq \gamma_j \stackrel{\text{def}}{=} \varepsilon_j/\sqrt{\ell_j}$ with probability at least $1 - \delta_j$. From our choice of parameters it can be verified that $\max_j m_j \leq m = O(\sqrt{k}/\varepsilon^2)$, hence we can use the same sample S as input to these testers for all $0 \leq j \leq j_0 - 1$. In fact, it is easy to see that $\sum_{j=0}^{j_0-1} m_j = O(m)$, which implies that the overall algorithm runs in sample-linear time. Since each tester in Step 3(b) has error probability δ_j, by a union bound over all $j \in \{0, \ldots, j_0 - 1\}$, the total error probability is at most $\sum_{j=0}^{j_0-1} \delta_j \leq (1/6) \cdot \sum_{j=0}^{\infty} 2^{-j} = 1/3$. Therefore, with probability at least $2/3$ all the testers in Step 3(b) succeed. We will henceforth condition on this "good" event, and establish the completeness and soundness properties of the overall algorithm under this conditioning.

We start by establishing completeness. If $q = p = U_n$, then for any partition $\mathcal{I}^{(j)}$, $0 \leq j \leq j_0 - 1$, we have that $q_r^{\mathcal{I}^{(j)}} = p_r^{\mathcal{I}^{(j)}} = U_{\ell_j}$. By our aforementioned conditioning, all testers in Step 3(b) will output "YES", hence the overall algorithm will also output "YES", as desired.

We now proceed to establish the soundness of our algorithm. Assuming that $\|q - p\|_{\mathcal{A}_k} \geq \varepsilon$, we want to show that the algorithm Test-Uniformity-$\mathcal{A}_k(q, n, \varepsilon)$ outputs "NO" with probability at least $2/3$. Towards this end, we prove the following structural lemma:

LEMMA 3.1. *There exists a constant* $C > 0$ *such that the following holds: If* $\|q - p\|_{\mathcal{A}_k} \geq \varepsilon$*, there exists* $j \in \mathbb{Z}_+$ *with* $0 \leq j \leq j_0 - 1$ *such that* $\|q_r^{\mathcal{I}^{(j)}} - U_{\ell_j}\|_2^2 \geq \gamma_j^2 \stackrel{\text{def}}{=} \varepsilon_j^2/\ell_j = C^2 \cdot (\varepsilon^2/k) \cdot 2^{-j/4}$.

Given the lemma, the soundness property of our algorithm follows easily. Indeed, since all testers Test-Uniformity-$L_2(q_r^{\mathcal{I}^{(j)}}, \ell_j, \varepsilon_j, \delta_j)$ of Step 3(b) are successful by our conditioning, Lemma 3.1 implies that at least one of them outputs "NO", hence the overall algorithm will output "NO".

The proof of Lemma 3.1 in its full generality is quite technical. For the sake of the intuition, in the following subsection (Section 3.1) we provide a proof of the lemma for the important special case that the unknown distribution q is promised to be k-*flat*, i.e., piecewise constant with k pieces. This setting captures many of the core ideas and, at the same time, avoids some of the necessary technical difficulties of the general case. Finally, in Section 3.2 we present our proof for the general case.

3.1 Proof of Structural Lemma: k-flat Case

For this special case we will prove the lemma for $C = 1/80$. Since q is k-flat there exists a partition $\mathcal{I}^* = (I_j^*)_{j=1}^k$ of $[n]$ into k intervals so that q is constant within each such interval. This in particular implies that $\|q - p\|_{\mathcal{A}_k} = \|q - p\|_1$, where $p \stackrel{\text{def}}{=} U_n$. For $J \in \mathcal{I}^*$ let us denote by q_J the value of q within interval J, that is, for all $j \in [k]$ and $i \in I_j^*$ we have $q_i = q_{I_j^*}$. For notational convenience, we sometimes use p_J to denote the value of $p = U_n$ within interval J. By assumption we have that $\|q - p\|_1 = \sum_{j=1}^k |I_j^*| \cdot |q_{I_j^*} - 1/n| \geq \varepsilon$.

Throughout the proof, we work with intervals $I_j^* \in \mathcal{I}^*$ such that $q_{I_j^*} < 1/n$. We will henceforth refer to such intervals as *troughs* and will denote by $\mathcal{T} \subseteq [k]$ the

corresponding set of indices, i.e.,

$$\mathcal{T} = \{j \in [k] \mid q_{I_j^*} < 1/n\}.$$

For each trough $J \in \{I_j^*\}_{j\in\mathcal{T}}$ we define its *depth* as

$$\mathbf{depth}(J) = (p_J - q_J)/p_J = n \cdot (1/n - q_J)$$

and its *width* as

$$\mathbf{width}(J) = p(J) = (1/n) \cdot |J|.$$

Note that the width of J is identified with the probability mass that the uniform distribution assigns to it. The *discrepancy* of a trough J is defined by

$$\mathbf{Discr}(J) = \mathbf{depth}(J) \cdot \mathbf{width}(J) = |J| \cdot (1/n - q_J)\,,$$

and corresponds to the contribution of J to the L_1 distance between q and p.

It follows from Scheffe's identity that half of the contribution to $\|q-p\|_1$ comes from troughs, namely

$$\|q-p\|_1^{\mathcal{T}} \stackrel{\text{def}}{=} \sum_{j\in\mathcal{T}} \mathbf{Discr}(I_j^*) = (1/2) \cdot \|q-p\|_1 \geq \varepsilon/2.$$

An important observation is that we may assume that all troughs have *width* at most $1/k$ at the cost of potentially doubling the total number of intervals. Indeed, it is easy to see that we can artificially subdivide "wider" troughs so that each new trough has width at most $1/k$. This process comes at the expense of at most doubling the number of troughs. Let us denote by $\{\widetilde{I}_j\}_{j\in\mathcal{T}'}$ this set of (new) troughs, where $|\mathcal{T}'| \leq 2k$ and each $\widetilde{I}_j$ is a subset of some I_i^*, $i \in \mathcal{T}$. We will henceforth deal with the set of troughs $\{\widetilde{I}_j\}_{j\in\mathcal{T}'}$ each of width at most $1/k$. By construction, it is clear that

$$\|q-p\|_1^{\mathcal{T}'} \stackrel{\text{def}}{=} \sum_{j\in\mathcal{T}'} \mathbf{Discr}(\widetilde{I}_j) = \|q-p\|_1^{\mathcal{T}} \geq \varepsilon/2. \tag{3.1}$$

At this point we note that we can essentially ignore troughs $J \in \{\widetilde{I}_j\}_{j\in\mathcal{T}'}$ with small discrepancy. Indeed, the total contribution of intervals $J \in \{\widetilde{I}_j\}_{j\in\mathcal{T}'}$ with $\mathbf{Discr}(J) \leq \varepsilon/20k$ to the LHS of (3.1) is at most $|\mathcal{T}'| \cdot (\varepsilon/20k) \leq 2k \cdot (\varepsilon/20k) = \varepsilon/10$. Let $\mathcal{T}^*$ be the subset of $\mathcal{T}'$ corresponding to troughs with discrepancy at least $\varepsilon/20k$, i.e., $j \in \mathcal{T}^*$ if and only if $j \in \mathcal{T}'$ and $\mathbf{Discr}(\widetilde{I}_j) \geq \varepsilon/20k$. Then, we have that

$$\|q-p\|_1^{\mathcal{T}^*} \stackrel{\text{def}}{=} \sum_{j\in\mathcal{T}^*} \mathbf{Discr}(\widetilde{I}_j) \geq 2\varepsilon/5. \tag{3.2}$$

Observe that for any interval J it holds $\mathbf{Discr}(J) \leq \mathbf{width}(J)$. Note that this part of the argument depends critically on considering only troughs. Hence, for $j \in \mathcal{T}^*$ we have that

$$\varepsilon/(20k) \leq \mathbf{width}(\widetilde{I}_j) \leq 1/k. \tag{3.3}$$

Thus far we have argued that a constant fraction of the contribution to $\|q-p\|_1$ comes from troughs whose width satisfies (3.3). Our next crucial claim is that each such trough must have a "large" overlap with one of the intervals $I_i^{(j)}$ considered by our algorithm Test-Uniformity-$\mathcal{A}_k$. In particular, consider a trough $J \in \{\widetilde{I}_j\}_{j\in\mathcal{T}^*}$. We claim that there exists $j \in \{0,\ldots,j_0-1\}$ and $i \in [\ell_j]$ such that $|I_i^{(j)}| \geq |J|/4$ and so that $I_i^{(j)} \subseteq J$. To see this we first pick a j so that $\mathbf{width}(J)/2 > 2^{-j}/k \geq \mathbf{width}(J)/4$. Since the $I_i^{(j)}$ have width less than half that of J, J must intersect at least three of these intervals. Thus, any but the two outermost such intervals will be entirely contained within J, and furthermore has width $2^j/k \geq \mathbf{width}(J)/4$.

Since the interval $L \in \mathcal{I}^{(j+1)}$ is a "domain point" for the reduced distribution $q_r^{\mathcal{I}^{(j+1)}}$, the L_1 error between $q_r^{\mathcal{I}^{(j+1)}}$ and $U_{\ell_{j+1}}$ incurred by this element is at least $(1/4) \cdot \mathbf{Discr}(J)$, and the corresponding L_2^2 error is at least $(1/16) \cdot (\mathbf{Discr}(J))^2 \geq \frac{\varepsilon}{320k} \cdot \mathbf{Discr}(J)$, where the inequality follows from the fact that $\mathbf{Discr}(J) \geq \varepsilon/(20k)$. Hence, we have that

$$\|q_r^{\mathcal{I}^{(j+1)}} - U_{\ell_{j+1}}\|_2^2 \geq \varepsilon/(320k) \cdot \mathbf{Discr}(J). \tag{3.4}$$

As shown above, for every trough $J \in \{\widetilde{I}_j\}_{j\in\mathcal{T}^*}$ there exists a level $j \in \{0,\ldots,j_0-1\}$ such that (3.4) holds. Hence, summing (3.4) over all levels we obtain

$$\begin{aligned}\sum_{j=0}^{j_0-1} \|q_r^{\mathcal{I}^{(j+1)}} - U_{\ell_{j+1}}\|_2^2 &\geq \varepsilon/(320k) \cdot \sum_{j\in\mathcal{T}^*} \mathbf{Discr}(\widetilde{I}_j) \\ &\geq \varepsilon^2/(800k),\end{aligned}$$

where the second inequality follows from (3.2). Note that

$$\begin{aligned}\sum_{j=0}^{j_0-1} \gamma_j^2 \leq \sum_{j=0}^{j_0-1} \frac{\varepsilon^2 \cdot 2^{3j/4}}{80^2 \cdot k 2^j} &= \frac{\varepsilon^2}{6400k} \sum_{j=0}^{j_0-1} 2^{-j/4} \\ &< \varepsilon^2/(800k).\end{aligned}$$

Therefore, by the above, we must have that

$$\|q_r^{\mathcal{I}^{(j+1)}} - U_{\ell_{j+1}}\|_2^2 > \gamma_j^2$$

for some $0 \leq j \leq j_0 - 1$. This completes the proof of Lemma 3.1 for the special case of q being k-flat.

3.2 Proof of Structural Lemma: General Case To prove the general version of our structural result for the $\mathcal{A}_k$ distance, we will need to choose an appropriate value

for the universal constant C. We show that it is sufficient to take $C \leq 5 \cdot 10^{-6}$. (While we have not attempted to optimize constant factors, we believe that a more careful analysis will lead to substantially better constants.)

A useful observation is that our Test-Uniformity-$\mathcal{A}_k$ algorithm only distinguishes which of the intervals of $\mathcal{I}^{(j_0-1)}$ each of our samples lies in, and can therefore equivalently be thought of as a uniformity tester for the reduced distribution $q_r^{\mathcal{I}^{(j_0-1)}}$. In order to show that it suffices to consider only this restricted sample set, we claim that

$$\|q_r^{\mathcal{I}^{(j_0-1)}} - U_{\ell_{j_0-1}}\|_{\mathcal{A}_k} \geq \|p-q\|_{\mathcal{A}_k} - \varepsilon/2.$$

In particular, these $\mathcal{A}_k$ distances would be equal if the dividers of the optimal partition for q were all on boundaries between intervals of $\mathcal{I}^{(j_0-1)}$. If this was not the case though, we could round the endpoint of each trough inward to the nearest such boundary (note that we can assume that the optimal partition has no two adjacent troughs). This increases the discrepancy of each trough by at most $2k \cdot 2^{-j_0}$, and thus for $j_0 - \log_2(1/\varepsilon)$ a sufficiently large universal constant, the total discrepancy decreases by at most $\varepsilon/2$.

Thus, we have reduced ourselves to the case where $n = 2^{j_0-1} \cdot k$ and have argued that it suffices to show that our algorithm works to distinguish $\mathcal{A}_k$-distance in this setting with $\varepsilon_j = 10^{-5} \cdot \varepsilon \cdot 2^{3j/8}$.

The analysis of the completeness and the soundness of the tester is identical to Proposition 3.1. The only missing piece is the proof of Lemma 3.1, which we now restate for the sake of convenience:

LEMMA 3.2. *If $\|q-p\|_{\mathcal{A}_k} \geq \varepsilon$, there exists some $j \in \mathbb{Z}_+$ with $0 \leq j \leq j_0 - 1$ such that*

$$\|q_r^{\mathcal{I}^{(j)}} - U_{\ell_j}\|_2^2 \geq \gamma_j^2 := \varepsilon_j^2/\ell_j = 10^{-10} 2^{-j/4} \varepsilon^2/k.$$

The analysis of the general case here is somewhat more complicated than the special case for q being k-flat case that was presented in the previous section. This is because it is possible for one of the intervals J in the optimal partition (i.e., the interval partition $\mathcal{I}^* \in \mathfrak{J}_k$ maximizing $\|q_r^{\mathcal{I}} - q_r^{\mathcal{I}}\|_1$ in the definition of the $\mathcal{A}_k$ distance) to have large overlap with an interval I that our algorithm considers – that is, $I \in \cup_{j=0}^{j_0-1} \mathcal{I}^{(j)}$ – without having $q(I)$ and $p(I)$ differ substantially. Note that the unknown distribution q is not guaranteed to be constant within such an interval J, and in particular the difference $q-p$ does not necessarily preserve its sign within J.

To deal with this issue, we note that there are two possibilities for an interval J in the optimal partition: Either one of the intervals $I_i^{(j)}$ (considered by our algorithm) of size at least $|J|/2$ has discrepancy comparable to J, or the distribution q differs from p even more substantially on one of the intervals separating the endpointss of $I_i^{(j)}$ from the endpoints of J. Therefore, either an interval contained within this will detect a large L_2 error, or we will need to again pass to a subinterval. To make this intuition rigorous, we will need a mechanism for detecting where this recursion will terminate. To handle this formally, we introduce the following definition: between q and p to be

$$\|q-p\|_{[k]}^2 \stackrel{\text{def}}{=} \max_{\mathcal{I}=(I_1,\ldots,I_r)\in \mathbf{W}_{1/k}} \sum_{i=1}^{r} \frac{\mathbf{Discr}^2(I_i)}{\mathbf{width}^{1/8}(I_i)}$$

where $\mathbf{W}_{1/k}$ is the collection of all interval partitions of $[n]$ into intervals of width at most $1/k$.

The notion of the scale-sensitive-L_2 distance will be a useful intermediate tool in our analysis. The rough idea of the definition is that the optimal partition will be able to detect the correctly sized intervals for our tester to notice. (It will act as an analogue of the partition into the intervals where q is constant for the k-flat case.)

The first thing we need to show is that if q and p have large $\mathcal{A}_k$ distance then they also have large scale-sensitive-L_2 distance. Indeed, we have the following lemma:

LEMMA 3.3. *For $p = U_n$ and q an arbitrary distribution over $[n]$, we have that*

$$\|q-p\|_{[k]}^2 \geq \frac{\|q-p\|_{\mathcal{A}_k}^2}{(2k)^{7/8}}.$$

Proof. Let $\varepsilon = \|q-p\|_{\mathcal{A}_k}^2$. Consider the optimal $\mathcal{I}^*$ in the definition of the $\mathcal{A}_k$ distance. As in our analysis for the k-flat case, by further subdividing intervals of width more than $1/k$ into smaller ones, we can obtain a new partition, $\mathcal{I}' = (I'_i)_{i=1}^s$, of cardinality $s \leq 2k$ all of whose parts have width at most $1/k$. Furthermore, we have that $\sum_i \mathbf{Discr}(I'_i) \geq \varepsilon$. Using this partition to bound from below $\|q-p\|_{[k]}^2$, by Cauchy-Schwarz we obtain that

$$\begin{aligned}
\|q-p\|_{[k]}^2 &\geq \sum_i \frac{\mathbf{Discr}^2(I'_i)}{\mathbf{width}(I'_i)^{1/8}} \\
&\geq \frac{\left(\sum_i \mathbf{Discr}(I'_i)\right)^2}{\sum_i \mathbf{width}(I'_i)^{1/8}} \\
&\geq \frac{\varepsilon^2}{2k(1/(2k))^{1/8}} \\
&= \frac{\varepsilon^2}{(2k)^{7/8}}.
\end{aligned}$$

The second important fact about the scale-sensitive-L_2 distance is that if it is large then one of the partitions considered in our algorithm will produce a large L_2 error.

PROPOSITION 3.2. *Let $p = U_n$ be the uniform distribution and q a distribution over $[n]$. Then we have that*

$$\|q-p\|_{[k]}^2 \leq 10^8 \sum_{j=0}^{j_0-1} \sum_{i=1}^{2^j \cdot k} \frac{\mathbf{Discr}^2(I_i^{(j)})}{\mathbf{width}^{1/8}(I_i^{(j)})}. \tag{3.5}$$

Proof. Let $\mathcal{J} \in \mathbf{W}_{1/k}$ be the optimal partition used when computing the scale-sensitive-L_2 distance $\|q-p\|_{[k]}$. In particular, it is a partition into intervals of width at most $1/k$ so that $\sum_i \frac{\mathbf{Discr}^2(J_i)}{\mathbf{width}(J_i)^{1/8}}$ is maximized. To prove Equation (3.5), we prove a notably stronger claim. In particular, we will prove that for each interval $J_\ell \in \mathcal{J}$

$$\frac{\mathbf{Discr}^2(J_\ell)}{\mathbf{width}^{1/8}(J_\ell)} \leq 10^8 \sum_{j=0}^{j_0-1} \sum_{i:I_i^{(j)} \subset J_\ell} \frac{\mathbf{Discr}^2(I_i^{(j)})}{\mathbf{width}^{1/8}(I_i^{(j)})}. \tag{3.6}$$

Summing over ℓ would then yield $\|q-p\|_{[k]}^2$ on the left hand side and a strict subset of the terms from Equation (3.5) on the right hand side. From here on, we will consider only a single interval J_ℓ. For notational convenience, we will drop the subscript and merely call it J.

First, note that if $|J| \leq 10^8$, then this follows easily from considering just the sum over $j = j_0 - 1$. Then, if $t = |J|$, J is divided into t intervals of size one. The sum of the discrepancies of these intervals equals the discrepancy of J, and thus, the sum of the squares of the discrepancies is at least $\mathbf{Discr}^2(J)/t$. Furthermore, the widths of these subintervals are all smaller than the width of J by a factor of t. Thus, in this case the sum of the right hand side of Equation (3.6) is at least $1/t^{7/8} \geq \frac{1}{10^7}$ of the left hand side.

Otherwise, if $|J| > 10^8$, we can find a j so that $\mathbf{width}(J)/10^8 < 1/(2^j \cdot k) \leq 2 \cdot \mathbf{width}(J)/10^8$. We claim that in this case Equation (3.6) holds even if we restrict the sum on the right hand side to this value of j. Note that J contains at most 10^8 intervals of $\mathcal{I}^{(j)}$, and that it is covered by these intervals plus two narrower intervals on the ends. Call these end-intervals R_1 and R_2. We claim that $\mathbf{Discr}(R_i) \leq \mathbf{Discr}(J)/3$. This is because otherwise it would be the case that

$$\frac{\mathbf{Discr}^2(R_i)}{\mathbf{width}^{1/8}(R_i)} > \frac{\mathbf{Discr}^2(J)}{\mathbf{width}^{1/8}(J)}.$$

(This is because $(1/3)^2 \cdot (2/10^8)^{-1/8} > 1$.) This is a contradiction, since it would mean that partitioning J into R_i and its complement would improve the sum defining $\|q-p\|_{[k]}$, which was assumed to be maximum. This in turn implies that the sum of the discrepancies of the $I_i^{(j)}$ contained in J must be at least $\mathbf{Discr}(J)/3$, so the sum of their squares is at least $\mathbf{Discr}^2(J)/(9 \cdot 10^8)$. On the other hand, each of these intervals is narrower than J by a factor of at least $10^8/2$, thus the appropriate sum of $\frac{\mathbf{Discr}^2(I_i^{(j)})}{\mathbf{width}^{1/8}(I_i^{(j)})}$ is at least $\frac{\mathbf{Discr}^2(J)}{10^8 \mathbf{width}^{1/8}(J)}$. This completes the proof.

We are now ready to prove Lemma 3.2.

Proof. If $\|q-p\|_{\mathcal{A}_k} \geq \varepsilon$ we have by Lemma 3.3 that

$$\|q-p\|_{[k]}^2 \geq \frac{\varepsilon^2}{(2k)^{7/8}}.$$

By Proposition 3.2, this implies that

$$\begin{aligned} \frac{\varepsilon^2}{(2k)^{7/8}} &\leq 10^8 \sum_{j=0}^{j_0-1} \sum_{i=1}^{2^j \cdot k} \frac{\mathbf{Discr}^2(I_i^{(j)})}{\mathbf{width}^{1/8}(I_i^{(j)})} \\ &= 10^8 \sum_{j=0}^{j_0-1} (2^j k)^{1/8} \|q^{\mathcal{I}^{(j)}} - U_{\ell_j}\|_2^2. \end{aligned}$$

Therefore,

$$\sum_{j=0}^{j_0-1} 2^{j/8} \|q^{\mathcal{I}^{(j)}} - U_{\ell_j}\|_2^2 \geq 5 \cdot 10^{-9} \varepsilon^2/k. \tag{3.7}$$

On the other hand, if $\|q^{\mathcal{I}^{(j)}} - U_{\ell_j}\|_2^2$ were at most $10^{-10} 2^{-j/4} \varepsilon^2/k$ for each j, then the sum above would be at most

$$10^{-10} \varepsilon^2/k \sum_j 2^{-j/8} < 5 \cdot 10^{-9} \varepsilon^2/k.$$

This would contradict Equation (3.7), thus proving that $\|q^{\mathcal{I}^{(j)}} - U_{\ell_j}\|_2^2 \geq 10^{-10} 2^{-j/4} \varepsilon^2/k$ for at least one j, proving Lemma 3.2.

4 Conclusions and Future Work

In this work we designed a computationally efficient algorithm for the problem of identity testing against a known distribution, which yields sample optimal bounds for a wide range of natural and important classes of structured distributions. A natural direction for future work is to generalize our results to the problem of identity testing between two unknown structured distributions. What is the optimal sample complexity in this more general setting? We emphasize that new ideas are required for this problem, as the algorithm and analysis in this work crucially exploit the a priori knowledge of the explicit distribution.

References

[ADJ+11] J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, and S. Pan. Competitive closeness testing. *Journal of Machine Learning Research - Proceedings Track*, 19:47–68, 2011.

[Bat01] T. Batu. *Testing Properties of Distributions*. PhD thesis, Cornell University, 2001.

[BBBB72] R.E. Barlow, D.J. Bartholomew, J.M. Bremner, and H.D. Brunk. *Statistical Inference under Order Restrictions*. Wiley, New York, 1972.

[BDKR02] T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld. The complexity of approximating entropy. In *ACM Symposium on Theory of Computing*, pages 678–687, 2002.

[BFF+01] T. Batu, E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White. Testing random variables for independence and identity. In *Proc. 42nd IEEE Symposium on Foundations of Computer Science*, pages 442–451, 2001.

[BFR+00] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing that distributions are close. In *IEEE Symposium on Foundations of Computer Science*, pages 259–269, 2000.

[BFR+13] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing closeness of discrete distributions. *J. ACM*, 60(1):4, 2013.

[Bir87a] L. Birgé. Estimating a density under order restrictions: Nonasymptotic minimax risk. *Annals of Statistics*, 15(3):995–1012, 1987.

[Bir87b] L. Birgé. On the risk of histograms for estimating decreasing densities. *Annals of Statistics*, 15(3):1013–1022, 1987.

[BKR04] T. Batu, R. Kumar, and R. Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *ACM Symposium on Theory of Computing*, pages 381–390, 2004.

[Bru58] H. D. Brunk. On the estimation of parameters restricted by inequalities. *The Annals of Mathematical Statistics*, 29(2):pp. 437–454, 1958.

[BRW09] F. Balabdaoui, K. Rufibach, and J. A. Wellner. Limit distribution theory for maximum likelihood estimation of a log-concave density. *The Annals of Statistics*, 37(3):pp. 1299–1331, 2009.

[BS10] M. Belkin and K. Sinha. Polynomial learning of distribution families. In *FOCS*, pages 103–112, 2010.

[BW07] F. Balabdaoui and J. A. Wellner. Estimation of a k-monotone density: Limit distribution theory and the spline connection. *The Annals of Statistics*, 35(6):pp. 2536–2564, 2007.

[BW10] F. Balabdaoui and J. A. Wellner. Estimation of a k-monotone density: characterizations, consistency and minimax lower bounds. *Statistica Neerlandica*, 64(1):45–70, 2010.

[CDSS13] S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Learning mixtures of structured distributions over discrete domains. In *SODA*, pages 1380–1394, 2013.

[CDSS14] S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Efficient density estimation via piecewise polynomial approximation. In *STOC*, pages 604–613, 2014.

[CDVV14] S. Chan, I. Diakonikolas, P. Valiant, and G. Valiant. Optimal algorithms for testing closeness of discrete distributions. In *SODA*, pages 1193–1203, 2014.

[CT04] K.S. Chan and H. Tong. Testing for multimodality with dependent data. *Biometrika*, 91(1):113–123, 2004.

[DDO+13] C. Daskalakis, I. Diakonikolas, R. O'Donnell, R.A. Servedio, and L. Tan. Learning Sums of Independent Integer Random Variables. In *FOCS*, pages 217–226, 2013.

[DDS12a] C. Daskalakis, I. Diakonikolas, and R.A. Servedio. Learning k-modal distributions via testing. In *SODA*, pages 1371–1385, 2012.

[DDS12b] C. Daskalakis, I. Diakonikolas, and R.A. Servedio. Learning Poisson Binomial Distributions. In *STOC*, pages 709–728, 2012.

[DDS+13] C. Daskalakis, I. Diakonikolas, R. Servedio, G. Valiant, and P. Valiant. Testing k-modal distributions: Optimal algorithms via reductions. In *SODA*, pages 1833–1852, 2013.

[DL01] L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. Springer Series in Statistics, Springer, 2001.

[DR09] L. D umbgen and K. Rufibach. Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency. *Bernoulli*, 15(1):40–68, 2009.

[FOS05] J. Feldman, R. O'Donnell, and R. Servedio. Learning mixtures of product distributions over discrete domains. In *Proc. 46th Symposium on Foundations of Computer Science (FOCS)*, pages 501–510, 2005.

[Fou97] A.-L. Fougères. Estimation de densités unimodales. *Canadian Journal of Statistics*, 25:375–387, 1997.

[GGR98] O. Goldreich, S. Goldwasser, and D. Ron. Property testing and its connection to learning and approximation. *Journal of the ACM*, 45:653–750, 1998.

[GR00] O. Goldreich and D. Ron. On testing expansion in bounded-degree graphs. Technical Report TR00-020, Electronic Colloquium on Computational Complexity, 2000.

[Gre56] U. Grenander. On the theory of mortality measurement. *Skand. Aktuarietidskr.*, 39:125–153, 1956.

[Gro85] P. Groeneboom. Estimating a monotone density. In *Proc. of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, pages 539–555, 1985.

[GW09] F. Gao and J. A. Wellner. On the rate of convergence of the maximum likelihood estimator of a k-monotone density. *Science in China Series A: Mathematics*, 52:1525–1538, 2009.

[Ham87] F. R. Hampel. Design, data & analysis. chapter Design, modelling, and analysis of some biological data sets, pages 93–128. John Wiley & Sons, Inc., New York, NY, USA, 1987.

[HP76] D. L. Hanson and G. Pledger. Consistency in concave regression. *The Annals of Statistics*, 4(6):pp. 1038–1050, 1976.

[ILR12] P. Indyk, R. Levi, and R. Rubinfeld. Approximating and Testing k-Histogram Distributions in Sub-linear Time. In *PODS*, pages 15–22, 2012.

[JW09] H. K. Jankowski and J. A. Wellner. Estimation of a discrete monotone density. *Electronic Journal of Statistics*, 3:1567–1605, 2009.

[KM10] R. Koenker and I. Mizera. Quasi-concave density estimation. *Ann. Statist.*, 38(5):2998–3027, 2010.

[KMR+94] M. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R. Schapire, and L. Sellie. On the learnability of discrete distributions. In *Proceedings of the 26th Symposium on*

Theory of Computing, pages 273–282, 1994.

[KMV10] A. T. Kalai, A. Moitra, and G. Valiant. Efficiently learning mixtures of two Gaussians. In *STOC*, pages 553–562, 2010.

[LRR11] R. Levi, D. Ron, and R. Rubinfeld. Testing properties of collections of distributions. In *ICS*, pages 179–194, 2011.

[MV10] A. Moitra and G. Valiant. Settling the polynomial learnability of mixtures of Gaussians. In *FOCS*, pages 93–102, 2010.

[NP33] J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.

[Pan08] L. Paninski. A coincidence-based test for uniformity given very sparsely-sampled discrete data. *IEEE Transactions on Information Theory*, 54:4750–4755, 2008.

[Pea00] K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, 50(302):157–175, 1900.

[Rao69] B.L.S. Prakasa Rao. Estimation of a unimodal density. *Sankhya Ser. A*, 31:23–36, 1969.

[Reb05] L. Reboul. Estimation of a function under shape restrictions. Applications to reliability. *Ann. Statist.*, 33(3):1330–1356, 2005.

[RS96] R. Rubinfeld and M. Sudan. Robust characterizations of polynomials with applications to program testing. *SIAM J. on Comput.*, 25:252–271, 1996.

[Rub12] R. Rubinfeld. Taming big probability distributions. *XRDS*, 19(1):24–28, 2012.

[Val11] P. Valiant. Testing symmetric properties of distributions. *SIAM J. Comput.*, 40(6):1927–1968, 2011.

[VV11] G. Valiant and P. Valiant. Estimating the unseen: an $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs. In *STOC*, pages 685–694, 2011.

[VV14] G. Valiant and P. Valiant. An automatic inequality prover and instance optimal identity testing. In *FOCS*, 2014.

[Wal09] G. Walther. Inference and modeling with log-concave distributions. *Statistical Science*, 24(3):319–327, 2009.

[Weg70] E.J. Wegman. Maximum likelihood estimation of a unimodal density. I. and II. *Ann. Math. Statist.*, 41:457–471, 2169–2174, 1970.

[WWW+05] X. Wang, M. Woodroofe, M. Walker, M. Mateo, and E. Olszewski. Estimating dark matter distributions. *The Astrophysical Journal*, 626:145–158, 2005.

Appendix: Omitted Proofs

A A Useful Primitive: Testing Uniformity in L_2 norm

In this section, we give an algorithm for uniformity testing with respect to the L_2 distance, thereby establishing Theorem 2.2. The algorithm Test-Uniformity-$L_2(q, n, \varepsilon)$ described below draws $O(\sqrt{n}/\varepsilon^2)$ samples from a distribution q over $[n]$ and distinguishes between the cases that $q = U_n$ versus $\|q - U_n\|_2 > \varepsilon/\sqrt{n}$ with probability at least $2/3$. Repeating the algorithm $O(\log(1/\delta))$ times and taking the majority answer results in a confidence probability of $1 - \delta$, giving the desired algorithm Test-Uniformity-$L_2(q, n, \varepsilon, \delta)$ of Theorem 2.2.

Our estimator is a variant of Pearson's chi-squared test [Pea00], and can be viewed as a special case of the recent "chi-square type" testers in [CDVV14, VV14]. We remark that, as follows from the Cauchy-Schwarz inequality, the same estimator distinguishes the uniform distribution from any distribution q such that $\|q - U_n\|_1 > \varepsilon$, i.e., algorithm Test-Uniformity-$L_2(q, n, \varepsilon)$ is an optimal uniformity tester for the L_1 norm. The L_2 guarantee we prove here is new, is strictly stronger than the aforementioned L_1 guarantee, and is crucial for our purposes in Section 3.

For $\lambda \geq 0$, we denote by $\mathrm{Poi}(\lambda)$ the Poisson distribution with parameter λ. In our algorithm below, we employ the standard "Poissonization" approach: namely, we assume that, rather than drawing m independent samples from a distribution, we first select m' from $\mathrm{Poi}(m)$, and then draw m' samples. This Poissonization makes the number of times different elements occur in the sample independent, with the distribution of the number of occurrences of the i-th domain element distributed as $\mathrm{Poi}(mq_i)$, simplifying the analysis. As $\mathrm{Poi}(m)$ is tightly concentrated about m, we can carry out this Poissonization trick without loss of generality at the expense of only subconstant factors in the sample complexity.

Algorithm Test-Uniformity-$L_2(q, n, \varepsilon)$
Input: sample access to a distribution q over $[n]$, and $\varepsilon > 0$.
Output: "YES" if $q = U_n$; "NO" if $\|q - U_n\|_2 \geq \varepsilon/\sqrt{n}$.

1. Draw $m' \sim \mathrm{Poi}(m)$ iid samples from q.
2. Let X_i be the number of occurrences of the ith domain elements in the sample from q
3. Define $Z = \sum_{i=1}^{n} (X_i - m/n)^2 - X_i$.
4. If $Z \geq 4m/\sqrt{n}$ return "NO"; otherwise, return "YES".

The following theorem characterizes the performance of the above estimator:

THEOREM A.1. *For any distribution q over $[n]$ the above algorithm distinguishes the case that $q = U_n$ from the case that $||q - U_n||_2 \geq \varepsilon/\sqrt{n}$ when given $O(\sqrt{n}/\varepsilon^2)$ samples from q with probability at least $2/3$.*

Proof. Define $Z_i = (X_i - m/n)^2 - X_i$. Since X_i is distributed as $\mathrm{Poi}(mq_i)$, $\mathbb{E}[Z_i] = m^2\Delta_i^2$, where $\Delta_i := 1/n - q_i$. By linearity of expectation we can write $\mathbb{E}[Z] = \sum_{i=1}^n \mathbb{E}[Z_i] = m^2 \cdot \sum_{i=1}^n \Delta_i^2$. Similarly we can calculate

$$\mathrm{Var}[Z_i] = 2m^2(\Delta_i - 1/n)^2 + 4m^3(1/n - \Delta_i)\Delta_i^2.$$

Since the X_i's (and hence the Z_i's) are independent, it follows that $\mathrm{Var}[Z] = \sum_{i=1}^n \mathrm{Var}[Z_i]$.

We start by establishing completeness. Suppose $q = U_n$. We will show that $\Pr[Z \geq 4m/\sqrt{n}] \leq 1/3$. Note that in this case $\Delta_i = 0$ for all $i \in [n]$, hence $\mathbb{E}[Z] = 0$ and $\mathrm{Var}[Z] = 2m^2/n$. Chebyshev's inequality implies that

$$\begin{aligned}\Pr[Z \geq 4m/\sqrt{n}] &= \Pr\left[Z \geq (2\sqrt{2})\sqrt{\mathrm{Var}[Z]}\right] \\ &\leq (1/8) < 2/3\end{aligned}$$

as desired.

We now proceed to prove soundness of the tester. Suppose that $\|q - U_n\|_2 \geq \frac{\varepsilon}{\sqrt{n}}$. In this case we will show that $\Pr[Z \leq 4m/\sqrt{n}] \leq 1/3$. Note that Chebyshev's inequality implies that

$$\Pr\left[Z \leq \mathbb{E}[Z] - 2\sqrt{\mathrm{Var}[Z]}\right] \leq 1/4.$$

It thus suffices to show that $\mathbb{E}[Z] \geq 8m/\sqrt{n}$ and $\mathbb{E}[Z]^2 \geq 16[Z]$. Establishing the former inequality is easy. Indeed,

$$\mathbb{E}[Z] = m^2 \cdot \|q - U_n\|_2^2 \geq m^2 \cdot (\varepsilon^2/n) \geq 8m/\sqrt{n}$$

for $m \geq 8\sqrt{n}/\varepsilon^2$.

Proving the latter inequality requires a more detailed analysis. We will show that for a sufficiently large constant $C > 0$, if $m \geq C\sqrt{n}/\varepsilon^2$ we will have

$$[Z] \ll \mathbb{E}[Z]^2.$$

Ignoring multiplicative constant factors, we equivalently need to show that

$$\begin{aligned}m^2 \cdot \left(\sum_{i=1}^n \left(\Delta_i^2 - 2\Delta_i/n\right) + 1/n\right) + \\ m^3 \sum_{i=1}^n \left(\Delta_i^2/n + \Delta_i^3\right) \\ \ll m^4 \left(\sum_{i=1}^n \Delta_i^2\right)^2.\end{aligned}$$

To prove the desired inequality, it suffices to bound from above the absolute value of each of the five terms of the LHS separately. For the first term we need to show that $m^2 \cdot \sum_{i=1}^n \Delta_i^2 \ll m^4 \cdot \left(\sum_{i=1}^n \Delta_i^2\right)^2$ or equivalently

$$m \gg 1/\|q - U_n\|_2. \tag{A.1}$$

Since $\|q - U_n\|_2 \geq \varepsilon/\sqrt{n}$, the RHS of (A.1) is bounded from above by $\sqrt{n}/\varepsilon$, hence (A.1) holds true for our choice of m.

For the second term we want to show that $\sum_{i=1}^n |\Delta_i| \ll m^2 n \cdot \left(\sum_{i=1}^n \Delta_i^2\right)^2$. Recalling that $\sum_{i=1}^n |\Delta_i| \leq \sqrt{n} \cdot \sqrt{\sum_{i=1}^n \Delta_i^2}$, as follows from the Cauchy-Schwarz inequality, it suffices to show that $m^2 \gg (1/\sqrt{n}) \cdot 1/(\sum_{i=1}^n \Delta_i^2)^{3/2}$ or equivalently

$$m \gg \frac{1}{n^{1/4}} \cdot \frac{1}{\|q - U_n\|_2^{3/2}}. \tag{A.2}$$

Since $\|q - U_n\|_2 \geq \varepsilon/\sqrt{n}$, the RHS of (A.2) is bounded from above by $\sqrt{n}/\varepsilon^{3/2}$, hence (A.2) is also satisfied.

For the third term we want to argue that $m^2/n \ll m^4 \cdot \left(\sum_{i=1}^n \Delta_i^2\right)^2$ or

$$m \gg \frac{1}{n^{1/2}} \cdot \frac{1}{\|q - U_n\|_2^2}, \tag{A.3}$$

which holds for our choice of m, since the RHS is bounded from above by $\sqrt{n}/\varepsilon^2$.

Bounding the fourth term amounts to showing that $(m^3/n)\sum_{i=1}^n \Delta_i^2 \ll m^4 \left(\sum_{i=1}^n \Delta_i^2\right)^2$ which can be rewritten as

$$m \gg \frac{1}{n} \cdot \frac{1}{\|q - U_n\|_2^2}, \tag{A.4}$$

and is satisfied since the RHS is at most $1/\varepsilon^2$.

Finally, for the fifth term we want to prove that $m^3 \cdot \sum_{i=1}^n |\Delta_i|^3 \ll m^4 \cdot \left(\sum_{i=1}^n \Delta_i^2\right)^2$ or that $\sum_{i=1}^n |\Delta_i|^3 \ll m \cdot \left(\sum_{i=1}^n \Delta_i^2\right)^2$. From Jensen's inequality it follows that $\sum_{i=1}^n |\Delta_i|^3 \leq \left(\sum_{i=1}^n \Delta_i|^2\right)^{3/2}$; hence, it is sufficient to show that $\left(\sum_{i=1}^n \Delta_i|^2\right)^{3/2} \ll m \cdot \left(\sum_{i=1}^n \Delta_i^2\right)^2$ or

$$m \gg \frac{1}{\|q - U_n\|_2}. \tag{A.5}$$

Since $\|q - U_n\|_2 \geq \varepsilon/\sqrt{n}$ the above RHS is at most $\sqrt{n}/\varepsilon$ and (A.5) is satisfied. This completes the soundness proof and the proof of Theorem A.1.

A.2 Optimal Algorithms and Lower Bounds for Testing Identity of Structured Distributions

Optimal Algorithms and Lower Bounds for Testing Closeness of Structured Distributions

Ilias Diakonikolas*
University of Edinburgh
ilias.d@ed.ac.uk.

Daniel M. Kane
University of California, San Diego†
dakane@cs.ucsd.edu.

Vladimir Nikishkin‡
University of Edinburgh
v.nikishkin@sms.ed.ac.uk.

August 22, 2015

Abstract

We give a general unified method that can be used for L_1 *closeness testing* of a wide range of univariate structured distribution families. More specifically, we design a sample optimal and computationally efficient algorithm for testing the equivalence of two unknown (potentially arbitrary) univariate distributions under the $\mathcal{A}_k$-distance metric: Given sample access to distributions with density functions $p, q : I \to \mathbb{R}$, we want to distinguish between the cases that $p = q$ and $\|p - q\|_{\mathcal{A}_k} \geq \epsilon$ with probability at least $2/3$. We show that for any $k \geq 2, \epsilon > 0$, the *optimal* sample complexity of the $\mathcal{A}_k$-closeness testing problem is $\Theta(\max\{k^{4/5}/\epsilon^{6/5}, k^{1/2}/\epsilon^2\})$. This is the first $o(k)$ sample algorithm for this problem, and yields new, simple L_1 closeness testers, in most cases with optimal sample complexity, for broad classes of structured distributions.

*Supported by EPSRC grant EP/L021749/1 and a Marie Curie Career Integration grant.
†Some of this work was performed while visiting the University of Edinburgh.
‡Supported by a University of Edinburgh PCD Scholarship.

1 Introduction

We study the problem of closeness testing (equivalence testing) between two unknown probability distributions. Given independent samples from a pair of distributions p, q, we want to determine whether the two distributions are the same versus significantly different. This is a classical problem in statistical hypothesis testing [NP33, LR05] that has received considerable attention by the TCS community in the framework of *property testing* [RS96, GGR98]: given sample access to distributions p, q, and a parameter $\epsilon > 0$, we want to distinguish between the cases that p and q are identical versus ϵ-far from each other in L_1 norm (statistical distance). Previous work on this problem focused on characterizing the sample size needed to test the identity of two arbitrary distributions of a given support size [BFR$^+$00, CDVV14]. It is now known that the optimal sample complexity (and running time) of this problem for distributions with support of size n is $\Theta(\max\{n^{2/3}/\epsilon^{4/3}, n^{1/2}/\epsilon^2\})$.

The aforementioned sample complexity characterizes worst-case instances, and one might hope that drastically better results can be obtained for most natural settings, in particular when the underlying distributions are known a priori to have some "nice structure". In this work, we focus on the problem of testing closeness for *structured* distributions. Let $\mathcal{C}$ be a family over univariate distributions. The problem of *closeness testing for* $\mathcal{C}$ is the following: Given sample access to two unknown distribution $p, q \in \mathcal{C}$, we want to distinguish between the case that $p = q$ versus $\|p - q\|_1 \geq \epsilon$. Note that the sample complexity of this testing problem depends on the underlying class $\mathcal{C}$, and we are interested in obtaining efficient algorithms that are *sample optimal* for $\mathcal{C}$.

We give a general algorithm that can be used for L_1 closeness testing of a wide range of structured distribution families. More specifically, we give a sample optimal and computationally efficient algorithm for testing the identity of two unknown (potentially arbitrary) distributions p, q under a different metric between distributions – the so called $\mathcal{A}_k$-distance (see Section 2 for a formal definition). Here, k is a positive integer that intuitively captures the number of "crossings" between the probability density functions p, q.

Our main result (see Theorem 1) says the following: *For any $k \in \mathbb{Z}_+, \epsilon > 0$, and sample access to arbitrary univariate distributions p, q, there exists a closeness testing algorithm under the $\mathcal{A}_k$-distance using $O(\max\{k^{4/5}/\epsilon^{6/5}, k^{1/2}/\epsilon^2\})$ samples. Moreover, this bound is information-theoretically optimal.* We remark that our $\mathcal{A}_k$-testing algorithm applies to *any* pair of univariate distributions (over both continuous and discrete domains). The main idea in using this general algorithm for testing closeness of structured distributions in L_1 distance is this: if the underlying distributions p, q belong to a structured distribution family $\mathcal{C}$, we can use the $\mathcal{A}_k$-distance as a proxy for the L_1 distance (for an appropriate value of the parameter k), and thus obtain an L_1 closeness tester for $\mathcal{C}$.

We note that $\mathcal{A}_k$-distance between distributions has been recently used to obtain sample optimal efficient algorithms for *learning* structured distributions [CDSS14, ADLS15], and for testing the identity of a structured distribution against an *explicitly known* distribution [DKN15] (e.g., uniformity testing). In both these settings, the sample complexity of the corresponding problem (learning/identity testing) with respect to the $\mathcal{A}_k$-distance is identified with the sample complexity of the problem under the L_1 distance *for distributions of support k*. More specifically, the sample complexity of learning an unknown univariate distribution (over a continuous or discrete domain) up to $\mathcal{A}_k$-distance ϵ is $\Theta(k/\epsilon^2)$ [CDSS14] (independent of the domain size), which is exactly the sample complexity of learning a *discrete distribution with support size k* up to L_1 error ϵ. Similarly, the sample complexity of uniformity testing of a univariate distribution (over a continuous or discrete domain) up to $\mathcal{A}_k$-distance ϵ is $\Theta(k^{1/2}/\epsilon^2)$ [DKN15] (again, independent of the domain size), which is identical to the sample complexity of uniformity testing of a discrete distribution with support size k up to L_1 error ϵ [Pan08].

Rather surprisingly, this analogy is *provably false* for the closeness testing problem: we prove that the sample complexity of the $\mathcal{A}_k$ closeness testing problem is $\Theta(\max\{k^{4/5}/\epsilon^{6/5}, k^{1/2}/\epsilon^2\})$, while L_1 closeness testing between distributions of support k can be achieved with $O(\max\{k^{2/3}/\epsilon^{4/3}, k^{1/2}/\epsilon^2\})$ samples [CDVV14]. More specifically, our upper bound for $\mathcal{A}_k$ closeness testing problem applies for all univariate probability distributions (both continuous and discrete). Our matching information–theoretic lower bound holds for continuous distributions, or discrete distributions of support size n sufficiently large as a function of k, which is the most interesting regime for our applications.

1.1 Related and Prior Work In this subsection we review the related literature and compare our results with previous work.

Distribution Property Testing Testing properties of distributions [BFR$^+$00, BFR$^+$13] has developed into a mature research area within theoretical computer science. The paradigmatic problem in this field is the following: given sample access to one or more unknown probability distributions, determine whether they satisfy some global property or are "far" from satisfying the property. The goal is to obtain an algorithm for this task that is both statistically and computationally efficient, i.e., an algorithm with (information–theoretically) optimal sample size and polynomial runtime. See [GR00, BFR$^+$00, BFF$^+$01, Bat01, BDKR02, BKR04, Pan08, Val11, VV11, DDS$^+$13, ADJ$^+$11, LRR11, ILR12, CDVV14, VV14, DKN15] for a sample of works, and [Rub12] for a survey.

Shape Restricted Estimation Statistical estimation under shape restrictions – i.e., inference about a probability distribution under the constraint that its probability density function satisfies certain qualitative properties – is a classical topic in statistics [BBBB72]. Various structural restrictions have been studied in the literature, starting from monotonicity, unimodality, convexity, and concavity [Gre56, Bru58, Rao69, Weg70, HP76, Gro85, Bir87a, Bir87b, Fou97, CT04, JW09], and more recently focusing on structural restrictions such as log-concavity and k-monotonicity [BW07, DR09, BRW09, GW09, BW10, KM10, Wal09, DW13, CS13, KS14, BD14, HW15]. The reader is referred to [GJ14] for a recent book on the topic.

Comparison with Prior Work Chan, Diakonikolas, Servedio, and Sun [CDSS14] proposed a general approach to L_1 learn univariate probability distributions whose densities are well approximated by piecewise polynomials. They designed an efficient agnostic learning algorithm for piecewise polynomial distributions, and as a corollary obtained efficient learners for various families of structured distributions. The approach of [CDSS14] uses the $\mathcal{A}_k$ distance metric between distributions, but is otherwise orthogonal to ours. Batu *et al.* [BKR04] gave algorithms for closeness testing between two monotone distributions with sample complexity $O(\log^3 n)$. Subsequently, Daskalakis *et al.* [DDS$^+$13] improved and generalized this result to t-modal distributions, obtaining a closeness tester with sample complexity $O((t\log(n))^{2/3}/\epsilon^{8/3} + t^2/\epsilon^4)$. We remark that the approach of [DDS$^+$13] inherently yields an algorithm with sample complexity $\Omega(t)$, which is sub-optimal.

The main ideas underlying this work are very different from those of [DDS$^+$13] and [DKN15]. The approach of [DDS$^+$13] involves constructing an adaptive interval decomposition of the domain followed by an application of a (known) closeness tester to the "reduced" distributions over those intervals. This approach incurs an extraneous term in the sample complexity, that is needed to construct the appropriate decomposition. The approach of [DKN15] considers several oblivious interval decompositions of the domain (i.e., without drawing any samples) and applies a "reduced" identity tester for each such decomposition. This idea yields sample–optimal bounds for $\mathcal{A}_k$ identity testing against a *known* distribution. However, it crucially exploits the knowledge of the explicit distribution, and unfortunately fails in the setting where both distributions are unknown. We elaborate on these points in Section 2.3.

2 Our Results and Techniques

2.1 Basic Definitions We will use p, q to denote the probability density functions (or probability mass functions) of our distributions. If p is discrete over support $[n] := \{1, \ldots, n\}$, we denote by p_i the probability of element i in the distribution. For two discrete distributions p, q, their L_1 and L_2 distances are $\|p-q\|_1 = \sum_{i=1}^{n} |p_i - q_i|$ and $\|p-q\|_2 = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$. For $I \subseteq \mathbb{R}$ and density functions $p, q : I \to \mathbb{R}_+$, we have $\|p-q\|_1 = \int_I |p(x) - q(x)| dx$.

Fix a partition of the domain I into disjoint intervals $\mathcal{I} := (I_i)_{i=1}^{\ell}$. For such a partition $\mathcal{I}$, the *reduced distribution* $p_r^{\mathcal{I}}$ corresponding to p and $\mathcal{I}$ is the discrete distribution over $[\ell]$ that assigns the i-th "point" the mass that p assigns to the interval I_i; i.e., for $i \in [\ell]$, $p_r^{\mathcal{I}}(i) = p(I_i)$. Let $\mathfrak{J}_k$ be the collection of all partitions of the domain I into k intervals. For $p, q : I \to \mathbb{R}_+$ and $k \in \mathbb{Z}_+$, we define the $\mathcal{A}_k$-distance between p and q by

$$\|p-q\|_{\mathcal{A}_k} \stackrel{\text{def}}{=} \max_{\mathcal{I}=(I_i)_{i=1}^{k} \in \mathfrak{J}_k} \sum_{i=1}^{k} |p(I_i) - q(I_i)| = \max_{\mathcal{I} \in \mathfrak{J}_k} \|p_r^{\mathcal{I}} - q_r^{\mathcal{I}}\|_1.$$

2.2 Our Results Our main result is an optimal algorithm and a matching information–theoretic lower bound for the problem of testing the equivalence between two unknown univariate distributions under the $\mathcal{A}_k$ distance metric:

Theorem 1 (Main)**.** *Given $\epsilon > 0$, an integer $k \geq 2$, and sample access to two distributions with probability density functions $p, q : [0,1] \to \mathbb{R}_+$, there is a computationally efficient algorithm which uses $O(\max\{k^{4/5}/\epsilon^{6/5}, k^{1/2}/\epsilon^2\})$ samples from p, q, and with probability at least $2/3$ distinguishes whether $q = p$ versus $\|q-p\|_{\mathcal{A}_k} \geq \epsilon$. Additionally, $\Omega(\max\{k^{4/5}/\epsilon^{6/5}, k^{1/2}/\epsilon^2\})$ samples are information-theoretically necessary for this task.*

Note that Theorem 1 applies to arbitrary univariate distributions (over both continuous and discrete domains). In particular, the sample complexity of the algorithm does not depend on the support size of the underlying distributions. We believe that the notion of testing under the $\mathcal{A}_k$ distance is very natural, and well suited for (arbitrary) continuous distributions, where the notion of L_1 testing is (provably) impossible.

As a corollary of Theorem 1, we obtain sample–optimal algorithms for the L_1 closeness testing of various structured distribution families $\mathcal{C}$ in a unified way. The basic idea is to use the $\mathcal{A}_k$ distance as a "proxy" for the L_1 distance for an appropriate value of k that depends on $\mathcal{C}$ and ϵ. We have the following simple fact:

Fact 2. *For a univariate distribution family $\mathcal{C}$ and $\epsilon > 0$, let $k = k(\mathcal{C}, \epsilon)$ be the smallest integer such that for any $f_1, f_2 \in \mathcal{C}$ it holds that $\|f_1 - f_2\|_1 \leq \|f_1 - f_2\|_{\mathcal{A}_k} + \epsilon/2$. Then there exists an L_1 closeness testing algorithm for $\mathcal{C}$ using $O(\max\{k^{4/5}/\epsilon^{6/5}, k^{1/2}/\epsilon^2\})$ samples.*

Indeed, given sample access to $q, p \in \mathcal{C}$, we apply the $\mathcal{A}_k$-closeness testing algorithm of Theorem 1 for the value of k in the statement of the fact, and error $\epsilon' = \epsilon/2$. If $q = p$, the algorithm will output "YES" with probability at least $2/3$. If $\|q-p\|_1 \geq \epsilon$, then by the condition of Fact 2 we have that $\|q-p\|_{\mathcal{A}_k} \geq \epsilon'$, and the algorithm will output "NO" with probability at least $2/3$.

We remark that the value of k in Fact 2 is a natural complexity measure for the difference between two probability density functions in the class $\mathcal{C}$. It follows from the definition of the $\mathcal{A}_k$ distance that this value corresponds to the number of "essential" crossings between f_1 and f_2 – i.e., the number of crossings between the functions f_1 and f_2 that significantly affect their L_1 distance. Intuitively, the number of essential crossings – as opposed to the domain size – is, in some sense, the "right" parameter to characterize the sample complexity of L_1 closeness testing for $\mathcal{C}$.

Distribution Family	Our upper bound	Previous work
t-piecewise constant	$O\big(\max\big\{\frac{t^{4/5}}{\epsilon^{6/5}}, \frac{t^{1/2}}{\epsilon^2}\big\}\big)$	$O\big(\frac{t}{\epsilon^2}\big)$ [CDSS14]
t-piecewise degree-d	$O\big(\max\big\{\frac{(t(d+1))^{4/5}}{\epsilon^{6/5}}, \frac{(t(d+1))^{1/2}}{\epsilon^2}\big\}\big)$	$O\big(\frac{t(d+1)}{\epsilon^2}\big)$ [CDSS14]
log-concave	$O\big(\frac{1}{\epsilon^{9/4}}\big)$	$O\big(\frac{1}{\epsilon^{5/2}}\big)$ [CDSS14]
k-mixture of log-concave	$O\big(\max\big\{\frac{k^{4/5}}{\epsilon^{8/5}}, \frac{k^{1/2}}{\epsilon^{9/4}}\big\}\big)$	$O\big(\frac{k}{\epsilon^{5/2}}\big)$ [CDSS14]
t-modal over $[n]$	$O\big(\max\big\{\frac{(t\log n)^{4/5}}{\epsilon^2}, \frac{(t\log n)^{1/2}}{\epsilon^{5/2}}\big\}\big)$	$O\big(\frac{(t\log n)^{2/3}}{\epsilon^{8/3}} + \frac{t^2}{\epsilon^4}\big)$ [DDS$^+$13]
MHR over $[n]$	$O(\max\big\{\frac{\log(n/\epsilon)^{4/5}}{\epsilon^2}, \frac{\log(n/\epsilon)^{1/2}}{\epsilon^{5/2}}\big\})$	$O\big(\frac{\log(n/\epsilon)}{\epsilon^3}\big)$ [CDSS14]

Table 1: Algorithmic results for closeness testing of selected families of structured probability distributions. The second column indicates the sample complexity of our general algorithm applied to the class under consideration. The third column indicates the sample complexity of the best previously known algorithm for the same problem.

The upper bound implied by the above fact is information-theoretically optimal for a wide range of structured distribution classes $\mathcal{C}$. In particular, our bounds apply to all the structured distribution families considered in [CDSS14, DKN15, ADLS15] including (arbitrary mixtures of) t-flat (i.e., piecewise constant with t pieces), t-piecewise degree-d polynomials, t-monotone, monotone hazard rate, and log-concave distributions. For t-flat distributions we obtain an L_1 closeness testing algorithm that uses $O(\max\{t^{4/5}/\epsilon^{6/5}, t^{1/2}/\epsilon^2\})$ samples, which is the first $o(t)$ sample algorithm for the problem. For log-concave distributions, we obtain a sample size of $O(\epsilon^{-9/4})$ matching the information–theoretic lower bound even for the case that one of the distributions is explicitly given [DKN15]. Table 1 summarizes our upper bounds for a selection of natural and well-studied distribution families. These results are obtained from Theorem 1 and Fact 2, via the appropriate structural approximation results [CDSS13, CDSS14].

We would like to stress that our algorithm and its analysis are very different than previous results in the property testing literature. We elaborate on this point in the following subsection.

2.3 Our Techniques In this subsection, we provide a high-level overview of our techniques in tandem with a comparison to prior work.

Our upper bound is achieved by an explicit, sample near-linear-time algorithm. A good starting point for considering this problem would be the testing algorithm of [DKN15], which deals with the case where p is an *explicitly known* distribution. The basic idea of the testing algorithm in this case [DKN15] is to partition the domain into intervals in several different ways, and run a known L_2 tester on the reduced distributions (with respect to the intervals in the partition) as a black-box. At a high-level, these intervals partitions can be constructed by exploiting our knowledge of p, in order to divide our domain into several equal mass intervals under p. It can be shown that if p and q have large $\mathcal{A}_k$ distance from each other, one of these partitions will be able to detect the difference.

Generalizing this algorithm to the case where p is *unknown* turns out to be challenging, because there seems to be no way to find the appropriate interval partitions with $o(k)$ samples. If we allowed ourselves to take $\Omega(k/\epsilon)$ samples from p, we would be able to approximate an appropriate interval partition, and make the aforementioned approach go through. Alas, this would not lead to an

$o(k)$ sample algorithm. If we can only draw m samples from our distributions, the best that we could hope to do would be to use our samples in order to partition the domain into $m+1$ interval regions. This, of course, is not going to be sufficient to allow an analysis along the lines of the above approach to work. In particular, if we partition our domain *deterministically* into $m = o(k)$ intervals, it may well be the case that the reduced distributions over those intervals are identical, despite the fact that the original distributions have large $\mathcal{A}_k$ distance. In essence, the differences between p and q may well cancel each other out on the chosen intervals.

However, it is important to note that our interval boundaries are *not* deterministic. This suggests that unless we get unlucky, the discrepancy between p and q will not actually cancel out in our partition. As a slight modification of this idea, instead of partitioning the domain into intervals (which we expect to have only $O(1)$ samples each) and comparing the number of samples from p versus q in each, we sort our samples and test how many of them came from the same distribution as their neighbors (with respect to the natural ordering on the real line).

We intuitively expect that, if $p = q$, the number of pairs of ordered samples drawn from the same distribution versus a different one will be the same. Indeed, this can be formalized and the completeness of this tester is simple to establish. The soundness analysis, however, is somewhat involved. We need to show that the expected value of the statistic that we compute is larger than its standard deviation. While the variance is easy to bound from above, bounding the expectation is quite challenging. To do so, we define a function, $f(t)$, that encodes how likely it is that the samples nearby point t come from one distribution or the other. It turns out that f satisfies a relatively nice differential equation, and relates in a clean way to the expectation of our statistic. From this, we can show that any discrepancy between p and q taking place on a scale too short to be detected by the above partitioning approach will yield a notable contribution to our expectation.

The analysis of our lower bound begins by considering a natural class of testers, namely those that take some number of samples from p and q, sort the samples (while keeping track of which distribution they came from) and return an output that depends only on the ordering of these samples. For such testers we exhibit explicit families of pairs of distributions that are hard to distinguish from being identical. There is a particular pattern that appears many times in these examples, where there is a small interval for which q has an appropriate amount of probability mass, followed by an interval of p, followed by another interval of q. When the parameters are balanced correctly, it can be shown that when at most two samples are drawn from this subinterval, the distribution on their orderings is indistinguishable from the case where $p = q$. By constructing distributions with many copies of the pattern, we essentially show that a tester of this form will not be able to be confident that $p \neq q$, unless there are many of these small intervals from which it draws three or more samples. On the other hand, a simple argument shows that this is unlikely to be the case.

The above lower bound provides explicit distributions that are hard to distinguish from being identical by any tester in this limited class. To prove a lower bound against general testers, we proceed via a reduction: we show that an order–based tester can be derived from any general tester. It should be noted that this makes our lower bound in a sense non-constructive, as we do not know of any explicit families of distributions that are hard to distinguish from uniform for general testers. In order to perform this reduction, we show that for a general tester we can find some large subset S of its domain such that if all samples drawn from p and q by the tester happen to lie in S, then the output of the tester will depend only on the ordering of the samples. This essentially amounts to a standard result from Ramsey theory. Then, by taking any other problem, we can embed it into our new sample space by choosing new p and q that are the same up to an order-preserving rearrangement of the domain (which will also preserve $\mathcal{A}_k$ distance), ensuring that they are supported only on S.

3 Algorithm for $\mathcal{A}_k$ Closeness Testing

In this section we provide the sample optimal closeness tester under the $\mathcal{A}_k$ distance.

3.1 An $O(k^{4/5}/\epsilon^{6/5})$-sample tester In this subsection we give a tester with sample complexity $O(k^{4/5}/\epsilon^{6/5})$ that applies for $\epsilon = \Omega(k^{-1/6})$. For simplicity, we focus on the case that we take samples from two unknown distributions with probability density functions $p, q : [0,1] \to \mathbb{R}_+$. Our results are easily seen to extend to discrete probability distributions.

Algorithm Simple-Test-Identity-$\mathcal{A}_k(p, q, \epsilon)$
Input: sample access to pdf's $p, q : [0,1] \to \mathbb{R}_+$, $k \in \mathbb{Z}_+$, and $\epsilon > 0$.
Output: "YES" if $q = p$; "NO" if $\|q - p\|_{\mathcal{A}_k} \geq \epsilon$.

1. Let $m = C \cdot (k^{4/5}/\epsilon^{6/5})$, for a sufficiently large constant C. Draw two sets of samples S_p, S_q each of size $\mathrm{Poi}(m)$ from p and from q respectively.
2. Merge S_p and S_q while remembering from which distribution each sample comes from. Let S be the union of S_p and S_q sorted in increasing order (breaking ties randomly).
3. Compute the statistic Z defined as follows:

$$Z \stackrel{\text{def}}{=} \# \text{ (pairs of successive samples in } S \text{ coming from the same distribution)} - \# \text{ (pairs of successive samples in } S \text{ coming from different distributions)}$$

4. If $Z > 3 \cdot (\sqrt{m})$ return "NO". Otherwise return "YES".

Proposition 3. *The algorithm* Simple-Test-Identity-$\mathcal{A}_k(p, q, \epsilon)$*, on input two samples each of size* $O(k^{4/5}/\epsilon^{6/5})$ *drawn from two distributions with densities* $p, q : [0,1] \to \mathbb{R}_+$*, an integer* $k > 2$*, and* $\epsilon = \Omega(k^{-1/6})$*, correctly distinguishes the case that* $q = p$ *from the case* $\|p - q\|_{\mathcal{A}_k} \geq \epsilon$*, with probability at least* $2/3$*.*

Proof. First, it is straightforward to verify the claimed sample complexity, since the algorithm only draws samples in Step 1. To simplify the analysis we make essential use of the following simple claim:

Claim 4. *We can assume without loss of generality that the pdf's* $p, q : [0,1] \to \mathbb{R}_+$ *are continuous functions bounded from above by* 2*.*

Proof. We start by showing we can assume that p, q are at most 2. Let $p, q : [0,1] \to \mathbb{R}_+$ be arbitrary pdf's. We consider the cumulative distribution function (CDF) Φ of the mixture $(p + q)/2$. Let $X \sim p$, $Y \sim q$, $W \sim (p + q)/2$ be random variables. Since Φ is non-decreasing, replacing X and Y by $\Phi(X)$ and $\Phi(Y)$ does not affect the algorithm (as the ordering on the samples remains the same). We claim that, after making this replacement, $\Phi(X)$ and $\Phi(Y)$ are continuous distributions with probability density functions bounded by 2. In fact, we will show that the sum of their probability density functions is exactly 2. This is because for any $0 \leqslant a \leqslant b \leqslant 1$,

$$\Pr[\Phi(X) \in [a, b]] + \Pr[\Phi(Y) \in [a, b]] = 2 \Pr[\Phi(W) \in [a, b]] = 2(b - a) \,,$$

where the second equality is by the definition of a CDF. Thus, we can assume that p and q are bounded from above by 2.

To show that we can assume continuity, note that p and q can be approximated by continuous density functions p' and q' so that the L_1 errors $\|p-p'\|_1, \|q-q'\|_1$ are each at most $1/(10m)$. If our algorithm succeeds with the continuous densities p' and q', it must also succeed for p and q. Indeed, since the L_1 distance between p and p' and q and q' is at most $1/(10m)$, a set of m samples taken from p or q are statistically indistinguishable to m samples taken from p' or q'. This proves that it is no loss of generality to assume that p and q are continuous. □

Note that the algorithm makes use of the well-known "Poissonization" approach. Namely, instead of drawing $m = O(k^{4/5}/\epsilon^{6/5})$ samples from p and from q, we draw $m' = \mathrm{Poi}(m)$ samples from p and $m'' = \mathrm{Poi}(m)$ sample from q. The crucial properties of the Poisson distribution are that it is sharply concentrated around its mean and it makes the number of times different elements occur in the sample independent.

We now establish completeness. Note that our algorithm draws $\mathrm{Poi}(2m)$ samples from p or q. If $p = q$, then our process equivalently selects $\mathrm{Poi}(2m)$ values from p and then randomly and independently with equal probability decides whether or not each sample came from p or from q. Making these decisions one at a time in increasing order of points, we note that each adjacent pair of elements in S randomly and independently contributes either a $+1$ or a -1 to Z. Therefore, the distribution of Z is exactly that of a sum of $\mathrm{Poi}(2m) - 1$ independent $\{\pm 1\}$ random variables. Therefore, Z has mean 0 and variance $2m-1$. By Chebyshev's inequality it follows that $|Z| \leqslant 3\sqrt{m}$ with probability at least $7/9$. This proves completeness.

We now proceed to prove the soundness of our algorithm. Assuming that $\|p-q\|_{\mathcal{A}_k} > \epsilon$, we want to show that the value of Z is at most $3 \cdot \sqrt{m}$ with probability at most $1/3$. To prove this statement, we will again use Chebyshev's inequality. In this case it suffices to show that $\mathbb{E}[Z] \gg \sqrt{\mathrm{Var}[Z]} + \sqrt{m}$ for the inequality to be applicable. We begin with an important definition.

Definition 5. Let $f : [0,1] \to [-1,1]$ equal

$$\begin{aligned} f(t) \stackrel{\text{def}}{=} \quad & \Pr\left[\text{largest sample in } S \text{ that is at most } t \text{ was drawn from } p\right] \\ & - \Pr\left[\text{largest sample in } S \text{ that is at most } t \text{ was drawn from } q\right] . \end{aligned}$$

The importance of this function is demonstrated by the following lemma.

Lemma 6. *We have that:* $\mathbb{E}[Z] = m \int_0^1 f(t)(p(t) - q(t))dt$.

Proof. Given an interval I, we let Z_I be the contribution to Z coming from pairs of consecutive points of S the larger of which is drawn from I. We wish to approximate the expectation of Z_I. We let $\tau(I) = m(p(I) + q(I))$ be the expected total number of points drawn from I. We note that the contribution coming from cases where more than one point is drawn from I is $O(\tau(I)^2)$. We next consider the contribution under the condition that only one sample is drawn from I. For this, we let EP_I and EQ_I be the events that the largest element of S preceding I comes from p or q respectively. We have that the expected contribution to Z_I coming from events where exactly one element of S is drawn from I is:

$$\begin{aligned} &(\Pr[\mathrm{EP}_I] - \Pr[\mathrm{QP}_I]) \Pr(\text{The only element drawn from } I \text{ is from } p) \\ -&(\Pr[\mathrm{EP}_I] - \Pr[\mathrm{QP}_I]) \Pr(\text{The only element drawn from } I \text{ is from } q). \end{aligned}$$

Letting x_I be the left endpoint of I, this is

$$f(x_I)(mp(I) - mq(I)) + O(\tau(I)^2).$$

Therefore,

$$\mathbb{E}[Z_I] = f(x_I)(mp(I) - mq(I)) + O(\tau(I)^2).$$

Letting $\mathcal{I}$ be a partition of our domain into intervals, we find that

$$\begin{aligned}\mathbb{E}[Z] &= \sum_{I\in\mathcal{I}} \mathbb{E}[Z_I] \\ &= \sum_{I\in\mathcal{I}} f(x_I)(mp(I) - mq(I)) + O(\tau(I)^2) \\ &= O(m \max_{I\in\mathcal{I}} \tau(I)) + \sum_{I\in\mathcal{I}} f(x_I)(mp(I) - mq(I)).\end{aligned}$$

As the partition $\mathcal{I}$ becomes iteratively more refined, these sums approach Riemann sums for the integral of

$$mf(x)(p(x) - q(x))dx.$$

Therefore, taking a limit over partitions $\mathcal{I}$, we have that

$$\mathbb{E}[Z] = m \int f(x)(p(x) - q(x))dx.$$

□

We will also make essential use of the following technical lemma:

Lemma 7. *The function f is differentiable with derivative $f'(t) = m\left(p(t) - q(t) - (p(t) + q(t))f(t)\right)$.*

Proof. Consider the difference between $f(t)$ and $f(t+h)$ for some small $h > 0$. We note that $f(t) = \mathbb{E}[F_t]$ where F_t is 1 if the sample of S preceding t came from p, -1 if the sample came from q, and 0 if no sample came before t. Note that

$$F_{t+h} = \begin{cases} F_t & \text{if no samples from } p \text{ nor } q \text{ are drawn from } [t, t+h] \\ 1 & \text{if one sample from } p \text{ and none from } q \text{ are drawn from } [t, t+h] \\ -1 & \text{if one sample from } q \text{ and none from } p \text{ are drawn from } [t, t+h] \\ \pm 1 & \text{if at least two samples from } p \text{ or } q \text{ are drawn from } [t, t+h]. \end{cases}$$

Since p and q are continuous at $t \in [0,1]$, these four events happen with probabilities $1 - mh(p(t) + q(t)) + o(h)$, $mhp(t) + o(h)$, $mhq(t) + o(h)$, $o(h)$, respectively. Therefore, taking an expectation we find that $f(t+h) = f(t)(1 - mh(p(t) + q(t))) + mh(p(t) - q(t)) + o(h)$. This, and a similar relation relating $f(t)$ to $f(t-h)$, proves that f is differentiable with the desired derivative. □

To analyze the desired expectation, $\mathbb{E}[Z]$, we consider the quantity $\int_0^1 f'(t)f(t)dt = (1/2)\left(f^2(1) - f^2(0)\right)$. Substituting f' from Lemma 7 above gives

$$\int_0^1 f'(t)f(t)dt = m\int_0^1 f(t)(p(t) - q(t))dt - m\int_0^1 f^2(t)(p(t) + q(t))dt.$$

Combining this with Lemma 6, we get

$$\mathbb{E}[Z] = m\int_0^1 f^2(t)(p(t) + q(t))dt + f^2(1)/2\,. \tag{1}$$

The second term in (1) above is $O(1)$, so we focus our attention to bound the first term from below. To do this, we consider intervals $I \subset [0,1]$ over which $|p(I) - q(I)|$ is "large" and show that they must produce some noticeable contribution to the first term. Fix such an interval I. We want to show that f^2 is large somewhere in I. Intuitively, we attempt to prove that on at least one of the endpoints of the interval, the value of f is big. Since f does not vary too rapidly, f^2 will be large on some large fraction of I. Formally, we have the following lemma:

Lemma 8. *For $\delta > 0$, let $I \subset [0,1]$ be an interval with $|p(I) - q(I)| = \delta$ and $p(I) + q(I) < 1/m$. Then, there exists an $x \in I$ such that $|f(x)| \geqslant \frac{m\delta}{3}$.*

Proof. Suppose for the sake of contradiction that $|f(x)| < m\delta/3$ for all $x \in I = [X, Y]$. Then, we have that

$$\begin{aligned} 2m\delta/3 &> |f(X) - f(Y)| = \left| \int_X^Y f'(t)dt \right| = \left| \int_X^Y \left(m(p(t) - q(t)) - mf(t)(p(t) + q(t)) \right) dt \right| \\ &= \left| m(p(I) - q(I)) - m \int_X^Y f(t)(p(t) + q(t))dt \right| \geqslant m|p(I) - q(I)| - m \left| \int_X^Y f(t)(p(t) + q(t))dt \right| \\ &> m\delta - m \int_X^Y (m\delta/3)\,(p(t) + q(t))dt = m\delta\,(1 - m(p(I) + q(I))/3) > 2m\delta/3\,, \end{aligned}$$

which yields the desired contradiction. □

We are now able to show that the contribution to $\mathbb{E}[Z]$ coming from such an interval is large.

Lemma 9. *Let I be an interval satisfying the hypotheses of Lemma 8. Then*

$$\int_I f^2(t)(p(t) + q(t))dt = \Omega(m^2\delta^3)\,.$$

Proof. By Lemma 8, f is large at some point x of the interval $I = [X, Y]$. Without loss of generality, we assume that $p([X, x]) + q([X, x]) \leqslant (p(I) + q(I))/2$. Let $I' = [x, Y']$ be the interval so that $p(I') + q(I') = \delta/9$. Note that $I' \subset I$ (since by assumption $|p(I) - q(I)| > \delta$ and thus $p(I) + q(I) > \delta$). Furthermore, note that since with probability at least $1 - m\delta/9$, no samples from S lie in I', we have that for all z in I' it holds $|f(x) - f(z)| \leqslant 2m\delta/9$, so $|f(z)| \geqslant m\delta/9$. Therefore,

$$\begin{aligned} \int_I f^2(t)(p(t) + q(t))dt &\geqslant \int_{I'} f^2(t)(p(t) + q(t))dt \geqslant \int_{I'} \left(\frac{m\delta}{9} \right)^2 (p(t) + q(t))dt \\ &= \frac{m^2\delta^2}{81}(p(I') + q(I')) = \frac{m^2\delta^3}{729}\,. \end{aligned}$$

□

Since $\|p - q\|_{\mathcal{A}_k} > \epsilon$, there is a partition $\mathcal{I}$ of $[0,1]$ into k intervals so that $\|p_r^{\mathcal{I}} - q_r^{\mathcal{I}}\|_1 > \epsilon$. By subdividing intervals further if necessary, we can guarantee that $\mathcal{I}$ has at most $3k$ intervals, $\|p_r^{\mathcal{I}} - q_r^{\mathcal{I}}\| > \epsilon$, and for each subinterval $I \in \mathcal{I}$ it holds $p(I), q(I) \leqslant 1/k$. For each such interval $I \in \mathcal{I}$, let $\delta_I = |p(I) - q(I)|$. Note that $\sum_{I \in \mathcal{I}} \delta_I \geqslant \epsilon$.

By (1) we have that

$$\begin{aligned} \mathbb{E}[Z] &= m \sum_{I \in \mathcal{I}} \int_I f^2(t)(p(t) + q(t))dt + O(1) \\ &= \Omega\left(m \sum_{I \in \mathcal{I}} m^2 \delta_I^3 \right) = \Omega\left(m^3 (\sum_{I \in \mathcal{I}} \delta_I)^3/(3k)^2 \right) \\ &= \Omega\left(m^3\epsilon^3/k^2 \right) = \Omega(C^{5/2}\sqrt{m})\,. \end{aligned}$$

We note that the second to last line above follows by Hölder's inequality. It remains to bound from above the variance of Z.

Lemma 10. *We have that* $\mathrm{Var}[Z] = O(m)$.

Proof. We divide the domain $[0,1]$ into m intervals I_i, $i = 1, \ldots, m$, each of total mass $2/m$ under the sum-distribution $p + q$. Consider the random variable X_i denoting the contribution to Z coming from pairs of adjacent samples in S such that the right sample is drawn from I_i. Clearly, $Z = \sum_{i=1}^{m} X_i$ and $\mathrm{Var}[Z] = \sum_{i=1}^{m} \mathrm{Var}[X_i] + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j)$.

To bound the first sum, note that the number of pairs of S in an interval I_i is no more than the number of samples drawn from I_i, and the variance of X_i is less than the expectation of the square of the number of samples from I_i. Since the number of samples from I_i is a Poisson random variables with parameter 2, we have $\mathrm{Var}[X_i] = O(1)$. This shows that $\sum_{i=1}^{m} \mathrm{Var}[X_i] = O(m)$.

To bound the sum of covariance, consider X_i and X_j conditioned on the samples drawn from intervals other than I_i and I_j. Note that if any sample is drawn from an intermediate interval, X_i and X_j are uncorrelated, and otherwise their covariance is at most $\sqrt{\mathrm{Var}(X_i)\mathrm{Var}(X_j)} = O(1)$. Since the probability that no sample is drawn from any intervening interval decreases exponentially with their separation, it follows that $\mathrm{Cov}(X_i, X_j) = O(1) \cdot e^{-\Omega(|j-i|)}$. This completes the proof. □

An application of Chebyshev's inequality completes the analysis of the soundness and the proof of Proposition 3. □

3.2 The General Tester In this section, we present a tester whose sample complexity is optimal (up to constant factors) for all values of ϵ and k, thereby establishing the upper bound part of Theorem 1. Our general tester (Algorithm Test-Identity-$\mathcal{A}_k$) builds on the tester presented in the previous subsection (Algorithm Simple-Test-Identity-$\mathcal{A}_k$). It is not difficult to see that the latter algorithm can fail once ϵ becomes sufficiently small, if the discrepancy between p and q is concentrated on intervals of mass larger than $1/m$. In this scenario, the tester Simple-Test-Identity-$\mathcal{A}_k$ will not take sufficient advantage of these intervals. To obtain our enhanced tester Test-Identity-$\mathcal{A}_k$, we will need to combine Simple-Test-Identity-$\mathcal{A}_k$ with an alternative tester when this is the case. Note that we can easily bin the distributions p and q into intervals of total mass approximately $1/m$ by taking m random samples. Once we do this, we can use an identity tester similar to that in our previous work [DKN15] to detect the discrepancy in these intervals. In particular we show the following:

Proposition 11. *Let p, q be discrete distributions over $[n]$ satisfying $\|p\|_2, \|q\|_2 = O(1/\sqrt{n})$. There exists a testing algorithm with the following properties: On input $k \in \mathbb{Z}_+$, $2 \leq k \leq n$, and $\delta, \epsilon > 0$, the algorithm draws $O\left((\sqrt{k}/\epsilon^2) \cdot \log(1/\delta)\right)$ samples from p and q and with probability at least $1 - \delta$ distinguishes between the cases $p = q$ and $\|p - q\|_{\mathcal{A}_k} > \epsilon$.*

The above proposition says that the identity testing problem under the $\mathcal{A}_k$ distance can be solved with $O(\sqrt{k}/\epsilon^2)$ samples when both distributions p and q are promised to be "nearly" uniform (in the sense that their L_2 norm is $O(1)$ times that of the uniform distribution). To prove Proposition 11 we follow a similar approach as in [DKN15]: Starting from the L_2 identity tester of [CDVV14], we consider several oblivious interval decompositions of the domain into intervals of approximately the same mass, and apply a "reduced" identity tester for each such decomposition. The details of the analysis establishing Proposition 11 are postponed to Appendix A.

We are now ready to present our general testing algorithm:

Algorithm Test-Identity-$\mathcal{A}_k(p, q, \epsilon)$
Input: sample access to distributions $p, q : [0, 1] \to \mathbb{R}_+$, $k \in \mathbb{Z}_+$, and $\epsilon > 0$.
Output: "YES" if $q = p$; "NO" if $\|q - p\|_{\mathcal{A}_k} \geq \epsilon$.

1. Let $m = Ck^{4/5}/\epsilon^{6/5}$, for a sufficiently large constant C. Draw two sets of samples S_p, S_q each of size $\mathrm{Poi}(m)$ from p and from q respectively.
2. Merge S_p and S_q while remembering from which distribution each sample comes from. Let S be the union of S_p and S_q sorted in increasing order (breaking ties randomly).
3. Compute the statistic Z defined as follows:

$$Z \stackrel{\text{def}}{=} \# \text{ (pairs of successive samples in } S \text{ coming from the same distribution)} - \# \text{ (pairs of successive samples in } S \text{ coming from different distributions)}$$

4. If $Z > 5\sqrt{m}$ return "NO".
5. Repeat the following steps $O(C)$ times:
   (a) Draw $\mathrm{Poi}(m)$ samples from $(p + q)/2$.
   (b) Split the domain into intervals with the interval endpoints given by the above samples. Let p' and q' be the reduced distributions with respect to these intervals.
   (c) Run the tester of Proposition 11 on p' and q' with error probability $1/C^2$ to determine if $\|p' - q'\|_{\mathcal{A}_{2k+1}} > \epsilon/C$. If the output of this tester is "NO", output "NO".
6. Output "YES".

Our main result for this section is the following:

Theorem 12. *Algorithm Test-Identity-$\mathcal{A}_k$ draws $O(\max\{k^{4/5}/\epsilon^{6/5}, k^{1/2}/\epsilon^2\})$ samples from p, q and with probability at least $2/3$ returns "YES" if $p = q$ and "NO" if $\|p - q\|_{\mathcal{A}_k} > \epsilon$.*

Proof. First, it is easy to see that the sample complexity of the algorithm is $O(m + k^{1/2}/\epsilon^2)$. Recall that we can assume that p, q are continuous pdf's bounded from above by 2.

We start by establishing completeness. If $p = q$, it is once again the case that $\mathbb{E}[Z] = 0$ and $\mathrm{Var}[Z] < 2m$, so by Chebyshev' s inequality, Step 4 will fail with probability at most $1/9$. Next when taking our samples in Step 5(a), note that the expected samples size is $O(m)$ and that the expected squared L_2 norms of the reduced distributions p' and q' are $O(1/m)$. Therefore, with probability at least $1 - 1/C^2$, p' and q' satisfy the hypothesis of Proposition 11. Hence, this holds for all C iterations with probability at least $8/9$.

Conditioning on this event, since $p' = q'$, the tester in Step 5(c) will return "YES" with probability at least $1 - 1/C^2$ on each iteration. Therefore, it returns "YES" on all iterations with probability at least $8/9$. By a union bound, it follows that if $p = q$, our algorithm returns "YES" with probability at least $2/3$.

We now proceed to establish soundness. Suppose that $\|p - q\|_{\mathcal{A}_k} \geqslant \epsilon$. Then there exists a partition $\mathcal{I}$ of the domain into k intervals such that $\|p_r^{\mathcal{I}} - q_r^{\mathcal{I}}\| \geqslant \epsilon$. For an interval $I \in \mathcal{I}$, let $\delta(I) = |p(I) - q(I)|$. We will call an $I \in \mathcal{I}$ small if there is a subinterval $J \subseteq I$ so that $p(J) + q(J) < 1/m$ and $|p(J) - q(J)| \geqslant \delta(I)/3$. We will call I large otherwise. Note that $\sum_{I \in \mathcal{I}, I \text{ small}} \delta(I) +$

$\sum_{I\in\mathcal{I},I \text{ large}} \delta(I) = \sum_{I\in\mathcal{I}} \delta(I) \geqslant \epsilon$. Therefore either $\sum_{I\in\mathcal{I},I \text{ small}} \delta(I) \geqslant \epsilon/2$, or $\sum_{I\in\mathcal{I},I \text{ large}} \delta(I) \geqslant \epsilon/2$. We analyze soundness separately in each of these cases.

Consider first the case that $\sum_{I\in\mathcal{I},I \text{ small}} \delta(I) \geqslant \epsilon/2$. The analysis in this case is very similar to the soundness proof of Proposition 3 which we describe for the sake of completeness.

By definition, for each small interval I, there exists a subinterval J so that $p(J)+q(J) < 1/m$ and $|p(J)-q(J)| > \delta(I)/2$. By Lemma 9, for such J we have that $\int_J f^2(t)(p(t)+q(t))dt = \Omega(m^2\delta^3(I))$, and therefore, that $\int_I f^2(t)(p(t)+q(t))dt = \Omega(m^2\delta^3(I))$. Hence, we have that

$$\begin{aligned}
\mathbb{E}[Z] &\geqslant m\int_0^1 f^2(t)(p(t)+q(t))dt \\
&\geqslant \sum_{I\in\mathcal{I},I \text{ small}} m\int_I f^2(t)(p(t)+q(t))dt \\
&\geqslant \sum_{I\in\mathcal{I},I \text{ small}} \Omega(m^3\delta^3(I)) \\
&\geqslant \Omega(m^3)\left(\sum_{I\in\mathcal{I},I \text{ small}} \delta(I)\right)^3/k^2 \\
&= \Omega(m^3\epsilon^3/k^2) \\
&= \Omega(C^{5/2}\sqrt{m}).
\end{aligned}$$

On the other hand, Lemma 10 gives that $\mathrm{Var}[Z] = O(m)$, so for C sufficiently large, Chebyshev's inequality implies that with probability at least $2/3$ it holds $Z > 5\sqrt{m}$. That is, our algorithm outputs "NO" with probability at least $2/3$.

Now consider that case that $\sum_{I\in\mathcal{I},I \text{ large}} \delta(I) \geqslant \epsilon/2$. We claim that the second part of our tester will detect the discrepancy between p and q with high constant probability. Once again, we can say that with probability at least $8/9$ the squared L_2 norms of the reduced distributions p' and q' are both $O(1/m)$ and that the size of the reduced domain is $O(m)$. Thus, the conditions of Proposition 11 are satisfied on all iterations with probability at least $8/9$. To complete the proof, we will show that with constant probability we have $\|p'-q'\|_{\mathcal{A}_{2k+1}} > \epsilon/C$. To do this, we construct an explicit partition $\mathcal{I}'$ of our reduced domain into at most $2k+1$ intervals so that with constant probability $\|p_r^{\mathcal{I}'} - q_r^{\mathcal{I}'}\|_1 > \epsilon/C$. This will imply that with probability at least $8/9$ that on at least one of our C trials that $\|p_r^{\mathcal{I}'} - q_r^{\mathcal{I}'}\|_1 > \epsilon/C$.

More specifically, for each interval $I \in \mathcal{I}$ we place interval boundaries at the smallest and largest sample points taken from I in Step 5(a) (ignoring them if fewer than two samples landed in I). Since we have selected at most $2k$ points, this process defines a partition $\mathcal{I}'$ of the domain into at most $2k+1$ intervals. We will show that the reduced distributions $p'' = p_r^{\mathcal{I}'}$ and $q'' = q_r^{\mathcal{I}'}$ have large expected L_1 error.

In particular, for each interval $I \in \mathcal{I}$ let I' be the interval between the first and last sample points of I. Note that I' is an interval in the partition $\mathcal{I}'$. We claim that if I is large, then with constant probability

$$|p(I') - q(I')| = \Omega(\delta(I)).$$

Let $I = [X, Y]$ and $I' = [x, y]$ (so x and y are the smallest and largest samples taken from I, respectively). We note that if $p([X,x]) + q([X,x]) < 1/m$ and $p([y,Y]) + q([y,Y]) < 1/m$ then

$$|p(I')-q(I')| \geqslant |p(I)-q(I)|-|p([X,x])-q([X,x])|-|p([y,Y])-q([y,Y])| \geqslant \delta(I)-\delta(I)/3-\delta(I)/3 = \delta(I)/3,$$

where the second inequality uses the fact that I is large. On the other hand, we note that $p([X,x])+q([X,x])$ and $p([y,Y])+q([y,Y])$ are exponential distributions with mean $1/m$, and thus, this event happens with constant probability. Let N_I be the indicator random variable for the event that $|p(I')-q(I')| \geqslant \delta(I)/3$. We have that

$$\|p''-q''\|_1 \geqslant \sum_I N_I\delta(I)/3 \geqslant \sum_{I\in\mathcal{I},I \text{ large}} N_I\delta(I)/3.$$

Thus, we have that

$$\|p''-q''\|_1 \geqslant \sum_{I\in\mathcal{I},I \text{ large}} \delta(I)/3 - \sum_{I\in\mathcal{I},I \text{ large}} (1-N_I)\delta(I)/3.$$

Therefore, since

$$\mathbb{E}\left[\sum_{I\in\mathcal{I},I \text{ large}} (1-N_I)\delta(I)/3\right] < \left(\sum_{I\in\mathcal{I},I \text{ large}} \delta(I)/3\right)(1-c)$$

for some fixed $c>0$, we have that with constant probability that

$$\|p''-q''\|_1 \geqslant c\sum_{I\in\mathcal{I},I \text{ large}} \delta(I)/3 \geqslant c\epsilon/6 \geqslant \epsilon/C.$$

This means that with probability at least $8/9$ for at least one iteration we will have that $\|p'-q'\|_{\mathcal{A}_{2k+1}} > \epsilon/C$, and therefore, with probability at least $2/3$, our algorithm outputs "NO".

□

4 Lower Bound for $\mathcal{A}_k$ Closeness Testing

Our upper bound from Section 3 seems potentially suboptimal. Instead of obtaining an upper bound of $O(\max\{k^{2/3}/\epsilon^{4/3}, k^{1/2}/\epsilon^2\})$, which would be analogical to the unstructured testing result of [CDVV14], we obtain a very different bound of $O(\max\{k^{4/5}/\epsilon^{6/5}, k^{1/2}/\epsilon^2\})$. In this section we show, surprisingly, that our upper bound is optimal for continuous distributions, or discrete distributions with support size n that is sufficiently large as a function of k.

Intuitively, our lower bound proof consists of two steps. In the first step, we show it is no loss of generality to assume that an optimal algorithm only considers the ordering of the samples, and ignores all other information. In the second step, we construct a pair of distributions which is hard to distinguish given the condition that the tester is only allowed to look at the ordering of the samples and nothing more.

Our first step is described in the following theorem. We note that unlike the arguments in the upper bound proofs, this part of our lower bound technique will work best for random variables of discrete support.

Theorem 13. *For all $n,k,m \in \mathbb{Z}_+$ there exists $N \in \mathbb{Z}_+$ such that the following holds: If there exists an algorithm A that for every pair of distributions p and q, supported over $[N]$, distinguishes the case $p=q$ from the case $\|p-q\|_{\mathcal{A}_k} \geqslant \epsilon$ drawing m samples, then there exists an algorithm A' that for every pair of distributions p' and q' supported on $[n]$ distinguishes the case $p'=q'$ versus $\|p'-q'\|_{\mathcal{A}_k} \geqslant \epsilon$ using the same number samples m. Moreover, A' only considers the ordering of the samples and ignores all other information.*

Proof. As a preliminary simplification, we assume that our algorithm, instead of taking m samples from any combination of p or q of its choosing, takes exactly m samples from p and m samples from q, as such algorithms are strictly more powerful. This also allows us to assume that the algorithm merely takes these random samples and applies some processing to determine its output.

As a critical tool of our proof, we will use the classical Ramsey theorem for hypergraphs. For completeness, we restate it here in a slightly adapted form.

Lemma 14 (Ramsey theorem for hypergraphs, [CFS10])**.** *Given a set S and an integer t let $\binom{S}{t}$ denote the set of subsets of S of cardinality t. For all positive integers, a, b and c, there exists a positive integer N so that for any function $f : \binom{[N]}{a} \to [b]$, there exists an $S \subset [N]$ with $|S| = c$ so that f is constant on $\binom{S}{a}$.*

In words, this means that if we color all subsets of size a of a size N set with at most b different colors, then for large enough N we will find a (bigger) subset T such that all its subsets are colored with the same color. Note that in our setting c from the theorem equals n.

The idea of our proof is as follows. Given an algorithm A, we will use it to implement the algorithm A'. Given A, we produce some monotonic function $f : [n] \to [N]$, and run A on the distributions $f(p)$ and $f(q)$. Since f is order preserving, $\|f(p) - f(q)\|_{\mathcal{A}_k} = \|p - q\|_{\mathcal{A}_k}$, so our algorithm is guaranteed to work. The tricky part will be to guarantee that the output of this new algorithm A' depends only on the ordering of the samples that it takes. Since we may assume that A is deterministic, once we pick which $2m$ samples are taken from $[N]$ the output will be some function of the ordering of these samples (and in particular which are from p and which are from q). For the algorithm A, this function may depend upon the values that the samples happened to have. Thus, for A' to depend only on order, we need it to be the case that A behaves the same way on any subset of $\mathrm{Im}(f)$ of size $2m$. Fortunately, we can find such a set using Lemma 14.

Since our sample set has size at most $2m$, it is clear that the total number of possible sample sets is at most N^{2m}. We color each of these subsets of $[N]$ of size $a = 2m$ one of a finite number of colors. The color associates to the sample set the function that A uses to obtain an output given $2m$ samples given by this set coming in a particular order (some of which are potentially equal). The total number of such functions is at most $b = 2^{2^{4m}}$. We let n be the proposed support size for p' and q'. By Lemma 14, for N sufficiently large, there are sets of size n such that the function has the same value in samples from these sets. Letting f be the unique monotonic function from $[n]$ to $[N]$ with this set as its image, causes the output A' to depend only on the ordering of the samples.

The above reduction works as long as the samples given to our algorithm A' are distinct. To deal with the case where samples are potentially non-distinct, we show that it is possible to reduce to the case where all $2m$ samples are distinct with $9/10$ probability. To do this, we divide each of our original bins into $200m^2$ sub-bins, and upon drawing a sample from a given bin, we assign it instead to a uniformly random sub-bin. This procedure maintains the $\mathcal{A}_k$ distance between our distributions, and guarantees that the probability of a collision is small. Now, our algorithm A' will depend only on the order of the samples so long as there is no collision. As this happens with probability $9/10$, we can also ensure that this is the case when collisions do occur without sacrificing correctness. This completes our proof. □

We will now give the "hard" instance of the testing problem for algorithms that only consider the ordering of the samples. We will first describe a construction that works for $\epsilon = \Omega(k^{-1/6})$. We define a mini-bucket to be a segment I, which can be divided into three subsegments I_1, I_2, I_3 in that order so that $p(I_1) = p(I_3) = \epsilon/(2k)$, $p(I_2) = 0$, and $q(I_1) = q(I_3) = 0, q(I_2) = \epsilon/k$. We define a bucket to be an interval consisting of a mini-bucket followed by an interval on which $p = q$ and on

which both p, q have total mass $(1-\epsilon)/k$. Our distributions for p and q will consist of k consecutive buckets. See Figure 1 for an illustration.

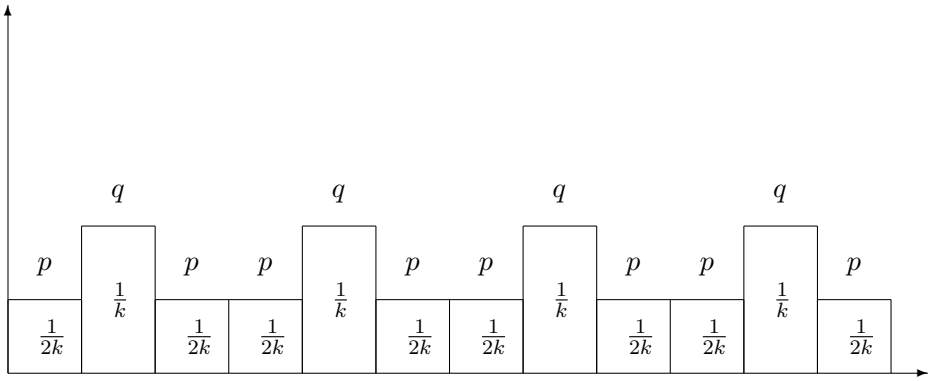

Figure 1: $\phi = \frac{1}{2}p + \frac{1}{2}q$ when $\epsilon = 1$

Next consider partitioning the domain into macro-buckets each of which is a union of buckets of total mass $\Theta(1/m)$. Note that these distributions have $\mathcal{A}_{2k+1}$ distance of 2ϵ. An important fact to note is the following:

Observation 15. *If zero, one or two draws are made randomly and independently from $(p+q)/2$ on a mini-bucket, then the distribution of which of p or q the samples came from and their relative ordering is indistinguishable from the case where $p = q$.*

To prove the lower bound for the algorithm A', which is only allowed to look at the ordering of samples. We let X be a random variable that is taken to be 0 or 1 each with probabilty 1/2. When $X = 0$ we define p and q as above with mini-buckets, macro-buckets and regular buckets as described. When $X = 1$, we let $p = q$ and define mini-buckets to have total mass ϵ/k for each of p and q, buckets to have total mass $1/k$ each, and we combine buckets into macro-buckets as in the $X = 0$ case.

Let Y be the distribution on the (ordered) sequences, obtained by drawing $m' = \mathrm{Poi}(m)$ samples from p and $m'' = \mathrm{Poi}(m)$ samples from q, with p and q given by X. We are interested in bounding the mutual information between X and Y, since it must be $\Omega(1)$ if the algorithm is going to succeed with probability bounded away from 1/2. We show the following:

Theorem 16. *We have that $I(X : Y) = O(m^5\epsilon^6/k^4)$.*

Proof. We begin with a couple of definitions. Let Y' denote (Y, α), where α is the information about which draws come from which macro-bucket. Y' consists of Y'_i, the sequence of samples coming from the i-th macro-bucket. Note that

$$I(X : Y) \leqslant I(X : Y') \leqslant \sum_{i=1}^{O(m)} I(X : Y'_i) \, .$$

We will now estimate $I(X : Y'_i)$. We claim that it is $O(\frac{m^4\epsilon^6}{k^4})$ for each i. This would cause the sum to be small enough and give our theorem. We have that,

$$I(X : Y'_i) = \mathbb{E}_y \left[O\left(1 - \frac{\Pr(Y'_i = y|X = 0)}{\Pr(Y'_i = y|X = 1)}\right)^2 \right].$$

We then have that

$$I(X : Y_i') = \sum_{\ell=0}^{\infty} \sum_{y:|y|=\ell} \frac{O(1)^\ell}{\ell!} O\left(1 - \frac{\Pr(Y_i' = y | X = 0, |y| = \ell)}{\Pr(Y_i' = y | X = 1, |y| = \ell)}\right)^2.$$

We note that if $X = 1, |y| = \ell$ that any of the 2^ℓ possible orderings are equally likely. On the other hand, if $X = 0$, this also holds in an approximate sense. To show this, first consider picking which mini-buckets our ℓ draws are from. If no three land in the same mini-bucket, then Observation 15 implies that all orderings are equally likely. Therefore, the statistical distance between $Y_i'|X = 0, |y| = \ell$ and $Y_i'|X = 1, |y| = \ell$ is at most the probability that some three draws come from the same mini-bucket. This is in turn at most the expected number of triples that land in the same mini-bucket, which is equal to $\binom{\ell}{3}$ times the probability that a particular triple does. The probability of landing in a particular mini-bucket is $O(m\epsilon/k)^3$. By definition, there are $O(m/k)$ mini-buckets in a macro-bucket, so this probability is $O(\ell^3\epsilon^3(m/k)^2)$. Therefore, we have that

$$\begin{aligned}
I(X : Y_i') &= \sum_{\ell} \frac{O(1)^\ell}{\ell!} \sum_{y:|y|=\ell} O(4^\ell) \left(\Pr(Y_i' = y | X = 0, |y| = \ell) - \Pr(Y_i' = y | X = 1, |y| = \ell)\right)^2 \\
&\leqslant \sum_{\ell} \frac{O(1)^\ell}{\ell!} \left(\sum_{y:|y|=\ell} \left|\Pr(Y_i' = y | X = 0, |y| = \ell) - \Pr(Y_i' = y | X = 1, |y| = \ell)\right| \right)^2 \\
&= \sum_{\ell} \frac{O(1)^\ell}{\ell!} O\left(\ell^6 \epsilon^6 m^4 / k^4\right) \\
&= \frac{m^4}{k^4} \sum_{\ell} \frac{O(1)^\ell \ell^6 \epsilon^6}{\ell!} \\
&= O\left(\frac{m^4 \epsilon^6}{k^4}\right).
\end{aligned}$$

This completes our proof. □

The above construction only works when $k \geqslant m$, or equivalently, when $\epsilon = \Omega(k^{-1/6})$. When ϵ is small, we need a slightly different construction. We will similarly split our domain into mini-buckets and macro-buckets and argue based on shared information. Once again we define two distributions p and q, though this time the distributions themselves will need to be randomized. Given k and ϵ, we begin by splitting the domain into k macro-buckets. Each macro-bucket will have mass $1/k$ under both p and q.

First pick a global variable X to be either 0 or 1 with equal probability. If $X = 1$ then we will have $p = q$ and if $X = 0$, $\|p - q\|_{\mathcal{A}_{2k+1}} = \epsilon$. For each macro-bucket, pick an x uniformly in $[0, (1 - \epsilon)/k]$. The macro-bucket will consist of an interval on which $p = q$ with mass x (for each of p, q), followed by a mini-bucket, followed by an interval of mass $(1 - \epsilon)/k - x$ on which $p = q$. The mini-bucket is an interval of mass ϵ/k under either p or q. If $X = 1$, we have $p = q$ on the mini-bucket. If $X = 0$, the mini-bucket consists of an interval of mass $\epsilon/(2k)$ under q and 0 under p, an interval of mass ϵ/k under p and 0 under q, and then another interval of mass $\epsilon/(2k)$ under q and 0 under p.

We let Y be the random variable associated with the ordering of elements from a set of $\mathrm{Poi}(m)$ draws from each of p and q. We show:

Theorem 17. *If $m\epsilon = O(k)$, $\log(mk/\epsilon) = O(\epsilon^{-1})$, and $k = O(m)$, with implied constants sufficiently small, then $I(X : Y) = O(m^5\epsilon^6/k^4)$.*

Note that the above statement differs from Theorem 16 in that X and Y are defined differently.

Proof. Once again, we let Y' be Y along with the information of which draws came from which macro-bucket, and let Y_i' be the information of the draws from the i-th macro-bucket along with their ordering. It suffices for us to show that $I(X : Y_i') = O(m^5\epsilon^6k^{-5})$ for each i (as now there are only k macro-buckets rather than m).

Let s be a string of ℓ ordered draws from p and q. In particular, we may consider s to be a string $s_1s_2\dots s_\ell$, where $s_i \in \{p, q\}$. We wish to consider the probability that $Y_i' = s$ under the conditions that $X = 0$ or that $X = 1$. In order to do this, we further condition on which elements of s were drawn from the mini-bucket. For $1 \leqslant a \leqslant b \leqslant \ell$ we consider the probability that not only did we obtain sequence s, but that the draws $s_a, \dots, s_b$ were exactly the ones coming from the mini-bucket within this macro-bucket. Let h denote the ordered string coming from elements drawn from the mini-bucket and M the ordered sequence of strings coming from elements not drawn from the mini-bucket. The probability of the event in question is then

$$\Pr(h = s_a \dots s_b)\Pr(M = s_1 \dots s_{a-1}s_{b+1} \dots s_\ell)\Pr(\text{the mini-bucket is placed between } s_{a-1} \text{ and } s_{b+1}).$$

Note that the mini-bucket can be thought of as being randomly and uniformly inserted within an interval of length $(1-\epsilon)/k$ and that this is equally likely to be inserted between any pair of elements of M. Thus, the probability of the third term in the product is exactly $1/(\ell + a - b)$. The second probability is the probability that $\ell + a - b - 1$ elements are drawn from the complement of the mini-bucket times $2^{-(\ell+a-b+1)}$, as draws from p and q are equally likely. Thus, letting $t = b - a + 1$ (i.e., the number of elements in the mini-bucket), we have that

$$\Pr(Y_i' = s) = e^{-m/k}\sum_{t=0}^{\ell}\left(\frac{\left(\frac{m(1-\epsilon)}{2k}\right)^{\ell-t}}{(\ell-t)!}\right)\left(\frac{\left(\frac{m\epsilon}{k}\right)^t}{t!}\right)\left(\frac{1}{\ell-t}\right)\sum_a \Pr(h = s_a \dots s_{a+t-1} : |h| = t).$$

Note that this equality holds even after conditioning upon X. We next simplify this expression further by grouping together terms in the last sum based upon the value of the substring $s_a \dots s_{a+t-1}$, which we call r. We get that

$$\Pr(Y_i' = s) = e^{-m/k}\sum_{t=0}^{\ell}\left(\frac{\left(\frac{m(1-\epsilon)}{2k}\right)^{\ell-t}}{(\ell-t)!}\right)\left(\frac{\left(\frac{m\epsilon}{k}\right)^t}{t!}\right)\left(\frac{1}{\ell-t}\right)\sum_{|r|=t}\Pr(h = r : |h| = t)N_{r,s},$$

where $N_{r,s}$ is the number of occurrences of r as a substring of s.

Next, we wish to bound

$$\sum_{|s|=\ell}|\Pr(Y_i' = s : X = 0) - \Pr(Y_i' = s : X = 1)|^2. \tag{2}$$

By the above formula this is at most

$$e^{-2m/k}\sum_{|s|=\ell}\sum_{t=0}^{\ell}\left(\frac{\left(\frac{m(1-\epsilon)}{2k}\right)^{\ell-t}}{(\ell-t)!}\right)\left(\frac{\left(\frac{m\epsilon}{k}\right)^t}{t!}\right)\left(\frac{1}{\ell-t}\right)\cdot$$

$$\left|\sum_{|r|=t}N_{r,s}\left(\Pr(h = r : |h| = t, X = 0) - \Pr(h = r : |h| = t, X = 1)\right)\right|^2.$$

For fixed values of t we consider the sum

$$\sum_{|s|=\ell}\left|\sum_{|r|=t} N_{r,s}(\Pr(h=r:|h|=t,X=0)-\Pr(h=r:|h|=t,X=1))\right|^2.$$

Note that if $t \leqslant 2$ then $\Pr(h=r:|h|=t,X=0)=\Pr(h=r:|h|=t,X=1)$, and so the above sum is 0. Otherwise, it is at most

$$\sum_{|s|=\ell}\sum_{|r|=t}|N_{r,s}-(\ell+1-t)/2^t|^2$$

because $\sum_r \Pr(h=r:|h|=t,X=0)=\sum_r \Pr(h=r:|h|=t,X=1)=1$. Note on the other hand that the expectation over random strings s of length ℓ of $N_{r,s}-(\ell+1-t)/2^t$ is 0. Furthermore, the variance of $N_{r,s}$ is easily bounded by $t\ell 2^{-t}$ as whether or not two disjoint substrings of s are equal to r are independent events. Therefore, the above sum is at most

$$2^\ell 2^t t\ell 2^{-t}=2^\ell t\ell.$$

Hence, by Cauchy-Schwartz, we have that

$$\sum_{|s|=\ell}\left|\sum_{|r|=t} N_{r,s}-(\ell+1-t)/2^t\right|^2 \leqslant 2^\ell 2^t t\ell.$$

Therefore, the expression in (2) is at most

$$e^{-2m/k}\left(\sum_{t=3}^{\ell}\left(\frac{\left(\frac{m(1-\epsilon)}{2k}\right)^{\ell-t}}{(\ell-t)!}\right)\left(\frac{O\left(\frac{m\epsilon}{k}\right)^t}{t!}\right)\left(\frac{1}{\ell-t}\right)\left(2^\ell 2^t\ell t\right)^{1/2}\right)^2.$$

Assuming that $\ell\epsilon$ is sufficiently small, these terms are decreasing exponentially with t, and thus this is

$$O\left(e^{-2m/k}\left(\frac{(m^2/(2k^2))^\ell}{(\ell!)^2}\right)\epsilon^6\ell^5\right).$$

Now we have that for N a sufficiently small constant times ϵ^{-1},

$$\begin{aligned}
I(X:Y_i') &= \sum_s \Pr(Y_i'=s:X=1)O\left(1-\frac{\Pr(Y_i'=s:X=0)}{\Pr(Y_i'=s:X=1)}\right)^2\\
&= \sum_\ell\sum_{s:|s|=\ell} e^{m/k}\left(\frac{(m/(2k))^\ell}{\ell!}\right)^{-1} O(\Pr(Y_i'=s:X=1)-\Pr(Y_i'=s:X=0))^2\\
&\leqslant \sum_\ell e^{m/k}\left(\frac{(m/(2k))^\ell}{\ell!}\right)^{-1} O\left(\sum_{s:|s|=\ell}|\Pr(Y_i'=s:X=1)-\Pr(Y_i'=s:X=0)|^2\right)\\
&\leqslant \sum_{\ell>N} O\left(\frac{(2m/k)^\ell}{\ell!}\right)+\sum_{\ell<N} e^{m/k}\left(\frac{(m/(2k))^\ell}{\ell!}\right)^{-1} O\left(e^{-2m/k}\left(\frac{(m^2/(2k^2))^\ell}{(\ell!)^2}\right)\epsilon^6\ell^5\right)\\
&\leqslant \sum_{\ell>N} O\left(\frac{m}{kN}\right)^\ell+\sum_\ell O\left(e^{-m/k}\frac{(m/k)^\ell}{\ell!}\epsilon^6\ell^5\right).
\end{aligned}$$

Since $\frac{m}{kN} \leq \frac{m\epsilon}{k}$ is sufficiently small, the first term is at most $(1/2)^N$ which is polynomially small in mk/ϵ, and thus negligible. The second term is the expectation of $\epsilon^6\ell^5$ for ℓ a Poisson random variable with mean m/k. Thus, it is easily seen to be $O((m/k)^5\epsilon^6)$. Therefore, we have that $I(X : Y_i') = O(m^5\epsilon^6k^{-5})$, and therefore, $I(X : Y) = O(m^5\epsilon^6k^{-4})$, as desired. □

We are now ready to complete the proof of our general lower bound.

Theorem 18. *For any $k > 2$, there exists an N so that any algorithm that is given sample access to two distributions, p and q over $[N]$, and can distinguish between the cases $p = q$ and $\|p - q\|_{\mathcal{A}_k}$ with probability at least $2/3$, requires at least $\Omega\left(\max\left\{k^{4/5}/\epsilon^{6/5}, k^{1/2}/\epsilon^2\right\}\right)$ samples.*

Proof. The lower bound of $k^{1/2}/\epsilon^2$ follows from the known lower bound [Pan08] even in the case where q is known and p and q have support of size k. It now suffices to consider the case that $\epsilon > k^{-1/2}$ and m a sufficiently small constant times $k^{4/5}\epsilon^{-6/5}$.

Note that by Theorem 13, we may assume that the algorithm in question takes m samples from each of p and q and determines its output based only on the ordering of the samples. We need to show that this is impossible for N sufficiently large.

We note that if we allow p and q to be continuous distributions instead of discrete ones we are already done. If $m < k$, we use our first counter-example construction, and if $m \geqslant k$ use the second one. If we let X be randomly 0 or 1, and set $p = q$ for $X = 1$ and p, q as described above when $X = 0$, then by Theorems 16 and 17, the shared information between X and the output of our algorithm is at most $O(m^5\epsilon^6k^{-4}) = o(1)$, and thus our algorithm cannot correctly determine X with constant probability.

In order to prove our Theorem, we will need to make this work for distributions p and q with finite support size as follows: By splitting our domain into m^3 intervals each of equal mass under $p + q$, we note that the $\mathcal{A}_k$ distance between the distributions is only negligibly affected. Furthermore, with high probability, m samples will have no pair chosen from the same bin. Thus, the distribution on orderings of samples from these discrete distributions are nearly identical to the continuous case, and thus our algorithm would behave nearly identically. This completes the proof. □

References

[ADJ+11] J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, and S. Pan. Competitive closeness testing. *Journal of Machine Learning Research - Proceedings Track*, 19:47–68, 2011.

[ADLS15] J. Acharya, I. Diakonikolas, J. Li, and L. Schmidt. Sample-optimal density estimation in nearly-linear time. *CoRR*, abs/1506.00671, 2015.

[Bat01] T. Batu. *Testing Properties of Distributions.* PhD thesis, Cornell University, 2001.

[BBBB72] R.E. Barlow, D.J. Bartholomew, J.M. Bremner, and H.D. Brunk. *Statistical Inference under Order Restrictions.* Wiley, New York, 1972.

[BD14] F. Balabdaoui and C. R. Doss. Inference for a Mixture of Symmetric Distributions under Log-Concavity. Available at http://arxiv.org/abs/1411.4708, 2014.

[BDKR02] T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld. The complexity of approximating entropy. In *ACM Symposium on Theory of Computing*, pages 678–687, 2002.

[BFF+01] T. Batu, E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White. Testing random variables for independence and identity. In *Proc. 42nd IEEE Symposium on Foundations of Computer Science*, pages 442–451, 2001.

[BFR+00] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing that distributions are close. In *IEEE Symposium on Foundations of Computer Science*, pages 259–269, 2000.

[BFR+13] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing closeness of discrete distributions. *J. ACM*, 60(1):4, 2013.

[Bir87a] L. Birgé. Estimating a density under order restrictions: Nonasymptotic minimax risk. *Annals of Statistics*, 15(3):995–1012, 1987.

[Bir87b] L. Birgé. On the risk of histograms for estimating decreasing densities. *Annals of Statistics*, 15(3):1013–1022, 1987.

[BKR04] T. Batu, R. Kumar, and R. Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *ACM Symposium on Theory of Computing*, pages 381–390, 2004.

[Bru58] H. D. Brunk. On the estimation of parameters restricted by inequalities. *The Annals of Mathematical Statistics*, 29(2):pp. 437–454, 1958.

[BRW09] F. Balabdaoui, K. Rufibach, and J. A. Wellner. Limit distribution theory for maximum likelihood estimation of a log-concave density. *The Annals of Statistics*, 37(3):pp. 1299–1331, 2009.

[BW07] F. Balabdaoui and J. A. Wellner. Estimation of a k-monotone density: Limit distribution theory and the spline connection. *The Annals of Statistics*, 35(6):pp. 2536–2564, 2007.

[BW10] F. Balabdaoui and J. A. Wellner. Estimation of a k-monotone density: characterizations, consistency and minimax lower bounds. *Statistica Neerlandica*, 64(1):45–70, 2010.

[CDSS13] S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Learning mixtures of structured distributions over discrete domains. In *SODA*, 2013.

[CDSS14] S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Efficient density estimation via piecewise polynomial approximation. In *STOC*, pages 604–613, 2014.

[CDVV14] S. Chan, I. Diakonikolas, P. Valiant, and G. Valiant. Optimal algorithms for testing closeness of discrete distributions. In *SODA*, pages 1193–1203, 2014.

[CFS10] D. Conlon, J. Fox, and B. Sudakov. Hypergraph ramsey numbers. *Journal of the American Mathematical Society*, 23(1):247–266, 2010.

[CS13] Y. Chen and R. J. Samworth. Smoothed log-concave maximum likelihood estimation with applications. *Statist. Sinica*, 23:1373–1398, 2013.

[CT04] K.S. Chan and H. Tong. Testing for multimodality with dependent data. *Biometrika*, 91(1):113–123, 2004.

[DDS+13] C. Daskalakis, I. Diakonikolas, R. Servedio, G. Valiant, and P. Valiant. Testing k-modal distributions: Optimal algorithms via reductions. In *SODA*, pages 1833–1852, 2013.

[DKN15] I. Diakonikolas, D. Kane, and V. Nikishkin. Testing identity of structured distributions. In *SODA*, pages 1841–1854, 2015.

[DR09] L. D umbgen and K. Rufibach. Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency. *Bernoulli*, 15(1):40–68, 2009.

[DW13] C. R. Doss and J. A. Wellner. Global Rates of Convergence of the MLEs of Log-concave and s-concave Densities. Available at http://arxiv.org/abs/1306.1438, 2013.

[Fou97] A.-L. Fougères. Estimation de densités unimodales. *Canadian Journal of Statistics*, 25:375–387, 1997.

[GGR98] O. Goldreich, S. Goldwasser, and D. Ron. Property testing and its connection to learning and approximation. *Journal of the ACM*, 45:653–750, 1998.

[GJ14] P. Groeneboom and G. Jongbloed. *Nonparametric Estimation under Shape Constraints: Estimators, Algorithms and Asymptotics*. Cambridge University Press, 2014.

[GR00] O. Goldreich and D. Ron. On testing expansion in bounded-degree graphs. Technical Report TR00-020, Electronic Colloquium on Computational Complexity, 2000.

[Gre56] U. Grenander. On the theory of mortality measurement. *Skand. Aktuarietidskr.*, 39:125–153, 1956.

[Gro85] P. Groeneboom. Estimating a monotone density. In *Proc. of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, pages 539–555, 1985.

[GW09] F. Gao and J. A. Wellner. On the rate of convergence of the maximum likelihood estimator of a k-monotone density. *Science in China Series A: Mathematics*, 52:1525–1538, 2009.

[HP76] D. L. Hanson and G. Pledger. Consistency in concave regression. *The Annals of Statistics*, 4(6):pp. 1038–1050, 1976.

[HW15] Q. Han and J. A. Wellner. Approximation and Estimation of s-Concave Densities via Renyi Divergences. Available at http://arxiv.org/abs/1505.00379, 2015.

[ILR12] P. Indyk, R. Levi, and R. Rubinfeld. Approximating and Testing k-Histogram Distributions in Sub-linear Time. In *PODS*, pages 15–22, 2012.

[JW09] H. K. Jankowski and J. A. Wellner. Estimation of a discrete monotone density. *Electronic Journal of Statistics*, 3:1567–1605, 2009.

[KM10] R. Koenker and I. Mizera. Quasi-concave density estimation. *Ann. Statist.*, 38(5):2998–3027, 2010.

[KS14] A. K. H. Kim and R. J. Samworth. Global rates of convergence in log-concave density estimation. Available at http://arxiv.org/abs/1404.2298, 2014.

[LR05] E. L. Lehmann and Joseph P. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, 2005.

[LRR11] R. Levi, D. Ron, and R. Rubinfeld. Testing properties of collections of distributions. In *ICS*, pages 179–194, 2011.

[NP33] J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.

[Pan08] L. Paninski. A coincidence-based test for uniformity given very sparsely-sampled discrete data. *IEEE Transactions on Information Theory*, 54:4750–4755, 2008.

[Rao69] B.L.S. Prakasa Rao. Estimation of a unimodal density. *Sankhya Ser. A*, 31:23–36, 1969.

[RS96] R. Rubinfeld and M. Sudan. Robust characterizations of polynomials with applications to program testing. *SIAM J. on Comput.*, 25:252–271, 1996.

[Rub12] R. Rubinfeld. Taming big probability distributions. *XRDS*, 19(1):24–28, 2012.

[Val11] P. Valiant. Testing symmetric properties of distributions. *SIAM J. Comput.*, 40(6):1927–1968, 2011.

[VV11] G. Valiant and P. Valiant. Estimating the unseen: an $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs. In *STOC*, pages 685–694, 2011.

[VV14] G. Valiant and P. Valiant. An automatic inequality prover and instance optimal identity testing. In *FOCS*, 2014.

[Wal09] G. Walther. Inference and modeling with log-concave distributions. *Statistical Science*, 24(3):319–327, 2009.

[Weg70] E.J. Wegman. Maximum likelihood estimation of a unimodal density. I. and II. *Ann. Math. Statist.*, 41:457–471, 2169–2174, 1970.

Appendix

A Proof of Proposition 11

In this section, we prove Proposition 11. We note that it suffices to attain confidence probability $2/3$ with $O(\sqrt{k}/\epsilon^2)$ samples, as we can then run $O(\log(1/\delta))$ independent iterations to boost the confidence to $1-\delta$. Our starting point is the following Theorem from [CDVV14]:

Theorem 19 ([CDVV14], Proposition 3.1)**.** *For any distributions p and q over $[n]$ such that $\|p\|_2 \leqslant \frac{O(1)}{\sqrt{n}}$ and $\|q\|_2 \leqslant \frac{O(1)}{\sqrt{n}}$ there is a testing algorithm that distinguishes with probability at least $2/3$ the case that $q = p$ from the case that $||q-p||_2 \geqslant \epsilon/\sqrt{n}$ when given $O(\sqrt{n}/\epsilon^2)$ samples from q and p.*

Our $\mathcal{A}_k$ testing algorithm for this regime is the following:

Algorithm Test-Identity-Flat-$\mathcal{A}_k(p, q, n, \epsilon)$
Input: sample access to distributions p and q over $[n]$ with $\|p\|_2, \|q\|_2 = O(1/\sqrt{n})$, $k \in \mathbb{Z}_+$ with $2 \le k \le n$, and $\epsilon > 0$.
Output: "YES" if $q = p$; "NO" if $\|q - p\|_{\mathcal{A}_k} \ge \epsilon$.

1. Draw samples S_1, S_2 of size $m = O(\sqrt{k}/\epsilon^2)$ from q and p.
2. By artificially increasing the support if necessary, we can guarantee that $n = k \cdot 2^{j_0}$, where $j_0 \stackrel{\text{def}}{=} \lceil \log_2(1/\epsilon) \rceil + O(1)$.
3. Consider the collection $\{\mathcal{I}^{(j)}\}_{j=0}^{j_0-1}$ of j_0 partitions of $[n]$ into intervals; the partition $\mathcal{I}^{(j)} = (I_i^{(j)})_{i=1}^{\ell_j}$ consists of $\ell_j = k \cdot 2^j$ many intervals with $I_i^{(j_0)}$ of length $n/\ell_j + O(1)$, and $I_i^{(j)}$ the union of two adjacent intervals of $I_i^{(j+1)}$.
4. For $j = 0, 1, \ldots, j_0 - 1$:
   (a) Consider the reduced distributions $q_r^{\mathcal{I}^{(j)}}$ and $p_r^{\mathcal{I}^{(j)}}$. Use the samples S_1, S_2 to simulate samples to $q_r^{\mathcal{I}^{(j)}}$ and $p_r^{\mathcal{I}^{(j)}}$.
   (b) Run Test-Identity-$L_2(q_r^{\mathcal{I}^{(j)}}, p_r^{\mathcal{I}^{(j)}}, \ell_j, \epsilon_j, \delta_j)$ for $\epsilon_j = C \cdot \epsilon \cdot 2^{3j/8}$ for $C > 0$ a sufficiently small constant and $\delta_j = 2^{-j}/6$, i.e., test whether $q_r^{\mathcal{I}^{(j)}} = p_r^{\mathcal{I}^{(j)}}$ versus $\|q_r^{\mathcal{I}^{(j)}} - p_r^{\mathcal{I}^{(j)}}\|_2 > \gamma_j \stackrel{\text{def}}{=} \epsilon_j/\sqrt{\ell_j}$.
5. If all the testers in Step 3(b) output "YES", then output "YES"; otherwise output "NO".

Note in the above that when $\epsilon_j > 1$, that the appropriate tester requires no samples. The following proposition characterizes the performance of the above algorithm.

Proposition 20. *The algorithm* Test-Identity-Flat-$\mathcal{A}_k(p, q, n, \epsilon)$*, on input a sample of size* $m = O(\sqrt{k}/\epsilon^2)$ *drawn from distributions* q *and* p *over* $[n]$ *with* $\|p\|_2, \|q\|_2 = O(1/\sqrt{n})$*,* $\epsilon > 0$*, and an integer* k *with* $2 \le k \le n$*, correctly distinguishes the case that* $q = p$ *from the case that* $\|q - p\|_{\mathcal{A}_k} \ge \epsilon$*, with probability at least* $2/3$*.*

Proof. First, it is straightforward to verify the claimed sample complexity, as the algorithm only draws samples in Step 1. Note that the algorithm uses the same set of samples S_1, S_2 for all testers in Step 4(b). Note that it is easy to see that $\|p_r^{\mathcal{I}^{(j)}}\|_2, \|q_r^{\mathcal{I}^{(j)}}\|_2 = O(1/\sqrt{\ell_j})$, and therefore, by Theorem 19, the tester Test-Identity-$L_2(q_r^{\mathcal{I}^{(j)}}, p_r^{\mathcal{I}^{(j)}}, \ell_j, \epsilon_j, \delta_j)$, on input a set of $m_j = O((\sqrt{\ell_j}/\epsilon_j^2) \cdot \log(1/\delta_j))$ samples from $q_r^{\mathcal{I}^{(j)}}$ and $p_r^{\mathcal{I}^{(j)}}$ distinguishes the case that $q_r^{\mathcal{I}^{(j)}} = p_r^{\mathcal{I}^{(j)}}$ from the case that $\|q_r^{\mathcal{I}^{(j)}} - p_r^{\mathcal{I}^{(j)}}\|_2 \ge \gamma_j \stackrel{\text{def}}{=} \epsilon_j/\sqrt{\ell_j}$ with probability at least $1 - \delta_j$. From our choice of parameters it can be verified that $\max_j m_j \le m = O(\sqrt{k}/\epsilon^2)$, hence we can use the same sample S_1, S_2 as input to these testers for all $0 \le j \le j_0 - 1$. In fact, it is easy to see that $\sum_{j=0}^{j_0-1} m_j = O(m)$, which implies that the overall algorithm runs in sample-linear time. Since each tester in Step 3(b) has error probability δ_j, by a union bound over all $j \in \{0, \ldots, j_0 - 1\}$, the total error probability is at most $\sum_{j=0}^{j_0-1} \delta_j \le (1/6) \cdot \sum_{j=0}^{\infty} 2^{-j} = 1/3$. Therefore, with probability at least $2/3$ all the testers in Step 4(b) succeed. We will henceforth condition on this "good" event, and establish the completeness and soundness properties of the overall algorithm under this conditioning.

We start by establishing completeness. If $q = p$, then for any partition $\mathcal{I}^{(j)}$, $0 \le j \le j_0 - 1$, we have that $q_r^{\mathcal{I}^{(j)}} = p_r^{\mathcal{I}^{(j)}}$. By our aforementioned conditioning, all testers in Step 3(b) will output "YES", hence the overall algorithm will also output "YES", as desired.

We now proceed to establish the soundness of our algorithm. Assuming that $\|q - p\|_{\mathcal{A}_k} \ge \epsilon$, we

want to show that the algorithm Test-Identity-$\mathcal{A}_k(q, n, \epsilon)$ outputs "NO" with probability at least 2/3. Towards this end, we prove the following structural lemma:

Lemma 21. *For $C > 0$ a sufficiently small constant, if $\|q - p\|_{\mathcal{A}_k} \geq \epsilon$, there exists $j \in \mathbb{Z}_+$ with $0 \leq j \leq j_0 - 1$ such that $\|q_r^{\mathcal{I}^{(j)}} - p_r^{\mathcal{I}^{(j)}}\|_2^2 \geq \gamma_j^2$.*

Given the lemma, the soundness property of our algorithm follows easily. Indeed, since all testers Test-Identity-$L_2(q_r^{\mathcal{I}^{(j)}}, \ell_j, \epsilon_j, \delta_j)$ of Step 4(b) are successful by our conditioning, Lemma 21 implies that at least one of them outputs "NO", hence the overall algorithm will output "NO". □

The proof of Lemma 21 is very similar to the analogous lemma in [DKN15]. For the same of completeness, it is given in the following subsection.

A.1 Proof of Lemma 21 We claim that it is sufficient to take $C \leqslant 5 \cdot 10^{-6}$. Thus, we are in the case where $n = 2^{j_0 - 1} \cdot k$ and have argued that it suffices to show that our algorithm works to distinguish $\mathcal{A}_k$-distance in this setting with $\epsilon_j = 10^{-5} \cdot \epsilon \cdot 2^{3j/8}$.

We make use of the following definition:

Definition 22. For p and q arbitrary distributions over $[n]$, we define the *scale-sensitive-L_2 distance* between q and p to be

$$\|q - p\|_{[k]}^2 \stackrel{\text{def}}{=} \max_{\mathcal{I}=(I_1,\ldots,I_r)\in \mathbf{W}_{1/k}} \sum_{i=1}^{r} \frac{\mathbf{Discr}^2(I_i)}{\mathbf{width}^{1/8}(I_i)}$$

where $\mathbf{W}_{1/k}$ is the collection of all interval partitions of $[n]$ into intervals of width at most $1/k$, $\mathbf{Discr}(I) = |p(I) - q(I)|$, and $\mathbf{width}(I)$ is the number of bins in I divided by n.

The first thing we need to show is that if q and p have large $\mathcal{A}_k$ distance then they also have large scale-sensitive-L_2 distance. Indeed, we have the following lemma:

Lemma 23. *For p and q an arbitrary distributions over $[n]$, we have that*

$$\|q - p\|_{[k]}^2 \geqslant \frac{\|q - p\|_{\mathcal{A}_k}^2}{(2k)^{7/8}}.$$

Proof. Let $\epsilon = \|q - p\|_{\mathcal{A}_k}^2$. Consider the optimal $\mathcal{I}^*$ in the definition of the $\mathcal{A}_k$ distance. By further subdividing intervals of width more than $1/k$ into smaller ones, we can obtain a new partition, $\mathcal{I}' = (I_i')_{i=1}^s$, of cardinality $s \leq 2k$ all of whose parts have width at most $1/k$. Furthermore, we have that $\sum_i \mathbf{Discr}(I_i') \geqslant \epsilon$. Using this partition to bound from below $\|q - p\|_{[k]}^2$, by Cauchy-Schwarz we obtain that

$$\begin{aligned}
\|q - p\|_{[k]}^2 &\geqslant \sum_i \frac{\mathbf{Discr}^2(I_i')}{\mathbf{width}(I_i')^{1/8}} \\
&\geqslant \frac{(\sum_i \mathbf{Discr}(I_i'))^2}{\sum_i \mathbf{width}(I_i')^{1/8}} \\
&\geqslant \frac{\epsilon^2}{2k(1/(2k))^{1/8}} \\
&= \frac{\epsilon^2}{(2k)^{7/8}}.
\end{aligned}$$

□

The second important fact about the scale-sensitive-L_2 distance is that if it is large then one of the partitions considered in our algorithm will produce a large L_2 error.

Proposition 24. *Let p and q be distributions over $[n]$. Then we have that*

$$\|q-p\|_{[k]}^2 \leqslant 10^8 \sum_{j=0}^{j_0-1} \sum_{i=1}^{2^j \cdot k} \frac{\mathbf{Discr}^2(I_i^{(j)})}{\mathbf{width}^{1/8}(I_i^{(j)})}. \tag{3}$$

Proof. Let $\mathcal{J} \in \mathbf{W}_{1/k}$ be the optimal partition used when computing the scale-sensitive-L_2 distance $\|q-p\|_{[k]}$. In particular, it is a partition into intervals of width at most $1/k$ so that $\sum_i \frac{\mathbf{Discr}^2(J_i)}{\mathbf{width}(J_i)^{1/8}}$ is as large as possible. To prove (3), we prove a notably stronger claim. In particular, we will prove that for each interval $J_\ell \in \mathcal{J}$

$$\frac{\mathbf{Discr}^2(J_\ell)}{\mathbf{width}^{1/8}(J_\ell)} \leqslant 10^8 \sum_{j=0}^{j_0-1} \sum_{i:I_i^{(j)} \subset J_\ell} \frac{\mathbf{Discr}^2(I_i^{(j)})}{\mathbf{width}^{1/8}(I_i^{(j)})}. \tag{4}$$

Summing over ℓ would then yield $\|q-p\|_{[k]}^2$ on the left hand side and a strict subset of the terms from (3) on the right hand side. From here on, we will consider only a single interval J_ℓ. For notational convenience, we will drop the subscript and merely call it J.

First, note that if $|J| \leqslant 10^8$, then this follows easily from considering just the sum over $j = j_0-1$. Then, if $t = |J|$, J is divided into t intervals of size 1. The sum of the discrepancies of these intervals equals the discrepancy of J, and thus, the sum of the squares of the discrepancies is at least $\mathbf{Discr}^2(J)/t$. Furthermore, the widths of these subintervals are all smaller than the width of J by a factor of t. Thus, in this case the sum of the right hand side of (4) is at least $1/t^{7/8} \geqslant \frac{1}{10^7}$ of the left hand side.

Otherwise, if $|J| > 10^8$, we can find a j so that $\mathbf{width}(J)/10^8 < 1/(2^j \cdot k) \leqslant 2 \cdot \mathbf{width}(J)/10^8$. We claim that in this case Equation (4) holds even if we restrict the sum on the right hand side to this value of j. Note that J contains at most 10^8 intervals of $\mathcal{I}^{(j)}$, and that it is covered by these intervals plus two narrower intervals on the ends. Call these end-intervals R_1 and R_2. We claim that $\mathbf{Discr}(R_i) \leqslant \mathbf{Discr}(J)/3$. This is because otherwise it would be the case that

$$\frac{\mathbf{Discr}^2(R_i)}{\mathbf{width}^{1/8}(R_i)} > \frac{\mathbf{Discr}^2(J)}{\mathbf{width}^{1/8}(J)}.$$

(This is because $(1/3)^2 \cdot (2/10^8)^{-1/8} > 1$.) This is a contradiction, since it would mean that partitioning J into R_i and its complement would improve the sum defining $\|q-p\|_{[k]}$, which was assumed to be maximum. This means that the sum of the discrepancies of the $I_i^{(j)}$ contained in J must be at least $\mathbf{Discr}(J)/3$, so the sum of their squares is at least $\mathbf{Discr}^2(J)/(9 \cdot 10^8)$. On the other hand, each of these intervals is narrower than J by a factor of at least $10^8/2$, thus the appropriate sum of $\frac{\mathbf{Discr}^2(I_i^{(j)})}{\mathbf{width}^{1/8}(I_i^{(j)})}$ is at least $\frac{\mathbf{Discr}^2(J)}{10^8 \mathbf{width}^{1/8}(J)}$. This completes the proof. □

We are now ready to prove Lemma 21.

Proof. If $\|q-p\|_{\mathcal{A}_k} \geqslant \epsilon$ we have by Lemma 23 that

$$\|q-p\|_{[k]}^2 \geqslant \frac{\epsilon^2}{(2k)^{7/8}}.$$

By Proposition 24, this implies that

$$\begin{aligned}\frac{\epsilon^2}{(2k)^{7/8}} &\leqslant 10^8 \sum_{j=0}^{j_0-1} \sum_{i=1}^{2^j \cdot k} \frac{\mathbf{Discr}^2(I_i^{(j)})}{\mathbf{width}^{1/8}(I_i^{(j)})} \\ &= 10^8 \sum_{j=0}^{j_0-1} (2^j k)^{1/8} \|q^{\mathcal{I}^{(j)}} - p^{\mathcal{I}^{(j)}}\|_2^2.\end{aligned}$$

Therefore,

$$\sum_{j=0}^{j_0-1} 2^{j/8} \|q^{\mathcal{I}^{(j)}} - p^{\mathcal{I}^{(j)}}\|_2^2 \geqslant 5 \cdot 10^{-9} \epsilon^2/k. \tag{5}$$

On the other hand, if $\|q^{\mathcal{I}^{(j)}} - p^{\mathcal{I}^{(j)}}\|_2^2$ were at most $10^{-10} 2^{-j/4} \epsilon^2/k$ for each j, then the sum above would be at most

$$10^{-10} \epsilon^2/k \sum_j 2^{-j/8} < 5 \cdot 10^{-9} \epsilon^2/k.$$

This would contradict Equation (5), thus proving that $\|q^{\mathcal{I}^{(j)}} - U_{\ell_j}\|_2^2 \geqslant 10^{-10} 2^{-j/4} \epsilon^2/k$ for at least one j, proving Lemma 21. □

A.3 Testing Closeness of Structured Distributions over Discrete Domains

This paper still awaits submission, thus the author names are not written.

Testing Closeness of Structured Distributions over Discrete Domains

Anonymous Author(s)
Affiliation
Address
email

Abstract

We investigate the problem of testing the equivalence between two structured discrete distributions. Let $\mathcal{D}$ be a family of distributions over a discrete support of size n. Given a set of samples from two distributions $p, q \in \mathcal{D}$, we want to distinguish (with high probability) between the cases that $p = q$ and $\|p - q\|_1 \geq \epsilon$. The main contribution of this paper is a new general algorithm for this testing problem and a nearly matching information-theoretic lower bound. Specifically, the sample complexity of our algorithm matches our lower bound up to logarithmic factors. As a corollary, we obtain new near-sample optimal equivalence testers for a wide range of discrete structured distributions in a unified way.

1 Introduction

1.1 Background and Motivation

The prototypical inference question in the area of *distribution property testing* [1] is the following: Given a set of samples from a collection of probability distributions, can we determine whether these distributions satisfy a certain property? During the past two decades, this broad question – whose roots lie in statistical hypothesis testing [2, 3] – has received considerable attention by the computer science community, see [4, 5] for two recent surveys. The majority of work in this field has focused on characterizing the sample size needed to test properties of arbitrary distributions of a given support size. After two decades of study, this "worst-case" regime is well-understood: for many properties of interest there exist sample-optimal testers (matched by information-theoretic lower bounds) [6, 7, 8, 9, 10].

In many settings of interest, we know a priori that the underlying distributions have some "nice structure" (exactly or approximately). For example, we may have some qualitative information about the shape of the underlying densities, e.g., they may be mixtures of a small number of log-concave distributions, or multi-modal distributions with a small number of modes, etc. The problem of *learning* a probability distribution under such shape constraints is a classical topic in statistics – starting with the pioneering work of Grenander [11]– that has recently attracted the interest of computer scientists [12, 13, 14, 15, 16]. See [17] for a classical book, and [18] for a recent book on the topic. Perhaps surprisingly, the theory of *distribution testing* under shape constraints is less fully developed: A recent sequence of works [19, 9, 20] leverages such structural assumptions to obtain more efficient testers for a number of natural settings. However, for many natural properties of interest either no non-trivial testers are yet known or there is a large gap between our sample upper and lower bounds.

In this work, we focus on the problem of testing equivalence (closeness) between two discrete *structured* distributions. Let $\mathcal{D}$ be a family over univariate distributions over $[n]$ (or $\mathbb{Z}$). The problem of *closeness testing for* $\mathcal{D}$ is the following: Given sample access to two unknown distribution $p, q \in \mathcal{D}$,

Submitted to 29th Conference on Neural Information Processing Systems (NIPS 2016). Do not distribute.

we want to distinguish between the case that $p = q$ versus $\|p - q\|_1 \geq \epsilon$. (Here, $\|p - q\|_1$ denotes the ℓ_1-distance between the distributions p, q.) The sample complexity of this problem depends on the underlying family $\mathcal{D}$. For example, if $\mathcal{D}$ is the class of *all* distributions over $[n]$, then it is known [7] that the optimal sample complexity is $\Theta(\max\{n^{2/3}/\epsilon^{4/3}, n^{1/2}/\epsilon^2\})$.

The aforementioned sample complexity cannot be improved only if the family $\mathcal{D}$ is arbitrary, and we may be able to obtain significantly improved upper bounds under shape constraints. For example, if both p, q are promised to be (approximately) log-concave over $[n]$, there is an algorithm to test equivalence between them using $O(1/\epsilon^{9/4})$ samples [20]. Note that this sample bound is independent of the support size n, and is dramatically better than the worst-case tight bound [7] when n is large.

More generally, [20] described a framework to obtain efficient equivalence testers for various families of structured distributions over both continuous and discrete domains. While the results of [20] are sample-optimal for some families of distributions (in particular, over continuous domains), it was not known whether they can be improved for natural families of discrete distributions. In this paper, we work in the framework of [20], and obtain new improved algorithms for the discrete case and nearly-matching lower bounds (within logarithmic factors). We elaborate on our contributions in the subsection below.

1.2 Our Results and Comparison to Prior Work

We work in the general framework introduced by [9, 20]. Instead of designing a different equivalence tester for any given family $\mathcal{D}$, the approach of [9, 20] proceeds by designing a generic equivalence tester under a *different metric* than the ℓ_1-distance. This metric, termed $\mathcal{A}_k$-distance [21], where $k \geq 2$ is a positive integer, interpolates between Kolmogorov distance (when $k = 2$) and the ℓ_1-distance (when $k = n$). It turns out that, for most structured distribution families $\mathcal{D}$, the $\mathcal{A}_k$-distance can be used as a proxy for the ℓ_1-distance for a value of $k \ll n$. We can obtain an ℓ_1 closeness tester for $\mathcal{D}$ by plugging in the right value of k in a general $\mathcal{A}_k$ closeness tester.

To explain our results in detail, we will need some terminology.

Notation. We will use p, q to denote the probability mass functions of our distributions. If p is discrete over support $[n] := \{1, \ldots, n\}$, we denote by p_i the probability of element i in the distribution. For two discrete distributions p, q, their ℓ_1 and ℓ_2 distances are $\|p - q\|_1 = \sum_{i=1}^n |p_i - q_i|$ and $\|p - q\|_2 = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$. Fix a partition of the domain I into disjoint intervals $\mathcal{I} := (I_i)_{i=1}^{\ell}$. For such a partition $\mathcal{I}$, the *reduced distribution* $p_r^{\mathcal{I}}$ corresponding to p and $\mathcal{I}$ is the discrete distribution over $[\ell]$ that assigns the i-th "point" the mass that p assigns to the interval I_i; i.e., for $i \in [\ell]$, $p_r^{\mathcal{I}}(i) = p(I_i)$. Let $\mathfrak{J}_k$ be the collection of all partitions of the domain I into k intervals. For $p, q : I \to \mathbb{R}_+$ and $k \in \mathbb{Z}_+$, we define the $\mathcal{A}_k$-distance between p and q by $\|p - q\|_{\mathcal{A}_k} \stackrel{\text{def}}{=} \max_{\mathcal{I}=(I_i)_{i=1}^k \in \mathfrak{J}_k} \sum_{i=1}^k |p(I_i) - q(I_i)| = \max_{\mathcal{I} \in \mathfrak{J}_k} \|p_r^{\mathcal{I}} - q_r^{\mathcal{I}}\|_1$.

In this context, [20] showed that gave a closeness testing algorithm under the $\mathcal{A}_k$-distance using $O(\max\{k^{4/5}/\epsilon^{6/5}, k^{1/2}/\epsilon^2\})$ samples. It was also shown that this sample bound is information–theoretically optimal (up to constant factors) for continuous distributions, or discrete distributions of support size n sufficiently large as a function of k. This naturally raises the following open questions:

- The sample complexity of $\mathcal{A}_k$ closeness testing on $[n]$ depends on three parameters: $n, k, 1/\epsilon$. [20] obtained an upper bound that is independent of n and is optimal when $n \to \infty$. This leaves open the case of finite n. Observe that the $O(\max\{k^{4/5}/\epsilon^{6/5}, k^{1/2}/\epsilon^2\})$ upper bound of [20] is not optimal in all regimes. For example, it is worse than the $\Theta(\max\{n^{2/3}/\epsilon^{4/3}, n^{1/2}/\epsilon^2\})$ upper bound for $n = O(k)$. What is the sample complexity of the problem as a function of $n, k, 1/\epsilon$?

- As mentioned above, [20] shows a sample lower bound of $\Omega(\max\{k^{4/5}/\epsilon^{6/5}, k^{1/2}/\epsilon^2\})$ for $\mathcal{A}_k$-closeness testing. However, the hard instances do not correspond to a natural family of structured distributions, and in particular do not immediately yield ℓ_1-closeness testing lower bounds for such natural families. Can we obtain tight sample lower bounds for natural families of structured distributions?

We resolve both these open problems. Our main algorithmic result is the following:

Theorem 1 *Given sample access to distributions p and q on $[n]$ and $\epsilon > 0$ there exists an algorithm that takes*

$$O\left(\max\left(\min\left(k^{4/5}/\epsilon^{6/5}, k^{2/3}\log^{4/3}(3+n/k)\log\log(3+n/k)/\epsilon^{4/3}\right), k^{1/2}/\epsilon^2\right)\right)$$

samples from each of p and q and distinguishes with $2/3$ probability between the cases that $p = q$ and $\|p - q\|_{\mathcal{A}_k} \geq \epsilon$.

On the lower bound side, we show:

Theorem 2 *Let p and q be distributions on $[n]$ and let $\epsilon > 0$ be sufficiently small. Any tester that distinguishes between $p = q$ and $\|p - q\|_{\mathcal{A}_k}$ for some $k \leq n$ must use $\Omega(m)$ samples for $m = \min(k^{2/3}\log^{1/3}(3+n/k)/\epsilon^{4/3}, k^{4/5}/\epsilon^{6/5})$.*

In fact, this lower bound holds even if p and q are both guaranteed to be piecewise constant distributions on $O(k+m)$ pieces.

1.3 Overview of Techniques

To prove our upper bound, we use a technique of iteratively reducing the number of bins (domain elements). In particular, we show that if we merge bins together in consecutive pairs, this does not significantly affect the $\mathcal{A}_k$ distance between the distributions, unless a large fraction of the discrepancy between our distributions is supported on $O(k)$ bins near the boundaries in the optimal partition. In order to take advantage of this, we provide a novel identity tester that requires few samples to distinguish between the cases where $p = q$ and the case where p and q have a large ℓ_1 distance supported on only k of the bins. We are able to take advantage of the small support essentially because having a discrepancy supported on few bins implies that the ℓ_2 distance between the distributions must be reasonably large.

Our new lower bounds are somewhat more complicated. We prove them by exhibiting explicit families of pairs of distributions, where in one case $p = q$ and in the other p and q have large $\mathcal{A}_k$ distance, but so that it is information-theoretically impossible to distinguish between these two families with a small number of samples. In both cases, p and q are explicit piecewise constant distributions with a small number of pieces. In both cases, our domain is partitioned into a small number of bins and the restrictions of the distributions to different bins are independent, making our analysis easier. In some bins we will have $p = q$ each with mass about $1/m$ (where m is the number of samples). These bins will serve the purpose of adding "noise" making harder to read the "signal" from the other bins. In the remaining bins, we will have either that $p = q$ being supported on some interval, or p and q will be supported on consecutive, non-overlapping intervals. If three samples are obtained from any one of these intervals, the order of the samples and the distributions that they come from will provide us with information about which family we came from. Unfortunately, since triple collisions are relatively uncommon, this will not be useful unless $m \gg \max(k^{4/5}/\epsilon^{6/5}, k^{1/2}/\epsilon^2)$. Bins from which we have one or zero samples will tell us nothing, but bins from which we have exactly two samples may provide information.

For these bins, it can be seen that we learn nothing from the ordering of the samples, but we may learn something from their spacing. In particular, in the case where p and q are supported on disjoint intervals, we would suspect that two samples very close to each other are far more likely to be taken from the same distribution rather than from opposite distributions. On the other hand, in order to properly interpret this information, we will need to know something about the scale of the distributions involved in order to know when two points should be considered to be "close". To overcome this difficulty, we will stretch each of our distributions by a random exponential amount. This will effectively conceal any information about the scales involved so long as the total support size of our distributions is exponentially large.

2 A Near-Optimal Closeness Tester over Discrete Domains

The basic idea of our algorithm is the following. From the distributions p and q construct new distributions p' and q' by merging pairs of consecutive buckets. Note that p' and q' each have much smaller domains (of size about $n/2$). Furthermore, note that the $\mathcal{A}_k$ distance between p and q is

$\sum_{I\in\mathcal{I}} |p(I) - q(I)|$ for some partition $\mathcal{I}$ into k intervals. By using essentially the same partition, we can show that $\|p' - q'\|_{\mathcal{A}_k}$ should be almost as large as $\|p - q\|_{\mathcal{A}_k}$. This will in fact hold unless much of the error between p and q is supported at points near the endpoints of intervals in $\mathcal{I}$. If this is the case, it turns out there is an easy algorithm to detect this discrepancy. We require the following definitions:

Definition 1 *For a discrete distribution p on $[n]$, the merged distribution obtained from p is the distribution p' on $\lceil n/2 \rceil$, so that $p'(i) \stackrel{\text{def}}{=} p(2i) + p(2i+1)$. For a partition $\mathcal{I}$ of $[n]$, define the divided partition $\mathcal{I}'$ of domain $\lceil n/2 \rceil$, so that $I'_i \in \mathcal{I}'$ has the points obtained by point-wise glueing together odd points and even points.*

Definition 2 *Let p and q be distributions on $[n]$. For integers $k \geq 1$, let $\|p - q\|_{1,k}$ be the sum of the largest k values of $|p(i) - q(i)|$ over $i \in [n]$.*

We begin by showing that either $\|p' - q'\|_{\mathcal{A}_k}$ is close to $\|p - q\|_{\mathcal{A}_k}$ or $|p - q|_{1,k}$ is large.

Lemma 1 *For any two distributions p and q on $[n]$, let $p' = \lceil p/2 \rceil$ and $q' = \lceil q/2 \rceil$ be the merged distributions. Then,*

$$\|p - q\|_{\mathcal{A}_k} \leq \|p' - q'\|_{\mathcal{A}_k} + 2\|p - q\|_{1,k} \ .$$

Proof: Let $\mathcal{I}$ be the partition of $[n]$ into k intervals so that $\|p - q\|_{\mathcal{A}_k} = \sum_{I\in\mathcal{I}} |p(I) - q(I)|$. Let $\mathcal{I}'$ be obtained from $\mathcal{I}$ by rounding each upper endpoint of each interval except for last down to the nearest even integer, and rounding the lower endpoint of each interval up to the nearest odd integer. Note that

$$\sum_{I\in\mathcal{I}'} |p(I) - q(I)| = \sum_{I\in\mathcal{I}'} |p'(I/2) - q'(I/2)| \leq \|p' - q'\|_{\mathcal{A}_k} \ .$$

The partition $\mathcal{I}'$ is obtained from $\mathcal{I}$ by taking at most k points and moving them from one interval to another. Therefore, the difference

$$\left| \sum_{I\in\mathcal{I}} |p(I) - q(I)| - \sum_{I\in\mathcal{I}'} |p(I) - q(I)| \right| \ ,$$

is at most twice the sum of $|p(i) - q(i)|$ over these k points, and therefore at most $2\|p - q\|_{1,k}$. Combing this with the above gives our result. ■

Next we need to show that if two distributions have $\|p - q\|_{1,k}$ large that this can be detected easily.

Lemma 2 *Let p and q be distributions on $[n]$. Let $k > 0$ be a positive integer, and $\epsilon > k^{-1/4}$. There exists an algorithm which takes $O(k^{2/3}/\epsilon^{4/3})$ samples from each of p and q and, with probability at least $2/3$, distinguishes between the cases that $p = q$ and $\|p - q\|_{1,k} > \epsilon$.*

We start by introducing some important terminology from [10]. We begin with the definition of a split distribution:

Definition 3 *Given a distribution p on $[n]$ and a multiset S of elements of $[n]$, define the* split distribution *p_S on $[n + |S|]$ as follows: For $1 \leq i \leq n$, let a_i denote 1 plus the number of elements of S that are equal to i. Thus, $\sum_{i=1}^{n} a_i = n + |S|$. We can therefore associate the elements of $[n + |S|]$ to elements of the set $B = \{(i, j) : i \in [n], 1 \leq j \leq a_i\}$. We now define a distribution p_S with support B, by letting a random sample from p_S be given by (i, j), where i is drawn randomly from p and j is drawn randomly from $[a_i]$.*

And we recall some basic facts about it:

Fact 1 *Let p and q be probability distributions on $[n]$, and S a given multiset of $[n]$. Then: (i) We can simulate a sample from p_S or q_S by taking a single sample from p or q, respectively. (ii) It holds $\|p_S - q_S\|_1 = \|p - q\|_1$.*

Lemma 3 ([10]) *Let p be a distribution on $[n]$. Then: (i) For any multisets $S \subseteq S'$ of $[n]$, $\|p_{S'}\|_2 \leq \|p_S\|_2$, and (ii) If S is obtained by taking $\mathrm{Poi}(m)$ samples from p, then $\mathbb{E}[\|p_S\|_2^2] \leq 1/m$.*

Proof of Lemma 2: We begin by presenting the algorithm:

Algorithm `Small-Support-Discrepancy-Tester`
Input: sample access to pdf's $p, q : [n] \to \mathbb{R}_+$, $k \in \mathbb{Z}_+$, and $\epsilon > 0$.
Output: "YES" if $q = p$; "NO" if $\|q - p\|_{1,k} \geq \epsilon$.

1. Let $m = k^{2/3}/\epsilon^{4/3}$.
2. Let S be the multiset obtained by taking m independent samples from p.
3. Use the ℓ_2 tester to distinguish between the cases $p_S = q_S$ and $\|p_S - q_S\|_2^2 \geq k^{-1}\epsilon^2/2$, and return the result.

The analysis is simple. We note that with 90% probability it holds $\|p_S\|_2 = O(1/m)$, and therefore the number of samples needed is $O(m + km^{-1/2}/\epsilon^{-2}) = O(k^{2/3}/\epsilon^{4/3})$. If $p = q$, then $p_S = q_S$ and the algorithm will return "YES" with appropriate probability. If $\|q - p\|_{1,k} \geq \epsilon$, then $\|p_S - q_S\|_{1,k+m} \geq \epsilon$. Since $k + m$ elements contribute to total L^1 error at least ϵ, by Cauchy-Schwartz, we have that $\|p_S - q_S\|_2^2 \geq \epsilon^2/(k+m) \geq k^{-1}\epsilon^2/2$. Therefore, in this case, the algorithm returns "NO" with appropriate probability. ∎

Proof of Theorem 1: Given the algorithms from [10], we only need an algorithm that distinguishes in $O(k^{2/3}\log^{4/3}(n/k)\log\log(n/k)/\epsilon^{4/3})$ samples when $\epsilon > k^{-1/4}$.

We present the algorithm here:

Algorithm `Small-Domain-`$\mathcal{A}_k$`-tester`
Input: sample access to pdf's $p, q : [n] \to \mathbb{R}_+$, $k \in \mathbb{Z}_+$, and $\epsilon > 0$.
Output: "YES" if $q = p$; "NO" if $\|q - p\|_{\mathcal{A}_k} \geq \epsilon$.

1. For $i := 0$ to $t \stackrel{\text{def}}{=} \lceil \log_2(n/k) \rceil$, let $p^{(i)}, q^{(i)}$ be distributions on $[\lceil 2^{-i}n \rceil]$ defined by $p^{(i)} = \lceil 2^{-i}p \rceil$ and $q^{(i)} = \lceil 2^{-i}q \rceil$.
2. Take $Ck^{2/3}\log^{4/3}(3 + n/k)\log\log(3 + n/k)/\epsilon^{4/3}$ samples, for C sufficiently large, and use these samples to distinguish between the cases $p^{(i)} = q^{(i)}$ and $\|p^{(i)} - q^{(i)}\|_{1,k} > \epsilon/(4\log_2(3 + n/k))$ with probability of error at most $1/(10\log_2(3 + n/k))$ for each i from 0 to t, using the same samples for each test.
3. If any test yields that $p^{(i)} \neq q^{(i)}$, return "NO". Otherwise, return "YES".

We now show correctness. In terms of sample complexity, we note that by taking a majority over $O(\log\log(3 + n/k))$ independent runs of the tester from Lemma 2 we can run this algorithm in an appropriate sample complexity. Taking a union bound, we can also assume that all tests performed in step 2 returned the correct answer. If $p = q$ then $p^{(i)} = q^{(i)}$ for all i and thus, our algorithm returns "YES". Otherwise, we have that $\|p - q\|_{\mathcal{A}_k} \geq \epsilon$. By repeated application of Lemma 1, we have that

$$\|p - q\|_{\mathcal{A}_k} \leq \sum_{i=0}^{t-1} 2\|p^{(i)} - q^{(i)}\|_{1,k} + \|p^{(t)} - q^{(t)}\|_{\mathcal{A}_k} \leq 2\sum_{i=0}^{t} 2\|p^{(i)} - q^{(i)}\|_{1,k} \,,$$

where the last step was because $p^{(t)}$ and $q^{(t)}$ have a support of size at most k and so $\|p^{(t)} - q^{(t)}\|_{\mathcal{A}_k} = \|p^{(t)} - q^{(t)}\|_1 = \|p^{(t)} - q^{(t)}\|_{1,k}$. Therefore, if this is at least ϵ, it must be the case that $\|p^{(i)} - q^{(i)}\|_{1,k} > \epsilon/(4\log_2(3 + n/k))$ for some $0 \leq i \leq t$, and thus our algorithm returns "NO".

This completes our proof. ∎

3 Nearly Matching Information-Theoretic Lower Bound

Here we prove a nearly matching sample lower bound. We begin by proving a bound for continuous distributions that are piecewise constant with few pieces. Our bound on discrete distributions will follow from taking the adversarial distribution from this example and rounding its values to the nearest integer. In order for this to work, we will need ensure to that our adversarial distribution

does not have its $\mathcal{A}_k$-distance decrease by too much when we apply this operation. To satisfy this requirement, we will guarantee that our distributions will be piecewise constant with all the pieces of length at least 1.

Proposition 1 *Let $k \in \mathbb{Z}_+$, $\epsilon > 0$ sufficiently small, and $W > 2$. Let $m = \min(k^{2/3}\log^{1/3}(W)/\epsilon^{4/3}, k^{4/5}/\epsilon^{6/5})$. There exist distributions $\mathcal{D}, \mathcal{D}'$ over pairs of distributions p and q on $[0, 2(m+k)W]$ so that p and q are $O(m+k)$-flat with pieces of length at least 1, so that when drawn from $\mathcal{D}$, $p = q$ deterministically, when drawn from $\mathcal{D}'$ $\|p - q\|_{\mathcal{A}_k} > \epsilon$ with 90% probability, and so that $o(m)$ samples are insufficient to distinguish whether or not the pair is drawn from $\mathcal{D}$ or $\mathcal{D}'$ with better than $2/3$ probability.*

The basic idea of the proof will be as follows. At a high-level, our construction will mimic the lower bound construction of [20]. We will divide our domain into $m + k$ bins so that no information about which distributions had samples drawn from a given bin or the ordering of these samples will help to distinguish between the cases of $p = q$ and otherwise, unless at least three samples are taken from the bin in question. Approximately k of these bins will each have mass ϵ/k and might convey this information if at least three samples are taken from the bin. However, the other m bins will each have mass approximately $1/m$, and will be used to add noise. In all, if we take s samples, we expect to see approximately $s^3\epsilon^3/k^2$ of the lighter bins with at least three samples. However, we will see approximately s^3/m^2 of our heavy bins with three samples. In order for the signal to overwhelm the noise we will need to ensure that we have $(s^3\epsilon^3/k^2)^2 > s^3/m^2$.

The above analysis assumes that we cannot obtain information from the bins in which only two samples are drawn. This naively should not be the case. If $p = q$ then the distance between two samples drawn from that bin will be independent of whether or not they are drawn from the same distribution. However, if p and q are supported on disjoint intervals, one would expect that points that are close to each other should be far more likely to be drawn from the same distribution than from different distributions. In order to disguise this, we will scale the length of the intervals by a random, exponential amount, essentially making it impossible to determine what is meant by two points being close to each other. In effect, this will imply that two points drawn from the same bin will only tell us $O(1/\log(W))$ bits of information about whether $p = q$ or not. Thus, in order for this information to be sufficient, we will need that $(s^2\epsilon^2/k)^2/\log(W) > (s^2/m)$.

Proof of Proposition 1: We use ideas from [10] to obtain this lower bound using an argument from information theory.

We may assume that $\epsilon > k^{1/2}$, because otherwise we may employ the standard lower bound that $\Omega(\sqrt{k}/\epsilon^2)$ samples are required to distinguish two distributions on a support of size k.

First, we note that it is sufficient to take $\mathcal{D}$ and $\mathcal{D}'$ distributions over pairs of non-negative, piecewise constant distributions with total mass $\Theta(1)$ with 90% probability so that running a Poisson process with parameter $o(m)$ is insufficient to distinguish a pair from $\mathcal{D}$ from a pair from $\mathcal{D}'$ [10].

We construct this distribution as follows: We divide the domain into $m + k$ bins of length $2W$. For each bin i, we independently generate a random ℓ_i, so that $\log(\ell_i/2)$ is uniformly distributed over $[0, 2\log(W)/3]$. We then produce an interval I_i within bin i of total length ℓ_i and with random offset. In all cases, we will have p and q supported on the union of the I_i's.

For each i with probability $m/(m+k)$, we have the restrictions of p and q to I_i both uniform with $p(I_i) = q(I_i) = 1/m$. The other $k/(m+k)$ of the time we have $p(I_i) = q(I_i) = \epsilon/k$. In this latter case, if p and q are being drawn from $\mathcal{D}$, p and q are each constant on this interval. If they are being drawn from $\mathcal{D}'$, then $p + q$ will be constant on the interval, with all of that mass coming from p on a random half and coming from q on the other half.

Note that in all cases p and q are piecewise constant with $O(m + k)$ pieces of length at least 1. It is easy to show that with high probability the total mass of each of p and q is $\Theta(1)$, and that if drawn from $\mathcal{D}'$ that $\|p - q\|_{\mathcal{A}_k} \gg \epsilon$ with at least 90% probability.

Now we will show that if one is given m samples from each of p and q, taken randomly from either $\mathcal{D}$ or $\mathcal{D}'$, that the shared information between the samples and the source family will be small. This implies that one is unable to consistently guess whether our pair was taken from $\mathcal{D}$ or $\mathcal{D}'$.

Let X be a random variable that is uniformly randomly either 0 or 1. Let A be obtained by applying a Poisson process with parameter $s = o(m)$ on the pair of distributions p, q drawn from $\mathcal{D}$ if $X = 0$

or from $\mathcal{D}'$ if $X = 1$. We note that it suffices to show that the shared information $I(X : A) = o(1)$. In particular, by Fano's inequality, we have:

Lemma 4 *If X is a uniform random bit and A is a correlated random variable, then if f is any function so that $f(A) = X$ with at least 51% probability, then $I(X : A) \geq 2 \cdot 10^{-4}$.*

Let A_i be the samples of A taken from the i^{th} bin. Note that the A_i are conditionally independent on X. Therefore, we have that $I(X : A) \leq \sum_i I(X : A_i) = (m + k)I(X : A_1)$. We will proceed to bound $I(X : A_1)$.

We note that $I(X : A_1)$ is at most the integral over pairs of multisets a (representing a set of samples from q and a set of samples from p), of

$$O\left(\frac{(\Pr(A_1 = a|X = 0) - \Pr(A_1 = a|X = 1))^2}{\Pr(A_1 = a)}\right).$$

Thus,

$$I(X : A_1) = \sum_{h=0}^{\infty} \int_{|a|=h} O\left(\frac{(\Pr(A_1 = a|X = 0) - \Pr(A_1 = a|X = 1))^2}{\Pr(A_1 = a)}\right).$$

We will split this sum up based on the h.

For $h = 0$, we note that the distributions for $p + q$ are the same for $X = 0$ and $X = 1$. Therefore, the probability of selecting no samples is the same. Therefore, this contributes 0.

For $h = 1$, we note that the distributions for $p + q$ are the same in both cases, and conditioning on I_1 and $(p + q)(I_1)$ that $\mathbb{E}[p]$ and $\mathbb{E}[q]$ are the same in each of the cases $X = 0$ and $X = 1$. Therefore, again in this case, we have no contribution.

For $h \geq 3$, we note that $I(X : A_1) \leq I(X : A_1, I_1) \leq I(X : A_1|I_1)$, since I_1 is independent of X. We note that $\Pr(A_1 = a|X = 0, p(I_1) = 1/m) = \Pr(A_1 = a|X = 1, p(I_1) = 1/m)$. Therefore, we have that

$$\Pr(A_1 = a|X = 0) - \Pr(A_1 = a|X = 1) = \Pr(A_1 = a|X = 0, p(I_1) = \epsilon/k) - \Pr(A_1 = a|X = 1, p(I_1) = \epsilon/k).$$

If $p(I_1) = \epsilon/k$, the probability that exactly h elements are selected in this bin is at most $k/(m + k)(2s\epsilon/k)^h/h!$, and if they are selected, they are uniformly distributed in I_1 (although which of the sets p and q they are taken from is non-uniform). However, the probability that h elements are taken from I_1 is at least $\Omega(m/(m + k)(sm)^{-h}/h!)$ from the case where $p(I_1) = 1/m$, and in this case the elements are uniformly distributed in I_1 and uniformly from each of p and q. Therefore, we have that this contribution to our shared information is at most $k^2/(m(m + k))O(s\epsilon^2 m/k^2)^h/h!$. We note that $\epsilon^2 m/k^2 < 1$. Therefore, the sum of this over all $h \geq 3$ is $k^2/(m(m + k))O(s\epsilon^2 m/k^2)^3$. Summing over all $m + k$ bins, this is $k^{-4}\epsilon^6 s^3 m^2 = o(1)$. It remains to analyze the case where $h = 2$. Once again, we have that ignoring which of p and q elements of A_1 came from that A_1 is identically distributed conditioned on $p(I_1) = 1/m$ and $|A_1| = 2$ as it is conditioned on $p(I_1) = \epsilon/k$ and $|A_1| = 2$. Since once again, the distributions $\mathcal{D}$ and $\mathcal{D}'$ are indistinguishable in the former case, we have that the contribution of the $h = 2$ terms to the shared information is at most

$$O\left(\frac{(k/(k + m)(\epsilon s/k)^2)^2}{m/(k + m)(s/m)^2}\right) d_{\mathrm{TV}}((A_1|X = 0, p(I_1)\epsilon/k, |A_1| = 2), (A_1|X = 1, p(I_1) = \epsilon/k, |A_1| = 2))$$

or

$$O\left(s^2 m k^{-2}\epsilon^4/(k + m)\right) d_{\mathrm{TV}}((A_1|X = 0, p(I_1) = \epsilon/k, |A_1| = 2), (A_1|X = 1, p(I_1) = \epsilon/k, |A_1| = 2)) .$$

It will suffice to show that conditioned upon $p(I_1) = \epsilon/k$ and $|A_1| = 2$ that $d_{\mathrm{TV}}((A_1|X = 0), (A_1|X = 1)) = O(1/\log(W))$.

Let f be the order preserving linear function from $[0, 2]$ to I_1. Notice that conditional on $|A_1| = 2$ and $p(I_1) = \epsilon/k$ that we may sample from A_1 as follows:

- Pick two points $x > y$ uniformly at random from $[0, 2]$.
- Assign the points to p and q as follows:
  - If $X = 0$ uniformly randomly assign these points to either distribution p or q.

- If $X = 1$ randomly do either:
  - Assign points in $[0, 1]$ to q and other points to p.
  - Assign points in $[0, 1]$ to p and other points to q.
- Randomly pick I_1 and apply f to x and y to get outputs $z = f(x), w = f(y)$.

Notice that the four cases: (i) both points coming from p, (ii) both points coming from q, (iii) a point from p preceding a point from q, (iv) a point from q preceding a point from p, are all equally likely conditioned on either $X = 0$ or $X = 1$. However, we will note that this ordering is no longer independent of the choice of x and y.

We note therefore that we can sample from A_1 subject to $X = 0$ and from A_1 subject to $X = 1$ in such a way that this ordering is the same deterministically. We consider running the above sampling algorithm to select (x, y) while sampling from $X = 0$ and (x', y') when sampling from $X = 1$ so that we are in the same one of the above four cases. We note that

$$d_{\mathrm{TV}}((A_1|X = 0), (A_1|X = 1)) \leq \mathbb{E}_{x,y,x',y'}[d_{\mathrm{TV}}((f(x), f(y)), (f(x'), f(y')))] \ ,$$

where variational distance is over the random choices of f.

To show that this is small, we note that $|f(x) - f(y)|$ is distributed like $\ell_1(x - y)$. This means that $\log(|f(x) - f(y)|)$ is uniform over $[\log(f(x) - f(y)), \log(f(x) - f(y)) + 2\log(W)/3]$. Similarly, $\log(|f'(x') - f'(y')|)$ is uniform over $[\log(f(x') - f(y')), \log(f(x') - f(y')) + 2\log(W)/3]$. These differ in total variation distance by

$$O\left(\frac{|\log(f(x) - f(y))| + |\log(f(x') - f(y'))|}{\log(W)}\right) \ .$$

Taking the expectation over x, y, x', y' we get $O(1/\log(W))$. Therefore, we may further correlate the choices made in selecting our two samples, so that the $z - w = z' - w'$ except with probability $O(1/\log(W))$. We note that after conditioning on this z and z' are both uniformly distributed over subintervals of $[0, 2W]$ of length at least $2(W - W^{2/3})$. Therefore, the distributions on z and z' differ by at most $O(W^{-1/3})$. Hence, the total variation distance between A_1 conditioned on $|A_1| = 2, p(I_1) = \epsilon/k, X = 0$ and conditioned on $|A_1| = 2, p(I_1) = \epsilon/k, X = 1$ is at most $O(1/\log(W)) + O(W^{-1/3}) = O(1/\log(W))$. This completes our proof. ■

We can now turn this into a lower bound for testing $\mathcal{A}_k$ distance on discrete domains.

Proof of Theorem 2: Assume for sake of contradiction that this is not the case, and that there exists a tester taking $o(m)$ samples. We use this tester to come up with a continuous tester that violates Proposition 1.

We begin by proving a few technical bounds on the parameters involved. Firstly, note that we already have a lower bound of $\Omega(k^{1/2}/\epsilon^2)$, so we may assume that this is much less than m. We now claim that $m = O(\min(k^{2/3}\log^{1/3}(3 + n/(m + k))/\epsilon^{4/3}, k^{4/5}/\epsilon^{6/5})$. If $m \leq k$, there is nothing to prove. Otherwise,

$$k^{2/3}\log^{1/3}(3 + n/(m + k))/\epsilon^{4/3} \geq m(m/k)^{-1/3}\log(3 + n/(m + k))^{1/3}.$$

Thus, there is nothing more to prove unless $\log(3 + n/(m + k)) \gg m/k$. But, in this case, $\log(3 + n/(m + k)) \gg \log(m/k)$ and thus $\log(3 + n/(m + k)) = \Theta(\log(3 + n/k))$, and we are done.

We now let $W = n/(6(m + k))$, and let $\mathcal{D}$ and $\mathcal{D}'$ be as specified in Proposition 1. We claim that we have a tester to distinguish a p, q from $\mathcal{D}$ from ones taken from $\mathcal{D}'$ in $o(m)$ samples. We do this as follows: By rounding p and q down to the nearest third of an integer, we obtain p',q' supported on set of size n. Since p and q were piecewise constant on pieces of size at least 1, it is not hard to see that $\|p' - q'\|_{\mathcal{A}_k} \geq \|p - q\|_{\mathcal{A}_k}/3$. Therefore, a tester to distinguish $p' = q'$ from $\|p' - q'\|_{\mathcal{A}_k} \geq \epsilon$ can be used to distinguish $p = q$ from $\|p - q\|_{\mathcal{A}_k} \geq 3\epsilon$. This is a contradiction and proves our lower bound. ■

References

[1] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing that distributions are close. In *IEEE Symposium on Foundations of Computer Science*, pages 259–269, 2000.

[2] J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.

[3] E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, 2005.

[4] R. Rubinfeld. Taming big probability distributions. *XRDS*, 19(1):24–28, 2012.

[5] C. L. Canonne. A survey on distribution testing: Your data is big. but is it blue? *Electronic Colloquium on Computational Complexity (ECCC)*, 22:63, 2015.

[6] L. Paninski. A coincidence-based test for uniformity given very sparsely-sampled discrete data. *IEEE Transactions on Information Theory*, 54:4750–4755, 2008.

[7] S. Chan, I. Diakonikolas, P. Valiant, and G. Valiant. Optimal algorithms for testing closeness of discrete distributions. In *SODA*, pages 1193–1203, 2014.

[8] G. Valiant and P. Valiant. An automatic inequality prover and instance optimal identity testing. In *FOCS*, 2014.

[9] I. Diakonikolas, D. M. Kane, and V. Nikishkin. Testing Identity of Structured Distributions. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015, San Diego, CA, USA, January 4-6, 2015*, 2015.

[10] I. Diakonikolas and D. M. Kane. A new approach for testing properties of discrete distributions. *CoRR*, abs/1601.05557, 2016.

[11] U. Grenander. On the theory of mortality measurement. *Skand. Aktuarietidskr.*, 39:125–153, 1956.

[12] C. Daskalakis, I. Diakonikolas, and R.A. Servedio. Learning k-modal distributions via testing. In *SODA*, pages 1371–1385, 2012.

[13] S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Learning mixtures of structured distributions over discrete domains. In *SODA*, pages 1380–1394, 2013.

[14] S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Efficient density estimation via piecewise polynomial approximation. In *STOC*, pages 604–613, 2014.

[15] S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Near-optimal density estimation in near-linear time using variable-width histograms. In *NIPS*, pages 1844–1852, 2014.

[16] J. Acharya, I. Diakonikolas, J. Li, and L. Schmidt. Sample-optimal density estimation in nearly-linear time. *CoRR*, abs/1506.00671, 2015.

[17] R.E. Barlow, D.J. Bartholomew, J.M. Bremner, and H.D. Brunk. *Statistical Inference under Order Restrictions*. Wiley, New York, 1972.

[18] P. Groeneboom and G. Jongbloed. *Nonparametric Estimation under Shape Constraints: Estimators, Algorithms and Asymptotics*. Cambridge University Press, 2014.

[19] C. Daskalakis, I. Diakonikolas, R. Servedio, G. Valiant, and P. Valiant. Testing k-modal distributions: Optimal algorithms via reductions. In *SODA*, pages 1833–1852, 2013.

[20] I. Diakonikolas, D. M. Kane, and V. Nikishkin. Optimal algorithms and lower bounds for testing closeness of structured distributions. In *56th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2015*, 2015.

[21] L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. Springer Series in Statistics, Springer, 2001.