Testing k-Modal Distributions: Optimal Algorithms via Reductions

Constantinos Daskalakis* MIT
Ilias Diakonikolas† UC Berkeley
Rocco A. Servedio ‡ Columbia University
Gregory Valiant§ UC Berkeley
Paul Valiant¶ UC Berkeley

December 23, 2011

Abstract

We give highly efficient algorithms, and almost matching lower bounds, for a range of basic statistical problems that involve testing and estimating the L_1 (total variation) distance between two k-modal distributions p and q over the discrete domain $\{1, \ldots, n\}$. More precisely, we consider the following four problems: given sample access to an unknown k-modal distribution p,

TESTING IDENTITY TO A KNOWN OR UNKNOWN DISTRIBUTION:

1. Determine whether $p = q$ (for an explicitly given k-modal distribution q) versus p is ϵ-far from q;
2. Determine whether $p = q$ (where q is available via sample access) versus p is ϵ-far from q;

ESTIMATING L_1 DISTANCE ("TOLERANT TESTING") AGAINST A KNOWN OR UNKNOWN DISTRIBUTION:

3. Approximate $d_{TV}(p, q)$ to within additive ϵ where q is an explicitly given k-modal distribution q;
4. Approximate $d_{TV}(p, q)$ to within additive ϵ where q is available via sample access.

For each of these four problems we give sub-logarithmic sample algorithms, that we show are tight up to additive $\mathrm{poly}(k)$ and multiplicative $\mathrm{polylog} \log n + \mathrm{polylog} k$ factors. Thus our bounds significantly improve the previous results of [BKR04], which were for testing identity of distributions (items (1) and (2) above) in the special cases $k = 0$ (monotone distributions) and $k = 1$ (unimodal distributions) and required $O((\log n)^3)$ samples.

As our main conceptual contribution, we introduce a new reduction-based approach for distribution-testing problems that lets us obtain all the above results in a unified way. Roughly speaking, this approach enables us to transform various distribution testing problems for k-modal distributions over $\{1, \ldots, n\}$ to the corresponding distribution testing problems for unrestricted distributions over a much smaller domain $\{1, \ldots, \ell\}$ where $\ell = O(k \log n)$.

*costis@csail.mit.edu. Research supported by NSF CAREER award CCF-0953960 and by a Sloan Foundation Fellowship.

†ilias@cs.berkeley.edu. Research supported by a Simons Foundation Postdoctoral Fellowship. Some of this work was done while at Columbia University, supported by NSF grant CCF-0728736, and by an Alexander S. Onassis Foundation Fellowship.

‡rocco@cs.columbia.edu. Supported by NSF grants CCF-0347282 and CCF-0523664.

§gregory.valiant@gmail.com. Supported by an NSF graduate research fellowship.

¶pvaliant@gmail.com. Supported by an NSF postdoctoral research fellowship.

1 Introduction

Given samples from a pair of unknown distributions, the problem of "identity testing"—that is, distinguishing whether the two distributions are *the same* versus significantly different—and, more generally, the problem of estimating the L_1 distance between the distributions, is perhaps *the* most fundamental statistical task. Despite a long history of study, by both the statistics and computer science communities, the sample complexities of these basic tasks were only recently established. Identity testing, given samples from a pair of distributions of support $[n]$, can be done using $\tilde{O}(n^{2/3})$ samples [BFR$^+$00], and this upper bound is optimal up to $\mathrm{polylog}(n)$ factors [Val08a]. Estimating the L_1 distance ("tolerant testing") between distributions of support $[n]$ requires $\Theta(n/\log n)$ samples, and this is tight up to constant factors [VV11a, VV11b]. The variants of these problems when one of the two distributions is explicitly given require $\widetilde{\Theta}(\sqrt{n})$ samples for identity testing [BFF$^+$01] and $\Theta(n/\log n)$ samples for L_1 distance estimation [VV11a, VV11b] respectively.

While it is surprising that these tasks can be performed using a sublinear number of samples, for many real-world applications using $\sqrt{n}$, $n^{2/3}$, or $\frac{n}{\log n}$ samples is still impractical. As these bounds characterize worst-case instances, one might hope that drastically better performance may be possible for many settings typically encountered in practice. Thus, a natural research direction, which we pursue in this paper, is to understand how structural properties of the distributions in question may be leveraged to yield improved sample complexities.

In this work we consider monotone, unimodal, and more generally k-modal distributions. Monotone, unimodal, and bimodal distributions abound in the natural world. The distribution of many measurements—heights or weights of members of a population, concentrations of various chemicals in cells, parameters of many atmospheric phenomena–often belong to this class of distributions. Because of their ubiquity, much work in the natural sciences rests on the analysis of such distributions (for example, on November 1, 2011 a Google Scholar search for the exact phrase "bimodal distribution" in the bodies of papers returned more than 90,000 hits). Though perhaps not as pervasive, k-modal distributions for larger values of k commonly arise as mixtures of unimodal distributions and are natural objects of study. On the theoretical side, motivated by the many applications, monotone, unimodal, and k-modal distributions have been intensively studied in the probability and statistics literatures for decades, see e.g. [Gre56, Rao69, BBBB72, CKC83, Gro85, Bir87a, Bir87b, Kem91, Fou97, CT04, JW09].

1.1 Our results. Our main results are algorithms, and nearly matching lower bounds, that give a complete picture of the sample complexities of identity testing and estimating L_1 distance for monotone and k-modal distributions. We obtain such results in both the setting where the two distributions are given via samples, and the setting where one of the distributions is given via samples and the other is described explicitly.

All our results have the nature of a reduction: performing these tasks on k-modal distributions over $[n]$ turns out to have almost exactly the same sample complexities as performing the corresponding tasks on *arbitrary* distributions over $[k \log n]$. For any small constant k (or even $k = O((\log n)^{1/3})$) and arbitrarily small constant ϵ, all our results are tight to within either $\mathrm{polylog} \log n$ or $\mathrm{polylog} \log \log n$ factors. See Table 1 for the new sample complexity upper and lower bounds for the monotone and k-modal tasks; see Section 2 for the (exponentially higher) sample complexities of the general-distribution tasks on which our results rely. While our main focus is on sample complexity rather than running time, we note that all of our algorithms run in $\mathrm{poly}(\log n, k, 1/\epsilon)$ bit operations (note that even reading a single sample from a distribution over $[n]$ takes $\log n$ bit operations).

We view the equivalence between the sample complexity of each of the above tasks on a monotone or unimodal distribution of domain $[n]$ and the sample complexity of the same task on an unrestricted distribution of domain $[\log n]$ as a surprising result, because such an equivalence *fails* to hold for related estimation tasks. For example, consider the task of distinguishing whether a distribution on $[n]$ is uniform versus far from uniform. For general distributions this takes $\Theta(\sqrt{n})$ samples, so one might expect the corresponding problem for monotone distributions to need $\sqrt{\log n}$ samples; in fact, however, one can test this with a *constant* number of samples, by simply comparing the empirically observed probability masses of the left and right halves of the domain. An example in the other direction is the problem of finding a constant additive estimate for the entropy of a distribution. On domains of size $[n]$ this can be done in $\frac{n}{\log n}$ samples, and thus one might expect to be able to estimate entropy for monotone distributions on $[n]$ using $\frac{\log n}{\log \log n}$ samples. Nevertheless, it is not hard to see that $\Omega(\log^2 n)$ samples are required.

Testing problem	Our upper bound	Our lower bound
p, q are both monotone:		
Testing identity, q is known:	$O\left((\log n)^{1/2}(\log\log n)\cdot\epsilon^{-5/2}\right)$	$\Omega\left((\log n)^{1/2}\right)$
Testing identity, q is unknown:	$O\left((\log n)^{2/3}\cdot(\log\log n)\cdot\epsilon^{-10/3}\right)$	$\Omega\left(\left(\frac{\log n}{\log\log n}\right)^{2/3}\right)$
Estimating L_1 distance, q is known:	$O\left(\frac{\log n}{\log\log n}\cdot\epsilon^{-3}\right)$	$\Omega\left(\frac{\log n}{\log\log n\cdot\log\log\log n}\right)$
Estimating L_1 distance, q is unknown:	$O\left(\frac{\log n}{\log\log n}\cdot\epsilon^{-3}\right)$	$\Omega\left(\frac{\log n}{\log\log n\cdot\log\log\log n}\right)$
p, q are both k-modal:		
Testing identity, q is known:	$O\left(\frac{k^2}{\epsilon^4}+\frac{(k\log n)^{1/2}}{\epsilon^3}\cdot\log\left(\frac{k\log n}{\epsilon}\right)\right)$	$\Omega\left((k\log n)^{1/2}\right)$
Testing identity, q is unknown:	$O\left(\frac{k^2}{\epsilon^4}+\frac{(k\log n)^{2/3}}{\epsilon^{10/3}}\cdot\log\left(\frac{k\log n}{\epsilon}\right)\right)$	$\Omega\left(\left(\frac{k\log n}{\log(k\log n)}\right)^{2/3}\right)$
Estimating L_1 distance, q is known:	$O\left(\frac{k^2}{\epsilon^4}+\frac{k\log n}{\log(k\log n)}\cdot\epsilon^{-4}\right)$	$\Omega\left(\frac{k\log n}{\log(k\log n)\cdot\log\log(k\log n)}\right)$
Estimating L_1 distance, q is unknown:	$O\left(\frac{k^2}{\epsilon^4}+\frac{k\log n}{\log(k\log n)}\cdot\epsilon^{-4}\right)$	$\Omega\left(\frac{k\log n}{\log(k\log n)\cdot\log\log(k\log n)}\right)$

Table 1: Our upper and lower bounds for identity testing and L_1 estimation. In the table we omit a "$\log(1/\delta)$" term which is present in all our upper bounds for algorithms which give the correct answer with probability $1-\delta$. For the "testing identity" problems, our lower bounds are for distinguishing whether $p = q$ versus $d_{\mathrm{TV}}(p,q) > 1/2$ with success probability $2/3$. For estimating L_1 distance, our bounds are for estimating $d_{\mathrm{TV}}(p,q)$ to within $\pm\epsilon$, for any $k = O(n^{1/2})$, with the lower bounds corresponding to success probability $2/3$.

The reduction-like techniques which we use to establish both our algorithmic results and our lower bounds (discussed in more detail in Section 1.2 below) reveal an unexpected relationship between the class of k-modal distributions of support $[n]$ and the class of general distributions of support $[k\log n]$. We hope that this reduction-based approach may provide a framework for the discovery of other relationships that will be useful in future work in the extreme sublinear regime of statistical property estimation and property testing.

Comparison with prior work. Our results significantly extend and improve upon the previous algorithmic results of Batu et al [BKR04] for identity testing of monotone or unimodal ($k = 1$) distributions, which required $O(\log^3 n)$ samples. More recently, [DDS11] established the sample complexity of *learning* k-modal distributions to be essentially $\Theta(k\log(n)\epsilon^{-3})$. Such a learning algorithm easily yields a testing algorithm with the same sample complexity for all four variants of the testing problem (one can simply run the learner twice to obtain hypotheses $\widehat{p}$ and $\widehat{q}$ that are sufficiently close to p and q respectively, and output accordingly).

While the [DDS11] result can be applied to our testing problems (though giving suboptimal results), we stress that the ideas underlying [DDS11] and this paper are quite different. The [DDS11] paper learns a k-modal distribution by using a known algorithm for learning monotone distributions [Bir87b] k times in a black-box manner; the notion of *reducing the domain size*—which we view as central to the results and contributions of this paper—is nowhere present in [DDS11]. By contrast, the focus in this paper is on introducing the use of reductions as a powerful (but surprisingly, seemingly previously unused) tool in the development of algorithms for basic statistical tasks on distributions, which, at least in this case, is capable of giving essentially optimal upper and lower bounds for natural restricted classes of distributions.

1.2 Techniques. Our main conceptual contribution is a new reduction-based approach that lets us obtain all our upper and lower bounds in a clean and unified way. The approach works by reducing the monotone and k-modal distribution testing problems to general distribution testing and estimation problems *over a much smaller domain*, and vice versa. For the monotone case this smaller domain is essentially of size $\log(n)/\epsilon$, and for the k-modal case the smaller domain is essentially of size $k\log(n)/\epsilon^2$. By solving the general distribution problems over the smaller

domain using known results we get a valid answer for the original (monotone or k-modal) problems over domain $[n]$. More details on our algorithmic reduction are given in Section A.

Conversely, our lower bound reduction lets us reexpress arbitrary distributions over a small domain $[\ell]$ by monotone (or unimodal, or k-modal, as required) distributions over an exponentially larger domain, while preserving many of their features with respect to the L_1 distance. Crucially, this reduction allows one to *simulate* drawing samples from the larger monotone distribution given access to samples from the smaller distribution, so that a known impossibility result for unrestricted distributions on $[\ell]$ may be leveraged to yield a corresponding impossibility result for monotone (or unimodal, or k-modal) distributions on the exponentially larger domain.

The inspiration for our results is an observation of Birgé [Bir87b] that given a monotone-decreasing probability distribution over $[n]$, if one subdivides $[n]$ into an exponentially increasing series of consecutive sub-intervals, the ith having size $(1+\epsilon)^i$, then if one replaces the probability mass on each interval with a uniform distribution on that interval, the distribution changes by only $O(\epsilon)$ in total variation distance. Further, given such a subdivision of the support into $\log_{1+\epsilon}(n)$ intervals, one may essentially treat the original monotone distribution as essentially a distribution over these intervals, namely a distribution of support $\log_{1+\epsilon}(n)$. In this way, one may hope to reduce monotone distribution testing or estimation on $[n]$ to general distribution testing or estimation on a domain of size $\log_{1+\epsilon}(n)$, and vice versa. See Section B for details.

For the monotone testing problems the partition into subintervals is constructed obliviously (without drawing any samples or making any reference to p or q of any sort) – for a given value of ϵ the partition is the same for all non-increasing distributions. For the k-modal testing problems, constructing the desired partition is significantly more involved. This is done via a careful procedure which uses $k^2 \cdot \mathrm{poly}(1/\epsilon)$ samples[1] from p and q and uses the oblivious decomposition for monotone distributions in a delicate way. This construction is given in Section C.

2 Notation and Preliminaries

2.1 Notation. We write $[n]$ to denote the set $\{1, \ldots, n\}$, and for integers $i \leq j$ we write $[i, j]$ to denote the set $\{i, i+1, \ldots, j\}$. We consider discrete probability distributions over $[n]$, which are functions $p : [n] \to [0, 1]$ such that $\sum_{i=1}^{n} p(i) = 1$. For $S \subseteq [n]$ we write $p(S)$ to denote $\sum_{i \in S} p(i)$. We use the notation P for the *cumulative distribution function (cdf)* corresponding to p, i.e. $P : [n] \to [0, 1]$ is defined by $P(j) = \sum_{i=1}^{j} p(i)$.

A distribution p over $[n]$ is non-increasing (resp. non-decreasing) if $p(i+1) \leq p(i)$ (resp. $p(i+1) \geq p(i)$), for all $i \in [n-1]$; p is *monotone* if it is either non-increasing or non-decreasing. Thus the "orientation" of a monotone distribution is either non-decreasing (denoted $\uparrow$) or non-increasing (denoted $\downarrow$).

We call a nonempty interval $I = [a, b] \subseteq [2, n-1]$ a *max-interval* of p if $p(i) = c$ for all $i \in I$ and $\max\{p(a-1), p(b+1)\} < c$. Analogously, a *min-interval* of p is an interval $I = [a, b] \subseteq [2, n-1]$ with $p(i) = c$ for all $i \in I$ and $\min\{p(a-1), p(b+1)\} > c$. We say that p is *k-modal* if it has at most k max-intervals and min-intervals. We note that according to our definition, what is usually referred to as a bimodal distribution is a 3-modal distribution.

Let p, q be distributions over $[n]$ with corresponding cdfs P, Q. The *total variation distance* between p and q is $d_{\mathrm{TV}}(p, q) := \max_{S \subseteq [n]} |p(S) - q(S)| = (1/2) \sum_{i \in [n]} |p(i) - q(i)|$. The *Kolmogorov distance* between p and q is defined as $d_{\mathrm{K}}(p, q) := \max_{j \in [n]} |P(j) - Q(j)|$. Note that $d_{\mathrm{K}}(p, q) \leq d_{\mathrm{TV}}(p, q)$.

Finally, a *sub-distribution* is a function $q : [n] \to [0, 1]$ which satisfies $\sum_{i=1}^{n} q(i) \leq 1$. For p a distribution over $[n]$ and $I \subseteq [n]$, the *restriction of p to I* is the sub-distribution p^I defined by $p^I(i) = p(i)$ if $i \in I$ and $p^I(i) = 0$ otherwise. Likewise, we denote by p_I the conditional distribution of p on I, i.e. $p_I(i) = p(i)/p(I)$ if $i \in I$ and $p_I(i) = 0$ otherwise.

2.2 Basic tools from probability. We will require the *Dvoretzky-Kiefer-Wolfowitz (DKW) inequality* ([DKW56]) from probability theory. This basic fact says that $O(1/\epsilon^2)$ samples suffice to learn any distribution within error ϵ with respect to the *Kolmogorov distance*. More precisely, let p be any distribution over $[n]$. Given m independent

[1] Intuitively, the partition must be finer in regions of higher probability density; for non-increasing distributions (for example) this region is at the left side of the domain, but for general k-modal distributions, one must draw samples to discover the high-probability regions.

samples $s_1, \ldots, s_m$ drawn from $p : [n] \to [0,1]$, the *empirical distribution* $\widehat{p}_m : [n] \to [0,1]$ is defined as follows: for all $i \in [n]$, $\widehat{p}_m(i) = |\{j \in [m] \mid s_j = i\}|/m$. The DKW inequality states that for $m = \Omega((1/\epsilon^2) \cdot \ln(1/\delta))$, with probability $1-\delta$ the empirical distribution $\widehat{p}_m$ will be ϵ-close to p in Kolmogorov distance. This sample bound is asymptotically optimal and independent of the support size.

Theorem 1 ([DKW56, Mas90])**.** *For all $\epsilon > 0$, it holds:* $\Pr[d_{\mathrm{K}}(p, \widehat{p}_m) > \epsilon] \leq 2e^{-2m\epsilon^2}$.

Another simple result that we will need is the following, which is easily verified from first principles:

Observation 1. *Let $I = [a,b]$ be an interval and let u_I denote the uniform distribution over I. Let p_I denote a non-increasing distribution over I. Then for every initial interval $I' = [a,b']$ of I, we have $u_I(I') \leq p_I(I')$.*

2.3 Testing and estimation for arbitrary distribution Our testing algorithms work by reducing to known algorithms for testing arbitrary distributions over an ℓ-element domain. We will use the following well known results:

Theorem 2 (testing identity, known distribution [BFF+01])**.** *Let q be an explicitly given distribution over $[\ell]$. Let p be an unknown distribution over $[\ell]$ that is accessible via samples. There is a testing algorithm* TEST-IDENTITY-KNOWN(p, q, ϵ, δ) *that uses $s_{IK}(\ell, \epsilon, \delta) := O(\ell^{1/2}\log(\ell)\epsilon^{-2}\log(1/\delta))$ samples from p and has the following properties:*

- *If $p \equiv q$ then with probability at least $1-\delta$ the algorithm outputs "accept;" and*
- *If $d_{\mathrm{TV}}(p,q) \geq \epsilon$ then with probability at least $1-\delta$ the algorithm outputs "reject."*

Theorem 3 (testing identity, unknown distribution [BFR+10])**.** *Let p and q both be unknown distributions over $[\ell]$ that are accessible via samples. There is a testing algorithm* TEST-IDENTITY-UNKNOWN(p, q, ϵ, δ) *that uses $s_{IU}(\ell, \epsilon, \delta) := O(\ell^{2/3}\log(\ell/\delta)\epsilon^{-8/3})$ samples from p and q and has the following properties:*

- *If $p \equiv q$ then with probability at least $1-\delta$ the algorithm outputs "accept;" and*
- *If $d_{\mathrm{TV}}(p,q) \geq \epsilon$ then with probability at least $1-\delta$ the algorithm outputs "reject."*

Theorem 4 (L_1 estimation [VV11b])**.** *Let p be an unknown distribution over $[\ell]$ that is accessible via samples, and let q be a distribution over $[\ell]$ that is either explicitly given, or accessible via samples. There is an estimator* L_1-ESTIMATE(p, q, ϵ, δ) *that, with probability at least $1-\delta$, outputs a value in the interval $(d_{\mathrm{TV}}(p,q) - \epsilon, d_{\mathrm{TV}}(p,q) + \epsilon)$. The algorithm uses $s_E(\ell, \epsilon, \delta) := O\left(\frac{\ell}{\log \ell} \cdot \epsilon^{-2}\log(1/\delta)\right)$ samples.*

3 Testing and Estimating Monotone Distributions

3.1 Oblivious decomposition of monotone distributions Our main tool for testing monotone distributions is an *oblivious decomposition* of monotone distributions that is a variant of a construction of Birgé [Bir87b]. As we will see it enables us to reduce the problem of testing a monotone distribution to the problem of testing an arbitrary distribution over a much smaller domain.

Before stating the decomposition, some notation will be helpful. Fix a distribution p over $[n]$ and a partition of $[n]$ into disjoint intervals $\mathcal{I} := \{I_i\}_{i=1}^{\ell}$. The *flattened distribution* $(p_f)^{\mathcal{I}}$ corresponding to p and $\mathcal{I}$ is the distribution over $[n]$ defined as follows: for $j \in [\ell]$ and $i \in I_j$, $(p_f)^{\mathcal{I}}(i) = \sum_{t \in I_j} p(t)/|I_j|$. That is, $(p_f)^{\mathcal{I}}$ is obtained from p by averaging the weight that p assigns to each interval over the entire interval. The *reduced distribution* $(p_r)^{\mathcal{I}}$ corresponding to p and $\mathcal{I}$ is the distribution over $[\ell]$ that assigns the ith point the weight p assigns to the interval I_i; i.e., for $i \in [\ell]$, we have $(p_r)^{\mathcal{I}}(i) = p(I_i)$. Note that if p is non-increasing then so is $(p_f)^{\mathcal{I}}$, but this is not necessarily the case for $(p_r)^{\mathcal{I}}$.

The following simple lemma, proved in Section A, shows why reduced distributions are useful for us:

Definition 1. *Let p be a distribution over $[n]$ and let $\mathcal{I} = \{I_i\}_{i=1}^{\ell}$ be a partition of $[n]$ into disjoint intervals. We say that $\mathcal{I}$ is a (p, ϵ, ℓ)*-flat decomposition of $[n]$ *if $d_{\mathrm{TV}}(p, (p_f)^{\mathcal{I}}) \leq \epsilon$.*

Lemma 2. *Let $\mathcal{I} = \{I_i\}_{i=1}^{\ell}$ be a partition of $[n]$ into disjoint intervals. Suppose that p and q are distributions over $[n]$ such that $\mathcal{I}$ is both a (p, ϵ, ℓ)-flat decomposition of $[n]$ and is also a (q, ϵ, ℓ)-flat decomposition of $[n]$. Then $\left|d_{\mathrm{TV}}(p, q) - d_{\mathrm{TV}}((p_r)^{\mathcal{I}}, (q_r)^{\mathcal{I}})\right| \leq 2\epsilon$. Moreover, if $p = q$ then $(p_r)^{\mathcal{I}} = (q_r)^{\mathcal{I}}$.*

We now state our oblivious decomposition result for monotone distributions:

Theorem 5 ([Bir87b])**.** *(oblivious decomposition) Fix any $n \in \mathbb{Z}^+$ and $\epsilon > 0$. The partition $\mathcal{I} := \{I_i\}_{i=1}^{\ell}$ of $[n]$, in which the jth interval has size $\lfloor (1+\epsilon)^j \rfloor$ has the following properties: $\ell = O\left((1/\epsilon) \cdot \log(\epsilon \cdot n + 1)\right)$, and $\mathcal{I}$ is a $(p, O(\epsilon), \ell)$-flat decomposition of $[n]$ for any non-increasing distribution p over $[n]$.*

There is an analogous version of Theorem 5, asserting the existence of an "oblivious" partition for non-decreasing distributions (which is of course different from the "oblivious" partition $\mathcal{I}$ for non-increasing distributions of Theorem 5); this will be useful later.

While our construction is essentially that of Birgé, we note that the version given in [Bir87b] is for non-increasing distributions over the continuous domain $[0, n]$, and it is phrased rather differently. Adapting the arguments of [Bir87b] to our discrete setting of distributions over $[n]$ is not conceptually difficult but requires some care. For the sake of being self-contained we provide a self-contained proof of the discrete version, stated above, that we require in Appendix E.

3.2 Efficiently testing monotone distributions Now we are ready to establish our upper bounds on testing monotone distributions (given in the first four rows of Table 1). All of the algorithms are essentially the same: each works by reducing the given monotone distribution testing problem to the same testing problem for arbitrary distributions over support of size $\ell = O(\log n/\epsilon)$ using the oblivious decomposition from the previous subsection. For concreteness we explicitly describe the tester for the "testing identity, q is known" case below, and then indicate the small changes that are necessary to get the testers for the other three cases.

TEST-IDENTITY-KNOWN-MONOTONE
Inputs: $\epsilon, \delta > 0$; sample access to non-increasing distribution p over $[n]$; explicit description of non-increasing distribution q over $[n]$

1. Let $\mathcal{I} := \{I_i\}_{i=1}^{\ell}$, with $\ell = \Theta(\log(\epsilon n + 1)/\epsilon)$, be the partition of $[n]$ given by Theorem 5, which is a $(p', \epsilon/8, \ell)$-flat decomposition of $[n]$ for any non-increasing distribution p'.
2. Let $(q_r)^{\mathcal{I}}$ denote the reduced distribution over $[\ell]$ obtained from q using $\mathcal{I}$ as defined in Section A.
3. Draw $m = s_{IK}(\ell, \epsilon/2, \delta)$ samples from $(p_r)^{\mathcal{I}}$, where $(p_r)^{\mathcal{I}}$ is the reduced distribution over $[\ell]$ obtained from p using $\mathcal{I}$ as defined in Section A.
4. Output the result of TEST-IDENTITY-KNOWN$((p_r)^{\mathcal{I}}, (q_r)^{\mathcal{I}}, \frac{\epsilon}{2}, \delta)$ on the samples from Step 3.

We now establish our claimed upper bound for the "testing identity, q is known" case. We first observe that in Step 3, the desired $m = s_{IK}(\ell, \epsilon/2, \delta)$ samples from $(p_r)^{\mathcal{I}}$ can easily be obtained by drawing m samples from p and converting each one to the corresponding draw from $(p_r)^{\mathcal{I}}$ in the obvious way. If $p = q$ then $(p_r)^{\mathcal{I}} = (q_r)^{\mathcal{I}}$, and TEST-IDENTITY-KNOWN-MONOTONE outputs "accept" with probability at least $1 - \delta$ by Theorem 2. If $d_{\mathrm{TV}}(p, q) \geq \epsilon$, then by Lemma 2, Theorem 5 and the triangle inequality, we have that $d_{\mathrm{TV}}((p_r)^{\mathcal{I}}, (q_r)^{\mathcal{I}}) \geq 3\epsilon/4$, so TEST-IDENTITY-KNOWN-MONOTONE outputs "reject" with probability at least $1 - \delta$ by Theorem 2. For the "testing identity, q is unknown" case, the the algorithm TEST-IDENTITY-UNKNOWN-MONOTONE is very similar to TEST-IDENTITY-KNOWN-MONOTONE. The differences are as follows: instead of Step 2, in Step 3 we draw $m = s_{IU}(\ell, \epsilon/2, \delta)$ samples from $(p_r)^{\mathcal{I}}$ and the same number of samples from $(q_r)^{\mathcal{I}}$; and in Step 4, we run TEST-IDENTITY-UNKNOWN$((p_r)^{\mathcal{I}}, (q_r)^{\mathcal{I}}, \frac{\epsilon}{2}, \delta)$ using the samples from Step 3. The analysis is exactly the same as above (using Theorem 3 in place of Theorem 2).

We now describe the algorithm L_1-ESTIMATE-KNOWN-MONOTONE for the "tolerant testing, q is known" case. This algorithm takes values ϵ and δ as input, so the partition $\mathcal{I}$ defined in Step 1 is a $(p', \epsilon/4, \ell)$-flat decomposition of $[n]$ for any non-increasing p'. In Step 3 the algorithm draws $m = s_E(\ell, \epsilon/2, \delta)$ samples and runs L_1-ESTIMATE$((p_r)^{\mathcal{I}}, (q_r)^{\mathcal{I}}, \epsilon/2, \delta)$ in Step 4. If $d_{\mathrm{TV}}(p, q) = c$ then by the triangle inequality we have that

$d_{\mathrm{TV}}((p_r)^{\mathcal{I}}, (q_r)^{\mathcal{I}}) \in [c - \epsilon/2, c + \epsilon/2]$ and L_1-ESTIMATE-KNOWN-MONOTONE outputs a value within the prescribed range with probability at least $1-\delta$, by Theorem 4. The algorithm L_1-ESTIMATE-UNKNOWN-MONOTONE and its analysis are entirely similar.

4 From Monotone to k-modal

In this section we establish our main positive testing results for k-modal distributions, the upper bounds stated in the final four rows of Table 1. In the previous section, we were able to use the oblivious decomposition to yield a partition of $[n]$ into relatively few intervals, with the guarantee that the corresponding flattened distribution is close to the true distribution. The main challenge in extending these results to unimodal or k-modal distributions, is that in order to make the analogous decomposition, one must first determine–by taking samples from the distribution–which regions are monotonically increasing vs decreasing. Our algorithm CONSTRUCT-FLAT-DECOMPOSITION(p, ϵ, δ) performs this task with the following guarantee:

Lemma 3. *Let p be a k-modal distribution over $[n]$. Algorithm* CONSTRUCT-FLAT-DECOMPOSITION(p, ϵ, δ) *draws $O(k^2\epsilon^{-4}\log(1/\delta))$ samples from p and outputs a (p, ϵ, ℓ)-flat decomposition of $[n]$ with probability at least $1 - \delta$, where $\ell = O(k\log(n)/\epsilon^2)$.*

The bulk of our work in Section C is to describe CONSTRUCT-FLAT-DECOMPOSITION(p, ϵ, δ) and prove Lemma 3, but first we show how Lemma 3 yields our claimed testing results for k-modal distributions. As in the monotone case all four algorithms are essentially the same: each works by reducing the given k-modal distribution testing problem to the same testing problem for arbitrary distributions over $[\ell]$. One slight complication is that the partition obtained for distribution p will generally differ from that for q. In the monotone distribution setting, the partition was oblivious to the distributions, and thus this concern did not arise. Naively, one might hope that the flattened distribution corresponding to any refinement of a partition will be at least as good as the flattened distribution corresponding to the actual partition. This hope is easily seen to be strictly false, but we show that it is true up to a factor of 2, which suffices for our purposes.

The following terminology will be useful: Let $\mathcal{I} = \{I_i\}_{i=1}^r$ and $\mathcal{I}' = \{I'_i\}_{i=1}^s$ be two partitions of $[n]$ into r and s intervals respectively. The *common refinement* of $\mathcal{I}$ and $\mathcal{I}'$ is the partition $\mathcal{J}$ of $[n]$ into intervals obtained from $\mathcal{I}$ and $\mathcal{I}'$ in the obvious way, by taking all possible nonempty intervals of the form $I_i \cap I'_j$. It is clear that $\mathcal{J}$ is both a refinement of $\mathcal{I}$ and of $\mathcal{I}'$ and that the number of intervals $|\mathcal{J}|$ in $\mathcal{J}$ is at most $r + s$. We prove the following lemma in Section A:

Lemma 4. *Let p be any distribution over $[n]$, let $\mathcal{I} = \{I_i\}_{i=1}^a$ be a (p, ϵ, a)-flat decomposition of $[n]$, and let $\mathcal{J} = \{J_i\}_{i=1}^b$ be a refinement of $\mathcal{I}$. Then $\mathcal{J}$ is a $(p, 2\epsilon, b)$-flat decomposition of $[n]$.*

We describe the TEST-IDENTITY-KNOWN-KMODAL algorithm below.

TEST-IDENTITY-KNOWN-KMODAL
Inputs: $\epsilon, \delta > 0$; sample access to k-modal distributions p, q over $[n]$

1. Run CONSTRUCT-FLAT-DECOMPOSITION$(p, \epsilon/2, \delta/4)$ and let $\mathcal{I} = \{I_i\}_{i=1}^{\ell}$, $\ell = O(k\log(n)/\epsilon^2)$, be the partition that it outputs. Run CONSTRUCT-FLAT-DECOMPOSITION$(p, \epsilon/2, \delta/4)$ and let $\mathcal{I}' = \{I'_i\}_{i=1}^{\ell'}$, $\ell' = O(k\log(n)/\epsilon^2)$, be the partition that it outputs. Let $\mathcal{J}$ be the common refinement of $\mathcal{I}$ and $\mathcal{I}$ and let $\ell_{\mathcal{J}} = O(k\log(n)/\epsilon^2)$ be the number of intervals in $\mathcal{J}$.
2. Let $(q_r)^{\mathcal{J}}$ denote the reduced distribution over $[\ell_{\mathcal{J}}]$ obtained from q using $\mathcal{J}$ as defined in Section A.
3. Draw $m = s_{IK}(\ell_{\mathcal{J}}, \epsilon/2, \delta/2)$ samples from $(p_r)^{\mathcal{J}}$, where $(p_r)^{\mathcal{J}}$ is the reduced distribution over $[\ell_{\mathcal{J}}]$ obtained from p using $\mathcal{J}$ as defined in Section A.
4. Run TEST-IDENTITY-KNOWN$((p_r)^{\mathcal{J}}, (q_r)^{\mathcal{J}}, \frac{\epsilon}{2}, \frac{\delta}{2})$ using the samples from Step 3 and output what it outputs.

We note that Steps 2, 3 and 4 of TEST-IDENTITY-KNOWN-KMODAL are the same as the corresponding steps of TEST-IDENTITY-KNOWN-MONOTONE. For the analysis of TEST-IDENTITY-KNOWN-KMODAL, Lemmas 3 and 4 give us that with probability $1-\delta/2$, the partition $\mathcal{J}$ obtained in Step 1 is both a $(p,\epsilon,\ell_{\mathcal{J}})$-flat and $(q,\epsilon,\ell_{\mathcal{J}})$-flat decomposition of $[n]$; we condition on this going forward. From this point on the analysis is essentially identical to the analysis for TEST-IDENTITY-KNOWN-MONOTONE and is omitted.

The modifications required to obtain algorithms TEST-IDENTITY-UNKNOWN-KMODAL, L_1-ESTIMATE-KNOWN-KMODAL and L_1-ESTIMATE-UNKNOWN-KMODAL, and the analysis of these algorithms, are completely analogous to the modifications and analyses of Section 3.2 and are omitted.

4.1 The CONSTRUCT-FLAT-DECOMPOSITION algorithm. We present CONSTRUCT-FLAT-DECOMPOSITION(p,ϵ,δ) followed by an intuitive explanation. Note that it employs a procedure ORIENTATION$(\widehat{p},I)$, which uses no samples and is presented and analyzed in Section 4.2.

CONSTRUCT-FLAT-DECOMPOSITION
INPUTS: $\epsilon,\delta>0$; sample access to k-modal distribution p over $[n]$

1. Initialize $\mathcal{I}:=\emptyset$.
2. Fix $\tau := \epsilon^2/(20000k)$. Draw $r=\Theta(\log(1/\delta)/\tau^2)$ samples from p and let $\widehat{p}$ denote the resulting empirical distribution (which by Theorem 1 has $d_{\mathrm{K}}(\widehat{p},p)\le\tau$ with probability at least $1-\delta$).
3. Greedily partition the domain $[n]$ into α *atomic intervals* $\{I_i\}_{i=1}^{\alpha}$ as follows: $I_1:=[1,j_1]$, where $j_1:=\min\{j\in[n]\mid \widehat{p}([1,j])\ge\epsilon/(100k)\}$. For $i\ge 1$, if $\cup_{j=1}^{i}I_j=[1,j_i]$, then $I_{i+1}:=[j_i+1,j_{i+1}]$, where j_{i+1} is defined as follows: If $\widehat{p}([j_i+1,n])\ge\epsilon/(100k)$, then $j_{i+1}:=\min\{j\in[n]\mid\widehat{p}([j_i+1,j])\ge\epsilon/(100k)\}$, otherwise, $j_{i+1}:=n$.
4. Construct a set of n_m *moderate intervals*, a set of n_h *heavy points*, and a set of n_n *negligible intervals* as follows: For each atomic interval $I_i=[a,b]$,
   (a) if $\widehat{p}([a,b])\le 3\epsilon/(100k)$ then I_i is declared to be a *moderate* interval;
   (b) otherwise we have $\widehat{p}([a,b])>3\epsilon/(100k)$ and we declare b to be a *heavy point*. If $a<b$ then we declare $[a,b-1]$ to be a *negligible* interval.

   For each interval I which is a heavy point, add I to $\mathcal{I}$. Add each negligible interval I to $\mathcal{I}$.
5. For each moderate interval I, run procedure ORIENTATION$(\widehat{p},I)$; let $\circ\in\{\uparrow,\downarrow,\perp\}$ be its output.

   If $\circ=\perp$ then add I to $\mathcal{I}$.

   If $\circ=\downarrow$ then let $\mathcal{J}_I$ be the partition of I given by Theorem 5 which is a $(p',\epsilon/4,O(\log(n)/\epsilon))$-flat decomposition of I for any non-increasing distribution p' over I. Add all the elements of $\mathcal{J}_I$ to $\mathcal{I}$.

   If $\circ=\uparrow$ then let $\mathcal{J}_I$ be the partition of I given by the dual version of Theorem 5, which is a $(p',\epsilon/4,O(\log(n)/\epsilon))$-flat decomposition of I for any non-decreasing distribution p' over I. Add all the elements of $\mathcal{J}_I$ to $\mathcal{I}$.
6. Output the partition $\mathcal{I}$ of $[n]$.

Roughly speaking, when CONSTRUCT-FLAT-DECOMPOSITION constructs a partition $\mathcal{I}$, it initially breaks $[n]$ up into two types of intervals. The first type are intervals that are "okay" to include in a flat decomposition, either because they have very little mass, or because they consist of a single point, or because they are close to uniform. The second type are intervals that are "not okay" to include in a flat decomposition – they have significant mass and are far from uniform – but the algorithm is able to ensure that almost all of these are monotone distributions with a known orientation. It then uses the oblivious decomposition of Theorem 5 to construct a flat decomposition of each such interval. (Note that it is crucial that the orientation is known in order to be able to use Theorem 5.)

In more detail, CONSTRUCT-FLAT-DECOMPOSITION(p,ϵ,δ) works as follows. The algorithm first draws a batch of samples from p and uses them to construct an estimate $\widehat{p}$ of the CDF of p (this is straightforward using the DKW inequality). Using $\widehat{p}$ the algorithm partitions $[n]$ into a collection of $O(k/\epsilon)$ disjoint intervals in the

following way:

- A small collection of the intervals are "negligible"; they collectively have total mass less than ϵ under p. Each negligible interval I will be an element of the partition $\mathcal{I}$.
- Some of the intervals are "heavy points"; these are intervals consisting of a single point that has mass $\Omega(\epsilon/k)$ under p. Each heavy point I will also be an element of the partition $\mathcal{I}$.
- The remaining intervals are "moderate" intervals, each of which has mass $\Theta(\epsilon/k)$ under p.

It remains to incorporate the moderate intervals into the partition $\mathcal{I}$ that is being constructed. This is done as follows: using $\widehat{p}$, the algorithm comes up with a "guess" of the correct orientation (non-increasing, non-decreasing, or close to uniform) for each moderate interval. Each moderate interval where the "guessed" orientation is "close to uniform" is included in the partition $\mathcal{I}$. Finally, for each moderate interval I where the guessed orientation is "non-increasing" or "non-decreasing", the algorithm invokes Theorem 5 on I to perform the oblivious decomposition for monotone distributions, and the resulting sub-intervals are included in $\mathcal{I}$. The analysis will show that the guesses are almost always correct, and intuitively this should imply that the $\mathcal{I}$ that is constructed is indeed a (p, ϵ, ℓ)-flat decomposition of $[n]$.

4.2 The ORIENTATION algorithm. The ORIENTATION algorithm takes as input an explicit distribution of a distribution $\widehat{p}$ over $[n]$ and an interval $I \subseteq [n]$. Intuitively, it assumes that $\widehat{p}_I$ is close (in Kolmogorov distance) to a monotone distribution p_I, and its goal is to determine the orientation of p_I: it outputs either $\uparrow$, $\downarrow$ or $\perp$ (the last of which means "close to uniform"). The algorithm is quite simple; it checks whether there exists an initial interval I' of I on which $\widehat{p}_I$'s weight is significantly different from $u_I(I')$ (the weight that the uniform distribution over I assigns to I') and bases its output on this in the obvious way. A precise description of the algorithm (which uses no samples) is given below.

ORIENTATION
INPUTS: explicit description of distribution $\widehat{p}$ over $[n]$; interval $I = [a, b] \subseteq [n]$

1. If $|I| = 1$ (i.e. $I = \{a\}$ for some $a \in [n]$) then return "$\perp$", otherwise continue.
2. If there is an initial interval $I' = [a, j]$ of I that satisfies $u_I(I') - (\widehat{p})_I(I') > \frac{\epsilon}{7}$ then halt and output "$\uparrow$". Otherwise,
3. If there is an initial interval $I' = [a, j]$ of I that satisfies $u_I(I') - (\widehat{p})_I(I') < -\frac{\epsilon}{7}$ then halt and output "$\downarrow$". Otherwise,
4. Output "$\perp$".

We proceed to analyze ORIENTATION. We show that if p_I is far from uniform then ORIENTATION outputs the correct orientation for it. We also show that whenever ORIENTATION does not output "$\perp$", whatever it outputs is the correct orientation of p_I. The proof is given in Section C.3.

Lemma 5. *Let p be a distribution over $[n]$ and let interval $I = [a, b] \subseteq [n]$ be such that p_I is monotone. Suppose $p(I) \geq 99\epsilon/(10000k)$, and suppose that for every interval $I' \subseteq I$ we have that $|\widehat{p}(I') - p(I')| \leq \frac{\epsilon^2}{10000k}$. Then*

1. *If p_I is non-decreasing and p_I is $\epsilon/6$-far from the uniform distribution u_I over I, then* ORIENTATION$(\widehat{p}, I)$ *outputs "$\uparrow$";*
2. *if* ORIENTATION$(\widehat{p}, I)$ *outputs "$\uparrow$" then p_I is non-decreasing;*
3. *if p_I is non-increasing and p_I is $\epsilon/6$-far from the uniform distribution u_I over I, then* ORIENTATION$(\widehat{p}, I)$ *outputs "$\downarrow$";*
4. *if* ORIENTATION$(\widehat{p}, I)$ *outputs "$\downarrow$" then p_I is non-increasing.*

5 Lower Bounds

Our algorithmic results follow from a reduction which shows how one can reduce the problem of testing properties of monotone or k-modal distributions to the task of testing properties of general distributions over a much smaller support. Our approach to proving lower bounds is complementary; we give a canonical scheme for transforming "lower bound instances" of general distributions to related lower bound instances of monotone distributions with much larger supports.

A generic lower bound instance for distance estimation has the following form: there is a distribution D over *pairs* of distributions, (p, p'), with the information theoretic guarantee that, given s independent samples from distributions p and p', with $(p, p') \leftarrow D$, it is impossible to distinguish the case that $d_{\mathrm{TV}}(p, p') \leq \epsilon_1$ versus $d_{\mathrm{TV}}(p, p') > \epsilon_2$ with any probability greater than $1 - \delta$, where the probability is taken over both the selection of $(p, p') \leftarrow D$ and the choice of samples. In general, such information theoretic lower bounds are difficult to prove. Fortunately, as mentioned above, we will be able to prove lower bounds for monotone and k-modal distributions by leveraging the known lower bound constructions in a black-box fashion.

Definitions 2 and 3, given below, define a two-stage transformation of a generic distribution into a related k-modal distribution over a much larger support. This transformation preserves total variation distance: for any pair of distributions, the variation distance between their transformations is identical to the variation distance between the original distributions. Additionally, we ensure that given access to s independent samples from an original input distribution, one can *simulate* drawing s samples from the related k-modal distribution yielded by the transformation. Given any lower–bound construction D for general distributions, the above transformation will yield a lower–bound instance D_k for $(k-1)$-modal distributions (so monotone distributions correspond to $k = 1$) defined by selecting a pair of distributions $(p, p') \leftarrow D$, then outputting the pair of transformed distributions. This transformed ensemble of distributions is a lower–bound instance, for if some algorithm could successfully test pairs of $(k-1)$-modal distributions from D_k, then that algorithm could be used to test pairs from D, by *simulating* samples drawn from the transformed versions of the distributions. The following proposition, proved in Section D, summarizes the above discussion:

Proposition 6. *Let D be a distribution over pairs of distributions supported on $[n]$ such that given s samples from distributions p, p' with $(p, p') \leftarrow D$, no algorithm can distinguish whether $d_{\mathrm{TV}}(p, p') \leq \epsilon_1$ versus $d_{\mathrm{TV}}(p, p') > \epsilon_2$ with probability greater than $1 - \delta$ (over both the draw of (p, p') from D and the draw of samples from p, p'). Let p_{max}, p_{min} be the respective maximum and minimum probabilities with which any element arises in distributions that are supported in D. Then there exists a distribution D_k over pairs of $(k-1)$-modal distributions supported on $[N] = [4ke^{\frac{8n}{k}(1+\log(p_{max}/p_{min}))}]$ such that no algorithm, when given s samples from distributions p_k, p'_k, with $(p_k, p'_k) \leftarrow D_k$, can distinguish whether $d_{\mathrm{TV}}(p_k, p'_k) \leq \epsilon_1$ versus $d_{\mathrm{TV}}(p_k, p'_k) > \epsilon_2$ with success probability greater than $1 - \delta$.*

Before proving this proposition, we state various corollaries which result from applying the Proposition to known lower-bound constructions for general distributions. The first is for the "testing identity, q is unknown" problem:

Corollary 7. *There exists a constant c such that for sufficiently large N and $1 \leq k = O(\log N)$, there is a distribution D_k over pairs of $2(k-1)$-modal distributions (p, p') over $[N]$, such that* no *algorithm, when given $c\left(\frac{k \log N}{\log \log N}\right)^{2/3}$ samples from a pair of distributions $(p, p') \leftarrow D$, can distinguish the case that $d_{\mathrm{TV}}(p, p') = 0$ from the case $d_{\mathrm{TV}}(p, p') > .5$ with probability at least .6.*

This Corollary gives the lower bounds stated in lines 2 and 6 of Table 1. It follows from applying Proposition 6 to a (trivially modified) version of the lower bound construction given in [BFR$^+$00, Val08b], summarized by the following theorem:

Theorem 6 ([BFR$^+$00, Val08b])**.** *There exists a constant c such that for sufficiently large n, there is a distribution D over pairs of distributions (p, p') over $[n]$, such that for any $(p, p') \leftarrow D$, the maximum probability with which*

any element occurs in p or p' is $\frac{1}{n^{2/3}}$, and the minimum probability is $\frac{1}{2n}$. Additionally, no algorithm, when given $cn^{2/3}$ samples from $(p, p') \leftarrow D$, can distinguish whether $d_{\mathrm{TV}}(p, p') = 0$, from $d_{\mathrm{TV}}(p, p') > .5$ with probability at least .6.

Our second corollary is for L_1 estimation, in the case that one of the distributions is explicitly given. This trivially also yields an equivalent lower bound for the setting in which both distributions are given via samples.

Corollary 8. *For any a, b with $0 < a < b < 1/2$, there exists a constant $c > 0$, such that for any sufficiently large N and $1 \leq k = O(\log N)$, there exists a $2(k-1)$-modal distribution q of support $[N]$, and a distribution D_k over $2(k-1)$-modal distributions over $[N]$, such that* no *algorithm, when given $c\frac{k \log N}{\log \log N \cdot \log \log \log N}$ samples from a distribution $p \leftarrow D$, can distinguish the case that $d_{\mathrm{TV}}(p, q) < a$ versus $d_{\mathrm{TV}}(p, p') > b$ with probability at least .6.*

This Corollary gives the lower bounds claimed in lines 3, 4, 7 and 8 of Table 1. It follows from applying Proposition 6 to the lower bound construction given in [VV11a], summarized by the following theorem:

Theorem 7 ([VV11a])**.** *For any a, b with $0 < a < b < 1/2$, there exists a constant $c > 0$, such that for any sufficiently large n, there is a distribution D over distributions with support $[n]$, such that for any $p \leftarrow D$, the maximum probability with which any element occurs in p is $O\left(\frac{\log n}{n}\right)$, and the minimum probability is $\frac{1}{2n}$. Additionally, no algorithm, when given $c\frac{n}{\log n}$ samples from $p \leftarrow D$ can distinguish whether $d_{\mathrm{TV}}(p, u_n) < a$ versus $d_{\mathrm{TV}}(p, u_n) > b$ with probability at least .6, where u_n denotes the uniform distribution over $[n]$.*

Note that the above theorem can be expressed in the language of Proposition 6 by defining the distribution D' over pairs of distributions which chooses a distribution according to D for the first distribution of each pair, and always selects u_n for the second distribution of each pair.

Our third corollary, which gives the lower bounds claimed in lines 1 and 5 of Table 1, is for the "testing identity, q is known" problem:

Corollary 9. *For any $\epsilon \in (0, 1/2]$, there is a constant c such that for sufficiently large N and $1 \leq k = O(\log m)$, there is a k-modal distribution p with support $[N]$, and a distribution D over $2(k-1)$-modal distributions of support $[N]$ such that* no *algorithm, when given $c(k \log m)^{1/2}$ samples from a distribution $p' \leftarrow D$, can distinguish the case that $d_{\mathrm{TV}}(p, p') = 0$ from the case $d_{\mathrm{TV}}(p, p') > \epsilon$ with probability at least .6.*

The above corollary follows from applying Proposition 6 to the following trivially verified lower bound construction:

Fact 10. *Let D be the ensemble of distributions of support n defined as follows: with probability $1/2$, $p \leftarrow D$ is the uniform distribution on support n, and with probability $1/2$, $p \leftarrow D$ assigns probability $1/2n$ to a random half of the domain elements, and probability $3/2n$ to the other half of the domain elements. No algorithm, when given fewer than $n^{1/2}/100$ samples from a distribution $p \leftarrow D$ can distinguish between $d_{\mathrm{TV}}(p, u_n) = 0$ versus $d_{\mathrm{TV}}(p, u_n) \geq .5$ with probability greater than .6.*

As noted previously (after Theorem 7), this fact can also be expressed in the language of Proposition 6.

6 Conclusions

We have introduced a simple new approach for tackling distribution testing problems for restricted classes of distributions, by reducing them to general-distribution testing problems over a smaller domain. We applied this approach to get new testing results for a range of distribution testing problems involving monotone and k-modal distributions, and established lower bounds showing that all our new algorithms are essentially optimal.

A general direction for future work is to apply our reduction method to obtain near-optimal testing algorithms for other interesting classes of distributions. This will involve constructing flat decompositions of various types of distributions using few samples, which seems to be a natural and interesting algorithmic problem. A specific goal is to develop a more efficient version of our CONSTRUCT-FLAT-DECOMPOSITION algorithm for k-modal distributions; is it possible to obtain an improved version of this algorithm that uses $o(k)$ samples?

References

[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 Conference on Foundations of Computer Science*, pages 442–451, 2001.

[BFR+00] Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing that distributions are close. In *IEEE Symposium on Foundations of Computer Science*, pages 259–269, 2000.

[BFR+10] T. Batu, L. Fortnow, R. Rubinfeld, W.D. Smith, and P. White. Testing closeness of discrete distributions, 2010.

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

[CKC83] L. Cobb, P. Koppstein, and N.H. Chen. Estimation and moment recursion relations for multimodal distributions of the exponential family. *J. American Statistical Association*, 78(381):124–130, 1983.

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

[DDS11] C. Daskalakis, I. Diakonikolas, and R.A. Servedio. Learning k-modal distributions via testing. Available at http://arxiv.org/abs/1107.2700, 2011.

[DKW56] A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Mathematical Statistics*, 27(3):642–669, 1956.

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

[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.

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

[Kem91] J.H.B. Kemperman. Mixtures with a limited number of modal intervals. *Annals of Statistics*, 19(4):2120–2144, 1991.

[Mas90] P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Annals of Probability*, 18(3):1269–1283, 1990.

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

[Val08a] P. Valiant. *Testing Symmetric Properties of Distributions*. PhD thesis, M.I.T., 2008.

[Val08b] Paul Valiant. Testing symmetric properties of distributions. In *STOC*, pages 383–392, 2008.

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

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

For simplicity, the appendix consists of a slightly expanded and self-contained version of the exposition in the body of the paper, following the "Notation and Preliminaries" section.

A Shrinking the domain size: Reductions for distribution-testing problems

In this section we present the general framework of our reduction-based approach and sketch how we instantiate this approach for monotone and k-modal distributions.

We denote by $|I|$ the cardinality of an interval $I \subseteq [n]$, i.e. for $I = [a, b]$ we have $|I| = b - a + 1$. Fix a distribution p over $[n]$ and a partition of $[n]$ into disjoint intervals $\mathcal{I} := \{I_i\}_{i=1}^{\ell}$. The *flattened distribution* $(p_f)^{\mathcal{I}}$ corresponding to p and $\mathcal{I}$ is the distribution over $[n]$ defined as follows: for $j \in [\ell]$ and $i \in I_j$, $(p_f)^{\mathcal{I}}(i) = \sum_{t \in I_j} p(t)/|I_j|$. That is, $(p_f)^{\mathcal{I}}$ is obtained from p by averaging the weight that p assigns to each interval over the entire interval. The *reduced distribution* $(p_r)^{\mathcal{I}}$ corresponding to p and $\mathcal{I}$ is the distribution over $[\ell]$ that assigns the ith point the weight p assigns to the interval I_i; i.e., for $i \in [\ell]$, we have $(p_r)^{\mathcal{I}}(i) = p(I_i)$. Note that if p is non-increasing then so is $(p_f)^{\mathcal{I}}$, but this is not necessarily the case for $(p_r)^{\mathcal{I}}$.

Definition 1. *Let p be a distribution over $[n]$ and let $\mathcal{I} = \{I_i\}_{i=1}^{\ell}$ be a partition of $[n]$ into disjoint intervals. We say that $\mathcal{I}$ is a* (p, ϵ, ℓ)-flat decomposition of $[n]$ *if $d_{\mathrm{TV}}(p, (p_f)^{\mathcal{I}}) \leq \epsilon$.*

The following useful lemma relates closeness of p and q to closeness of the reduced distributions:

Lemma 2 *Let $\mathcal{I} = \{I_i\}_{i=1}^{\ell}$ be a partition of $[n]$ into disjoint intervals. Suppose that p and q are distributions over $[n]$ such that $\mathcal{I}$ is both a (p, ϵ, ℓ)-flat decomposition of $[n]$ and is also a (q, ϵ, ℓ)-flat decomposition of $[n]$. Then $\left|d_{\mathrm{TV}}(p, q) - d_{\mathrm{TV}}((p_r)^{\mathcal{I}}, (q_r)^{\mathcal{I}})\right| \leq 2\epsilon$. Moreover, if $p = q$ then $(p_r)^{\mathcal{I}} = (q_r)^{\mathcal{I}}$.*

Proof. The second statement is clear by the definition of a reduced distribution. To prove the first statement, we first observe that for any pair of distributions p, q and any partition $\mathcal{I}$ of $[n]$ into disjoint intervals, we have that $d_{\mathrm{TV}}((p_r)^{\mathcal{I}}, (q_r)^{\mathcal{I}}) = d_{\mathrm{TV}}((p_f)^{\mathcal{I}}, (q_f)^{\mathcal{I}})$. We thus have that $\left|d_{\mathrm{TV}}(p, q) - d_{\mathrm{TV}}((p_r)^{\mathcal{I}}, (q_r)^{\mathcal{I}})\right|$ is equal to

$$\left|d_{\mathrm{TV}}(p, q) - d_{\mathrm{TV}}((p_f)^{\mathcal{I}}, (q_f)^{\mathcal{I}})\right| = d_{\mathrm{TV}}(p, q) - d_{\mathrm{TV}}((p_f)^{\mathcal{I}}, (q_f)^{\mathcal{I}}) \leq d_{\mathrm{TV}}(p, (p_f)^{\mathcal{I}}) + d_{\mathrm{TV}}(q, (q_f)^{\mathcal{I}}),$$

where the equality above is equivalent to $d_{\mathrm{TV}}(p, q) \geq d_{\mathrm{TV}}((p_f)^{\mathcal{I}}, (q_f)^{\mathcal{I}})$ (which is easily verified by considering each interval $I_i \in \mathcal{I}$ separately and applying triangle inequality) and the inequality is the triangle inequality. Since $\mathcal{I}$ is both a (p, ϵ, ℓ)-flat decomposition of $[n]$ and a (q, ϵ, ℓ)-flat decomposition of $[n]$, we have that $d_{\mathrm{TV}}(p, (p_f)^{\mathcal{I}}) \leq \epsilon$ and $d_{\mathrm{TV}}(q, (q_f)^{\mathcal{I}}) \leq \epsilon$. The RHS above is thus bounded by 2ϵ and the lemma follows. □

Lemma 2, while simple, is at the heart of our reduction-based approach; it lets us transform a distribution-testing problem over the large domain $[n]$ to a distribution-testing problem over the much smaller "reduced" domain $[\ell]$. At a high level, all our testing algorithms will follow the same basic approach: first they run a procedure which, with high probability, constructs a partition $\mathcal{I}$ of $[n]$ that is both a (p, ϵ, ℓ)-flat decomposition of $[n]$ and a (q, ϵ, ℓ)-flat decomposition of $[n]$. Next they run the appropriate general-distribution tester over the ℓ-element distributions $(p_r)^{\mathcal{I}}, (q_r)^{\mathcal{I}}$ and output what it outputs; Lemma 2 guarantees that the distance between $(p_r)^{\mathcal{I}}$ and $(q_r)^{\mathcal{I}}$ faithfully reflects the distance between p and q, so this output is correct.

We now provide a few more details that are specific to the various different testing problems that we consider. For the monotone distribution testing problems the construction of $\mathcal{I}$ is done obliviously (without drawing any samples or any reference to p or q of any sort) and there is no possibility of failure – the assumption that p and q

are both (say) non-decreasing guarantees that the $\mathcal{I}$ that is constructed is both a (p, ϵ, ℓ)-flat decomposition of $[n]$ and a (q, ϵ, ℓ)-flat decomposition of $[n]$. We describe this decomposition procedure in Section 3.1 and present our monotone distribution testing algorithms that are based on it in Section 3.2.

For the k-modal testing problems it is not so straightforward to construct the desired decomposition $\mathcal{I}$. This is done via a careful procedure which uses $k^2 \cdot \mathrm{poly}(1/\epsilon)$ samples from p and q. This procedure has the property that with probability $1-\delta/2$, the $\mathcal{I}$ it outputs is both a (p, ϵ, ℓ)-flat decomposition of $[n]$ and a (q, ϵ, ℓ)-flat decomposition of $[n]$, where $\ell = O(k \log(n)/\epsilon^2)$. Given this, by running a testing algorithm (which has success probability $1 - \delta/2$) on the pair $(p_r)^{\mathcal{I}}, (q_r)^{\mathcal{I}}$ of distributions over $[\ell]$, we will get an answer which is with probability $1 - \delta$ a legitimate answer for the original testing problem. The details are given in Section C.

We close this section with a result about partitions and flat decompositions which will be useful later. Let $\mathcal{I} = \{I_i\}_{i=1}^a, \mathcal{I}' = \{I'_j\}_{j=1}^b$ be two partitions of $[n]$. We say that $\mathcal{I}'$ is a *refinement* of $\mathcal{I}$ if for every $i \in [a]$ there is a subset S_i of $[b]$ such that $\cup_{j \in S_i} I'_j = I_i$ (note that for this to hold we must have $a \leq b$). Note that $\{S_i\}_{i=1}^a$ forms a partition of $[b]$. We prove the following useful lemma:

Lemma 4. *Let p be any distribution over $[n]$, let $\mathcal{I} = \{I_i\}_{i=1}^a$ be a (p, ϵ, a)-flat decomposition of $[n]$, and let $\mathcal{J} = \{J_i\}_{i=1}^b$ be a refinement of $\mathcal{I}$. Then $\mathcal{J}$ is a $(p, 2\epsilon, b)$-flat decomposition of $[n]$.*

Proof. Fix any $i \in [\ell]$ and let $S_i \subseteq [b]$ be such that $I_i = \cup_{j \in S_i} J_j$. To prove the lemma it suffices to show that

$$2 \sum_{t \in I_i} |p(t) - (p_f)^{\mathcal{I}}(t)| \geq \sum_{j \in S} \sum_{t \in J_j} |p(t) - (p_f)^{\mathcal{J}}(t)|, \tag{1}$$

since the sum on the LHS is the contribution that I_i makes to $d_{\mathrm{TV}}(p, (p_f)^{\mathcal{I}})$ and the sum on the RHS is the corresponding contribution I_i makes to $d_{\mathrm{TV}}(p, (p_f)^{\mathcal{J}})$. It may seem intuitively obvious that the sum on the LHS (which corresponds to approximating the sub-distribution p^{I_i} using a "global average") must be smaller than the sum on the RHS (which corresponds to using separate "local averages"). However, this intuition is not quite correct, and it is necessary to have the factor of two. To see this, consider a distribution p over $[n]$ such that $p(1) = (1/2) \cdot (1/n)$; $p(i) = 1/n$ for $i \in [2, n-1]$; and $p(n) = (3/2) \cdot (1/n)$. Taking $I_1 = [1, n/2]$ and $I_2 = [n/2 + 1, n]$, it is easy to check that inequality (1) is essentially tight (up to a $o(1)$ factor).

We now proceed to establish (1). Let $T \subseteq [n]$ and consider a partition of T into k nonempty sets T_i, $i \in [k]$. Denote $\mu \stackrel{\text{def}}{=} p(T)/|T|$ and $\mu_i \stackrel{\text{def}}{=} p(T_i)/|T_i|$. Then, (1) can be re-expressed as follows

$$2 \sum_{t \in T} |p(t) - \mu| \geq \sum_{i=1}^{k} \sum_{t \in T_i} |p(t) - \mu_i|. \tag{2}$$

We shall prove the above statement for all sequences of numbers $p(1), \ldots, p(n)$. Since adding or subtracting the same quantity from each number $p(t)$ does not change the validity of (2), for the sake of convenience we may assume all the numbers average to 0, that is, $\mu = 0$. Consider the i-th term on the right hand side, $\sum_{t \in T_i} |p(t) - \mu_i|$. We can bound this quantity from above as follows:

$$\sum_{t \in T_i} |p(t) - \mu_i| \leq \sum_{t \in T_i} |p(t)| + |T_i| \cdot |\mu_i| = \sum_{t \in T_i} |p(t)| + |p(T_i)| = 2 \sum_{t \in T_i} |p(t)| = 2 \sum_{t \in T_i} |p(t) - \mu|,$$

where the inequality follows from the triangle inequality (applied term by term), the first equality is by the definition of μ_i, the second equality is trivial, and the final equality uses the assumption that $\mu = 0$. The lemma follows by summing over $i \in [k]$, using the fact that the T_i's form a partition of T. □

B Efficiently Testing Monotone Distributions

B.1 Oblivious decomposition of monotone distributions Our main tool for testing monotone distributions is an *oblivious decomposition* of monotone distributions that is a variant of a construction of Birgé [Bir87b]. As we

will see it enables us to reduce the problem of testing a monotone distribution to the problem of testing an arbitrary distribution over a much smaller domain. The decomposition result is given below:

Theorem 5 ([Bir87b]). *(oblivious decomposition) Fix any $n \in \mathbb{Z}^+$ and $\epsilon > 0$. The partition $\mathcal{I} := \{I_i\}_{i=1}^{\ell}$ of $[n]$ described below has the following properties: $\ell = O((1/\epsilon) \cdot \log(\epsilon \cdot n + 1))$, and for any non-increasing distribution p over $[n]$, $\mathcal{I}$ is a $(p, O(\epsilon), \ell)$-flat decomposition of $[n]$.*

There is a dual version of Theorem 5, asserting the existence of an "oblivious" partition for non-decreasing distributions (which is of course different from the "oblivious" partition $\mathcal{I}$ for non-increasing distributions of Theorem 5); this will be useful later.

While our construction is essentially that of Birgé, we note that the version given in [Bir87b] is for non-increasing distributions over the continuous domain $[0, n]$, and it is phrased rather differently. Adapting the arguments of [Bir87b] to our discrete setting of distributions over $[n]$ is not conceptually difficult but requires some care. For the sake of being self-contained we provide a self-contained proof of the discrete version, stated above, that we require in Appendix E.

B.2 Efficiently testing monotone distributions Now we are ready to establish our upper bounds on testing monotone distributions (given in the first four rows of Table 1). All of the algorithms are essentially the same: each works by reducing the given monotone distribution testing problem to the same testing problem for arbitrary distributions over support of size $\ell = O(\log n/\epsilon)$ using the oblivious decomposition from the previous subsection. For concreteness we explicitly describe the tester for the "testing identity, q is known" case below, and then indicate the small changes that are necessary to get the testers for the other three cases.

TEST-IDENTITY-KNOWN-MONOTONE
Inputs: $\epsilon, \delta > 0$; sample access to non-increasing distribution p over $[n]$; explicit description of non-increasing distribution q over $[n]$

1. Let $\mathcal{I} := \{I_i\}_{i=1}^{\ell}$, with $\ell = \Theta(\log(\epsilon n + 1)/\epsilon)$, be the partition of $[n]$ given by Theorem 5, which is a $(p', \epsilon/8, \ell)$-flat decomposition of $[n]$ for any non-increasing distribution p'.
2. Let $(q_r)^{\mathcal{I}}$ denote the reduced distribution over $[\ell]$ obtained from q using $\mathcal{I}$ as defined in Section A.
3. Draw $m = s_{IK}(\ell, \epsilon/2, \delta)$ samples from $(p_r)^{\mathcal{I}}$, where $(p_r)^{\mathcal{I}}$ is the reduced distribution over $[\ell]$ obtained from p using $\mathcal{I}$ as defined in Section A.
4. Run TEST-IDENTITY-KNOWN$((p_r)^{\mathcal{I}}, (q_r)^{\mathcal{I}}, \frac{\epsilon}{2}, \delta)$ using the samples from Step 3 and output what it outputs.

We now establish our claimed upper bound for the "testing identity, q is known" case. We first observe that in Step 3, the desired $m = s_{IK}(\ell, \epsilon/2, \delta)$ samples from $(p_r)^{\mathcal{I}}$ can easily be obtained by drawing m samples from p and converting each one to the corresponding draw from $(p_r)^{\mathcal{I}}$ in the obvious way. If $p = q$ then by Lemma 2 we have that $(p_r)^{\mathcal{I}} = (q_r)^{\mathcal{I}}$, and TEST-IDENTITY-KNOWN-MONOTONE outputs "accept" with probability at least $1 - \delta$ by Theorem 2. If $d_{\mathrm{TV}}(p, q) \geq \epsilon$, then by Lemma 2 and Theorem 5 we have that $d_{\mathrm{TV}}((p_r)^{\mathcal{I}}, (q_r)^{\mathcal{I}}) \geq 3\epsilon/4$, so TEST-IDENTITY-KNOWN-MONOTONE outputs "reject" with probability at least $1 - \delta$ by Theorem 2. For the "testing identity, q is unknown" case, the the algorithm TEST-IDENTITY-UNKNOWN-MONOTONE is very similar to TEST-IDENTITY-KNOWN-MONOTONE. The differences are as follows: instead of Step 2, in Step 3 we draw $m = s_{IU}(\ell, \epsilon/2, \delta)$ samples from $(p_r)^{\mathcal{I}}$ and the same number of samples from $(q_r)^{\mathcal{I}}$; and in Step 4, we run TEST-IDENTITY-UNKNOWN$((p_r)^{\mathcal{I}}, (q_r)^{\mathcal{I}}, \frac{\epsilon}{2}, \delta)$ using the samples from Step 3. The analysis is exactly the same as above (using Theorem 3 in place of Theorem 2).

We now describe the algorithm L_1-ESTIMATE-KNOWN-MONOTONE for the "tolerant testing, q is known" case. This algorithm takes values ϵ and δ as input, so the partition $\mathcal{I}$ defined in Step 1 is a $(p', \epsilon/4, \ell)$-flat decomposition of $[n]$ for any non-increasing p'. In Step 3 the algorithm draws $m = s_E(\ell, \epsilon/2, \delta)$ samples and runs L_1-ESTIMATE$((p_r)^{\mathcal{I}}, (q_r)^{\mathcal{I}}, \epsilon/2, \delta)$ in Step 4. If $d_{\mathrm{TV}}(p, q) = c$ then by Lemma 2 we have that $d_{\mathrm{TV}}((p_r)^{\mathcal{I}}, (q_r)^{\mathcal{I}}) \in$

$[c-\epsilon/2, c+\epsilon/2]$ and L_1-ESTIMATE-KNOWN-MONOTONE outputs a value within the prescribed range with probability at least $1-\delta$, by Theorem 4. The algorithm L_1-ESTIMATE-UNKNOWN-MONOTONE case and its analysis are entirely similar.

C Efficiently Testing k-modal Distributions

In this section we establish our main positive testing results for k-modal distributions, the upper bounds stated in the final four rows of Table 1. The key to all these results is an algorithm CONSTRUCT-FLAT-DECOMPOSITION(p, ϵ, δ). We prove the following performance guarantee about this algorithm:

Lemma 3. *Let p be a k-modal distribution over $[n]$. Algorithm* CONSTRUCT-FLAT-DECOMPOSITION(p, ϵ, δ) *draws $O(k^2\epsilon^{-4}\log(1/\delta))$ samples from p and outputs a (p, ϵ, ℓ)-flat decomposition of $[n]$ with probability at least $1-\delta$, where $\ell = O(k\log(n)/\epsilon^2)$.*

The bulk of our work in Section C is to describe CONSTRUCT-FLAT-DECOMPOSITION(p, ϵ, δ) and prove Lemma 3, but first we show how Lemma 3 easily yields our claimed testing results for k-modal distributions. As in the monotone case all four algorithms are essentially the same: each works by reducing the given k-modal distribution testing problem to the same testing problem for arbitrary distributions over $[\ell]$. We describe the TEST-IDENTITY-KNOWN-KMODAL algorithm below, and then indicate the necessary changes to get the other three testers.

The following terminology will be useful: Let $\mathcal{I} = \{I_i\}_{i=1}^r$ and $\mathcal{I}' = \{I'_i\}_{i=1}^s$ be two partitions of $[n]$ into r and s intervals respectively. The *common refinement* of $\mathcal{I}$ and $\mathcal{I}'$ is the partition $\mathcal{J}$ of $[n]$ into intervals obtained from $\mathcal{I}$ and $\mathcal{I}'$ in the obvious way, by taking all possible nonempty intervals of the form $I_i \cap I'_j$. It is clear that $\mathcal{J}$ is both a refinement of $\mathcal{I}$ and of $\mathcal{I}'$ and that the number of intervals $|\mathcal{J}|$ in $\mathcal{J}$ is at most $r+s$.

TEST-IDENTITY-KNOWN-KMODAL
Inputs: $\epsilon, \delta > 0$; sample access to k-modal distributions p, q over $[n]$

1. Run CONSTRUCT-FLAT-DECOMPOSITION$(p, \epsilon/2, \delta/4)$ and let $\mathcal{I} = \{I_i\}_{i=1}^{\ell}$, $\ell = O(k\log(n)/\epsilon^2)$, be the partition that it outputs. Run CONSTRUCT-FLAT-DECOMPOSITION$(p, \epsilon/2, \delta/4)$ and let $\mathcal{I}' = \{I'_i\}_{i=1}^{\ell'}$, $\ell' = O(k\log(n)/\epsilon^2)$, be the partition that it outputs. Let $\mathcal{J}$ be the common refinement of $\mathcal{I}$ and $\mathcal{I}$ and let $\ell_{\mathcal{J}} = O(k\log(n)/\epsilon^2)$ be the number of intervals in $\mathcal{J}$.
2. Let $(q_r)^{\mathcal{J}}$ denote the reduced distribution over $[\ell_{\mathcal{J}}]$ obtained from q using $\mathcal{J}$ as defined in Section A.
3. Draw $m = s_{IK}(\ell_{\mathcal{J}}, \epsilon/2, \delta/2)$ samples from $(p_r)^{\mathcal{J}}$, where $(p_r)^{\mathcal{J}}$ is the reduced distribution over $[\ell_{\mathcal{J}}]$ obtained from p using $\mathcal{J}$ as defined in Section A.
4. Run TEST-IDENTITY-KNOWN$((p_r)^{\mathcal{J}}, (q_r)^{\mathcal{J}}, \frac{\epsilon}{2}, \frac{\delta}{2})$ using the samples from Step 3 and output what it outputs.

We note that Steps 2, 3 and 4 of TEST-IDENTITY-KNOWN-KMODAL are the same as the corresponding steps of TEST-IDENTITY-KNOWN-MONOTONE. For the analysis of TEST-IDENTITY-KNOWN-KMODAL, Lemmas 3 and 4 give us that with probability $1-\delta/2$, the partition $\mathcal{J}$ obtained in Step 1 is both a $(p, \epsilon, \ell_{\mathcal{J}})$-flat and $(q, \epsilon, \ell_{\mathcal{J}})$-flat decomposition of $[n]$; we condition on this going forward. From this point on the analysis is essentially identical to the analysis for TEST-IDENTITY-KNOWN-MONOTONE and is omitted.

The modifications required to obtain algorithms TEST-IDENTITY-UNKNOWN-KMODAL, L_1-ESTIMATE-KNOWN-KMODAL and L_1-ESTIMATE-UNKNOWN-KMODAL, and the analysis of these algorithms, are completely analogous to the modifications and analyses of Appendix B.2 and are omitted.

C.1 The CONSTRUCT-FLAT-DECOMPOSITION algorithm. We present CONSTRUCT-FLAT-DECOMPOSITION(p, ϵ, δ) followed by an intuitive explanation. Note that it employs a procedure ORIENTATION$(\widehat{p}, I)$, which uses no samples and is presented and analyzed in Section 4.2.

CONSTRUCT-FLAT-DECOMPOSITION
INPUTS: $\epsilon, \delta > 0$; sample access to k-modal distribution p over $[n]$

1. Initialize $\mathcal{I} := \emptyset$.
2. Fix $\tau := \epsilon^2/(20000k)$. Draw $r = \Theta(\log(1/\delta)/\tau^2)$ samples from p and let $\widehat{p}$ denote the resulting empirical distribution (which by Theorem 1 has $d_{\mathrm{K}}(\widehat{p}, p) \leq \tau$ with probability at least $1 - \delta$).
3. Greedily partition the domain $[n]$ into α *atomic intervals* $\{I_i\}_{i=1}^{\alpha}$ as follows: $I_1 := [1, j_1]$, where $j_1 := \min\{j \in [n] \mid \widehat{p}([1, j]) \geq \epsilon/(100k)\}$. For $i \geq 1$, if $\cup_{j=1}^{i} I_j = [1, j_i]$, then $I_{i+1} := [j_i + 1, j_{i+1}]$, where j_{i+1} is defined as follows: If $\widehat{p}([j_i + 1, n]) \geq \epsilon/(100k)$, then $j_{i+1} := \min\{j \in [n] \mid \widehat{p}([j_i + 1, j]) \geq \epsilon/(100k)\}$, otherwise, $j_{i+1} := n$.
4. Construct a set of n_m *moderate intervals*, a set of n_h *heavy points*, and a set of n_n *negligible intervals* as follows: For each atomic interval $I_i = [a, b]$,
   (a) if $\widehat{p}([a, b]) \leq 3\epsilon/(100k)$ then I_i is declared to be a *moderate* interval;
   (b) otherwise we have $\widehat{p}([a, b]) > 3\epsilon/(100k)$ and we declare b to be a *heavy point*. If $a < b$ then we declare $[a, b - 1]$ to be a *negligible* interval.

   For each interval I which is a heavy point, add I to $\mathcal{I}$. Add each negligible interval I to $\mathcal{I}$.
5. For each moderate interval I, run procedure ORIENTATION$(\widehat{p}, I)$; let $\circ \in \{\uparrow, \downarrow, \perp\}$ be its output.

   If $\circ = \perp$ then add I to $\mathcal{I}$.

   If $\circ = \downarrow$ then let $\mathcal{J}_I$ be the partition of I given by Theorem 5 which is a $(p', \epsilon/4, O(\log(n)/\epsilon))$-flat decomposition of I for any non-increasing distribution p' over I. Add all the elements of $\mathcal{J}_I$ to $\mathcal{I}$.

   If $\circ = \uparrow$ then let $\mathcal{J}_I$ be the partition of I given by the dual version of Theorem 5, which is a $(p', \epsilon/4, O(\log(n)/\epsilon))$-flat decomposition of I for any non-decreasing distribution p' over I. Add all the elements of $\mathcal{J}_I$ to $\mathcal{I}$.
6. Output the partition $\mathcal{I}$ of $[n]$.

Roughly speaking, when CONSTRUCT-FLAT-DECOMPOSITION constructs a partition $\mathcal{I}$, it initially breaks $[n]$ up into two types of intervals. The first type are intervals that are "okay" to include in a flat decomposition, either because they have very little mass, or because they consist of a single point, or because they are close to uniform. The second type are intervals that are "not okay" to include in a flat decomposition – they have significant mass and are far from uniform – but the algorithm is able to ensure that almost all of these are monotone distributions with a known orientation. It then uses the oblivious decomposition of Theorem 5 to construct a flat decomposition of each such interval. (Note that it is crucial that the orientation is known in order to be able to use Theorem 5.)

In more detail, CONSTRUCT-FLAT-DECOMPOSITION(p, ϵ, δ) works as follows. The algorithm first draws a batch of samples from p and uses them to construct an estimate $\widehat{p}$ of the CDF of p (this is straightforward using the DKW inequality). Using $\widehat{p}$ the algorithm partitions $[n]$ into a collection of $O(k/\epsilon)$ disjoint intervals in the following way:

- A small collection of the intervals are "negligible"; they collectively have total mass less than ϵ under p. Each negligible interval I will be an element of the partition $\mathcal{I}$.
- Some of the intervals are "heavy points"; these are intervals consisting of a single point that has mass $\Omega(\epsilon/k)$ under p. Each heavy point I will also be an element of the partition $\mathcal{I}$.
- The remaining intervals are "moderate" intervals, each of which has mass $\Theta(\epsilon/k)$ under p.

It remains to incorporate the moderate intervals into the partition $\mathcal{I}$ that is being constructed. This is done as follows: using $\widehat{p}$, the algorithm comes up with a "guess" of the correct orientation (non-increasing, non-decreasing, or close to uniform) for each moderate interval. Each moderate interval where the "guessed" orientation is "close to uniform" is included in the partition $\mathcal{I}$. Finally, for each moderate interval I where the guessed orientation is "non-increasing" or "non-decreasing", the algorithm invokes Theorem 5 on I to perform the oblivious decomposition for

monotone distributions, and the resulting sub-intervals are included in $\mathcal{I}$. The analysis will show that the guesses are almost always correct, and intuitively this should imply that the $\mathcal{I}$ that is constructed is indeed a (p, ϵ, ℓ)-flat decomposition of $[n]$.

C.2 Performance of CONSTRUCT-FLAT-DECOMPOSITION: Proof of Lemma 3. The claimed sample bound is obvious from inspection of the algorithm, as the only step that draws any samples is Step 2. The bound on the number of intervals in the flat decomposition follows directly from the upper bounds on the number of heavy points, negligible intervals and moderate intervals shown below, using also Theorem 5. It remains to show that the output of the algorithm is a valid flat decomposition of p. First, by the DKW inequality (Theorem 1) we have that with probability at least $1 - \delta$ it is the case that

$$|\widehat{p}(I) - p(I)| \leq \frac{\epsilon^2}{10000k}, \text{ for every interval } I \subseteq [n]. \tag{3}$$

We make some preliminary observations about the weight that p has on the intervals constructed in Steps 4 and 5. Since every atomic interval I_i constructed in Step 4 has $\widehat{p}(I) \geq \epsilon/(100k)$ (except potentially the rightmost one), it follows that the number α of atomic intervals constructed in Step 3 satisfies

$$\alpha \leq \lceil 100k/\epsilon \rceil.$$

We now establish bounds on the probability mass that p assigns to the moderate intervals, heavy points, and negligible intervals that are constructed in Step 4. Using (3), each interval I_i that is declared to be a moderate interval in Step 4(a) must satisfy

$$99\epsilon/(10000k) \leq p([a,b]) \leq 301\epsilon/(10000k) \quad \text{(for all moderate intervals } [a,b]). \tag{4}$$

By virtue of the greedy process that is used to construct atomic intervals in Step 3, each point b that is declared to be a heavy point in Step 4(b) must satisfy $\widehat{p}(b) \geq 2\epsilon/(100k)$ and thus using (3) again

$$p(b) \geq 199\epsilon/(10000k) \quad \text{(for all heavy points } b). \tag{5}$$

Moreover, each interval $[a, b-1]$ that is declared to be a negligible interval must satisfy $\widehat{p}([a,b-1]) < \epsilon/(100k)$ and thus using (3) again

$$p([a,b-1]) \leq 101\epsilon/(10000k) \quad \text{(for all negligible intervals } [a,b-1]). \tag{6}$$

It is clear that n_m (the number of moderate intervals) and n_h (the number of heavy points) are each at most α. Next we observe that the number of negligible intervals n_n satisfies

$$n_n \leq k.$$

This is because at the end of each negligible interval $[a, b-1]$ we have (observing that each negligible interval must be nonempty) that $p(b-1) \leq p([a,b-1]) \leq 101\epsilon/(10000k)$ while $p(b) \geq 199\epsilon/(10000k)$. Since p is k-modal, there can be at most $\lceil (k+1)/2 \rceil \leq k$ points $b \in [n]$ satisfying this condition. Since each negligible interval I satisfies $p(I) \leq 101\epsilon/(10000k)$ we have that the total probability mass under p of all the negligible intervals is at most $101\epsilon/10000$.

Thus far we have built a partition of $[n]$ into a collection of $n_m \leq \lceil 100k/\epsilon \rceil$ moderate intervals (which we denote $M_1, \ldots, M_{n_m}$), a set of $n_h \leq \lceil 100k/\epsilon \rceil$ heavy points (which we denote $h_1, \ldots, h_{n_h}$) and a set of $n_n \leq k$ negligible intervals (which we denote $N_1, \ldots, N_{n_n}$). Let $A \subseteq \{1, \ldots, n_m\}$ denote the set of those indices i such that ORIENTATION$(\widehat{p}, M_i)$ outputs $\perp$ in Step 6. The partition $\mathcal{I}$ that CONSTRUCT-FLAT-DECOMPOSITION constructs consists of $\{h_1\}, \ldots, \{h_{n_h}\}$, $N_1, \ldots, N_{n_n}$, $\{M_i\}_{i \in A}$, and $\bigcup_{i \in ([n_m] \setminus A)} \mathcal{J}_{M_i}$. We can thus write p as

$$p = \sum_{j=1}^{n_h} p(h_j) \cdot \mathbf{1}_{h_j} + \sum_{j=1}^{n_n} p(N_j) p_{N_j} + \sum_{j \in A} p(M_j) p_{M_j} + \sum_{j \in ([n_m] \setminus A)} \sum_{I \in \mathcal{J}_{M_j}} p(I) p_I. \tag{7}$$

Using Lemma 15 (proved in Appendix F) we can bound the total variation distance between p and $(p_f)^{\mathcal{I}}$ by

$$\begin{aligned} d_{\mathrm{TV}}(p,(p_f)^{\mathcal{I}}) \;\leq\; & \frac{1}{2}\sum_{j=1}^{n_h}|p(h_j)-(p_f)^{\mathcal{I}}(h_j)| + \frac{1}{2}\sum_{j=1}^{n_n}|p(N_j)-(p_f)^{\mathcal{I}}(N_j)| + \sum_{j=1}^{n_n} p(N_j)\cdot d_{\mathrm{TV}}(p_{N_j},((p_f)^{\mathcal{I}})_{N_j}) \\ & + \frac{1}{2}\sum_{j\in A}|p(M_j)-(p_f)^{\mathcal{I}}(M_j)| + \sum_{j\in A} p(M_j)\cdot d_{\mathrm{TV}}(p_{M_j},((p_f)^{\mathcal{I}})_{M_j}) \\ & + \frac{1}{2}\sum_{j\in([n_m]\setminus A)}\sum_{I\in J_{M_j}}|p(I)-(p_f)^{\mathcal{I}}(I)| + \sum_{j\in([n_m]\setminus A)}\sum_{I\in J_{M_j}} p(I)\cdot d_{\mathrm{TV}}(p_I,((p_f)^{\mathcal{I}})_I). \end{aligned} \tag{8}$$

Since $p(I)=(p_f)^{\mathcal{I}}(I)$ for every $I\in\mathcal{I}$, this simplifies to

$$\begin{aligned} d_{\mathrm{TV}}(p,(p_f)^{\mathcal{I}}) \;\leq\; & \sum_{j=1}^{n_n} p(N_j)\cdot d_{\mathrm{TV}}(p_{N_j},((p_f)^{\mathcal{I}})_{N_j}) + \sum_{j\in A} p(M_j)\cdot d_{\mathrm{TV}}(p_{M_j},((p_f)^{\mathcal{I}})_{M_j}) \\ & + \sum_{j\in([n_m]\setminus A)}\sum_{I\in J_{M_j}} p(I)\cdot d_{\mathrm{TV}}(p_I,((p_f)^{\mathcal{I}})_I). \end{aligned} \tag{9}$$

which we now proceed to bound.

Recalling from (6) that $p(N_j)\leq 101\epsilon/(10000k)$ for each negligible interval N_j, and recalling that $n_n\leq k$, the first summand in (9) is at most $101\epsilon/10000$.

To bound the second summand, fix any $j\in A$ so M_j is a moderate interval such that $\textsc{Orientation}(\widehat{p},M_j)$ returns $\perp$. If p_{M_j} is non-decreasing then by Lemma 5 it must be the case that $d_{\mathrm{TV}}(p_{M_j},((p_f)^{\mathcal{I}})_{M_j})\leq\epsilon/6$ (note that $((p_f)^{\mathcal{I}})_{M_j}$ is just u_{M_j}, the uniform distribution over M_j). Lemma 5 gives the same bound if p_{M_j} is non-increasing. If p_{M_j} is neither non-increasing nor non-decreasing then we have no nontrivial bound on $d_{\mathrm{TV}}(p_{M_j},((p_f)^{\mathcal{I}})_{M_j})$, but since p is k-modal there can be at most k such values of j in A. Recalling (4), overall we have that

$$\sum_{j\in A} p(M_j)\cdot d_{\mathrm{TV}}(p_{M_j},((p_f)^{\mathcal{I}})_{M_j}) \leq \frac{301\epsilon k}{10000k}+\frac{\epsilon}{6}\leq\frac{1968\epsilon}{10000},$$

and we have bounded the second summand.

It remains to bound the final summand of (9). For each $j\in([n_m]\setminus A)$, we know that $\textsc{Orientation}(\widehat{p},M_j)$ outputs either $\uparrow$ or $\downarrow$. If p_{M_j} is monotone, then by Lemma 5 we have that the output of $\textsc{Orientation}(\widehat{p},M_j)$ gives the correct orientation of p_{M_j}. Consequently $\mathcal{J}_{M_j}$ is a $(p_{M_j},\epsilon/4,O(\log(n)/\epsilon))$-flat decomposition of M_j, by Theorem 5. This means that $d_{\mathrm{TV}}(p_{M_j},((p_f)^{\mathcal{I}})_{M_j})\leq\epsilon/4$, which is equivalent to

$$\frac{1}{p(M_j)}\sum_{I\in\mathcal{J}_{M_j}} p(I)d_{\mathrm{TV}}(p_I,((p_f)^{\mathcal{I}})_I)\leq\frac{\epsilon}{4},\quad\text{i.e.}\quad\sum_{I\in\mathcal{J}_{M_j}} p(I)d_{\mathrm{TV}}(p_I,((p_f)^{\mathcal{I}})_I)\leq p(M_j)\cdot\frac{\epsilon}{4}.$$

Let $B\subset[n_m]\setminus A$ be such that, for all $j\in B$, p_{M_j} is monotone. Summing the above over all $j\in B$ gives:

$$\sum_{j\in B}\sum_{I\in\mathcal{J}_{M_j}} p(I)d_{\mathrm{TV}}(p_I,((p_f)^{\mathcal{I}})_I)\leq\sum_{j\in B} p(M_j)\cdot\frac{\epsilon}{4}\leq\frac{\epsilon}{4}.$$

Given that p is k-modal, the cardinality of the set $[n_m]\setminus(A\cup B)$ is at most k. So we have the bound:

$$\sum_{j\in[n_m]\setminus(A\cup B)}\sum_{I\in\mathcal{J}_{M_j}} p(I)d_{\mathrm{TV}}(p_I,((p_f)^{\mathcal{I}})_I)\leq\sum_{j\in[n_m]\setminus(A\cup B)} p(M_j)\leq\frac{301\epsilon k}{10000k}.$$

So the third summand of (9) is at most $\epsilon/4+301\epsilon/10000$, and overall we have that $(9)\leq\frac{\epsilon}{2}$. Hence, we have shown that $\mathcal{I}$ is a (p,ϵ,ℓ)-flat decomposition of $[n]$, and Lemma 3 is proved.

C.3 The ORIENTATION algorithm. The ORIENTATION algorithm takes as input an explicit distribution of a distribution $\widehat{p}$ over $[n]$ and an interval $I \subseteq [n]$. Intuitively, it assumes that $\widehat{p}_I$ is close (in Kolmogorov distance) to a monotone distribution p_I, and its goal is to determine the orientation of p_I: it outputs either $\uparrow, \downarrow$ or $\perp$ (the last of which means "close to uniform"). The algorithm is quite simple; it checks whether there exists an initial interval I' of I on which $\widehat{p}_I$'s weight is significantly different from $u_I(I')$ (the weight that the uniform distribution over I assigns to I') and bases its output on this in the obvious way. A precise description of the algorithm (which uses no samples) is given below.

ORIENTATION
INPUTS: explicit description of distribution $\widehat{p}$ over $[n]$; interval $I = [a, b] \subseteq [n]$

1. If $|I| = 1$ (i.e. $I = \{a\}$ for some $a \in [n]$) then return "$\perp$", otherwise continue.
2. If there is an initial interval $I' = [a, j]$ of I that satisfies $u_I(I') - (\widehat{p})_I(I') > \frac{\epsilon}{7}$ then halt and output "$\uparrow$". Otherwise,
3. If there is an initial interval $I' = [a, j]$ of I that satisfies $u_I(I') - (\widehat{p})_I(I') < -\frac{\epsilon}{7}$ then halt and output "$\downarrow$". Otherwise,
4. Output "$\perp$".

We proceed to analyze ORIENTATION. We show that if p_I is far from uniform then ORIENTATION outputs the correct orientation for it. We also show that whenever ORIENTATION does not output "$\perp$", whatever it outputs is the correct orientation of p_I. For ease of readability, for the rest of this subsection we use the following notation:

$$\Delta := \frac{\epsilon^2}{10000k}$$

Lemma 5. *Let p be a distribution over $[n]$ and let interval $I = [a, b] \subseteq [n]$ be such that p_I is monotone. Suppose $p(I) \geq 99\epsilon/(10000k)$, and suppose that for every interval $I' \subseteq I$ we have that*

$$|\widehat{p}(I') - p(I')| \leq \Delta. \tag{10}$$

Then

1. *If p_I is non-decreasing and p_I is $\epsilon/6$-far from the uniform distribution u_I over I, then* ORIENTATION$(\widehat{p}, I)$ *outputs "$\uparrow$";*
2. *if* ORIENTATION$(\widehat{p}, I)$ *outputs "$\uparrow$" then p_I is non-decreasing;*
3. *if p_I is non-increasing and p_I is $\epsilon/6$-far from the uniform distribution u_I over I, then* ORIENTATION$(\widehat{p}, I)$ *outputs "$\downarrow$";*
4. *if* ORIENTATION$(\widehat{p}, I)$ *outputs "$\downarrow$" then p_I is non-increasing.*

Proof. Let $I' = [a, j] \subseteq I$ be any initial interval of I. We first establish the upper bound

$$|p_I(I') - (\widehat{p})_I(I')| \leq \epsilon/49 \tag{11}$$

as this will be useful for the rest of the proof. Using (10) we have

$$\begin{aligned} p_I(I') - (\widehat{p})_I(I') &= \frac{p(I')}{p(I)} - \frac{\widehat{p}(I')}{\widehat{p}(I)} \geq \frac{p(I')}{p(I)} - \frac{p(I') + \Delta}{p(I) - \Delta} \\ &= -\Delta \cdot \frac{p(I') + p(I)}{p(I)(p(I) - \Delta)}. \end{aligned} \tag{12}$$

Now using the fact that $p(I') \leq p(I)$ and $p(I) \geq 99\epsilon/(10000k)$, we get that (12) is at least

$$-\Delta \cdot \frac{2p(I)}{(98/99)p(I)^2} = -\frac{2 \cdot 99\Delta}{98p(I)} \geq -\frac{2 \cdot 99\Delta \cdot 10000k}{98 \cdot 99\epsilon} = -\frac{\epsilon}{49}.$$

So we have established the lower bound $p_I(I') - (\widehat{p})_I(I') \geq -\epsilon/49$. For the upper bound, similar reasoning gives

$$\begin{aligned} p_I(I') - (\widehat{p})_I(I') &\leq \Delta \cdot \frac{p(I') + p(I)}{p(I)(p(I) + \Delta)} \leq \Delta \cdot \frac{2p(I)}{p(I)^2 \cdot (100/99)} \\ &\leq \Delta \cdot \frac{2 \cdot 10000k \cdot 99}{99\epsilon \cdot 100} = \frac{\epsilon}{50} \end{aligned}$$

and so we have shown that $|p_I(I') - (\widehat{p})_I(I')| \leq \epsilon/49$ as desired. Now we proceed to prove the lemma.

We first prove Part 1. Suppose that p_I is non-decreasing and $d_{\mathrm{TV}}(p_I, u_I) > \epsilon/6$. Since p_I is monotone and u_I is uniform and both are supported on I, we have that the pdfs for p_I and u_I have exactly one crossing. An easy consequence of this is that $d_{\mathrm{K}}(p_I, u_I) = d_{\mathrm{TV}}(p_I, u_I) > \epsilon/6$. By the definition of d_K and the fact that p_I is non-decreasing, we get that there exists a point $j \in I$ and an interval $I' = [a, j]$ which is such that

$$d_{\mathrm{K}}(p_I, u_I) = u_I(I') - p_I(I') > \frac{\epsilon}{6}.$$

Using (11) we get from this that

$$u_I(I') - (\widehat{p})_I(I') > \frac{\epsilon}{6} - \frac{\epsilon}{49} > \frac{\epsilon}{7}$$

and thus ORIENTATION outputs "↑" in Step 3 as claimed.

Now we turn to Part 2 of the lemma. Suppose that ORIENTATION$(\widehat{p}, I)$ outputs "↑". Then it must be the case that there is an initial interval $I' = [a, j]$ of I that satisfies $u_I(I') - (\widehat{p})_I(I') > \frac{\epsilon}{7}$. By (11) we have that $u_I(I') - p_I(I') > \frac{\epsilon}{7} - \frac{\epsilon}{49} = \frac{6\epsilon}{49}$. But Observation 1 tells us that if p_I were non-increasing then we would have $u_I(I') - p_I(I') \leq 0$; so p_I cannot be non-increasing, and therefore it must be non-decreasing.

For Part 3, suppose that p_I is non-increasing and $d_{\mathrm{TV}}(p_I, u_I) > \epsilon/6$. First we must show that ORIENTATION does *not* output "↑" in Step 3. Since p_I is non-increasing, Observation 1 gives us that $u_I(I') - p_I(I') \leq 0$ for every initial interval I' of I. Inequality (11) then gives $u_I(I') - (\widehat{p})_I(I') \leq \epsilon/49$, so ORIENTATION indeed does not output "↑" in Step 3 (and it reaches Step 4 in its execution). Now arguments exactly analogous to the arguments for part 1 (but using now the fact that p_I is non-increasing rather than non-decreasing) give that there is an initial interval I' such that $(\widehat{p})_I(I') - u_I(I') > \frac{\epsilon}{6} - \frac{\epsilon}{49} > \frac{\epsilon}{7}$, so ORIENTATION outputs "↓" in Step 4 and Part 3 of the lemma follows.

Finally, Part 4 of the lemma follows from analogous arguments as Part 2. □

D Proof of Proposition 6

We start by defining the transformation, and then prove the necessary lemmas to show that the transformation yields k-modal distributions with the specified increase in support size, preserves L_1 distance between pairs, and has the property that samples from the transformed distributions can be simulated given access to samples from the original distributions.

The transformation proceeds in two phases. In the first phase, the input distribution p is transformed into a related distribution f with larger support; f has the additional property that the ratio of the probabilities of consecutive domain elements is bounded. Intuitively the distribution f corresponds to a "reduced distribution" from Section A. In the second phase, the distribution f is transformed into the final $2(k-1)$-modal distribution g. Both stages of the transformation consist of subdividing each element of the domain of the input distribution into a set of elements of the output distribution; in the first stage, the probabilities of each element of the set are chosen according to a geometric sequence, while in the second phase, all elements of each set are given equal probabilities.

We now define this two-phase transformation and prove Proposition 6.

Definition 2. *Fix $\epsilon > 0$ and a distribution p over $[n]$ such that $p_{min} \leq p(i) \leq p_{max}$ for all $i \in [n]$. We define the distribution $f_{p,\epsilon,p_{max},p_{min}}$ in two steps. Let q be the distribution on support $[c]$ with $c = 1 + \lceil \log_{1+\epsilon} p_{max} - \log_{1+\epsilon} p_{min} \rceil$ that is defined by $q(i) = (1+\epsilon)^{i-1} \frac{\epsilon}{(1+\epsilon)^c - 1}$. The distribution $f_{p,\epsilon,p_{max},p_{min}}$ has support $[cn]$, and for $i \in [n]$ and $j \in [c]$ it assigns probability $p(i)q(j)$ to domain element $c(i-1)+j$.*

It is convenient for us to view the $\mod r$ operator as giving an output in $[r]$, so that "$r \mod r$" equals r.

Definition 3. *We define the distribution $g_{k,p,\epsilon,p_{max},p_{min}}$ from distribution $f_{p,\epsilon,p_{max},p_{min}}$ of support $[m]$ via the following process. Let $r = \lceil \frac{m}{k} \rceil$, and let $a_1 := 1$, and for all $i \in \{2, \ldots, r\}$, let $a_i := \lceil (1+\epsilon) a_{i-1} \rceil$. For each $i \in [m]$, we assign probability $\frac{f_{p,\epsilon,p_{max},p_{min}}(i)}{a_{i \mod r}}$ to each of the a_j support elements in the set $\{1+t, 2+t, \ldots, a_{i \mod r} + t\}$, where $t = \sum_{\ell=1}^{i-1} a_{(\ell \mod r)}$.*

Lemma 11. *Given $\epsilon, p_{min}, p_{max}$, and access to independent samples from distribution p, one can generate independent samples from $f_{p,\epsilon,p_{max},p_{min}}$ and from $g_{k,p,\epsilon,p_{max},p_{min}}$.*

Proof. To generate a sample according to $f_{p,\epsilon,p_{max},p_{min}}$, one simply takes a sample $i \leftarrow p$ and then draws $j \in [c]$ according to the distribution q as defined in Definition 2 (note that this draw according to q only involves ϵ, p_{min} and p_{max}). We then output the value $c(i-1)+j$. It follows immediately from the above definition that the distribution of the output value is $f_{p,\epsilon,p_{max},p_{min}}$.

To generate a sample according to $g_{k,p,\epsilon,p_{max},p_{min}}$ given a sample $i \leftarrow f_{p,\epsilon,p_{max},p_{min}}$, one simply outputs (a uniformly random) one of the $a_{(i \mod r)}$ support elements of $g_{k,p,\epsilon,p_{max},p_{min}}$ corresponding to the element i of $f_{p,\epsilon,p_{max},p_{min}}$. Specifically, if the support of $f_{p,\epsilon,p_{max},p_{min}}$ is $[m]$, then we output a random element of the set $\{1+t, 2+t, \ldots, a_{i \mod r} + t\}$, where $t = \sum_{\ell=1}^{i-1} a_{(\ell \mod r)}$, with a_j as defined in Definition 3, and $r = \lceil \frac{m}{k} \rceil$. □

Lemma 12. *If $p_{min} \leq p(i) \leq p_{max}$ for all $i \in [n]$, then the distribution $f_{p,\epsilon,p_{max},p_{min}}$ of Definition 2, with density $f : [cn] \to \mathbb{R}$, has the property that $\frac{f(i)}{f(i-1)} \leq 1+\epsilon$ for all $i > 1$, and the distribution $g_{k,p,\epsilon,p_{max},p_{min}}$ of Definition 3 is $2(k-1)$-modal.*

Proof. Note that the distribution q, with support $[c]$ as defined in Definition 2, has the property that $q(i)/q(i-1) = 1+\epsilon$ for all $i \in \{2, \ldots, c\}$, and thus $f(\ell)/f(\ell-1) = 1+\epsilon$ for any ℓ satisfying $(\ell \mod c) \neq 1$. For values ℓ that are 1 mod c, we have

$$\frac{f(\ell)}{f(\ell-1)} = \frac{p(i+1)}{p(i)(1+\epsilon)^{c-1}} \leq \frac{p(i+1)p_{min}}{p(i)p_{max}} \leq 1.$$

Given this property of $f_{p,\epsilon,p_{max},p_{min}}$, we now establish that $g_{k,p,\epsilon,p_{max},p_{min}}$ is monotone decreasing on each of the k equally sized contiguous regions of its domain. First consider the case $k = 1$; given a support element j, let i be such that $j \in \{1 + \sum_{\ell=1}^{i-1} a_\ell, \ldots, a_i + \sum_{\ell=1}^{i-1} a_\ell\}$. We thus have that

$$g_{1,p,\epsilon,p_{max},p_{min}}(j) = \frac{f_{p,\epsilon,p_{max},p_{min}}(i)}{a_i} \leq \frac{(1+\epsilon) f_{p,\epsilon,p_{max},p_{min}}(i-1)}{a_i} \leq \frac{f_{p,\epsilon,p_{max},p_{min}}(i-1)}{a_{i-1}} \leq g_{1,p,\epsilon,p_{max},p_{min}}(j-1),$$

and thus $g_{1,p,\epsilon,p_{max},p_{min}}$ is indeed 0-modal since it is monotone non-increasing. For $k > 1$ the above arguments apply to each of the k equally-sized contiguous regions of the support, so there are $2(k-1)$ modes, namely the local maxima occurring at the right endpoint of each region, and the local minima occurring at the left endpoint of each region. □

Lemma 13. *For any distributions p, p' with support $[n]$, and any $\epsilon, p_{max}, p_{min}$, we have that*

$$d_{\mathrm{TV}}(p, p') = d_{\mathrm{TV}}\left(f_{p,\epsilon,p_{max},p_{min}}, f_{p',\epsilon,p_{max},p_{min}}\right) = d_{\mathrm{TV}}\left(g_{k,p,\epsilon,p_{max},p_{min}}, g_{k,p',\epsilon,p_{max},p_{min}}\right).$$

Proof. Both equalities follow immediately from the fact that the transformations of Definitions 2 and 3 partition each element of the input distribution in a manner that is oblivious to the probabilities. To illustrate, letting $c = 1 + \lceil \log_{1+\epsilon} p_{max} - \log_{1+\epsilon} p_{min} \rceil$, and letting q be as in Definition 2, we have the following:

$$\begin{aligned} d_{\mathrm{TV}}\left(f_{p,\epsilon,p_{max},p_{min}}, f_{p',\epsilon,p_{max},p_{min}}\right) &= \sum_{i\in[n],j\in[c]} q(j)|p(i) - p'(i)| \\ &= \sum_{i\in[n]} |p(i) - p'(i)|. \end{aligned}$$

□

Lemma 14. *If p has support $[n]$, then for any $\epsilon < 1/2$, the distribution $g_{k,p,\epsilon,p_{max},p_{min}}$ is supported on $[N]$, where N is at most $k\frac{e^{\frac{8n}{k}(1+\log(p_{max}/p_{min}))}}{\epsilon^2}$.*

Proof. The support of $f_{p,\epsilon,p_{max},p_{min}}$ is $n(1 + \lceil \log_{1+\epsilon} p_{max} - \log_{1+\epsilon} p_{min} \rceil) \leq n\left(2 + \frac{\log(p_{max}/p_{min})}{\log(1+\epsilon)}\right)$. Letting $a_1 := 1$ and $b_1 := \lceil \frac{1}{\epsilon} \rceil$, and defining $a_i := \lceil a_{i-1}(1+\epsilon) \rceil$, and $b_i := \lceil b_{i-1}(1+\epsilon) \rceil$, we have that $a_i \leq b_i$ for all i. Additionally, $b_{i+1}/b_i \leq 1 + 2\epsilon$, since all $b_i \geq 1/\epsilon$, and thus the ceiling operation can increase the value of $(1+\epsilon)b_i$ by at most ϵb_i. Putting these two observations together, we have

$$\sum_{i=1}^{m} a_i \leq \sum_{i=1}^{m} b_i \leq \frac{(1+2\epsilon)^{m+1}}{2\epsilon^2}.$$

For any $\epsilon \leq 1/2$, we have that the support of $g_{k,p,1/2,p_{max},p_{min}}$ is at most

$$\begin{aligned} k\frac{(1+2\epsilon)^{\left\lceil \frac{n}{k}\left(2+\frac{\log(p_{max}/p_{min})}{\log(1+\epsilon)}\right)\right\rceil}}{\epsilon^2} &\leq k\frac{(1+2\epsilon)^{2\frac{n}{k}\left(2+4\frac{\log(p_{max}/p_{min})}{2\epsilon}\right)}}{\epsilon^2} \\ &\leq k\frac{(1+2\epsilon)^{\frac{1}{2\epsilon}\left(\frac{8n}{k}(1+\log(p_{max}/p_{min}))\right)}}{\epsilon^2} \\ &\leq k\frac{e^{\frac{8n}{k}(1+\log(p_{max}/p_{min}))}}{\epsilon^2}. \end{aligned}$$

□

Proof of Proposition 6. The proof is now a simple matter of assembling the above parts. Given a distribution D over pairs of distributions of support $[n]$, as specified in the proposition statement, the distribution D_k is defined via the process of taking $(p, p') \leftarrow D$, then applying the transformation of Definitions 2 and 3 with $\epsilon = 1/2$ and to yield a pair $\left(g_{k,p,1/2,p_{max},p_{min}}, g_{k,p',1/2,p_{max},p_{min}}\right)$. We claim that this D_k satisfies all the properties claimed in the proposition statement. Specifically, Lemmas 12 and 14, respectively, ensure that every distribution in the support of D_k has at most $2(k-1)$ modes, and has support size at most $4ke^{\frac{8n}{k}(1+\log(p_{max}/p_{min}))}$. Additionally, Lemma 13 guarantees that the transformation preserves L_1 distance, namely, for two distributions p, p' with support $[n]$, we have $L_1(p, p') = L_1(g_{k,p,1/2,p_{max},p_{min}}, g_{k,p',1/2,p_{max},p_{min}})$. Finally, Lemma 11 guarantees that, given s independent samples from p, one can simulate drawing s independent samples according to $g_{k,p,1/2,p_{max},p_{min}}$. Assuming for the sake of contradiction that one had an algorithm that could distinguish whether $L_1(g_{k,p,1/2,p_{max},p_{min}}, g_{k,p',1/2,p_{max},p_{min}})$ is less than ϵ_1 versus greater than ϵ_2 with the desired probability given s samples, one could take s samples from distributions $(p, p') \leftarrow D$, *simulate* having drawn them from $g_{k,p,1/2,p_{max},p_{min}}$ and $g_{k,p',1/2,p_{max},p_{min}}$, and then run the hypothesized tester algorithm on those samples, and output the answer, which will be the same for (p, p') as for $(g_{k,p,1/2,p_{max},p_{min}}, g_{k,p',1/2,p_{max},p_{min}})$. This contradicts the assumption that no algorithm with these success parameters exists for $(p, p') \leftarrow D$. □

E Proof of Theorem 5

We first note that we can assume that $\epsilon > 1/n$. Otherwise, the decomposition of $[n]$ into singleton intervals $I_i = \{i\}$, $i \in [n]$, trivially satisfies the statement of the theorem. Indeed, in this case we have that $(1/\epsilon) \cdot \log n > n$ and $p_f \equiv p$.

We first describe the oblivious decomposition and then show that it satisfies the statement of the theorem. The decomposition $\mathcal{I}$ will be a partition of $[n]$ into ℓ nonempty consecutive intervals $I_1, \ldots, I_\ell$. In particular, for $j \in [\ell]$, we have $I_j = [n_{j-1} + 1, n_j]$ with $n_0 = 0$ and $n_\ell = n$. The *length* of interval I_i, denoted by l_i, is defined to be the cardinality of I_i, i.e., $l_i = |I_i|$. (Given that the intervals are disjoint and consecutive, to fully define them it suffices to specify their lengths.)

We can assume wlog that n and $1/\epsilon$ are each at least sufficiently large universal constants. The interval lengths are defined as follows. Let $\ell \in \mathbb{Z}^+$ be the smallest integer such that

$$\sum_{i=1}^{\ell} \lfloor (1+\epsilon)^i \rfloor \geq n.$$

For $i = 1, 2, \ldots, \ell - 1$ we define

$$l_i := \lfloor (1+\epsilon)^i \rfloor.$$

For the ℓ-th interval, we set

$$l_\ell := n - \sum_{i=1}^{\ell-1} l_i.$$

It follows from the aforementioned definition that the number ℓ of intervals in the decomposition is at most

$$O\left((1/\epsilon) \cdot \log(1 + \epsilon \cdot n)\right).$$

Let p be any non-increasing distribution over $[n]$. We will now show that the above described decomposition satisfies

$$d_{\mathrm{TV}}(p_f, p) = O(\epsilon)$$

where p_f is the flattened distribution corresponding to p and the partition $\mathcal{I} = \{I_i\}_{i=1}^{\ell}$. We can write

$$d_{\mathrm{TV}}(p_f, p) = (1/2) \cdot \sum_{i=1}^{n} |p_f(i) - p(i)| = \sum_{j=1}^{\ell} d_{\mathrm{TV}}\left((p_f)^{I_j}, p^{I_j}\right)$$

where p^I denotes the (sub-distribution) restriction of p over I.

Let $I_j = [n_{j-1} + 1, n_j]$ with $l_j = |I_j| = n_j - n_{j-1}$. Then we have that

$$d_{\mathrm{TV}}\left((p_f)^{I_j}, p^{I_j}\right) = (1/2) \cdot \sum_{i=n_{j-1}+1}^{n_j} |p_f(i) - p(i)|.$$

Recall that p_f is by definition constant within each I_j and in particular equal to $\bar{p}_f^j = \sum_{i=n_{j-1}+1}^{n_j} p(i)/l_j$. Also recall that p is non-increasing, hence $p(n_{j-1}) \geq p(n_{j-1}+1) \geq \bar{p}_f^j \geq p(n_j)$. Therefore, we can bound from above the variation distance within I_j as follows

$$d_{\mathrm{TV}}\left((p_f)^{I_j}, p^{I_j}\right) \leq l_j \cdot (p(n_{j-1}+1) - p(n_j)) \leq l_j \cdot (p(n_{j-1}) - p(n_j)).$$

So, we have

$$d_{\mathrm{TV}}(p_f, p) \leq \sum_{j=1}^{\ell} l_j \cdot (p(n_{j-1}) - p(n_j)). \tag{13}$$

To bound the above quantity we analyze summands with $l_j < 1/\epsilon$ and with $l_j \geq 1/\epsilon$ separately.

Formally, we partition the set of intervals $I_1, \ldots, I_\ell$ into "short" intervals and "long intervals" as follows: If any interval I_j satisfies $l_j \geq 1/\epsilon$, then let $j_0 \in \mathbb{Z}^+$ be the largest integer such that $l_{j_0} < 1/\epsilon$; otherwise we have that every interval I_j satisfies $l_j < 1/\epsilon$, and in this case we let $j_0 = \ell$. If $j_0 < \ell$ then we have that $j_0 = \Theta((1/\epsilon) \cdot \log_2(1/\epsilon))$. Let $S = \{I_i\}_{i=1}^{j_0}$ denote the set of *short* intervals and let L denote its complement $L = \mathcal{I} \setminus S$.

Consider the short intervals and cluster them into *groups* according to their length; that is, a group contains all intervals in S of the same length. We denote by G_i the ith group, which by definition contains all intervals in S of length i; note that these intervals are consecutive. The *cardinality* of a group (denoted by $|\cdot|$) is the number of intervals it contains; the *length* of a group is the number of elements it contains (i.e. the sum of the lengths of the intervals it contains).

Note that G_1 (the group containing all singleton intervals) has $|G_1| = \Omega(1/\epsilon)$ (this follows from the assumption that $1/\epsilon < n$). Hence G_1 has length $\Omega(1/\epsilon)$. Let $j^* < 1/\epsilon$ be the maximum length of any short interval in S. It is easy to verify that each group G_j for $j \leq j^*$ is nonempty, and that for all $j \leq j^* - 1$, we have $|G_j| = \Omega\left((1/\epsilon) \cdot (1/j)\right)$, which implies that the length of G_j is $\Omega(1/\epsilon)$.

To bound the contribution to (13) from the short intervals, we consider the corresponding sum for each group, and use the fact that G_1 makes no contribution to the error. In particular, the contribution of the short intervals is

$$\sum_{l=2}^{j^*} l \cdot \left(p_l^- - p_l^+\right) \tag{14}$$

where p_l^- (resp. p_l^+) is the probability mass of the leftmost (resp. rightmost) point in G_l. Given that p is non-increasing, we have that $p_l^+ \geq p_{l+1}^-$. Therefore, we can upper bound (14) by

$$2 \cdot p_1^+ + \sum_{l=2}^{j^*-1} p_l^+ - j^* \cdot p_{j^*}^+.$$

Now note that $p_1^+ = O(\epsilon) \cdot p(G_1)$, since G_1 has length (total number of elements) $\Omega(1/\epsilon)$ and p is non-increasing. Similarly, for $l < j^*$, we have that $p_l^+ = O(\epsilon) \cdot p(G_l)$, since G_l has length $\Omega(1/\epsilon)$. Therefore, the above quantity can be upper bounded by

$$O(\epsilon) \cdot p(G_1) + O(\epsilon) \cdot \sum_{l=2}^{j^*-1} p(G_l) - j^* \cdot p_{j^*}^+ = O(\epsilon) \cdot p(S) - j^* \cdot p_{j^*}^+. \tag{15}$$

We consider two cases: The first case is that $L = \emptyset$. In this case, we are done because the above expression (15) is $O(\epsilon)$. The second case is that $L \neq \emptyset$ (we note in passing that in this case the total number of elements in all short intervals is $\Omega(1/\epsilon^2)$, which means that we must have $\epsilon = \Omega(1/\sqrt{n})$). In this case we bound the contribution of the long intervals using the same argument as Birgé. In particular, the contribution of the long intervals is

$$\sum_{j=j_0+1}^{\ell} l_j \cdot (p(n_{j-1}) - p(n_j)) \leq (j^* + 1) \cdot p_{j^*}^+ + \sum_{j=j_0+1}^{\ell-1} (l_{j+1} - l_j) \cdot p(n_j). \tag{16}$$

Given that $l_{j+1} - l_j \leq (2\epsilon) \cdot l_j$ and $\sum_j l_j \cdot p(n_j) \leq p(L)$, it follows that the second summand in (16) is at most $O(\epsilon) \cdot p(L)$. Therefore, the total variation distance between p and p_f is at most (15) + (16), i.e.

$$O(\epsilon) \cdot p(S) + O(\epsilon) \cdot p(L) + p_{j^*}^+. \tag{17}$$

Finally, note that $p(L) + p(S) = 1$ and $p_{j^*}^+ = O(\epsilon)$. (The latter holds because $p_{j^*}^+$ is the probability mass of the rightmost point in S; recall that S has length at least $1/\epsilon$ and p is decreasing.) This implies that (17) is at most $O(\epsilon)$, and this completes the proof of Theorem 5.

F Bounding variation distance

As noted above, our tester will work by decomposing the interval $[n]$ into sub-intervals. The following lemma will be useful for us; it bounds the variation distance between two distributions p and q in terms of how p and q behave on the sub-intervals in such a decomposition.

Lemma 15. *Let $[n]$ be partitioned into $I_1, \ldots, I_r$. Let p, q be two distributions over $[n]$. Then*

$$d_{\mathrm{TV}}(p,q) \leq \frac{1}{2}\sum_{j=1}^{r} |p(I_j) - q(I_j)| + \sum_{j=1}^{r} p(I_j) \cdot d_{\mathrm{TV}}(p_{I_j}, q_{I_j}). \tag{18}$$

Proof. Recall that $d_{\mathrm{TV}}(p,q) = \frac{1}{2}\sum_{i=1}^{n} |p(i) - q(i)|$. To prove the claim it suffices to show that

$$\frac{1}{2}\sum_{i \in I_1} |p(i) - q(i)| \leq \frac{1}{2}|p(I_1) - q(I_1)| + p(I_1) \cdot d_{\mathrm{TV}}(p_{I_1}, q_{I_1}). \tag{19}$$

We assume that $p(I_1) \leq q(I_1)$ and prove (19) under this assumption. This gives the bound in general since if $p(I_1) > q(I_1)$ we have

$$\frac{1}{2}\sum_{i \in I_1} |p(i) - q(i)| \leq |p(I_1) - q(I_1)| + q(I_1) \cdot d_{\mathrm{TV}}(p_{I_1}, q_{I_1}) < |p(I_1) - q(I_1)| + p(I_1) \cdot d_{\mathrm{TV}}(p_{I_1}, q_{I_1})$$

where the first inequality is by (19). The triangle inequality gives us

$$|p(i) - q(i)| \leq \left| p(i) - q(i) \cdot \frac{p(I_1)}{q(I_1)} \right| + \left| q(i) \cdot \frac{p(I_1)}{q(I_1)} - q(i) \right|.$$

Summing this over all $i \in I_1$ we get

$$\frac{1}{2}\sum_{i \in I_1} |p(i) - q(i)| \leq \frac{1}{2}\sum_{i \in I_1} \left| p(i) - q(i) \cdot \frac{p(I_1)}{q(I_1)} \right| + \frac{1}{2}\sum_{i \in I_1} \left| q(i) \cdot \frac{p(I_1)}{q(I_1)} - q(i) \right|.$$

We can rewrite the first term on the RHS as

$$\begin{aligned} \frac{1}{2}\sum_{i \in I_1} \left| p(i) - q(i) \cdot \frac{p(I_1)}{q(I_1)} \right| &= p(I_1) \cdot \frac{1}{2}\sum_{i \in I_1} \left| \frac{p(i)}{p(I_1)} - \frac{q(i)}{q(I_1)} \right| = p(I_1) \cdot \frac{1}{2}\sum_{i \in I_1} |p_{I_1}(i) - q_{I_1}(i)| \\ &= p(I_1) \cdot d_{\mathrm{TV}}(p_{I_1}, q_{I_1}) \end{aligned}$$

so to prove the desired bound it suffices to show that

$$\frac{1}{2}\sum_{i \in I_1} \left| q(i) \cdot \frac{p(I_1)}{q(I_1)} - q(i) \right| \leq |p(I_1) - q(I_1)|. \tag{20}$$

We have

$$\left| q(i) \cdot \frac{p(I_1)}{q(I_1)} - q(i) \right| = q(i) \cdot \left| \frac{p(I_1)}{q(I_1)} - 1 \right|$$

and hence we have

$$\frac{1}{2}\sum_{i \in I_1} \left| q(i) \cdot \frac{p(I_1)}{q(I_1)} - q(i) \right| = \frac{1}{2}\sum_{i \in I_1} q(i) \cdot \left| \frac{p(I_1)}{q(I_1)} - 1 \right| = \frac{1}{2} q(I_1) \cdot \left| \frac{p(I_1)}{q(I_1)} - 1 \right| = \frac{1}{2}|p(I_1) - q(I_1)|.$$

So we indeed have (20) as required, and the lemma holds. □