Are Few Bins Enough: Testing Histogram Distributions

Clément L. Canonne*

January 29, 2016

Abstract

A probability distribution over an ordered universe $[n] = \{1, \dots, n\}$ is said to be a k-histogram if it can be represented as a piecewise-constant function over at most k contiguous intervals. We study the following question: given samples from an arbitrary distribution D over $[n]$, one must decide whether D is a k-histogram, or is far in ℓ_1 distance from any such succinct representation. We obtain a sample and time-efficient algorithm for this problem, complemented by a nearly-matching information-theoretic lower bound on the number of samples required for this task. Our results significantly improve on the previous state-of-the-art, due to Indyk, Levi, and Rubinfeld [ILR12] and Canonne, Diakonikolas, Gouleakis, and Rubinfeld [CDGR16].

1 Introduction

1.1 Motivation and background

Large datasets have become the norm over recent decades, a trend that if anything has been hastening lately – and most likely will for the foreseeable future. This rapid increase in the amount of information to store, analyze, and process comes with many challenges; and in particular calls for succinct ways of *representing* the data, as well as (very) fast algorithms to operate on it.

One of the oldest and most widely used representations is that of *histograms*, where the range of possible values the data can take is divided into groups, or "bins" [Pea95]. The number of records from the dataset falling in each bin is then recorded, and serves as summary of the records themselves. Whenever the dataset can be well-approximated by histograms with few bins, this provides a space-efficient and flexible way of storing, querying, and analyzing the data and its distribution; specifically, whenever the number of bins k is much smaller than the size n of the universe. For these reasons, the study of histograms and algorithms that operate on them has received significant interest in databases [Koo80, PIHS96, GMP97, CMN98, JKM+98, WJLY04, XZX+13] and many other fields, such as statistics [Sco79, FD81, Bir87], streaming [GGI+02, TGIK02, GKS06], and learning theory [ILR12, CDSS13, CDSS14, GSW04, ADH+15] (see also [Ioa03] for a survey).

In this work we will be concerned with the framework of *property testing of distributions*, as first introduced in the seminal work of Batu et al. [BFR+00] (see also [Ron08, Ron10, Can15] for surveys on property and distribution testing). In this setting, access to the data is provided *via* random samples drawn

*Columbia University. Email: ccanonne@cs.columbia.edu. Research supported by NSF CCF-1115703 and NSF CCF-1319788.

independently from the dataset (that is, from the probability distribution that underlies it).[1] The algorithm must then decide, after looking at as a few samples as possible, whether this probability distribution satisfies some fixed property of interest – e.g., if the records are uniformly distributed. This setting is particularly relevant when confronted to massive datasets, whose sheer size makes the perspective of reading the whole input impractical, or even impossible. In this case, a standard approach to meet the memory and computational constraints is by random sampling of the data, which enables one to instead only perform computations on a small representative portion of the dataset. In view of this, we consider here the following testing question: *given some input parameters k and ε, can the distribution of the data be represented as a histogram on at most k bins, or is it significantly different (at distance at least ε) from any such "k-histogram" representation?*

An efficient primitive answering this question could then be used to represent or sketch the dataset as compactly as possible, when invoked as a subroutine to perform model selection. In more detail, given a bound ε on the desired approximation error, one can iteratively run such an algorithm (e.g., by doubling search) to look for the "smallest corresponding k," that is the smallest number of bins needed to accurately capture the statistical properties of the dataset (within error ε). Once this parameter identified, calling an agnostic learning algorithm as that of e.g. [ADLS15] with this k will yield a succinct approximation of the dataset, achieving an optimal tradeoff between accuracy and conciseness. (The efficiency of the testing procedure, notably the number of samples required, is in this case crucial. Indeed, this approach is only advantageous as long as the subroutine only takes $o(n)$ samples – if not, then the overhead it brings completely eclipses the savings made in the learning stage, as one can always approximate the whole dataset and compute the closest histogram "offline" from $O(n)$ data points.)

1.2 Our results

We obtain an efficient algorithm, complemented by a nearly matching lower bound, that together settle the question of testing whether an unknown probability distribution can be represented as a k-histogram, for any k in the range of interest (with regard to the size of the universe $\Omega = \{1, \dots, n\}$). Specifically, we prove the following theorems:

Theorem 1.1 (Upper Bound). *For any $1 \leq k \leq n$, there exists an efficient testing algorithm for the class of k-histograms with sample complexity $O\left(\frac{\sqrt{n}}{\varepsilon^2}\log k + \frac{k}{\varepsilon^3}\log^2 k\right)$.*

Theorem 1.2 (Lower Bound). *For any $1 \leq k \leq \frac{n}{120}$, any (non-necessarily efficient) testing algorithm for the class of k-histograms must have sample complexity $\Omega\left(\frac{\sqrt{n}}{\varepsilon^2} + \frac{1}{\varepsilon}\frac{k}{\log k}\right)$.*

Indeed, this essentially resolves the sample complexity of testing k-histograms, up to polylogarithmic factors in k and the dependence on ε of the second term. Moreover, we note that the proof of Theorem 1.2 implies the same lower bound on the sample complexity of testing *k-modal distributions*, that is the class of distributions whose probability mass function is allowed to go "up and down" or "down and up" at most k times.

Comparison with previous work. Our results significantly improve upon the previous algorithmic results of Indyk, Levi, and Rubinfeld [ILR12], which required $O\left(\frac{\sqrt{kn}}{\varepsilon^5}\log n\right)$ samples; as well as on later work by Canonne, Diakonikolas, Gouleakis, and Rubinfeld [CDGR16], where this upper bound is brought down to $O\left(\frac{\sqrt{kn}}{\varepsilon^3}\log n\right)$. Moreover, these results crucially left open the question of the interplay between the domain size n and the parameter k of the class to be tested.

[1] See Section 2 for the formal definition of the model.

At a high level, our results (almost) answer this question, by "decoupling" these two parameters. In particular, ignoring ε in the statement of Theorem 1.1 one can see the first term as capturing the (sublinear) dependence on the domain size, while the other second only depends on the complexity of the class to be tested.

Turning to the negative results, prior to our work the best lower bounds for this question were due to Paninski [Pan08], who establishes an $\Omega(\sqrt{n}/\varepsilon^2)$ sample lower bound for testing *uniformity* (that is, the case $k = 1$), and to Indyk, Levi, and Rubinfeld [ILR12] where a lower bound of $\Omega\left(\sqrt{kn}\right)$ samples is proven for $k \leq 1/\varepsilon$. Theorem 1.2 unifies and extends both results, obtaining a nearly-tight lower bound featuring the same decoupling between n and k as in our upper bound.

Importantly, the question we consider is the setting where one does *not* know beforehand the decomposition of the domain in k intervals on which the unknown distribution D "should be piecewise-constant," but instead must decide if such a decomposition exists. The (easier) problem of testing, given as input an explicit partition Π of the domain in k intervals, if D is indeed a histogram with regard to this *specific* Π has been recently considered in [DK16], where the authors obtain tight upper and lower bounds on the question.

1.3 Techniques

Upper bound. To obtain our algorithmic result, we follow an approach similar to that of Acharya, Daskalakis, and Kamath [ADK15], who show how to apply the "testing by learning" paradigm to the setting of distribution testing. At a high-level, the idea is to first learn an approximation of the unknown distribution *as if it satisfied the property of interest* (which can usually be achieved with relatively few samples); before verifying that the output of this learning stage is both (a) close to having the property and (b) close to the unknown distribution. While standard in property testing of *functions*, this method was believed to inherently fail in the case of probability distributions, due to the hardness of efficiently estimating the distance between distributions from samples [VV10] – as required for (b). Namely, the result of [VV10] implies that this last step would by itself cost a prohibitive number of samples, almost linear in the domain size n. The main idea of [ADK15] is to circumvent this impossibility result by first choosing to learn the unknown distribution not with regard to the total variation, but instead in χ^2 distance; and showing that the corresponding variant of distance estimation (deciding whether two distributions are close in χ^2 distance, versus far in total variation) *can* be achieved with only $\sqrt{n}$ samples.[2]

In order to establish Theorem 1.1, we adapt the above approach, with several crucial modifications. Namely, applying the ideas of [ADK15] out-of-the-box would require an efficient algorithm to learn the class of k-histograms in χ^2 distance, i.e. one with sample complexity $\mathrm{poly}(k, 1/\varepsilon)$ (independent of n). To the best of our knowledge, such learning algorithm does not appear in the literature, and it is not clear whether one can even exist. Instead, we settle for a weaker guarantee: that of learning a good approximation of an unknown k-histogram except on a small (but unknown) portion of the domain, *where the accuracy can be arbitrarily poor.* To handle this, we then need to adapt the second stage (testing in χ^2 vs. total variation) to identify and discard this small portion of the domain. This is done by iteratively applying (a modification of) the testing algorithm of [ADK15] several times, removing "bad chunks of the domain" one at a time. The challenge here is to do this in a careful and controlled manner, in order to keep the number of such iterations (and therefore samples) as small as possible. (Intuitively, this is where the $\log k$ factor in the first term of the sample complexity stems from – a union bound over k outcomes of the testing subroutine.)

[2]The use and analysis of a χ^2-based statistic to obtain optimal testing algorithms already appears in [CDVV14], where the authors consider the (related) question of distinguishing whether two distributions are *equal* vs. far in total variation distance.

Lower bound. Turning now to the converse result, we split the proof of Theorem 1.2 in two parts, establishing separately each term of the lower bound. The first one, $\Omega(\sqrt{n}/\varepsilon^2)$, is essentially a direct modification of the lower bound of Paninski on testing uniformity (i.e., 1-histograms). The second term, however, proves to be much less straightforward: the main ingredient in our $\Omega(k/\log k)$ bound is a reduction from a seemingly unrelated question, that of *estimating the support size* [VV10]. A key aspect of this reduction is to lift the corresponding lower bound of Valiant and Valiant – which heavily relies on the support size to be a symmetric property,[3] to our setting – a property that is clearly *not* symmetric, and thus at first glance intrinsically different. Perhaps surprisingly, we manage to connect these two questions in a black-box and conceptually simple way; moreover, we believe our reduction to be of independent interest, and applicable to other properties as well.

1.4 Organization

After introducing the required definitions and notations in Section 2, we establish our algorithmic result, Theorem 1.1, in Section 3. Finally, Section 4 contains the details of our lower bound, Theorem 1.2.

2 Notations and Preliminaries

All throughout this paper, we denote by $[n]$ the set $\{1, \ldots, n\}$, and by $\log$ the logarithm in base 2. We write $\Delta([n])$ for the set of discrete probability distributions over domain $[n]$, i.e. the set of all real-valued functions $D\colon [n] \to [0,1]$ such that $\sum_{i=1}^n D(i) = 1$. A *property* of distributions over $[n]$ is a subset $\mathcal{P} \subseteq \Delta([n])$, consisting of all distributions that have the property.

For any fixed $1 \leq k \leq n$, we let $\mathcal{H}_k \subseteq \Delta([n])$ denote the class of k-*histograms*, i.e. the property of being piecewise-constant with at most k "pieces." Formally, $D \in \mathcal{H}_k$ if and only if there exists a partition $\mathcal{I} = (I_1, \ldots, I_k)$ of $[n]$ into k intervals such that D is constant on each I_j.

In this work, we will measure the distance between two distributions D_1, D_2 on $[n]$ by their *total variation distance*

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

which takes value in $[0,1]$. (Note that this metric, sometimes referred to as *statistical distance* and equivalent to the ℓ_1 distance, is one of the most stringent ones and the standard distance measure in distribution testing.) To prove some of our results, we will also require as a tool the definition of the (asymmetric) χ^2-*distance* between two distributions $D_1, D_2 \in \Delta([n])$,

$$\mathrm{d}_{\chi^2}(D_1 \mid\mid D_2) = \sum_{i=1}^n \frac{(D_1(i) - D_2(i))^2}{D_2(i)} = -1 + \sum_{i=1}^n \frac{D_1(i)^2}{D_2(i)}.$$

Finally, recall that a *testing algorithm* for a fixed property $\mathcal{P}$ is a randomized algorithm TESTER which takes as input n, $\varepsilon \in (0,1]$, and is granted access to independent samples from an unknown distribution D:

(i) if $D \in \mathcal{P}$, the algorithm outputs accept with probability at least $2/3$;

(ii) if $\mathrm{d_{TV}}(D, D') \geq \varepsilon$ for every $D' \in \mathcal{P}$, it outputs reject with probability at least $2/3$.

That is, TESTER must accept if the unknown distribution has the property, and reject if it is ε-*far* from having it. The *sample complexity* of the algorithm is the number of samples it draws from the distribution in the worst case.

[3] That is, a property invariant to relabeling of the domain elements: D has a property if and only if, for every permutation σ of the domain, $D \circ \sigma$ has the property.

Poissonization. We hereafter follow, for our upper bounds, the standard "Poissonization trick": specifically, we assume that instead of drawing a fixed number m of samples from a distribution D, the algorithms instead randomly select $m' \sim \mathrm{Poisson}(m)$, and then draw m' independent samples from D. While greatly simplifying the analysis by making the number of times different domain elements appear in the samples independent, this can be achieved at only a negligible cost in both the sample complexity and the probability of error, due to the tight concentration of the Poisson distribution around its mean.

On discrete domains. Although the setting we consider is that of *discrete* domains, our techniques can be easily extended to continuous ones by suitably gridding the range of values. This approach may not in general be optimal, and will depend on the step chosen for the discretization; however, it is not clear how to even phrase distribution testing questions in a continuous setting without either additional assumptions on the unknown distribution or changing the metric.

3 Upper bound: an efficient testing algorithm

In this section, we prove our upper bound, Theorem 1.1. More specifically, we establish the following, more detailed, result:

Theorem 3.1. *For any $k \geq 1$, there exists a testing algorithm for $\mathcal{H}_k$ with sample complexity*

$$O\left(\frac{\sqrt{n}}{\varepsilon^2}\log k + \frac{k}{\varepsilon^3}\log k + \frac{k}{\varepsilon}\log^2\frac{k}{\varepsilon}\right).$$

Moreover, its running time is $\sqrt{n}\,\mathrm{poly}(\log k, 1/\varepsilon) + \mathrm{poly}(k, 1/\varepsilon)$.

We first state in the next subsection some results from the literature we shall rely upon, before delving into the proof of the theorem.

3.1 Tools from previous work

Our starting point will be a recent result of Acharya, Daskalakis, and Kamath, which shows how to efficiently perform a specific relaxation of *tolerant identity testing*,[4] with regard to a χ^2 guarantee:

Theorem 3.2 ([ADK15], Rephrased)**.** *There exists an algorithm* TESTER *that, on input n, $\varepsilon \in (0,1]$ as well as the explicit description of a distribution $D^* \in \Delta([n])$, takes $O(\sqrt{n}/\varepsilon^2)$ samples from an unknown distribution $D \in \Delta([n])$ and satisfies the following.*

(i) If $\mathrm{d}_{\chi^2}(D \mid\mid D^*) \leq \frac{\varepsilon^2}{500}$, then TESTER *outputs* **accept** *with probability at least 2/3;*

(ii) If $\mathrm{d}_{\mathrm{TV}}(D, D^*) \geq \varepsilon$, then TESTER *outputs* **reject** *with probability at least 2/3.*

Moreover, TESTER *runs in time $O(\sqrt{n}/\varepsilon^2)$.*

For our purpose, instead of invoking this result as a blackbox we will rely on the following refinement (which already appears in the section of [ADK15] dealing with unimodality):[5] given an explicit partition of $[n]$ on K intervals $I_1, \dots, I_K$ and a fully specified distribution D^*, the algorithm from Theorem 3.2 takes

[4] In tolerant identity testing, the goal is, provided the full description of a distribution D^* and samples from an unknown distribution D, to distinguish between $\mathrm{d}_{\mathrm{TV}}(D, D^*) \leq \varepsilon$ and $\mathrm{d}_{\mathrm{TV}}(D, D^*) \geq 2\varepsilon$. Valiant and Valiant [VV10] showed that even in the case of D^* being the uniform distribution, $\Omega\left(\frac{n}{\log n}\right)$ samples were required for this task.

[5] See Appendix D of the full version of [ADK15].

(Poisson) $m = O(\sqrt{n}/\varepsilon^2)$ samples from D, and computes the K (independent) statistics $Z_1, \dots, Z_K$ defined as

$$Z_j = \sum_{i \in I_j \cap \mathcal{A}_\varepsilon} \frac{(N_i - mD^*(i))^2 - N_i}{mD^*(i)}$$

where $\mathcal{A}_\varepsilon = \{\ i \in [n] : \ D^*(i) \geq \frac{\varepsilon}{50n}\ \}$ and N_i is the number of occurrences of $i \in [n]$ among the samples drawn from D. Observing that N_i is distributed according to $\operatorname{Poisson}(mD(i))$, standard computations yield that $\mathbb{E}Z_j = m \sum_{i \in I_j \cap \mathcal{A}_\varepsilon} \frac{(D(i)-D^*(i))^2}{D^*(i)}$. Letting $Z = \sum_{j=1}^K Z_j$, we get the quantity analyzed by Acharya, Daskalakis, and Kamath, for which they show the following:

Proposition 3.3 ([ADK15, Lemmata 1 and 2]). *The statistic Z above has the following guarantees, provided that $m \geq 20000\sqrt{n}/\varepsilon^2$.*

- *If $\mathrm{d}_{\chi^2}(D \mid\mid D^*) \leq \frac{\varepsilon^2}{500}$, then $\mathbb{E}Z \leq \frac{m\varepsilon^2}{500}$, which implies $\operatorname{Var} Z \leq \frac{m^2\varepsilon^4}{500000}$.*
- *If $\mathrm{d}_{\mathrm{TV}}(D, D^*) \geq \varepsilon$, then $\mathbb{E}Z \geq \frac{m\varepsilon^2}{5}$, which implies $\operatorname{Var} Z \leq \frac{\mathbb{E}[Z]^2}{100}$.*

Moreover, for any $j \in [K]$ such that $\mathbb{E}Z_j \geq \frac{m\varepsilon^2}{5}$, we have $\operatorname{Var} Z_j \leq \frac{\mathbb{E}[Z_j]^2}{100}$ (as per the second item).

We will also leverage another characteristic of the tester of Theorem 3.2; specifically, that it also works for *subdistributions* (i.e., considering only a portion of the domain, on which the two distributions do not necessarily sum to one nor to the same value), considering the natural restrictions of χ^2 and total variation to intervals (the latter as half the ℓ_1 norm, as defined above).[6]

Finally, we will make use of the fact below, which can be shown by a standard application of Chernoff bounds.

Proposition 3.4 ([ADK15, Claim 1 (Full version)]). *There exists an algorithm* ApproxPart *that, given a parameter $b > 1$, takes $O(b \log b)$ samples from a distribution D and, with probability at least $9/10$, outputs a partition of $[n]$ in $K \leq 2b + 2$ intervals $I_1, \dots, I_K$ such that the following holds:*

(i) For each $i \in [n]$ such that $D(i) \geq 1/b$, there exists $\ell \in [K]$ such that $I_\ell = \{i\}$;

(ii) There are at most two intervals I such that $D(I) < 1/(2b)$;

(iii) Every other interval is such that $D(I) \in [1/(2b), 2/b]$.

Moreover, the algorithm runs in time $O(b \log b)$.

3.2 Proof of Theorem 3.1

As described in Section 1.3, our algorithm relies on two main components: the first is an (almost) learning procedure for k-histograms which outputs an approximation $\hat{D}$ of an unknown distribution D, with the guarantee that *if* $D \in \mathcal{H}_k$, then $\hat{D}$ is close to D in χ^2 distance *except possibly on a small but unknown portion S of the domain*. The second is a testing procedure, inspired by the work of [ADK15], which takes this $\hat{D}$ as input and iteratively "sieves" the domain, in order to discard a set S' (the algorithm's "guess" for S); and eventually checks if $\hat{D}$ and D are indeed close in χ^2 distance on the sieved domain $[n] \setminus S'$.

A learning lemma. Let $\mathcal{I}$ be a partition of $[n]$ into intervals. For a subset of intervals $\mathcal{J} \subseteq \mathcal{I}$ define $\widetilde{D}^{\mathcal{J}}$ as follows. For every $i \notin \cup_{J \in \mathcal{J}} J$, $\widetilde{D}^{\mathcal{J}}(i) = D(i)$ and otherwise $\widetilde{D}^{\mathcal{J}}(i) = D(I)/\lvert I \rvert$ where I is such that $I \in \mathcal{I}$ and $i \in I$. Given a histogram $D \in \mathcal{H}_k$, we say that $i \in [n]$ is a *breakpoint* of D if $D(i) \neq D(i+1)$.

[6] Namely, for an interval I define $\mathrm{d}^I_{\chi^2}(D_1 \mid\mid D_2) = \sum_{i \in I} \frac{(D_1(i)-D_2(i))^2}{D_2(i)}$ and $\mathrm{d}^I_{\mathrm{TV}}(D_1, D_2) = \frac{1}{2}\sum_{i \in I} \lvert D_1(i) - D_2(i) \rvert$.

Algorithm 1

Require: Parameters k and $\varepsilon \in (0,1]$; sample access to a distribution D over $[n]$

Set $b \stackrel{\text{def}}{=} \frac{20k\log k}{\varepsilon}$, and $\varepsilon' \stackrel{\text{def}}{=} \frac{13\varepsilon}{30}$.

Learning

  Run ApproxPart (from Proposition 3.4) with parameter b; let $\mathcal{I}$ be the partition of $[n]$ into K intervals it outputs.

  Run Learner (from Lemma 3.5) with parameters K, $\frac{\varepsilon}{60}$, and $\mathcal{I}$; let $\hat{D} \in \mathcal{H}_K$ be its output.

Sieving

  Identify $O(k\log k)$ intervals from $\mathcal{I}$ to discard (with regard to $D, \hat{D}$), as detailed in Section 3.2.1. Let $\mathcal{I}' \subseteq \mathcal{I}$ be the set of remaining intervals, and $G = \cup_{I\in\mathcal{I}'} I$.

Checking

  Check there exists $D^* \in \mathcal{H}_k$ such that $\mathrm{d}_{\mathrm{TV}}^{G}\left(\hat{D}, D^*\right) \leq \frac{\varepsilon}{60}$; otherwise, **return** reject. ▷ Can be done in time $\mathrm{poly}(k, 1/\varepsilon)$ by dynamic programming, as in [CDGR16, Lemma 4.11]

Testing

  Run Tester (from Theorem 3.2) on D with parameters n, ε', and $\hat{D}$, restricted to the subdomain G; if the tester rejects, **return** reject.

return accept

Similarly, an interval $I \in \mathcal{I}$ is a *breakpoint interval* of D (with respect to $\mathcal{I}$) if it contains a breakpoint of D. (Note that there can be at most $k-1$ such breakpoints and breakpoint intervals).

Lemma 3.5. *There exists an algorithm* Learner *that, on input n, a partition of $[n]$ into ℓ intervals $\mathcal{I} = \{I_1, \dots, I_\ell\}$ and $\varepsilon \in (0,1]$, takes $O(\ell/\varepsilon^2)$ samples from an unknown distribution $D \in \Delta([n])$ and outputs (the succinct description of) a distribution $\hat{D} \in \mathcal{H}_\ell$ that satisfies the following. For every $k \leq \ell$, if $D \in \mathcal{H}_k$ and $\mathcal{J} = \{J_1, \dots, J_{r-1}\} \subseteq \mathcal{I}$ (with $r \leq k$) are the breakpoints intervals of D, then $\mathrm{d}_{\chi^2}\left(\widetilde{D}^{\mathcal{J}} \mid\mid \hat{D}\right) \leq \varepsilon^2$ with probability at least $9/10$. Moreover, the algorithm runs in time $O(\ell/\varepsilon^2)$.*

Proof. We follow the analysis of the Laplace estimator from [KOPS15], first defining a modified estimator (from m independent samples $s_1, \dots, s_m$ from a distribution D on $[n]$) by

$$\hat{D}(j) \stackrel{\text{def}}{=} \frac{m_{I_i}+1}{m+\ell} \cdot \frac{1}{|I_i|}, \quad i \in [\ell], j \in I_i$$

where $m_{I_i} \stackrel{\text{def}}{=} \sum_{j\in I_i} m_j$ and $m_i \stackrel{\text{def}}{=} |\{\, j \in [m] \;:\; s_j = i \,\}|$.

Suppose D is a k-histogram. The expected value of $\mathrm{d}_{\chi^2}\left(\widetilde{D}^{\mathcal{J}} \mid\mid \hat{D}\right)$, over the draws of the m samples, can be written

$$\mathbb{E}\left[\mathrm{d}_{\chi^2}\left(\widetilde{D}^{\mathcal{J}} \mid\mid \hat{D}\right)\right] = -1 + \sum_{I\in\mathcal{I}} |I| \cdot \mathbb{E}\left[\frac{\left(\frac{D(I)}{|I|}\right)^2}{\frac{m_I+1}{m+\ell}\cdot\frac{1}{|I|}}\right] = -1 + \sum_{I\in\mathcal{I}} \mathbb{E}\left[\frac{D(I)^2(m+\ell)}{m_I+1}\right]$$

Now, for a fixed $I \in \mathcal{I}$, we have

$$\mathbb{E}\left[\frac{1}{m_I+1}\right] = \sum_{s=0}^{m} \frac{1}{s+1}\binom{m}{s} D(I)^s \left(1 - D(I)\right)^{m-s} \leq \frac{1}{D(I)(m+1)}.$$

Plugging it back, this implies

$$\mathbb{E}\Big[\mathrm{d}_{\chi^2}\Big(\tilde{D}^{\mathcal{J}} \mid\mid \hat{D}\Big)\Big] \leq -1 + \sum_{I\in\mathcal{I}} \frac{D(I)(m+\ell)}{m+1} \leq \frac{\ell}{m}$$

Letting $m \geq c \cdot \frac{\ell}{\varepsilon^2}$ (where $c > 0$ is an absolute constant), this together with Markov's inequality yields the result. $\square$

Outline and correctness. The idea is to first run the algorithm APPROXPART of Proposition 3.4 on D with parameter b set to $(20k\log k)/\varepsilon$, getting $K = O((k\log k)/\varepsilon)$ intervals $I_1, \dots, I_K$ (meeting the stated guarantee with probability at least $9/10$), after taking $O(K\log K)$ samples. We then run the χ^2 learner of Lemma 3.5 with parameter $\frac{\varepsilon}{60}$ (which requires $O\Big(\frac{K}{\varepsilon^2}\Big)$ samples from D) to output a histogram $\hat{D}$ on this partition.

- In the completeness case, $D \in \mathcal{H}_k$: meaning that there exists a fixed and fully determined, albeit unknown, subset $\mathcal{B}$ of at most k intervals among the K which contain the "breakpoints intervals" for the piecewise-constant D. (Note that the only possible intervals that can be "bad" must be non-singletons.) Conditioning on the learning algorithm to meet its guarantees on $G \stackrel{\text{def}}{=} \bigcup_{j\in[K]\setminus\mathcal{B}} I_j$ (which happens with probability at least $9/10$), we obtain that $\mathrm{d}^G_{\chi^2}\Big(D \mid\mid \hat{D}\Big) \leq \frac{\varepsilon^2}{3600}$.
- In the soundness case, $\mathrm{d}_{\mathrm{TV}}(D, \mathcal{H}_k) \geq \varepsilon$. Since $D(I) \leq \frac{2}{b}$ for any non-singleton interval, this implies that no matter which set $\mathcal{B}$ of at most $k\log k$ intervals we discard, it amounts for no more than $\frac{\varepsilon}{10}$ total probability weight under D, and we can safely ignore it in the rest of the procedure. Indeed, for any such $\mathcal{B}$ and the corresponding remaining domain $G = \bigcup_{j\in[K]\setminus\mathcal{B}} I_j$, $\mathrm{d}^G_{\mathrm{TV}}(D, D') \geq \frac{9\varepsilon}{20}$ for any $D' \in \mathcal{H}_k$.

The goal is therefore to remove $k\log k$ (non-singleton) intervals, out of the K intervals, which together contribute the maximum amount to Z; that is, to remove $Z_{i_1}, \dots, Z_{i_k}$ such that $\sum_{\ell=1}^{k} \mathbb{E}Z_{i_\ell}$ is maximized (call this stage (‡)). Indeed, assuming this has been done (which corresponds to identifying a good restricted domain G), the two items above together ensure correctness of Algorithm 1, conditioning on all subroutines meeting their specification (which by a union bound happens with probability at least $2/3$). In more detail, assuming the sieving stage to have gone through, the algorithm will check that (a) $\hat{D}$ is $\frac{\varepsilon}{60}$-close in "total variation restricted to G" to some k-histogram D' (as it should if $D \in \mathcal{H}_k$); and then (b) run the tester of Theorem 3.2 on $D, \hat{D}$ (on G) with parameter $\varepsilon' \stackrel{\text{def}}{=} \frac{13\varepsilon}{30}$.

Completeness. if $D \in \mathcal{H}_k$ then the learning algorithm of Step 4 outputs $\hat{D}$ such that $\mathrm{d}^G_{\chi^2}\Big(D \mid\mid \hat{D}\Big) \leq \frac{\varepsilon^2}{3600}$, so that TESTER will accept in Step 13; and as this also implies $\mathrm{d}^G_{\mathrm{TV}}\Big(\hat{D}, D\Big) \leq \frac{\varepsilon}{60}$, Step 10 accepts as well.

Soundness. Conversely, suppose $\mathrm{d}_{\mathrm{TV}}(D, \mathcal{H}_k) \geq \varepsilon$ and the algorithm accepts in Step 10 there is a $D^* \in \mathcal{H}_k$ such that $\mathrm{d}^G_{\mathrm{TV}}\Big(\hat{D}, D^*\Big) \leq \frac{\varepsilon}{60}$. But by the above discussion we must have $\mathrm{d}^G_{\mathrm{TV}}(D, D^*) \geq \frac{9\varepsilon}{20}$, and from the triangle inequality $\mathrm{d}^G_{\mathrm{TV}}\Big(D, \hat{D}\Big) \geq \mathrm{d}^G_{\mathrm{TV}}(D, D^*) - \mathrm{d}^G_{\mathrm{TV}}\Big(\hat{D}, D^*\Big) \geq \frac{13\varepsilon}{30}$: the algorithm will output reject in Step 13.

The correctness having been established, the main question is therefore *how* to perform the "sieving stage" (‡), which we detail next.

3.2.1 Sieving: removing up to $k \log k$ possible bad intervals.

In what follows, we will compute the statistics Z_j from Proposition 3.3 several times, computed independently each time. Furthermore, by standard arguments (repeating the test, and taking the median value), we can assume the probability of success/correctness of this test to be $1 - \delta$, at the price of an extra $\log \frac{1}{\delta}$ factor in the sample complexity. (In particular, we shall take δ to depend on k, in order to apply a union bound over many tests.)

For simplicity, we deal with the following scenario (where the constants have been changed): among the K indices, there is a fixed but unknown subset $\mathcal{B} = \{i_1, \dots, i_k\}$ of k indices such that

1. $\sum_{j \notin \mathcal{B}} \mathbb{E} Z_j \leq m\alpha^2$;
2. $\sum_{j \in \mathcal{B}} \mathbb{E} Z_j > 100 m\alpha^2$

and we want to remove a subset $\mathcal{B}'$ of $2k$ indices such that $\sum_{j \notin \mathcal{B}'} \mathbb{E} Z_j \leq 100 m\alpha^2$. (This will deal with the completeness case, and setting $\alpha = \frac{\varepsilon}{C}$ for some big enough constant C in the learning stage will give us what we want.)

Discarding the heavy ones: Let $\mathcal{B}^+ \subseteq \mathcal{B}$ be the indices such that $\mathbb{E} Z_j \geq 100 m\alpha^2$. By assumption, $|\mathcal{B}^+| \leq k$, and this is a fixed (albeit unknown) set of indices fully determined by D. In particular, if we compute the statistics as in Proposition 3.3 with failure probability $\delta = \frac{1}{10(k+1)}$, by a union bound we can condition on (i) each Z_j, $j \in \mathcal{B}^+$ behaving as expected: $Z_j > 10 m\alpha^2$, and (ii) the fixed set $[K] \setminus \mathcal{B}$ also behaving as expected, so that $\sum_{j \notin \mathcal{B}} Z_j \leq 10 m\alpha^2$. By removing all j's such that $Z_j > 10 m\alpha^2$, and outputting reject if there are more than k, we thus have filtered all intervals from $\mathcal{B}^+$ (with success probability at least $9/10$), and no other. Let ℓ' be the number of elements removed, and $k' = k - \ell$ the number of "possible remaining bad elements."

Iteratively removing the rest: we therefore can now assume we have at most k' indices to remove (call this set $\mathcal{B}^-$), such that for each $\mathbb{E} Z_j < 100 m\alpha^2$. In particular, $\sum_{j \in \mathcal{B}^-} \mathbb{E} Z_j < 100 mk\alpha^2$. We repeat the following at most $\log k$ times, until either the test accepts at some step, or we performed more than $O(\log k)$ such steps (in which cases we proceed to the last stage, the sieving part being over); or we removed more than k' indices in total (in the latter case, we output reject and stop)

- compute the statistics Z_j for all remaining indices, and check the value of their sum Z.
- if $Z < 10 m\alpha^2$, accept.
- otherwise, sort the Z_j's by decreasing order, and remove the first ℓ indices, where $\ell \leq k'$ is the smallest integer such that $\sum_{j > \ell} Z_j \leq 2 m\alpha^2$.

Define $\mathcal{B}^-_{\text{rem}} \subseteq \mathcal{B}^-$ as the set of bad indices remaining at the current step. By a conditioning on the two subsets of indices $\mathcal{B}^-_{\text{rem}}$ and $[K] \setminus \mathcal{B}$, we have that $\sum_{j \notin \mathcal{B}} Z_j \leq 2 m\alpha^2$, and if $\sum_{j \in \mathcal{B}^-_{\text{rem}}} \mathbb{E} Z_j > 100 m\alpha^2$ then $\sum_{j \in \mathcal{B}^-_{\text{rem}}} Z_j > \frac{1}{2} \sum_{j \in \mathcal{B}^-_{\text{rem}}} \mathbb{E} Z_j$.

By assumption, we remove at least $\frac{1}{2} \sum_{j \in \mathcal{B}^-_{\text{rem}}} \mathbb{E} Z_j - 2 m\alpha^2 > \frac{1}{3} \sum_{j \in \mathcal{B}^-_{\text{rem}}} \mathbb{E} Z_j$ of the "bad weight" as long as $\sum_{j \in \mathcal{B}^-_{\text{rem}}} \mathbb{E} Z_j > 100 m\alpha^2$ and we know that at the beginning $\sum_{j \in \mathcal{B}^-} \mathbb{E} Z_j < 100 mk\alpha^2$. This implies that after $O(\log k)$ such steps, we have that the sum of $\mathbb{E} Z_j$ for the remaining Z_j's is at most $101 m\alpha^2$ (from what remains in the "bad" intervals, and the contribution of the "good" ones). Moreover, in total we removed at most $O(\log k) \cdot k' = O(k \log k)$ intervals, and ran $O(\log k)$ "tests" with $\delta = \Theta\left(\frac{1}{\log k}\right)$ (which costs $O\left(\frac{\sqrt{n}}{\alpha^2} \log \log k\right)$ samples).

Overall, over these two stages we end up paying $O\left(\frac{\sqrt{n}}{\alpha^2} \log k\right) + O\left(\frac{\sqrt{n}}{\alpha^2} \log \log k\right) = O\left(\frac{\sqrt{n}}{\alpha^2} \log k\right)$ samples, and perform the "sieving" (‡). This concludes the proof of Theorem 3.1: the total sample complexity is

$$O\left(\frac{\sqrt{n}}{\varepsilon^2}\log k\right)+O\left(\frac{k}{\varepsilon^3}\log k\right)+O\left(\frac{k\log k}{\varepsilon}\log\frac{k\log k}{\varepsilon}\right)=O\left(\frac{\sqrt{n}}{\varepsilon^2}\log k\right)+O\left(\frac{k}{\varepsilon^3}\log k\right)+O\left(\frac{k}{\varepsilon}\log^2\frac{k}{\varepsilon}\right)$$

as stated. The running time of the overall algorithm is easily seen to be as claimed, as each of the learning and testing subroutines runs in time linear in the number of samples. □

4 An information-theoretic lower bound

In this section, we prove Theorem 1.2, that is both an $\Omega(\sqrt{n}/\varepsilon^2)$ and an $\Omega(k/(\varepsilon\log k))$ lower bound on testing k-histograms (the latter for $k=\omega(1)$):

Proposition 4.1. *There exists an absolute constant $\varepsilon_0>0$ such that the following holds. For any $1\leq k<\frac{n}{3}$ and $\varepsilon\in(0,\varepsilon_0]$, any (non-necessarily efficient) testing algorithm for $\mathcal{H}_k$ must take $\Omega\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ samples.*

Proposition 4.2. *There exists an absolute constant $\varepsilon_1>0$ such that the following holds. For any large enough $k\leq\frac{n}{120}$ and $\varepsilon\in(0,\varepsilon_1]$, any (non-necessarily efficient) testing algorithm for $\mathcal{H}_k$ must take $\Omega\left(\frac{1}{\varepsilon}\frac{k}{\log k}\right)$ samples.*

As mentioned earlier, the first lower bound builds on a proof of Paninski [Pan08] on testing uniformity; while our argument for the second will rely on a result of Valiant and Valiant [VV10], namely a lower bound on estimating a *symmetric* property: support size. While $\mathcal{H}_k$ is clearly *not* a symmetric class (i.e., it is not invariant by permutation of the support), we show how one can still leverage this lower bound for our purpose.

4.1 Proof of Proposition 4.1

The result follows from adapting the proof of [Pan08], intended for the case of uniformity testing, or equivalently $\mathcal{H}_1$. In this argument, Paninski defines a family of distributions $\mathcal{Q}$, parameterized as follows. A distribution $D\in\mathcal{Q}_\varepsilon$ is defined by $\frac{n}{2}$ bits $z_1,\dots,z_{n/2}\in\{0,1\}$, and

$$D(2i)=\frac{1+(-1)^{z_i}\cdot c\varepsilon}{n},\qquad D(2i-1)=\frac{1-(-1)^{z_i}\cdot c\varepsilon}{n}$$

for $i\in[n/2]$, where $c>0$ is a suitably chosen constant. The result then follows from observing that any distribution in $\mathcal{Q}_\varepsilon$ is ε-far from uniform, and yet that $\Omega(\sqrt{n}/\varepsilon^2)$ samples are necessary to distinguish a uniformly chosen $D\sim\mathcal{Q}_\varepsilon$ from the uniform distribution with probability at least $2/3$.

To apply this argument to our case, it is sufficient to observe that for $k<\frac{n}{3}$ (and the right choice of the constant c), a random $D\sim\mathcal{Q}_\varepsilon$ will be ε-far from $\mathcal{H}_k$ as well. To see why, fix $D\in\mathcal{Q}_\varepsilon$, and let $D^*\in\mathcal{H}_k$ be a k-histogram minimizing $\mathrm{d_{TV}}(D,D^*)$. Define $S\subseteq[n/2]$ as the set of indices such that $D^*(2i-1)=D^*(2i)$; note that by the triangle inequality, for all $i\in S$ we have $|D(2i-1)-D^*(2i-1)|+|D(2i)-D^*(2i)|\geq|D(2i)-D(2i-1)|=\frac{2c\varepsilon}{n}$. Since one must have $|S|\geq\frac{n}{2}-k+1>\frac{n}{6}$ as $D^*\in\mathcal{H}_k$, this implies that

$$\begin{aligned}2\mathrm{d_{TV}}(D,D^*)&=\sum_{i=1}^{\frac{n}{2}}\left(|D(2i-1)-D^*(2i-1)|+|D(2i)-D^*(2i)|\right)\\&\geq\sum_{i\in S}\left(|D(2i-1)-D^*(2i-1)|+|D(2i)-D^*(2i)|\right)\\&\geq\frac{n}{6}\cdot\frac{2c\varepsilon}{n}=\frac{c\varepsilon}{3}\end{aligned}$$

so that taking $c \geq 6$ (and $\varepsilon_0 \leq 1/c$) yields the result. $\square$

Remark 4.3. We observe that a simpler proof of this lower bound, albeit restricted to the range $k = o(\sqrt{n})$, can be obtained by applying the framework of [CDGR16]. Specifically, one can invoke [CDGR16, Theorem 6.1], using as a blackbox the uniformity testing lower bound of Paninski along with the fact that k-histograms can be learned agnostically from $O(k/\varepsilon^2)$ samples ([ADLS15]).

4.2 Proof of Proposition 4.2

Outline. We start by considering a scalar *symmetric* property, support size. The corresponding problem $\textsc{SuppSize}_m$ is as follows: given sample access to an unknown distribution $D \in \Delta([m])$ with the promise that $D(i) \in \{0\} \cup [\frac{1}{m}, 1]$ for all $i \in [m]$,[7] one must distinguish between (i) $\operatorname{supp}(D) \leq \frac{2m}{3} + 1$ and (ii) $\operatorname{supp}(D) \geq \frac{7m}{8}$. This problem is known to require $c \cdot \frac{m}{\log m}$ samples, where $c > 0$ is an absolute constant, for m sufficiently large ([VV10, Theorem 1]).

We then argue that any tester for the property of being a k-histogram can be used to solve this problem, with only a constant factor blowup in the sample complexity. Indeed, if $\textsc{Tester}$ is a *bona fide* $q(n, k, \varepsilon)$-sample tester for testing k-histograms (with probability of success $2/3$), then it can be converted to a symmetric tester $\textsc{Tester}'$ for the weak support size problem as follows: first, "enlarge" the domain $[m]$ of D by embedding it in $[n]$, for some $n > m$ (that is, setting $D(i) = 0$ for all $m + 1 \leq i \leq n$). Second, pick uniformly at random a permutation $\sigma \in \mathcal{S}_n$ of the "enlarged domain", where $\mathcal{S}_n$ denote the set of all permutations of $[n]$. Then, given samples of a distribution D it remains to feed $\textsc{Tester}$ with q samples from a distribution $D_\sigma = D \circ \sigma^{-1}$ ("re-building" the identity of the samples according to σ: for $i \in [n]$, $D_\sigma(i) = D(\sigma^{-1}(i))$). The key point is to argue that with high constant probability over the choice of σ:

- If D has support size at most $\frac{m}{3}$, then D_σ is a k-histogram for $k \stackrel{\text{def}}{=} 2 \cdot \frac{m}{3} + 1$ (with probability one);
- If D has support size in $[\frac{7m}{8}, m]$, then D_σ is far from any k-histogram, as with high constant probability its support is "sprinkled" over many *isolated* points – say at least $\frac{3m}{4}$. Whenever this happens, D_σ needs at least $\frac{6m}{4} - 1$ intervals to be a histogram, and incurs constant distance ε_1 (where $\varepsilon_1 = \left(\frac{3}{4}m - k + 1\right)\frac{1}{2m} = \frac{1}{24}$) from any k-histogram, from a similar argument as in Proposition 4.1 and the lower bound $1/m$ on any non-zero probability.

Independently repeating a constant number of times this procedure (that is, drawing a new permutation σ, and applying $\textsc{Tester}$ on D_σ using fresh samples from D) and taking the majority vote then allows the test to succeed with probability at least $5/9$. But this in turn implies a lower bound on q, as otherwise it would contradict the lower bound on the number of samples required to tolerantly test the symmetric property $\textsc{SuppSize}_m$.

The last piece we need in our reduction is the guarantee that, when permuting the domain at random, (a) a distribution with support size at most ℓ will be a $(2\ell + 1)$ histogram (this point is obvious); but also (b) with high probability over the permutation, a distribution with support size $\ell \ll n$ will keep its support "sprinkled" over the domain, and therefore need much more than $(2\ell + 1)$ pieces to be represented as a histogram. The following lemma makes this intuition precise, showing that for reasonable values of ℓ a random permutation will keep the points of the support isolated with constant probability:

Lemma 4.4. *Let $\ell \leq \frac{n}{70}$. For any set $S \subseteq [n]$, define $s = \operatorname{cover}(S)$ as the minimum number of disjoint intervals $I_1, \dots, I_s \subseteq S$ necessary to cover S. (That is, $\operatorname{cover}(S)$ is the number of disjoint "chunks" S*

[7] That is, $1/m$ is a lower bound on the probability weight of any element in the support.

induces in $[n]$). Then, fixing $S \subseteq [n]$ of size ℓ, we have

$$\Pr_{\sigma\sim\mathcal{S}_n}\left[\operatorname{cover}(\sigma(S)) \leq \frac{6\ell}{7}\right] \leq \frac{7\ell}{n} \leq \frac{1}{10}$$

where the probability is taken over a uniform choice of permutation $\sigma \in \mathcal{S}_n$.

Proof. Let $X_1, \dots, X_{n-1}$ be the $n-1$ (identically distributed, but non-independent) indicator random variables defined as follows. X_i is 1 if $\sigma^{-1}(i) \leq \ell$, but $\sigma^{-1}(i+1) > \ell$ (that is, one of the ℓ "good" points ends up on i, but one of the $n-\ell$ "bad points" ends up on $i+1$).

Let $X = \sum_{i=1}^{n-1} X_i$ be their sum: note that X is a lower bound on the number of clusters, up to an additive one (X counts the number of "right borders," and may only be off if the last cluster-interval ends at n). Moreover,

$$\mathbb{E}X_i = \frac{\ell}{n} \cdot \frac{n-\ell}{n-1}$$

so that $\mathbb{E}X = \ell \cdot \frac{n-\ell}{n} = \ell\left(1 - \frac{\ell}{n}\right)$ by linearity. Define $Y = \ell - X \geq 0$ (with $\mathbb{E}Y = \frac{\ell^2}{n}$); by Markov's inequality

$$\Pr\left[X \leq \frac{6\ell}{7}\right] = \Pr\left[Y \geq \frac{\ell}{7}\right] \leq \frac{\mathbb{E}Y}{\ell/7} = \frac{\ell/n}{1/7} = \frac{7\ell}{n}.$$

□

Now, this in particular imply that for $m \leq \frac{n}{70}$, a distribution D with support size in $[\frac{7m}{8}, m]$ will, after a random permutation σ of the larger domain $[n]$, have at least $\frac{6}{7} \cdot \frac{7m}{8} = \frac{3m}{4}$ isolated "chunks." But that also implies that $D_\sigma \notin \mathcal{H}_{\frac{3m}{4}-2}$ (i.e., needs a partition of at least $\frac{3m}{4} - 1$ intervals to be a histogram).

Details. We can now make precise the reduction outlined above: assume we have a tester TESTER for the property of being a histogram, which takes as input n, k, ε as well as $q(n, k, \varepsilon)$ independent samples from an unknown distribution D; and distinguishes with success probability at least $2/3$ between (a) $D \in \mathcal{H}_k$ and (b) $\ell_1(D, \mathcal{H}_k) > \varepsilon_1$.

Given sufficiently large integer n, and k satisfying $k \leq \frac{n}{120}$, we define $m \stackrel{\text{def}}{=} \left\lceil \frac{3}{2}(k-1) \right\rceil \leq \frac{n}{70}$. Now, we can embed any instance D' of SUPPSIZE$_m$ (i.e., a distribution $D' \in \Delta([m])$ meeting the promise of the problem) by seeing it as a distribution on $[n]$, and use TESTER to solve SUPPSIZE$_m$ as follows:

1. Draw uniformly a random a permutation $\sigma \in \mathcal{S}_n$;
2. Run TESTER on $D'_\sigma \in \Delta([n])$ with parameters n, k, and $\varepsilon_1 \stackrel{\text{def}}{=} \frac{1}{24}$;
3. accept if and only if TESTER accepted.

By the foregoing discussion and Lemma 4.4, the above test succeeds in solving SUPPSIZE$_m$ with probability at least $1 - \frac{1}{10} - \frac{1}{3} = \frac{17}{30}$; repeating constantly many times independently and taking a majority vote brings this success probability to $2/3$. The overall sample complexity being $O(q(n,k))$, the lower bound of [VV10, Theorem 1] implies that, for some absolute constant $c > 0$ and k large enough, $q(n, k, \varepsilon_1) \geq c \cdot \frac{k}{\log k}$, as claimed.

Finally, using a standard "trick" (embedding the hard instance by adding an extra element with weight $1 - \frac{\varepsilon}{\varepsilon_1}$), this yields an $\Omega\left(\frac{1}{\varepsilon}\frac{k}{\log k}\right)$ lower bound on testing $\mathcal{H}_k$, for any $\varepsilon \leq \varepsilon_1$. □

Acknowledgements. We would like to thank Reut Levi for many insightful discussions and comments.

References

[ADH+15] Jayadev Acharya, Ilias Diakonikolas, Chinmay Hegde, Jerry Zheng Li, and Ludwig Schmidt. Fast and near-optimal algorithms for approximating distributions by histograms. In Tova Milo and Diego Calvanese, editors, *Proceedings of the 34th ACM Symposium on Principles of Database Systems, PODS 2015, Melbourne, Victoria, Australia, May 31 - June 4, 2015*, pages 249–263. ACM, 2015. 1.1

[ADK15] Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. In C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, R. Garnett, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3577–3598. Curran Associates, Inc., 2015. Full version available on arXiv at abs/1507.05952. 1.3, 3.2, 3.1, 5, 3.3, 3.4, 3.2

[ADLS15] Jayadev Acharya, Ilias Diakonikolas, Jerry Zheng Li, and Ludwig Schmidt. Sample-optimal density estimation in nearly-linear time. *CoRR*, abs/1506.00671, 2015. 1.1, 4.3

[AKSX04] Rakesh Agrawal, Jerry Kiernan, Ramakrishnan Srikant, and Yirong Xu. Order-preserving encryption for numeric data. In Gerhard Weikum, Arnd Christian König, and Stefan Deßloch, editors, *Proceedings of the ACM SIGMOD International Conference on Management of Data, Paris, France, June 13-18, 2004*, pages 563–574. ACM, 2004.

[BFR+00] Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing that distributions are close. In *Proceedings of FOCS*, pages 189–197, 2000. 1.1

[Bir87] Lucien Birgé. On the risk of histograms for estimating decreasing densities. *The Annals of Statistics*, 15(3):pp. 1013–1022, 1987. 1.1

[Can15] Clément L. Canonne. A Survey on Distribution Testing: your data is Big. But is it Blue? *Electronic Colloquium on Computational Complexity (ECCC)*, 22:63, April 2015. 1.1

[CDGR16] Clément L. Canonne, Ilias Diakonikolas, Themis Gouleakis, and Ronitt Rubinfeld. Testing Shape Restrictions of Discrete Distributions. In *33rd International Symposium on Theoretical Aspects of Computer Science (STACS)*, 2016. To appear. Also available on arXiv at abs/1507.03558. (document), 1.2, 10, 4.3

[CDSS13] Siu-on Chan, Ilias Diakonikolas, Rocco A. Servedio, and Xiaorui Sun. Learning mixtures of structured distributions over discrete domains. In *Proceedings of SODA*, pages 1380–1394, 2013. 1.1

[CDSS14] Siu-on Chan, Ilias Diakonikolas, Rocco A. Servedio, and Xiaorui Sun. Near-optimal density estimation in near-linear time using variable-width histograms. In *Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1844–1852, 2014. 1.1

[CDVV14] Siu-on Chan, Ilias Diakonikolas, Gregory Valiant, and Paul Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of SODA*, pages 1193–1203, 2014. 2

[CMN98] Surajit Chaudhuri, Rajeev Motwani, and Vivek R. Narasayya. Random sampling for histogram construction: How much is enough? In Laura M. Haas and Ashutosh Tiwary, editors, *SIGMOD*

1998, Proceedings ACM SIGMOD International Conference on Management of Data, June 2-4, 1998, Seattle, Washington, USA., pages 436–447. ACM Press, 1998. 1.1

[DK16] Ilias Diakonikolas and Daniel M. Kane. A New Approach for Testing Properties of Discrete Distributions. *ArXiv e-prints*, January 2016. 1.2

[DKN15] Ilias Diakonikolas, Daniel M. Kane, and Vladimir Nikishkin. Testing Identity of Structured Distributions. In *Proceedings of SODA*, pages 1841–1854. Society for Industrial and Applied Mathematics (SIAM), 2015.

[FD81] David Freedman and Persi Diaconis. On the histogram as a density estimator: L_2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 57(4):453–476, 1981. 1.1

[GGI+02] Anna C. Gilbert, Sudipto Guha, Piotr Indyk, Yannis Kotidis, S. Muthukrishnan, and Martin Strauss. Fast, small-space algorithms for approximate histogram maintenance. In John H. Reif, editor, *Proceedings on 34th Annual ACM Symposium on Theory of Computing, May 19-21, 2002, Montréal, Québec, Canada*, pages 389–398. ACM, 2002. 1.1

[GKS06] Sudipto Guha, Nick Koudas, and Kyuseok Shim. Approximation and streaming algorithms for histogram construction problems. *ACM Trans. Database Syst.*, 31(1):396–438, 2006. 1.1

[GMP97] Phillip B. Gibbons, Yossi Matias, and Viswanath Poosala. Fast incremental maintenance of approximate histograms. In Matthias Jarke, Michael J. Carey, Klaus R. Dittrich, Frederick H. Lochovsky, Pericles Loucopoulos, and Manfred A. Jeusfeld, editors, *VLDB'97, Proceedings of 23rd International Conference on Very Large Data Bases, August 25-29, 1997, Athens, Greece*, pages 466–475. Morgan Kaufmann, 1997. 1.1

[GSW04] Sudipto Guha, Kyuseok Shim, and Jungchul Woo. REHIST: relative error histogram construction algorithms. In Mario A. Nascimento, M. Tamer Özsu, Donald Kossmann, Renée J. Miller, José A. Blakeley, and K. Bernhard Schiefer, editors, *(e)Proceedings of the Thirtieth International Conference on Very Large Data Bases, Toronto, Canada, August 31 - September 3 2004*, pages 300–311. Morgan Kaufmann, 2004. 1.1

[ILR12] Piotr Indyk, Reut Levi, and Ronitt Rubinfeld. Approximating and Testing k-Histogram Distributions in Sub-linear Time. In *Proceedings of PODS*, pages 15–22, 2012. (document), 1.1, 1.2

[Ioa03] Yannis E. Ioannidis. The history of histograms (abridged). In *VLDB*, pages 19–30, 2003. 1.1

[JKM+98] H. V. Jagadish, Nick Koudas, S. Muthukrishnan, Viswanath Poosala, Kenneth C. Sevcik, and Torsten Suel. Optimal histograms with quality guarantees. In Ashish Gupta, Oded Shmueli, and Jennifer Widom, editors, *VLDB'98, Proceedings of 24rd International Conference on Very Large Data Bases, August 24-27, 1998, New York City, New York, USA*, pages 275–286. Morgan Kaufmann, 1998. 1.1

[Koo80] Robert Kooi. *The Optimization of Queries in Relational Databases*. PhD thesis, Case Western Reserve University, September 1980. 1.1

[KOPS15] Sudeep Kamath, Alon Orlitsky, Dheeraj Pichapati, and Ananda Theertha Suresh. On learning distributions from their samples. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, volume 40 of *JMLR Proceedings*, pages 1066–1100. JMLR.org, 2015. 3.2

[Pan08] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008. 1.2, 4, 4.1

[Pea95] Karl Pearson. Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 186:343–414, 1895. 1.1

[PIHS96] Viswanath Poosala, Yannis E. Ioannidis, Peter J. Haas, and Eugene J. Shekita. Improved histograms for selectivity estimation of range predicates. In H. V. Jagadish and Inderpal Singh Mumick, editors, *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Montreal, Quebec, Canada, June 4-6, 1996.*, pages 294–305. ACM Press, 1996. 1.1

[Ron08] Dana Ron. Property Testing: A Learning Theory Perspective. *Foundations and Trends in Machine Learning*, 1(3):307–402, 2008. 1.1

[Ron10] Dana Ron. Algorithmic and Analysis Techniques in Property Testing. *Foundations and Trends in Theoretical Computer Science*, 5:73–205, 2010. 1.1

[Rub12] Ronitt Rubinfeld. Taming Big Probability Distributions. *XRDS*, 19(1):24–28, September 2012.

[Sco79] David W. Scott. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, 1979. 1.1

[TGIK02] Nitin Thaper, Sudipto Guha, Piotr Indyk, and Nick Koudas. Dynamic multidimensional histograms. In Michael J. Franklin, Bongki Moon, and Anastassia Ailamaki, editors, *Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data, Madison, Wisconsin, June 3-6, 2002*, pages 428–439. ACM, 2002. 1.1

[Val11] Paul Valiant. Testing symmetric properties of distributions. *SIAM Journal on Computing*, 40(6):1927–1968, 2011.

[VV10] Gregory Valiant and Paul Valiant. A CLT and tight lower bounds for estimating entropy. *Electronic Colloquium on Computational Complexity (ECCC)*, 17:179, 2010. 1.3, 4, 4, 4.2, 4.2

[WJLY04] Wei Wang, Haifeng Jiang, Hongjun Lu, and Jeffrey Xu Yu. Bloom histogram: Path selectivity estimation for XML data with updates. In *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30*, VLDB '04, pages 240–251. VLDB Endowment, 2004. 1.1

[XZX$^+$13] Jia Xu, Zhenjie Zhang, Xiaokui Xiao, Yin Yang, Ge Yu, and Marianne Winslett. Differentially private histogram publication. *VLDB J.*, 22(6):797–822, 2013. 1.1