# Approximating and Testing k-Histogram Distributions in Sub-linear Time

Piotr Indyk*
CSAIL, MIT, Cambridge MA 02139.
indyk@theory.lcs.mit.edu.

Reut Levi†
School of Computer Science, Tel Aviv University.
reuti.levi@gmail.com.

Ronitt Rubinfeld‡
CSAIL, MIT, Cambridge MA 02139 and the Blavatnik School of Computer Science, Tel Aviv University.
ronitt@csail.mit.edu.

## ABSTRACT

A discrete distribution $p$, over $[n]$, is a $k$-histogram if its probability distribution function can be represented as a piece-wise constant function with $k$ pieces. Such a function is represented by a list of $k$ intervals and $k$ corresponding values. We consider the following problem: given a collection of samples from a distribution $p$, find a $k$-histogram that (approximately) minimizes the $\ell_2$ distance to the distribution $p$. We give time and sample efficient algorithms for this problem.

We further provide algorithms that distinguish distributions that have the property of being a $k$-histogram from distributions that are $\epsilon$-far from any $k$-histogram in the $\ell_1$ distance and $\ell_2$ distance respectively.

## Categories and Subject Descriptors

F.2 [**Theory of Computation**]: ANALYSIS OF ALGORITHMS AND PROBLEM COMPLEXITY

## General Terms

Algorithms

## 1. INTRODUCTION

The ubiquity of massive data sets is a phenomenon that began over a decade ago, and is becoming more and more pervasive. As a result, there has been recently a significant interests in constructing *succinct representations* of the data. Ideally, such representations should take little space and computation time to operate on, while (approximately) preserving the desired properties of the data.

One of the most natural and useful succinct representations of the data are *histograms*. For a data set $D$ whose elements come from the universe $[n]$, a $k$-histogram $H$ is a piecewise constant function defined over $[n]$ consisting of $k$ pieces. Note that a $k$-histogram can be described using $O(k)$ numbers. A "good" $k$-histogram is such that (a) the value $H(i)$ is a "good" approximation of the total number of times an element $i$ occurs in the data set (denoted by $P(i)$) and (b) the value of $k$ is small. Histograms are a popular and flexible way to approximate the distribution of data attributes (e.g., employees age or salary) in databases. They can be used for data visualization, analysis and approximate query answering. As a result, computing and maintaining histograms of the data has attracted a substantial amount of interests in databases and beyond, see e.g., [GMP97, JPK+98, GKS06, CMN98, TGIK02, GGI+02], or the survey [Ioa03].

A popular criterion for fitting a histogram to a distribution $P$ is the "least-squares" criterion. Specifically, the goal is to find $H$ that minimizes the $\ell_2$ norm $\|P - H\|_2^2$. Such histograms are often called *v-optimal histograms*, with "v" standing for "variance". There has been a substantial amount of work on algorithms, approximate or exact, that compute the optimal $k$-histogram $H$ given $P$ and $k$ by taking the dynamic programming approach [JPK+98, GKS06]. However, since these algorithms need to read the whole input to compute $H$, their running times are at least linear in $n$.

A more efficient way to construct data histograms is to use random samples from data set $D$. There have been some results on this front as well [CMN98, GMP97]. However, they have been restricted to so-called *equi-depth histograms* (which are essentially approximate quantiles of the data distribution) or *compressed histograms*. Although the name by which they are referred to sounds similar, both of these representations are quite different from the representations considered in this paper. We are not aware of any work on constructing v-optimal histograms from random samples with provable guarantees.

The problem of constructing an approximate histogram from random samples can be formulated in the framework of distribution property testing and estimation (see surveys [Rub06, Ron08]). In this framework, an algorithm is given access to

*This material is based upon work supported by David and Lucille Packard Fellowship, MADALGO (Center for Massive Data Algorithmics, funded by the Danish National Research Association) and NSF grant CCF-0728645

†Research supported by the Israel Science Foundation grant nos. 1147/09 and 246/08

‡Research supported by NSF grants 0732334 and 0728645, Marie Curie Reintegration grant PIRG03-GA-2008-231077 and the Israel Science Foundation grant nos. 1147/09 and 1675/09.


*PODS'12,* May 21–23, 2012, Scottsdale, Arizona, USA.

i.i.d. samples from an unknown probability distribution $p$ , and its goal is to to characterize or estimate various properties of $p$. In our case we define $p = P/\|p\|_1$. Then choosing a random element from the data set $D$ corresponds to choosing $i \in [n]$ according the distribution $p$.

In this paper we propose several algorithms for constructing and testing for the existence of good histograms approximating a given distribution $p$.

## 1.1 Histogram taxonomy

Formally a histogram is a function $H : [n] \to [0, 1]$ that is defined by a sequence of intervals $I_1, \ldots, I_k$ and a corresponding sequence of values $v_1, \ldots, v_k$. For $t \in [n]$, $H(t)$ represents an estimate to $p(t)$. We consider the following classes of histograms (see [TGIK02] for a full list of classes):

1. *Tiling histograms:* the intervals form a tiling of $[n]$ (i.e., they are disjoint and cover the whole domain). For any $t$ we have $H(t) = v_i$, where $t \in I_i$. In practice we represent a tiling $k$-histogram as a sequence $\{(I_1, v_1) \ldots (I_k, v_k)\}$.

2. *Priority histograms:* the intervals can overlap. For any $t$ we have $H(t) = v_i$, where $i$ is the largest index such that $t \in I_i$; if none exists $H(t) = 0$. In practice we represent a priority $k$-histogram as $\{(I_1, v_1, r_1) \ldots (I_k, v_k, r_k)\}$ where $r_1, \ldots, r_k$ correspond to the priority of the intervals.

Note that if a function has a tiling $k$-histogram representation then it has a priority $k$-histogram representation. Conversely if it has a priority $k$-histogram representation then it has a tiling $2k$-histogram representation.

## 1.2 Results

The following algorithms receive as input a distribution over $[n]$, $p$, an accuracy parameter $\epsilon$ and an integer $k$.

In Section 3, we describe an algorithm which outputs a priority $k$-histogram that is closest to $p$ in the $\ell_2$ distance up to $\epsilon$-additive error. The algorithm is a greedy algorithm, at each step it enumerates over all possible intervals and adds the interval which minimizes the approximated $\ell_2$ distance. The sample complexity of the algorithm is $\tilde{O}((k/\epsilon)^2 \ln n)$ and the running time is $\tilde{O}((k/\epsilon)^2 n^2)$. We then improve the running time substantially to $\tilde{O}((k/\epsilon)^2 \ln n)$ by enumerating on a partial set of intervals.

In Section 4, we provide a testing algorithm for the property of being a tiling $k$-histogram with respect to the $\ell_1$ norm. The sample complexity of the algorithm is $\tilde{O}(\epsilon^{-5}\sqrt{kn})$. We provide a similar test for the $\ell_2$ norm that has sample complexity of $O(\epsilon^{-4} \ln^2 n)$. We prove that testing if a distribution is a tiling $k$-histogram in the $\ell_1$-norm requires $\Omega(\sqrt{kn})$ samples for every $k \leq 1/\epsilon$.

## 1.3 Related Work

Our formulation of the problem falls within the framework of property testing [RS96, GGR98, BFR$^+$00]. Properties of single and pairs of distributions has been studied quite extensively in the past (see [BFR$^+$10, BFF$^+$01, AAK$^+$07, BDKR05, GMP97, BKR04, RRSS09, Val08, VV11]). One question that has received much attention in property testing is to determine whether or not two distributions are similar. A problem referred to as *Identity testing* assumes that the algorithm is given access to samples of distribution $\mathbf{p}$ and an explicit description of distribution $\mathbf{q}$. The goal is to distinguish a pair of distributions that are identical from a pair of distributions that are far from each other. A special case of Identity testing is Uniformity Testing, where the fixed distribution, $\mathbf{q}$, is the uniform distribution. A uniform distribution can be represented by a tiling 1-histogram and therefore the study of uniformity testing is closely related to our study. Goldreich and Ron [GR00] study Uniformity Testing in the context of approximating graph expansion. They show that counting pairwise collisions in a sample can be used to approximate the $\ell_2$-norm of the probability distribution from which the sample was drawn from. Several more recent works, including this one, make use of this technical tool. Batu et al. [BFR$^+$10] note that running the [GR00] algorithm with $\tilde{O}(\sqrt{n})$ samples yields an algorithms for uniformity testing in the $\ell_1$-norm. Paninski [Pan08] gives an optimal algorithm in this setting that takes a sample of size $O(\sqrt{n})$ and proves a matching lower bound of $\Omega(\sqrt{n})$. Valiant [Val08] shows that a tolerant tester for uniformity (for constant precision) would require $n^{1-o(1)}$ samples. Several works in property testing of distributions approximate the distribution by a small histogram distribution and use this representation as an essential way in their algorithm [BKR04], [BFF$^+$01].

Histograms were subject of extensive research in data stream literature, see [TGIK02, GGI$^+$02] and the references therein. Our algorithm in Section 3 is inspired by streaming algorithm in [TGIK02].

# 2. PRELIMINARIES

Denote by $\mathcal{D}_n$ the set of all discrete distributions over $[n]$. A *property* of a discrete distributions is a subset $\mathcal{P} \subseteq \mathcal{D}_n$. We say that a distribution $p \in \mathcal{D}_n$ is $\epsilon$-far from $p' \in \mathcal{D}_n$ in the $\ell_1$ distance ($\ell_2$ distance) if $\|p - p'\|_1 > \epsilon$ ($\|p - p'\|_2 > \epsilon$).

We say that an algorithm, $\mathcal{A}$, is a *testing algorithm* for the property $\mathcal{P}$ if given an accuracy parameter $\epsilon$ and a distribution $p$:

1. if $p \in \mathcal{P}$, $\mathcal{A}$ accepts $p$ with probability at least 2/3
2. if $p$ is $\epsilon$-far (according to any specified distance measure) from every distribution in $\mathcal{P}$, $\mathcal{A}$ rejects $p$ with probability at least 2/3.

Let $p \in \mathcal{D}_n$, then for every $\ell \in [n]$, denote by $p_\ell$ the probability of the $\ell$-th element. For every $I \subseteq [n]$, let $p(I)$ denote the weight of $I$, i.e. $\sum_{\ell \in I} p_\ell$. For every $I \subseteq [n]$ such that $p(I) \neq 0$, let $p_I$ denote the distribution of $p$ restricted to $I$ i.e. $p_I(\ell) = \frac{p_\ell}{p(I)}$. Call an interval $I$ *flat* if $p_I$ is uniform or $p(I) = 0$.

Given a set of $m$ samples from $p$, $S$, denote by $S_I$ the samples that fall in the interval $I$. For interval $I$ such that $|S_I| > 0$, define the *observed collision probability* of $I$ as $\frac{coll(S_I)}{\binom{|S_I|}{2}}$ where $coll(S_I) \stackrel{\text{def}}{=} \sum_{i \in I} \binom{occ(i,S_I)}{2}$ and $occ(i, S_I)$ is the number of occurrences of $i$ in $S_I$. In [GR00], in the proof of Lemma 1, it was shown that $\mathrm{E}\left[\frac{coll(S_I)}{\binom{|S_I|}{2}}\right] = \|p_I\|_2^2$ and that

$$\Pr\left[\left|\frac{coll(S_I)}{\binom{|S_I|}{2}} - \|p_I\|_2^2\right| > \delta \|p_I\|_2^2\right] < \frac{2}{\delta^2 \cdot \left(\binom{|S_I|}{2} \cdot \|p_I\|_2^2\right)^{1/2}} < \frac{4}{\delta^2 |S_I| \, \|p_I\|_2} . \quad (1)$$

In particular, since $\|p_I\|_2 \leq 1$, we also have that

$$\Pr\left[\left|\frac{coll(S_I)}{\binom{|S_I|}{2}} - \|p_I\|_2^2\right| > \epsilon\right] < \left(\frac{1}{\epsilon}\right)^2 \cdot \frac{1}{|S_I|} . \quad (2)$$

In a similar fashion we prove the following lemma.

LEMMA 1 (BASED ON [GR00]). *If we take $m \geq \frac{24}{\epsilon^2}$ samples, $S$, then, for every interval $I$,*

$$\Pr\left[\left|\frac{coll(S_I)}{\binom{|S|}{2}} - \sum_{\ell \in I} p_\ell^2\right| \leq \epsilon p(I)\right] > \frac{3}{4} \quad (3)$$

PROOF. For every $i < j$ define an indicator variable $C_{i,j}$ so that $C_{i,j} = 1$ if the $i$th sample is equal to the $j$th sample and is in the interval $I$. For every $i < j$, $\mu \stackrel{\text{def}}{=} \mathrm{E}[C_{i,j}] = \sum_{\ell \in I} p_\ell^2$. Let $P \stackrel{\text{def}}{=} \{(i,j) : 1 \leq i < j \leq m\}$. By Chebyshev's inequality:

$$\Pr\left[\left|\frac{\sum_{(i,j)\in P} C_{i,j}}{|P|} - \sum_{\ell \in I} p_\ell^2\right| > \epsilon p(I)\right] \leq \frac{\mathbf{Var}[\sum_{(i,j)\in P} C_{i,j}]}{(\epsilon \cdot p(I) \cdot |P|)^2}$$

From [GR00] we know that

$$\mathbf{Var}\left[\sum_{(i,j)\in P} C_{i,j}\right] \leq |P| \cdot \mu + |P|^{3/2} \cdot \mu^{3/2} \quad (4)$$

and since $\mu \leq p^2(I)$ we have $\mathbf{Var}\left[\sum_{(i,j)\in P} C_{i,j}\right] \leq p(I)^2 \cdot (|P| + |P|^{3/2} \cdot \mu^{1/2})$, thus

$$\begin{aligned}\Pr\left[\left|\frac{\sum_{(i,j)\in P} C_{i,j}}{|P|} - \sum_{\ell \in I} p_\ell^2\right| > \epsilon p(I)\right] &< \frac{|P| + |P|^{3/2} \cdot \mu^{1/2}}{\epsilon^2 |P|^2} \\ &\leq \frac{2}{\epsilon^2 |P|^{1/2}} \quad (5) \\ &\leq \frac{6}{\epsilon^2 m} \leq \frac{1}{4} \quad (6)\end{aligned}$$

□

## 3. NEAR-OPTIMAL PRIORITY $K$-HISTOGRAM

In this section we give an algorithm that given $p \in \mathcal{D}_n$, outputs a priority $k$-histogram which is close in the $\ell_2$ distance to an optimal tiling $k$-histogram that describes $p$. The algorithm, based on a sketching algorithm in [TGIK02], takes a greedy strategy. Initially the algorithm starts with an empty priority histogram. It then proceed by doing $k \ln \epsilon^{-1}$ iterations, where in each iteration it goes over all $\binom{n}{2}$ possible intervals and adds the best one, i.e the interval $I \subseteq [n]$ which minimizes the distance between $p$ and $H$ when added to the currently constructed priority histogram $H$. The algorithm has an efficient sample complexity of only logarithmic dependence on $n$ but the running time has polynomial dependence on $n$. This polynomial dependency is due to the exhaustive search for the interval which minimizes the distance between $p$ and $H$. We note that it is not clear that a logarithmic dependence, or any dependence at all, on the domain size, $n$, is needed. Furthermore, we suspect that a linear dependence on $k$, and not quadratic, is sufficient.

THEOREM 1. *Let $p \in \mathcal{D}_n$ be the distribution and let $H^*$ be the tiling $k$-histogram which minimizes $\|p - H^*\|_2^2$. The priority histogram $H$ reported by Algorithm 1 satisfies $\|p - H\|_2^2 \leq \|p - H^*\|_2^2 + 5\epsilon$. The sample complexity of Algorithm 1 is $\tilde{O}((k/\epsilon)^2 \ln n)$. The running time complexity of Algorithm 1 is $\tilde{O}((k/\epsilon)^2 n^2)$.*

**Algorithm 1:** Greedy algorithm for priority k-histogram

1. Obtain $\ell = \frac{\ln(12n^2)}{2\xi^2}$ samples, $S$, from $p$, where $\xi = \epsilon/(k \ln \frac{1}{\epsilon})$;
2. For each interval $I \subseteq [n]$ set $y_I := \frac{|S_I|}{\ell}$;
3. Obtain $r = \ln(6n^2)$ sets of samples, $S^1, \ldots, S^r$, each of size $m = \frac{24}{\xi^2}$ from $p$;
4. For each interval $I \subseteq [n]$ let $z_I$ be the median of $\frac{coll(S_I^1)}{\binom{|S^1|}{2}}, \ldots, \frac{coll(S_I^r)}{\binom{|S^r|}{2}}$;
5. Initialize the priority histogram $H$ to empty;
6. **for** $i := 1$ *to* $(k \ln \epsilon^{-1})$ **do**
7. **foreach** *interval* $J \subseteq [n]$ **do**
8. Create $H_{J,y_J}$ obtained by:
   - Adding $(J, y_J, r)$ to $H$, where $r = r_{\max} + 1$ and $r_{\max}$ is the maximal priority in $H$;
   - Recomputing the interval to the left (resp. right) of $J$, $I_L$ (resp. $I_R$) so it would not intersect with $J$;
   - Adding $(I_L, y_{I_L}, r)$ and $(I_R, y_{I_R}, r)$ to $H$;

   $c_J := \sum_{I \in H_{J,y_J}} \left(z_I - \frac{y_I^2}{|I|}\right)$;
9. Let $J_{\min}$ be the interval with the smallest value of $c_J$;
10. Update $H$ to be $H_{J_{\min}, y_{J_{\min}}}$;
11. **return** $H$

PROOF. By Chernoff's bound and union bound over the intervals in $[n]$, with high constant probability, for every $I$,

$$|y_I - p(I)| \leq \xi . \quad (7)$$

By Lemma 1 and Chernoff's bound, with high constant probability, for every $I$,

$$|z_I - \sum_{i \in I} p_i^2| \leq \xi p(I) . \quad (8)$$

Henceforth, we assume that the estimations obtained by the algorithm are good, namely, Equations (7) and (8) hold for every interval. It is clear that any function $f$ that has a representation as a tiling $k$-histogram, $H^*$, has a representation as a priority histogram $H$. Moreover, we can transform $H$ to represent $f$ in $k$ steps, simply by adding the $k$ intervals of $H^*$, $(I_1, v_1), \ldots, (I_k, v_k)$, to $H$, as $(I_1, v_1, r), \ldots, (I_k, v_k, r)$, where $r = r_{\max} + 1$ and $r_{\max}$ is the maximal priority over all intervals in $H$. This implies that there exists an interval $J$ and a value $y_J$ such that adding them to $H$ (as described in Algorithm 1) decreases the error in the following way

$$\|p - H_{J,y_J}\|_2^2 - \|p - H^*\|_2^2 \leq \quad (9)$$

$$\left(1 - \frac{1}{k}\right) \cdot \left(\|p - H\|_2^2 - \|p - H^*\|_2^2\right) . \quad (10)$$

where $H_{J,y_J}$ is defined in Algorithm 1 in Step (8). Next, we would like to write the distance between $H_{J,y_J}$ and $p$

as a function of $\sum_{i\in I} p_i^2$ and $p(I)$, for $I \in H_{J,y_J}$. We note that the value of $x$ that minimizes the sum $\sum_{i\in I}(p_i - x)^2$ is $x = \frac{p(I)}{|I|}$, therefore

$$\begin{aligned}
\|p - H_{J,y_J}\|_2^2 &\geq \sum_{I\in H_{J,y_J}} \sum_{i\in I} \left(p_i - \frac{p(I)}{|I|}\right)^2 &(11)\\
&= \sum_{I\in H_{J,y_J}} \sum_{i\in I} \left(p_i^2 - 2p_i\frac{p(I)}{|I|} + \left(\frac{p(I)}{|I|}\right)^2\right)\\
&= \sum_{I\in H_{J,y_J}} \left(\left(\sum_{i\in I} p_i^2\right) - \frac{p(I)^2}{|I|}\right) . &(12)
\end{aligned}$$

Since $c_J = \sum_{I\in H_{J,y_J}} \left(z_I - \frac{y_I^2}{|I|}\right)$, by applying the triangle inequality twice we get that

$$\begin{aligned}
c_J &\leq \sum_{I\in H_{J,y_J}} \left(|z_I - \sum_{i\in I} p_i^2| + |\sum_{i\in I} p_i^2 - \frac{y_I^2}{|I|}|\right) &(13)\\
&\leq \sum_{I\in H_{J,y_J}} |z_I - \sum_{i\in I} p_i^2| &(14)\\
&\quad + \sum_{I\in H_{J,y_J}} \left(|\left(\sum_{i\in I} p_i^2\right) - \frac{p(I)^2}{|I|}| + |\frac{p(I)^2}{|I|} - \frac{y_I^2}{|I|}|\right) ,
\end{aligned}$$

After reordering, we obtain that

$$\begin{aligned}
c_J &\leq \sum_{I\in H_{J,y_J}} \left(\left(\sum_{i\in I} p_i^2\right) - \frac{p(I)^2}{|I|}\right) &(15)\\
&\quad + \sum_{I\in H_{J,y_J}} \left(|z_I - \sum_{i\in I} p_i^2| + \frac{|y_I^2 - p(I)^2|}{|I|}\right) . &(16)
\end{aligned}$$

From the fact that $|y_I^2 - p(I)^2| = |y_I - p(I)| \cdot (y_I + p(I))$ and Equation (7) it follows that

$$|y_I^2 - p(I)^2| \leq \xi(\xi + 2p(I)) . \quad (17)$$

Therefore we obtain from Equations (8), (12), (16) and (17) that

$$\begin{aligned}
c_J &\leq \|p - H_{J,y_J}\|_2^2 + \sum_{I\in H_{J,y_J}} \left(\xi p(I) + \frac{\xi(\xi + 2p(I))}{|I|}\right)\\
&\leq \|p - H_{J,y_J}\|_2^2 + 3\xi + |\{I \in H_{J,y_J}\}|\xi^2 . &(18)
\end{aligned}$$

Since the algorithm calculates $c_J$ for every interval $J$, we derive from Equations (10) and (18) that at the $q$-th step

$$\begin{aligned}
&\left\|p - H_{J_{\min},y_{J_{\min}}}\right\|_2^2 - \|p - H^*\|_2^2 \leq &(19)\\
&\left(1 - \frac{1}{k}\right) \cdot \left(\|p - H\|_2^2 - \|p - H^*\|_2^2\right) + 3\xi + q\xi^2 &(20)
\end{aligned}$$

So for $H$ obtained by the algorithm after $q$ steps we have $\|p - H\|_2^2 - \|p - H^*\|_2^2 \leq \left(1 - \frac{1}{k}\right)^q + q(3\xi + q\xi^2)$. Setting $q = k\ln\frac{1}{\epsilon}$ we obtain that $\|p - H\|_2^2 \leq \|p - H^*\|_2^2 + 5\epsilon$ as desired. □

## 3.1 Improving the Running Time

We now turn to improving the running time complexity to match the sample complexity. Instead of going over all possible intervals in $[n]$ in search for an interval $I \subseteq [n]$ to add to the constructed priority histogram $H$. We search for $I$ over a much smaller subset of intervals, in particular, only those intervals whose endpoints are samples or neighbors of samples. In Lemma 2 we prove that if we decrease the value a histogram $H$ assigns to an interval $I$, then the square of the distance between $H$ and $p$ in the $\ell_2$-norm can grow by at most $2p(I)$. The lemma implies that we can treat light weight intervals as atomic components in our search because they do not affect the distance between $H$ and $p$ by much. While the running time is reduced significantly, we prove that the histogram this algorithm outputs is still close to being optimal.

LEMMA 2. *Let $p \in \mathcal{D}_n$ and let $I$ be an interval in $[n]$. For $0 \leq \beta_1 < \beta_2 \leq 1$,*

$$\sum_{i\in I}(p_i - \beta_1)^2 - \sum_{i\in I}(p_i - \beta_2)^2 \leq 2p(I) \quad (21)$$

THEOREM 2. *Let $p$ and $H^*$ be as in Theorem 1. There is an algorithm that outputs a priority histogram $H$ that satisfies $\|p - H\|_2^2 \leq \|p - H^*\|_2^2 + 8\epsilon$. The sample complexity of the algorithm and the running time complexity of the algorithm is $\tilde{O}((k/\epsilon)^2 \ln n)$.*

PROOF. In the improved algorithm, as in Algorithm 1, we take $\ell = \frac{\ln(12n^2)}{2\xi^2}$ samples, $T$. Instead of going over all $J \subseteq [n]$ in Step (7) we consider only a small subset of intervals as candidates. We denote this subset of intervals by $\mathcal{T}$. Let $T'$ be the set of all elements in $T$ and those that are distance one away, i.e. $T' = \{\min\{i+1, n\}, i, \max\{i-1, 0\} | i \in T\}$. Then $\mathcal{T}$ is the set of all intervals between pairs of elements in $T'$, i.e. $[a, b] \in \mathcal{T}$ if and only if $a \leq b$ and $a, b \in T'$. Thus, the size of $\mathcal{T}$ is bounded above by $\binom{3\ell+1}{2}$. Therefore we decrease the number of iterations in Step (7) from $\binom{n}{2}$ to at most $\binom{3\ell+1}{2}$.

It is easy to see that intervals which are not in $\mathcal{T}$ have small weight. Formally, let $I$ be an intervals such that $p(I) > \xi$. The probability that $I$ has no hits after taking $\ell$ samples is at most $(1-\xi)^\ell < 1/(2n^2)$. Therefore by union bound over all the intervals $I \subseteq [n]$, with high constant probability, for every interval which has no hits after taking $\ell$ samples, the weight of the interval is at most $\xi$.

Next we see why in Step (7) we can ignore intervals which have small weight. Consider a single run of the loop in Step (7) in Algorithm 1. Let $H$ be the histogram constructed by the algorithm so far and let $J_{\min}$ be the interval added to $H$ at the end of the run. We shall see that there is an interval $J \in \mathcal{T}$ such that

$$\left\|p - H_{J,\frac{p(J)}{|J|}}\right\|_2^2 - \left\|p - H_{J_{\min},y_{J_{\min}}}\right\|_2^2 \leq 4\xi . \quad (22)$$

Denote the endpoints of $J_{\min}$ by $a$ and $b$ where $a < b$. Let $I_1 = [a_1, b_1]$ be the largest interval in $\mathcal{T}$ such that $I_1 \subseteq J_{\min}$ and let $I_2 = [a_2, b_2]$ be the smallest interval in $\mathcal{T}$ such that $J_{\min} \subseteq I_2$. Therefore for every interval $J = [x, y]$ where $x \in \{a_1, a_2\}$ and $y \in \{b_1, b_2\}$ we have that $\sum_{i\in J\Delta J_{\min}} p_i \leq 2\xi$ where $J\Delta J_{\min}$ is the symmetric difference of $J$ and $J_{\min}$. Let $\beta_1$, $\beta_2$ the value assigned to $i \in [a_2, a_1]$, $i \in [a_2, a_1]$ by $H_{J_{\min}}, y_{J_{\min}}$, respectively. Notice that the algorithm only assigns values to intervals in $\mathcal{T}$, therefore $\beta_1$ and $\beta_2$ are well defined. Take $J$ to be as follows. If $\beta_1 > y_J$ then take the start-point of $J$ to be $a_1$ otherwise take it to be $a_2$. If $\beta_2 > y_J$ then take the end-point of $J$ to be $b_1$ otherwise take

it to be $b_2$. By lemma 2 it follows that

$$\left\|p - H_{J,y_{J_{\min}}}\right\|_2^2 - \left\|p - H_{J_{\min}, y_{J_{\min})}\right\|_2^2 \leq 2 \sum_{i \in J \Delta J_{\min}} p_i \leq 4\xi \ . \tag{23}$$

Thus, we obtain Equation (22) from the fact that

$$\left\|p - H_{J,\frac{p(J)}{|J|}}\right\|_2^2 = \min_{\delta} \|p - H_{J,\delta}\|_2^2 \tag{24}$$

Thus, by similar calculations as in the proof of theorem 1, after $q$ steps, $\|p - H\|_2^2 - \|p - H^*\|_2^2 \leq \left(1 - \frac{1}{k}\right)^q + q(3\xi + q\xi^2 + 4\xi)$; Setting $q = k \ln \frac{1}{\epsilon}$ we obtain that $\|p - H\|_2^2 - \|p - H^*\|_2^2 \leq 8\epsilon$. □

**Proof of Lemma 2:**

$$\sum_{i \in I}(p_i - \beta_1)^2 - \sum_{i \in I}(p_i - \beta_2)^2 = \tag{25}$$

$$\sum_{i \in I}(p_i^2 - 2\beta_1 p_i + \beta_1^2) - \sum_{i \in I}(p_i^2 - 2\beta_2 p_i + \beta_2^2) \leq \tag{26}$$

$$2p(I)(\beta_2 - \beta_1) + |I|(\beta_1^2 - \beta_2^2) \leq 2p(I) \tag{27}$$

□

# 4. TESTING WHETHER A DISTRIBUTION IS A TILING $k$-HISTOGRAM

In this section we provide testing algorithms for the property of being a tiling $k$-histogram. The testing algorithms attempt to partition $[n]$ into $k$ intervals which are flat according to $p$ (recall that an interval is flat if it has uniform conditional distribution or it has no weight). If it fails to do so then it rejects $p$. Intervals that are close to being flat can be detected because either they have light weight, in which case they can be found via sampling, or they are not light weight, in which case they have small $\ell_2$-norm. Small $\ell_2$-norm can in turn be detected via estimations of the collision probability. Thus an interval that has overall small number of samples or alternatively small number of pairwise collisions is considered by the algorithm to be a flat interval. The search of the flat intervals' boundaries is performed in a similar manner to a search of a value in a binary search. The efficiency of our testing algorithm is stated in the following theorems:

THEOREM 3. *Algorithm 2 is a testing algorithm for the property of being a tiling $k$-histogram for the $\ell_2$ distance measure. The sample complexity of the algorithm is $O(\epsilon^{-4} \ln^2 n)$. The running time complexity of the algorithm is $O(\epsilon^{-4} k \ln^3 n)$.*

THEOREM 4. *There exists a testing algorithm for the property of being a tiling $k$-histogram for the $\ell_1$ distance measure. The sample complexity of the algorithm is $\tilde{O}(\epsilon^{-5}\sqrt{kn})$. The running time complexity of the algorithm is $\tilde{O}(\epsilon^{-5} k\sqrt{kn})$.*

**Proof of Theorem 3:** Let $I$ be an interval in $[n]$ we first show that

$$\Pr\left[\left|z_I - \|p_I\|_2^2\right| \leq \max_i\{\frac{\epsilon^2}{2\hat{p}^i(I)}\}\right] > 1 - \frac{1}{6n^2} \ . \tag{28}$$

where $z_I$ is the median of $\frac{coll(S_I^1)}{\binom{|S_I^1|}{2}}, \ldots, \frac{coll(S_I^r)}{\binom{|S_I^r|}{2}}$. Recall that $\hat{p}^i(I) = \frac{2|S_I^i|}{m}$, hence, due to the facts that $m \geq \frac{64}{\epsilon^4}$ and

**Algorithm 2:** Test Tiling $k$-histogram

```
Obtain r = 16 ln(6n^2) sets of samples, S^1, ..., S^r, each
  of size m = 64 ln n · ε^{-4} from p;
Set previous := 1, low := 1, high := n;
for i := 1 to k do
    while high ≥ low do
        mid := low + (high - low) /2;
        if testFlatness-ℓ2 ([previous,mid], S^1, ..., S^r, ε)
        then
            low := mid+1;
        else
            high := mid−1;
    previous := low;
    high := n;
    If (previous = n) then return ACCEPT;
return REJECT
```

**Algorithm 3:** testFlatness-$\ell_2(I, S^1, \ldots, S^r, \epsilon)$

```
For each i ∈ [r] set p̂^i(I) := 2|S_I^i| / m;
If there exists i ∈ [r] such that |S_I^i|/m < ε^2/2 then return
  ACCEPT ;
Let z_I be the median of coll(S_I^1)/C(|S_I^1|,2), ..., coll(S_I^r)/C(|S_I^r|,2);
If z_I ≤ 1/|I| + max_i{ε^2/(2p̂^i(I))} then return ACCEPT ;
return REJECT;
```

$m \geq |S_I^i|$ we get that $|S_I^i| \geq |S_I^i| \cdot \frac{64}{\epsilon^4 m} \cdot \frac{|S_I^i|}{m} \geq \frac{16\hat{p}^i(I)^2}{\epsilon^4}$. By Equation 2, for each $i \in [r]$,

$$\Pr\left[\left|\frac{coll(S_I^i)}{\binom{|S_I^i|}{2}} - \|p_I\|_2^2\right| \leq \frac{\epsilon^2}{2\hat{p}^i(I)}\right] > \frac{3}{4}. \tag{29}$$

Since each estimate $\frac{coll(S_I^i)}{\binom{|S_I^i|}{2}}$ is close to $\|p_I\|_2^2$ with high constant probability, we get from Chernoff's bound that for $r = 16\ln(6n^2)$ the median of $r$ results is close to $\|p_I\|_2^2$ with very high probability as stated in Equation (28). By union bound over all the intervals in $[n]$, with high constant probability, the following holds for everyone of the at most $n^2$ intervals in $[n]$, $I$,

$$\left|z_I - \|p_I\|_2^2\right| \leq \max_i\{\frac{\epsilon^2}{2\hat{p}^i(I)}\} \ . \tag{30}$$

So henceforth we assume that this is the case.

Assume the algorithm rejects. When this occurs it implies that there are at least $k$ distinct intervals such that for each interval the test testFlatness-$\ell_2$ returned REJECT. For each of these intervals $I$ we have $p(I) \neq 0$ and $z_I > \frac{1}{|I|} + \max_i\{\frac{\epsilon^2}{2\hat{p}^i(I)}\}$. In this case $\|p_I\|_2^2 \geq \frac{1}{|I|}$, and so $I$ is not flat and contains at least one bucket boundary. Thus, there are at least $k$ internal bucket boundaries. Therefore $p$ is not a tiling $k$-histogram.

Assume the algorithm accepts $p$. When this occurs there is a partition of $[n]$ to $k$ intervals, $\mathcal{I}$, such that for each interval $I \in \mathcal{I}$, testFlatness-$\ell_2$ returned ACCEPT. Define $p'$ to be $\frac{p(I)}{|I|}$ on the intervals obtained by the algorithm. For

every $I \in \mathcal{I}$, If is the case that there exists $i \in [r]$, such that $\frac{|S_I^i|}{m} < \frac{\epsilon^2}{2}$, then by fact 1 (below), $p(I) < \epsilon^2$. Therefore, from the fact that $\sum_{i\in I}(p_i - x)^2$ is minimized by $x = \frac{p(I)}{|I|}$ and the Cauchy-Schwarz inequality we get that

$$\sum_{i\in I}\left(p_i - \frac{p(I)}{|I|}\right)^2 \leq \sum_{i\in I} p_i^2 \quad (31)$$

$$\leq p(I)^2 \leq \epsilon^2 p(I)\,. \quad (32)$$

Otherwise, if $\frac{|S_I^i|}{m} \geq \frac{\epsilon^2}{2}$ for every $i \in [r]$ then by the second item in fact 1, $p(I) \geq \frac{\epsilon^2}{4}$. By the first item in fact 1, it follows that $\hat{p}^i(I) = \frac{2|S_I^i|}{m} \geq p(I)$ and therefore

$$z_I \leq \frac{1}{|I|} + \frac{\epsilon^2}{2p(I)}\,. \quad (33)$$

where $z_I$ is the median of $\frac{coll(S_I^1)}{\binom{|S_I^1|}{2}}, \ldots, \frac{coll(S_I^r)}{\binom{|S_I^r|}{2}}$. This implies that $\|p_I\|_2^2 \leq \frac{1}{|I|} + \frac{\epsilon^2}{p(I)}$. Thus, $\|p_I - u\|_2^2 \leq \frac{\epsilon^2}{p(I)}$ and since $\|p_I - u\|_2^2 = \sum_{i\in I}\left(\frac{p_i}{p(I)} - \frac{1}{|I|}\right)^2$ we get that $\sum_{i\in I}\left(p_i - \frac{p(I)}{|I|}\right)^2 \leq \epsilon^2 p(I)$. Hence $\sum_{I\in\mathcal{I}}\sum_{i\in I}\left(p_i - \frac{p(I)}{|I|}\right)^2 \leq \epsilon^2$, thus, $p$ is $\epsilon$-close to $p'$ in the $\ell_2$-norm. □

FACT 1. *If we take $m \geq \frac{48\ln(2n^2\gamma)}{\epsilon^2}$ samples, $S$, then with probability greater than $1 - \frac{1}{\gamma}$:*

1. *For any $I$ such that $p(I) \geq \frac{\epsilon^2}{4}$, $\frac{p(I)}{2} \leq \frac{|S_I|}{m} \leq \frac{3p(I)}{2}$*
2. *For any $I$ such that $\frac{|S_I|}{m} \geq \frac{\epsilon^2}{2}$, $p(I) > \frac{\epsilon^2}{4}$*
3. *For any $I$ such that $\frac{|S_I|}{m} < \frac{\epsilon^2}{2}$, $p(I) < \epsilon^2$*

PROOF. Fix $I$, if $p(I) \geq \frac{\epsilon^2}{4}$, by Chernoff's bound with probability greater than $1 - 2e^{-\frac{m\epsilon^2}{48}}$,

$$\frac{p(I)}{2} \leq \frac{|S_I|}{m} \leq \frac{3p(I)}{2}. \quad (34)$$

In particular, if $p(I) = \frac{\epsilon^2}{4}$, then $\frac{|S_I^i|}{m} \leq \frac{3\epsilon^2}{8}$, thus if $\frac{|S_I|}{m} \geq \frac{\epsilon^2}{2} > \frac{3\epsilon^2}{8}$ then $p(I) > \frac{\epsilon^2}{4}$. If $\frac{|S_I|}{m} < \frac{\epsilon^2}{2}$ then either $p(I) \leq \frac{\epsilon^2}{4}$ or $p(I) > \frac{\epsilon^2}{4}$ but then $p(I) \leq \frac{2|S_I|}{m} < \epsilon^2$. By the union bound, with probability greater than $1 - n^2 \cdot 2e^{-\frac{m\epsilon^2}{48}} > 1 - \frac{1}{\gamma}$, the above is true for every $I$. □

**Algorithm 4:** testFlatness-$\ell_1(I, S^1, \ldots, S^r, \epsilon)$

1 If there exists $i \in [r]$ such that $|S_I^i| < \frac{16^3\sqrt{|I|}}{\epsilon^4}$ then **return** ACCEPT;
2 Let $z_I$ be the median of $\frac{coll(S_I^1)}{\binom{|S^1|}{2}}, , \ldots, \frac{coll(S_I^r)}{\binom{|S^r|}{2}}$;
3 If $z_I \leq \frac{1}{|I|}(1 + \frac{\epsilon^2}{4})$ then **return** ACCEPT ;
4 **return** REJECT;

**Proof of Theorem 4:** Apply Algorithm 2 with the following changes: take each set of samples $S^i$ to be of size $m = 2^{13}\sqrt{kn}\epsilon^{-5}$ and replace testFlatness-$\ell_2$ with testFlatness-$\ell_1$ . By Equation 1

$$\Pr\left[\left|\frac{coll(S_I)}{\binom{|S_I|}{2}} - \|p_I\|_2^2\right| > \delta\|p_I\|_2^2\right] < \frac{4}{\delta^2|S_I|\,\|p_I\|_2}. \quad (35)$$

Thus, if $S_I$ is such that $|S_I| \geq \frac{16\sqrt{|I|}}{\delta^2} \geq \frac{16}{\delta^2\|p_I\|_2}$, then

$$\Pr\left[\left|\frac{coll(S_I)}{\binom{|S_I|}{2}} - \|p_I\|_2^2\right| > \delta\|p_I\|_2^2\right] > \frac{3}{4}. \quad (36)$$

By additive Chernoff's bound and the union bound for $r = 16\ln(6n^2)$ and $\delta = \frac{\epsilon^2}{16}$, with high constant probability for every interval $I$ that passes Step 1 in Algorithm 4 it holds that $\left|\frac{coll(S_I)}{\binom{|S_I|}{2}} - \|p_I\|_2^2\right| \leq \delta\|p_I\|_2^2$ (the total number of intervals in $[n]$ is less than $n^2$). So from from this point on we assume that the algorithm obtains a $\delta$-multiplicative approximation of $\|p_I\|_2^2$ for every $I$ that passes Step 1.

Assume the algorithm rejects $p$, then there are at least $k$ distinct intervals such that for each interval the test testFlatness-$\ell_1$ returned REJECT. By our assumption each of these intervals is not flat and thus contains at least one bucket boundary. Thus, there are at least $k$ internal buckets boundaries, therefore $p$ is not a tiling $k$-histogram.

Assume the algorithm accepts $p$, then there is a partition of $[n]$ to $k$ intervals, $\mathcal{I}$, such that for each interval $I \in \mathcal{I}$, testFlatness-$\ell_1$ returned ACCEPT. Define $p'$ to be $\frac{p(I)}{|I|}$ on the intervals obtained by the algorithm. For any interval $I$ for which testFlatness-$\ell_1$ returned ACCEPT and passes Step 1 it holds that $\|p_I - u\|_2 < \frac{\epsilon}{2\sqrt{|I|}}$ thus $\sum_{i\in I}\left|p_i - \frac{p(I)}{|I|}\right| \leq \frac{\epsilon}{2}p(I)$. Denote by $\mathcal{L}$ the set of intervals for which testFlatness-$\ell_1$ returned ACCEPT on Step 1. By Chernoff's bound, for every $I \in \mathcal{L}$, with probability greater than $1 - e^{-\frac{m\epsilon}{32k}}$, either $p(I) \leq \frac{\epsilon}{4k}$ or $p(I) \leq \frac{2\cdot 16^3\sqrt{|I|}}{m\epsilon^4}$. Hence, with probability greater than $1 - n^2 \cdot r \cdot e^{-\frac{m\epsilon}{32k}} > 1 - \frac{1}{6}$, the total weight of the intervals in $\mathcal{L}$:

$$\sum_{I\in\mathcal{L}}\max\{\frac{2\cdot 16^3\sqrt{|I|}}{m\epsilon^4}, \frac{\epsilon}{4k}\} \leq \frac{\epsilon}{4} + \sum_{I\in\mathcal{L}}\frac{2\cdot 16^3\sqrt{|I|}}{m\epsilon^4} \quad (37)$$

$$= \frac{\epsilon}{4}\left(1 + \sum_{I\in\mathcal{L}}\frac{\sqrt{|I|}}{\sqrt{kn}}\right) \quad (38)$$

$$\leq \frac{\epsilon}{2}\,, \quad (39)$$

where the last inequality follows from the fact that $|\mathcal{L}| \leq k$ which implies that $\sum_{I\in\mathcal{L}}\sqrt{|I|/n} \leq \sqrt{k}$. Therefore, $p$ is $\epsilon$-close to $p'$ in $\ell_1$-norm. □

## 4.1 Lower Bound

We prove that for every $k \leq 1/\epsilon$, the upper bound in Theorem 4 is tight in term of the dependence in $k$ and $n$. We note that for $k = n$, testing tiling $k$-histogram is trivial, i.e. every distribution is a tiling $n$-histogram. Hence, we can not expect to have a lower bound for any $k$. We also note that the testing lower bound is also an approximation lower bound.

THEOREM 5. *Given a distribution $D$ testing if $D$ is a tiling $k$-histogram in the $\ell_1$-norm requires $\Omega(\sqrt{kn})$ samples for every $k \leq 1/\epsilon$.*

PROOF. Divide $[n]$ into $k$ intervals of equal size (up to $\pm 1$). In the YES instance the total probability of each interval alternates between 0 and $\lfloor 2/k \rfloor$ and within each interval the elements have equal probability. The NO instance is defined similarly with one exception, randomly pick one of the intervals that have total probability $\lfloor 2/k \rfloor$, $I$, and within $I$ randomly pick half of the elements to have probability 0 and the other half of the elements to have twice the probability of the corresponding elements in the YES instance. In the proof of the lower bound for testing uniformity it is shown that distinguishing a uniform distribution from a distribution that is uniform on a random half of the elements (and has 0 weight on the other half) requires $\Omega(\sqrt{n})$. Since the number of elements in $I$ is $\Theta(n/k)$, by a similar argument we know that at least $\Omega(\sqrt{n/k})$ samples are required from $I$ in order to distinguish the YES instance from the NO instance. From the fact that the total probability of $I$ is $\Theta(1/k)$ we know that in order to obtain $\Theta(\sqrt{n/k})$ hits in $I$ we are required to take a total number of samples which is of order $\sqrt{nk}$, thus we obtain a lower bound of $\Omega(\sqrt{nk})$. □

# 5. REFERENCES


[AAK+07] N. Alon, A. Andoni, T. Kaufman, K. Matulef, R. Rubinfeld, and N. Xie. Testing $k$-wise and almost $k$-wise independence. In *Proceedings of the Thirty-Ninth Annual ACM Symposium on the Theory of Computing (STOC)*, pages 496–505, 2007.

[BDKR05] T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld. The complexity of approximating the entropy. *SIAM Journal on Computing*, 35(1):132–150, 2005.

[BFF+01] T. Batu, L. Fortnow, E. Fischer, R. Kumar, R. Rubinfeld, and P. White. Testing random variables for independence and identity. In *Proceedings of the Forty-Second Annual Symposium on Foundations of Computer Science (FOCS)*, pages 442–451, 2001.

[BFR+00] T. Batu, L. Fortnow, R. Rubinfeld, W.D. Smith, and P. White. Testing that distributions are close. In *Proceedings of the Forty-First Annual Symposium on Foundations of Computer Science (FOCS)*, pages 259–269, Los Alamitos, CA, USA, 2000. IEEE Computer Society.

[BFR+10] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing closeness of discrete distributions. *CoRR*, abs/1009.5397, 2010. This is a long version of [BFR+00].

[BKR04] T. Batu, R. Kumar, and R. Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *Proceedings of the Thirty-Sixth Annual ACM Symposium on the Theory of Computing (STOC)*, pages 381–390, 2004.

[CMN98] S. Chaudhuri, R. Motwani, and V. Narasayya. Random sampling for histogram construction: how much is enough? *SIGMOD*, 1998.

[GGI+02] A. Gilbert, S. Guha, P. Indyk, Y. Kotidis, M. Muthukrishnan, and M. Strauss. Fast, small-space algorithms for approximate histogram maintenance. *STOC*, 2002.

[GGR98] O. Goldreich, S. Goldwasser, and D. Ron. Property testing and its connection to learning and approximation. *Journal of the ACM*, 45(4):653–750, 1998.

[GKS06] S. Guha, N. Koudas, and K. Shim. Approximation and streaming algorithms for histogram construction problems. *ACM Transactions on Database Systems (TODS)*, 31(1), 2006.

[GMP97] P.B. Gibbons, Y Matias, and V. Poosala. Fast incremental maintenance of approximate histograms. *VLDB*, 1997.

[GR00] O. Goldreich and D. Ron. On testing expansion in bounded-degree graphs. *Electronic Colloqium on Computational Complexity*, 7(20), 2000.

[Ioa03] Y. Ioannidis. The history of histograms (abridged). *VLDB*, 2003.

[JPK+98] H. V. Jagadish, V. Poosala, N. Koudas, K. Sevcik, S. Muthukrishnan, and T. Suel. Optimal histograms with quality guarantees. *VLDB*, 1998.

[Pan08] L. Paninski. Testing for uniformity given very sparsely-sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008.

[Ron08] D. Ron. Property testing: A learning theory perspective. *Foundations and Trends in Machine Learning*, 3:307–402, 2008.

[RRSS09] S. Raskhodnikova, D. Ron, A. Shpilka, and A. Smith. Strong lower bonds for approximating distributions support size and the distinct elements problem. *SIAM Journal on Computing*, 39(3):813–842, 2009.

[RS96] R. Rubinfeld and M. Sudan. Robust characterization of polynomials with applications to program testing. *SIAM Journal on Computing*, 25(2):252–271, 1996.

[Rub06] R. Rubinfeld. Sublinear time algorithms. In *Proc. International Congress of Mathematicians*, volume 3, pages 1095–1111, 2006.

[TGIK02] Nitin Thaper, Sudipto Guha, Piotr Indyk, and Nick Koudas. Dynamic multidimensional histograms. In *SIGMOD Conference*, pages 428–439, 2002.

[Val08] P. Valiant. Testing symmetric properties of distributions. In *Proceedings of the Fourtieth Annual ACM Symposium on the Theory of Computing (STOC)*, pages 383–392, 2008.

[VV11] G. Valiant and P. Valiant. Estimating the unseen: an $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs. In *Proceedings of the Fourty-Third Annual ACM Symposium on the Theory of Computing*, pages 685–694, 2011. See also ECCC TR10-179 and TR10-180.