Approximating and testing k-histogram distributions in sub-linear time

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.

**Citation**

**As Published**
http://dx.doi.org/10.1145/2213556.2213561

**Publisher**
Association for Computing Machinery (ACM)

**Version**
Author's final manuscript

**Accessed**
Sat Jan 05 04:55:04 EST 2019

**Citable Link**
http://hdl.handle.net/1721.1/87005

**Terms of Use**
Creative Commons Attribution-Noncommercial-Share Alike

**Detailed Terms**
http://creativecommons.org/licenses/by-nc-sa/4.0/
Approximating and Testing k-Histogram Distributions in Sub-linear Time

Piotr Indyk‡  Reut Levi†  Ronitt Rubinfeld‡

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.

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 interest 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 representa-

Copyright 2012 ACM 978-1-4503-1248-6/12/05 ...$10.00.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
PPOD'S 12, May 21–23, 2012, Scottsdale, Arizona, USA.
Copyright 2012 ACM 978-1-4503-1248-6/12/05 ...$10.00.
ing 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 \( i \in I_t \). 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 \( \tilde{O}(\epsilon^{-2} / \sqrt{k} n) \). We provide a similar test for the \( \ell_2 \) norm that has sample complexity of \( \tilde{O}(\epsilon^{-2} n^2) \). We prove that testing if a distribution is a tiling \( k \)-histogram in the \( \ell_1 \)-norm requires \( \Omega(\sqrt{k} n) \) 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, RRS09, 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 \( p \) and an explicit description of distribution \( q \). 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, \( q \), 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. [BBFR+01] note that running the [GR00] algorithm with \( \tilde{O}(\sqrt{n}) \) samples yields an algorithm for uniformity testing in the \( \ell_1 \)-norm. Paninski [Pan08] gives an optimal algorithm in this setting that takes a sample of size \( \tilde{O}(\sqrt{n}) \) 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, BFR+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 algorithms [TGIK02].

### 2. PRELIMINARIES

Denote by \( D_n \), the set of all discrete distributions over \([n]\). A **property** of a discrete distributions is a subset \( \mathcal{P} \subseteq D_n \).

We say that a distribution \( p \in D_n \) is \( \epsilon \)-far from \( p' \in D_n \) in the \( \ell_2 \) distance (\( \ell_2 \) distance) if \( \|p - p'\|_2 > \epsilon \).

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

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

Let \( p \in D_n \), 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. \( \sum_{i \in I} p_i \). For every \( I \subseteq [n] \) such that \( p(I) > 0 \), let \( p_r(I) = \frac{p(I)}{\sum_{i \in I} p_i} \). Call an interval \( I \) **flat** if \( p_r(I) \) is uniform or \( p(I) = 0 \).

Given a set of \( m \) samples from \( p, S \), denote by \( S_t \) the samples that fall in \( I_t \). For interval \( I \) such that \( |S_t| > 0 \), define the observed collision probability of \( I \) as \( \frac{\text{coll}(S_t)}{|S_t|^2} \) where \( \text{coll}(S_t) = \sum_i \epsilon(S_i, S_t) \) and \( \epsilon(S_i, S_t) \) is the number of occurrences of \( i \) in \( S_t \). In [GR00], in the proof of Lemma 1, it was shown that \( E \left[ \frac{\text{coll}(S_t)}{|S_t|^2} \right] = \|p_r\|_2^2 \) and that

\[
\Pr \left[ \frac{\text{coll}(S_t)}{|S_t|^2} - \|p_r\|_2^2 > \delta \|p_r\|_2^2 \right] < \frac{2}{\delta^2 \cdot \left( \|S_t\|_2 \cdot \|p_r\|_2 \right)^{3/2}} < \frac{4}{\delta^2 |S_t|^2 \|p_r\|_2}.
\]

(1)

In particular, since \( \|p_r\|_2 \leq 1 \), we also have that

\[
\Pr \left[ \frac{\text{coll}(S_t)}{|S_t|^2} - \|p_r\|_2^2 > \epsilon \right] < \left( \frac{1}{\delta^2} \right)^{1/2} \cdot \frac{1}{|S_t|}.
\]

(2)
Lema 1 (Based on [GR00]). If we take $m \geq \frac{24}{\epsilon^2}$ samples, $S$, then, for every interval $I$,
\[
\Pr \left[ \frac{\text{coll}(S_t)}{\left(\frac{\log(1+2)}{2}\right)} - \sum_{t \in I} p_t^2 \leq \epsilon p(I) \right] > \frac{3}{4} \tag{3}
\]

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 $i < j$, $\mu \equiv \mathbb{E}[C_{i,j}] = \sum_{t \in I} p_t^2$. Let $P_{i,j} \equiv \{(i,j) : 1 \leq i < j \leq m\}$. By Chebyshev’s inequality:
\[
\Pr \left[ \sum_{(i,j) \in P} C_{i,j} - \sum_{t \in I} p_t^2 > \epsilon p(I) \right] \leq \mathbb{V} \mathbb{A} \left[ \sum_{(i,j) \in P} C_{i,j} \right] \leq \frac{P_{i,j}}{|P|} \cdot \mu + |P|^{3/2} \cdot \mu^{3/2} \tag{4}
\]
and since $\mu \leq p^2(I)$ we have $\mathbb{V} \mathbb{A} \left[ \sum_{(i,j) \in P} C_{i,j} \right] \leq p(I)^2$. \(\sum_{(i,j) \in P} C_{i,j} / |P| - \sum_{t \in I} p_t^2 > \epsilon p(I)\) \(\epsilon p(I) + |P|^{3/2} \cdot \mu^{3/2} / \epsilon^2 p(I)^2 \leq \frac{2}{\epsilon^2 m} = \frac{1}{4} \tag{5}
\]

3. NEAR-OPTIMAL PRIORITY K-HISTOGRAM

In this section we give an algorithm that given $p \in D_n$, 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 [TGGK02], 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 $\binom{n}{k}$ 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 dependency, 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 $p \in D_n$ 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 $O((k/\epsilon)^2 \ln n)$. The running time complexity of Algorithm 1 is $O((k/\epsilon)^2 n^2)$.

Algorithm 1: Greedy algorithm for priority k-histogram

1. Obtain $\ell = \frac{\ln(12n^2)}{2\epsilon^2}$ samples, $S$, from $p$, where $\xi = \epsilon/(k \ln 1/\epsilon)$;
2. For each interval $I \subseteq [n]$ set $y_I := \frac{\||p|}{|I|}$;
3. Obtain $r = \ln(n/2)$ sets of samples, $S^1, \ldots, S^r$, each of size $m = \frac{24}{\epsilon^2}$ from $p$;
4. For each interval $I \subseteq [n]$ let $z_I$ be the median of $\text{coll}(S^1_I), \ldots, \text{coll}(S^r_I)$;
5. Initialize the priority histogram $H$ to empty;
6. For $i := 1$ to $(k \ln \epsilon^{-1})$ do
7. For each interval $J \subseteq [n]$ do
8. Let $H_{J,y_I}$ obtained by:
   - Adding $(J, y_I)$ to $H$, where $r = r_{\text{max}} + 1$ and $r_{\text{max}}$ 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 $(I_L, y_{IR})$ and $(I_R, y_{IR})$ to $H$;
9. Let $s_{\text{min}}$ be the interval with the smallest value of $c_{s_I}$;
10. Update $H$ to be $H_{s_{\text{min}}, y_{s_{\text{min}}}}$;
11. Return $H$.

Proof. By Chernoff’s bound and union bound over the intervals in $[n]$, with high constant probability, for every $I$,
\[
|y_I - p(I)| \leq \xi \cdot |I| \tag{7}
\]
By Lemma 1 and Chernoff’s bound, with high constant probability, for every $I$,
\[
|z_I - \sum_{t \in I} p_t^2| \leq \xi p(I) \tag{8}
\]
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 we can transform any priority histogram $H$ to any tiling $k$-histogram, $H^*$, by adding the $k$ intervals of $H^*$, $(I_1, v_1), \ldots, (I_k, v_k)$, to $H$, as $(I_1, v_1), \ldots, (I_k, v_k)$, where $v = r_{\text{max}} + 1$ and $r_{\text{max}}$ 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
\[
\|p - H_{J,y_J}\|_2^2 - \|p - H^*\|_2^2 \leq \left(1 - \frac{1}{k}\right) \cdot \left(\|p - H\|_2^2 - \|p - H^*\|_2^2 - \|p - H^*\|_2^2\right) \tag{9}
\]
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 $\sum_{t \in I} p_t^2$ and $p(I)$, for $I \in H_{J,y_J}$. We note that the value of $x$ that minimizes the sum $\sum_{t \in I} (p_t - x)^2$
is \( x = \frac{\nu(I)}{|I|} \), therefore

\[
\|p - H_{J,y,I}\|_2^2 \geq \sum_{t \in H_{J,y,I}} \left( p_t - \frac{p(I)}{|I|} \right)^2 \tag{11}
\]

\[
= \sum_{t \in H_{J,y,I}} \left( \sum_{i \in t} p_i^2 - 2p_t p(I) + \left( \frac{p(I)}{|I|} \right)^2 \right)
\]

\[
= \sum_{t \in H_{J,y,I}} \left( \sum_{i \in t} p_i^2 - \frac{p(I)^2}{|I|} \right). \tag{12}
\]

Since \( c_J = \sum_{t \in H_{J,y,I}} \left( z_i - \frac{\nu(I)}{2} \right) \), by applying the triangle inequality twice we get that

\[
c_J \leq \sum_{t \in H_{J,y,I}} \left( \sum_{i \in t} p_i^2 \right) + \sum_{i \in t} \left( \frac{p(I)^2}{|I|} \right), \tag{13}
\]

\[
\leq \sum_{t \in H_{J,y,I}} \left( |z_t - \sum_{i \in t} p_i^2| + \sum_{i \in t} \left( |\frac{p(I)^2}{|I|} + |\frac{p(I)^2}{|I|} - \frac{\nu(I)}{2} | \right) \right), \tag{14}
\]

After reordering, we obtain that

\[
c_J \leq \sum_{t \in H_{J,y,I}} \left( \sum_{i \in t} p_i^2 - \frac{p(I)^2}{|I|} \right)
\]

\[+ \sum_{t \in H_{J,y,I}} \left( |z_t - \sum_{i \in t} p_i^2| + \left| \frac{\nu(I)}{2} - p(I) \right| \right), \tag{16}
\]

From the fact that \( |\frac{\nu(I)}{2} - p(I)| \leq \xi (\xi + 2p(I)) \) and Equation (7) it follows that

\[
|\frac{\nu(I)}{2} - p(I)| \leq \xi (\xi + 2p(I)). \tag{17}
\]

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

\[
c_J \leq \|p - H_{J,y,I}\|_2^2 + \sum_{t \in H_{J,y,I}} \left( \xi p(I) + \xi (\xi + 2p(I)) \right)
\]

\[\leq \|p - H_{J,y,I}\|_2^2 + 3\xi + \|I \in H_{J,y,I}\|_2^2. \tag{18}\]

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

\[
\|p - H_{J_{\min,y,J_{\min}}}\|_2^2 - \|p - H^{*}\|_2^2 \leq \left( 1 - \frac{1}{k} \right) \cdot \left( \|p - H^{*}\|_2^2 - \|p - H^{*}\|_2^2 + 3\xi + 4\xi^2 \right) \tag{20}\]

So for \( H \) obtained by the algorithm after \( q \) steps we have

\[
\|p - H_{J_{\min,y,J_{\min}}}\|_2^2 - \|p - H^{*}\|_2^2 \leq \left( 1 - \frac{1}{k} \right) \cdot \left( \|p - H^{*}\|_2^2 - \|p - H^{*}\|_2^2 + 3\xi + 4\xi^2 \right) \leq 4\xi. \tag{22}\]

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 \( T \) such that \( I_1 \subseteq J_{\min} \) and let \( I_2 = [a_2, b_2] \) be the smallest interval in \( T \) such that \( J_{\min} \subseteq I_2 \). Therefore for every interval \( I = [x, y] \) where \( x \in \{a_1, a_2\} \) and \( y \in \{b_1, b_2\} \) we have that \( \sum_{t \in J \setminus J_{\min}} p_t \leq \delta \) where \( \sum_{t \in J \setminus J_{\min}} p_t \) is the symmetric difference of \( J \) and \( J_{\min} \). Let \( \beta_1, \beta_2 \) be the value assigned to \( i \in [a_2, a_1] \) and \( i \in [a_2, a_1] \) by \( H_{J_{\min,y,J_{\min}}} \), respectively. Notice that the algorithm only assigns values to intervals in \( T \), 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
\[ \left\| p - H_{y,\lesssim_{\min}} \right\|_2 \leq 2 \sum_{i \in J, y_{\lesssim_{\min}}} p_i \leq 4 \xi . \tag{23} \]
Thus, we obtain Equation (22) from the fact that
\[ \left\| p - H_j \right\|_2 = \min \left\{ \left\| p - H_{y,\lesssim_{\min}} \right\|_2 \right\} . \tag{24} \]
Thus, by similar calculations as in the proof of theorem 1, after $q$ steps, $\left\| p - H \right\|_2 - \left\| p - H^* \right\|_2 \leq (1 - \frac{1}{q})^q + q(3\xi + Q\xi^2 + \xi)$; Setting $q = k \ln \frac{1}{\delta}$ we obtain that $\left\| p - H \right\|_2 - \left\| p - H^* \right\|_2 \leq 8\delta$. □

Proof of Lemma 2:
\[ \sum_{i \in I} (p_i - \beta_1)^2 - \sum_{i \in I} (p_i - \beta_2)^2 = \sum_{i \in I} (p_i^2 - 2\beta_1 p_i + \beta_1^2) - \sum_{i \in I} (p_i^2 - 2\beta_2 p_i + \beta_2^2) \leq 2p(I)(\beta_2 - \beta_1) + |I|(|\beta_1^2 - \beta_2^2|) \leq 2p(I) \]
\[ \square \]

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 $O(e^{-\epsilon \ln^3 n})$. The running time complexity of the algorithm is $O(e^{-\epsilon k \ln^3 n})$.

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 $O(e^{-\epsilon \sqrt{k} \ln n})$. The running time complexity of the algorithm is $O(e^{-\epsilon \sqrt{k} \ln^3 n})$.

Proof of Theorem 3: Let $I$ be an interval in $[n]$ we first show that
\[ \Pr \left[ \left| z_I - \left\| p_I \right\|_2 \right| \leq \max_i \frac{\epsilon^2}{2p'(I)} \right] > 1 - \frac{1}{6n^2} . \tag{28} \]
where $z_I$ is the median of $\frac{\text{coll}(S_I^1)}{(\frac{|S_I^1|}{m})^2}, \ldots, \frac{\text{coll}(S_I^r)}{(\frac{|S_I^r|}{m})^2}$. Recall that $p'(I) = \frac{2|S_I^r|}{m}$, hence, due to the facts that $m \geq \frac{64}{n^2}$ and $n \geq 64 \ln(n)$.

Algorithm 2: Test Tiling $k$-histogram
1. Obtain $r = 16 \ln(6n^2)$ sets of samples, $S^1, \ldots, S^r$, each of size $m = 64 \ln n \cdot e^{-\epsilon}$ from $p$;
2. Set previous := 1, low := 1, high := $n$;
3. for $i := 1$ to $k$ do
4. while high $\geq$ low do
5. mid := low + (high - low) / 2;
6. if testFlatness-$\ell_2$ ((previous, mid), $S^1, \ldots, S^r, \epsilon$) then
7. low := mid + 1;
8. else
9. high := mid - 1;
10. previous := low;
11. high := n;
12. if (previous = n) then return ACCEPT;
13. return REJECT

Algorithm 3: testFlatness-$\ell_2$ ($I, S^1, \ldots, S^r, \epsilon$)
1. For each $i \in [r]$ set $p'(I) := \frac{2|S_I^i|}{m}$;
2. If there exists $i \in [r]$ such that $|S_I^i| < \epsilon^2$ then return ACCEPT;
3. Let $z_I$ be the median of $\frac{\text{coll}(S_I^1)}{(\frac{|S_I^1|}{m})^2}, \ldots, \frac{\text{coll}(S_I^r)}{(\frac{|S_I^r|}{m})^2}$;
4. If $z_I \leq \frac{\epsilon}{|I|} + \max_i \frac{\epsilon^2}{2p'(I)}$ then return ACCEPT;
5. return REJECT.

$m \geq |S_I^r|$ we get that $|S_I^r| \geq |S_I^1| \cdot \frac{64}{n^2} \cdot \frac{|S_I^r|}{m} \geq \frac{16\epsilon^4(t)^2}{t^2}$. By Equation 2, for each $i \in [r]$,\[ \Pr \left[ \left| \frac{\text{coll}(S_I^i)}{(\frac{|S_I^i|}{m})^2} - \left\| p_I \right\|_2^2 \right| \leq \frac{\epsilon^2}{2p'(I)} \right] > \frac{3}{4} . \tag{29} \]
Since each estimate $\frac{\text{coll}(S_I^i)}{(\frac{|S_I^i|}{m})^2}$ is close to $\left\| p_I \right\|_2^2$ 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 $\left\| p_I \right\|_2^2$ 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 $[n]$, $I$.
\[ |z_I - \left\| p_I \right\|_2^2| \leq \max_i \frac{\epsilon^2}{2p'(I)} . \tag{30} \]
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 testFlatness-$\ell_2$ returned REJECT. For each of these intervals $I$ we have $p(I) \neq 0$ and $z_I \geq \frac{|I|}{\epsilon^2}$. In this case $\left\| p_I \right\|_2 \geq \frac{1}{\epsilon^2}$, 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, $I$, such that for each interval $I \in I$, testFlatness-$\ell_2$ returned accept. Define $p'$ to be $\frac{\text{coll}(I)}{|I|^2}$ on the intervals obtained by the algorithm. For
every \( I \in \mathcal{I} \). If it is the case that there exists \( i \in [r] \), such that \( \frac{|S_i|}{m} < \frac{e^2}{4} \), then by fact 1 (below), \( p(I) < e^2 \). 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
\[
\sum_{i \in I} \left( p_i - \frac{p(I)}{|I|} \right)^2 \leq \sum_{i \in I} p_i^2 \leq p(I)^2 \leq e^2 p(I). \tag{32}
\]
Otherwise, if \( \frac{|S_i|}{m} \geq \frac{e^2}{4} \) for every \( i \in [r] \) then by the second item in fact 1, \( p(I) \geq \frac{e^2}{4} \). By the first item in fact 1, it follows that \( p'(I) = \frac{2|S_i|}{m} \geq p(I) \) and therefore
\[
z_I \leq 1 - \frac{e^2}{4p(I)}. \tag{33}
\]
where \( z_I \) is the median of \( \frac{\|\text{coll}(S_i')\|_2}{\|S_i'\|_2}, \ldots, \frac{\|\text{coll}(S_i')\|_2}{\|S_i'\|_2} \). This implies that \( \|p_I\|_2^2 \leq \frac{1}{4} + \frac{e^2}{2p(I)} \). Thus, \( \|p_I - u\|_2^2 \leq \frac{e^2}{2} \) and since \( \|p_I - u\|_2^2 = \sum_{i \in I} \left( p_i - \frac{p(I)}{|I|} \right)^2 \) we get that \( \sum_{i \in I} \left( p_i - \frac{p(I)}{|I|} \right)^2 \leq e^2 p(I) \). Hence \( \sum_{i \in I} \sum_{i \in I} \left( p_i - \frac{p(I)}{|I|} \right)^2 \leq e^2, \) thus, \( p \) is \( \epsilon \)-close to \( p' \) in the \( \ell_2 \)-norm.

FACT 1. If we take \( m \geq \frac{8k \ln (2e^2 k)}{|\mathcal{I}|} \) samples, \( S \), then with probability greater than \( 1 - \frac{1}{e^4} \):

\[\begin{align*}
1. & \text{ For any } I \text{ such that } p(I) \geq \frac{e^2}{4}, \quad p'(I) \leq \frac{|S_I|}{m} \leq \frac{\ln(|\mathcal{I}|)}{2}.
2. & \text{ For any } I \text{ such that } \frac{|S_I|}{m} \geq \frac{e^2}{4}, \quad p(I) > \frac{e^2}{4}.
3. & \text{ For any } I \text{ such that } \frac{|S_I|}{m} < \frac{e^2}{4}, \quad p(I) < e^2.
\end{align*}\]

PROOF. Fix \( I, \) if \( p(I) \geq \frac{e^2}{4} \), by Chernoff’s bound with probability greater than \( 1 - 2e^{-\frac{m^2e^2}{2m^2}} \),
\[
p(I) \leq \frac{|S_I|}{m} \leq \frac{3p(I)}{2} \tag{34}
\]
In particular, if \( p(I) = \frac{e^2}{4} \), then \( \frac{|S_I|}{m} \leq \frac{e^2}{8} \), thus if \( \frac{|S_I|}{m} \geq \frac{e^2}{4} \) then \( p(I) > \frac{e^2}{4} \). If \( \frac{|S_I|}{m} < \frac{e^2}{8} \), then either \( p(I) \leq \frac{e^2}{4} \) or \( p(I) > \frac{e^2}{8} \) but then \( p(I) \leq \frac{|S_I|}{m} < e^2 \). By the union bound, with probability greater than \( 1 - n^2 \cdot 2e^{-\frac{m^2e^2}{8m}} > 1 - \frac{1}{e^4} \), 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 \( |S_i'| < \frac{8k \sqrt{|\mathcal{I}|}}{me^4} \) then \textbf{return ACCEPT}.
2. Let \( z_I \) be the median of \( \frac{\|\text{coll}(S_i')\|_2}{\|S_i'\|_2}, \ldots, \frac{\|\text{coll}(S_i')\|_2}{\|S_i'\|_2} \).
3. If \( z_I \leq \frac{1}{4}(1 + \frac{e^2}{8}) \) then \textbf{return ACCEPT}.
4. \textbf{return REJECT}.

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

By Equation 1
\[
\Pr \left[ \left( \| \text{coll}(S_I) \|_2 \right)^2 - \| p_I \|_2^2 \geq \frac{4}{\delta^2 |S_I|} \| p_I \|_2 \right) < \frac{1}{2} \tag{35}
\]
Thus, if \( S_I \) is such that \( |S_I| \geq \frac{16 \sqrt{|\mathcal{I}|}}{\delta^2} \geq \frac{16}{\epsilon^2} \), then
\[
\Pr \left[ \left( \| \text{coll}(S_I) \|_2 \right)^2 - \| p_I \|_2^2 \geq \frac{4}{\delta^2 |S_I|} \| p_I \|_2 \right) > \frac{3}{4} \tag{36}
\]
By additive Chernoff’s bound and the union bound for \( r = 16 \ln(n^2) \) and \( \delta = \frac{e^2}{8} \), with high constant probability for every interval \( I \) that passes Step 1 in Algorithm 4 it holds that
\[
\| \text{coll}(S_I) \|_2 - \| p_I \|_2^2 \leq \| p_I \|_2^2 \] (the total number of intervals in \([n]\) is less than \( n^2 \)). So 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 accepts \( p \), then there is a partition of \([n]\) to \( |\mathcal{I}| \) intervals, \( I \), such that for each interval \( I \in \mathcal{I} \), testFlatness-\( \ell_1 \) returned ACCEPT. Define \( p' \) to be \( \frac{u}{|I|} \) 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
\[
\| p_I - u \|_2 < \frac{\sqrt{|I|}}{2\sqrt{|\mathcal{I}|}} \text{ thus } \| \sum_{I \in \mathcal{I}} (p_i - \frac{p(I)}{|I|}) \|_2 \leq \frac{2}{\sqrt{|\mathcal{I}|}} p(I).
\]
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 \( 1 - e^{-\frac{\ln|\mathcal{I}|}{8}} \), either \( p(I) \leq \frac{16}{\epsilon^2} \) or \( p(I) \leq \frac{16 \sqrt{|\mathcal{I}|}}{\epsilon^2} \). Hence, with probability greater than \( 1 - n^2 \cdot e^{-\frac{\ln|\mathcal{I}|}{8}} \), the total weight of the intervals in \( \mathcal{L} \):
\[
\sum_{I \in \mathcal{L}} \max\left( \frac{2}{4\epsilon^2} \sqrt{|I|}, \frac{\sqrt{|I|}}{2\epsilon^2} \right) \leq \frac{\epsilon}{4} + \sum_{I \in \mathcal{L}} \frac{2^{16} \sqrt{|I|}}{\epsilon^2} \tag{38}
\]
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 the \( \ell_1 \)-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 \([2/k]\) 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 \([2/k]\), \(I\), 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 \(\sqrt{n/k}\), thus we obtain a lower bound of \(\Omega(\sqrt{n/k})\).

5. REFERENCES


