A Competitive Test for Uniformity of Monotone Distributions

Jayadev Acharya **Ashkan Jafarpour** **Alon Orlitsky** **Ananda Theertha Suresh**
University of California San Diego

Abstract

We propose a test that takes random samples drawn from a monotone distribution and decides whether or not the distribution is uniform. The test is nearly optimal in that it uses at most $\mathcal{O}(n\sqrt{\log n})$ samples, where n is the number of samples that a genie who knew all but one bit about the underlying distribution would need for the same task. Conversely, we show that any such test would require $\Omega(n\sqrt{\log n})$ samples for some distributions.

1 INTRODUCTION

Many a lively debate rages over whether some quantity grows with a possibly related parameter, or is independent of it: cancer prevalence vs. radiation exposure; diabetes onset rate vs. age; flu frequency vs. (negative) temperature; heart attacks vs. stress; recovery speed vs. drug dosage; product failures vs. time from manufacture; taxes or education level or (negative) incarceration rate vs. income; gender vs. salary; party affiliation vs. income or age; grades vs. preparation time; quality vs. cost; and finally, age before beauty.

We propose a test that takes random samples generated by a monotone distribution and decides whether or not the distribution is uniform. The test is near optimal in the following sense. A genie who knew the underlying distribution, could clearly decide if it is uniform or not without any samples. We show that if a genie who knew all but one (specific) bit about the underlying distribution would require n samples to decide on uniformity, then our test will require at most $\mathcal{O}(n\sqrt{\log n})$ samples to make the same decision. We also show that $\Omega(n\sqrt{\log n})$ samples are necessary.

In general, the parameter of interest can be discrete (income) or continuous (temperature) and can assume different ranges. However, every monotone distribution over a finite continuous interval, and every monotone discrete distribution over a finite domain, can be converted to an equivalent monotone distribution over $[0, 1)$. For example, the monotone distribution $p(2) = 0.6$, $p(4) = 0.3$, $p(7) = 0.1$, can be converted to $0.6{\cdot}1_{[0,1/3)}(x)+0.3{\cdot}1_{[1/3,2/3)}(x)+0.1{\cdot}1_{[2/3,1)}(x)$. Hence we consider *probability density functions*, or *distributions* for short, over $[0, 1)$. Monotone distributions over infinite ranges can be similarly analyzed and will be addressed in the full version of this paper.

Over $[0, 1)$, the distribution $u(x) = 1$ is *uniform*. A distribution f is *monotone* if for all $x < y < z$, $(f(y) - f(x)) \cdot (f(z) - f(y)) \geq 0$. Note that monotone distributions can be either increasing or decreasing.

Let $\mathcal{M}$ be the collection of monotone distributions, and $\mathcal{M}^- = \mathcal{M} - \{u\}$, the collection of monotone non-uniform distributions. We would like to decide whether a distribution $f \in \mathcal{M}$ is uniform or not based on independent samples it generates.

A *test* is a mapping $t : [0, 1]^* \to \{\texttt{uni}, \texttt{non}\}$ declaring whether the observed samples are believed to be generated by a uniform or a non-uniform distribution. The error probability of t for $f \in \mathcal{M}$ based on n samples it generates is

$$P_e^t(f, n) \stackrel{\text{def}}{=} \begin{cases} P(t(X^n) = \texttt{non}), & \text{if } f = u, \\ P(t(X^n) = \texttt{uni}), & \text{if } f \in \mathcal{M}^-, \end{cases}$$

where $X^n = X_1, \ldots, X_n$ are samples distributed independently $\sim f$.

As in several recent works, we are mostly interested in the *sample complexity*, the number of samples required to achieve a certain error. For a distribution $f \in \mathcal{M}^-$, test t, and error $\epsilon > 0$, let

$$N_\epsilon^t(f) \stackrel{\text{def}}{=} \min\{n : P_e^t(f, n) < \epsilon \text{ and } P_e^t(u, n) < \epsilon\}$$

be the smallest number of samples for which t has error $< \epsilon$ for both f and the uniform distribution. We require t's error to be small for both f and u as a trivial test achieves 0 error for just one of them.

Appearing in Proceedings of the 16^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2013, Scottsdale, AZ, USA. Volume 31 of JMLR: W&CP 31.

The smallest number of samples required by any test to achieve error probability $< \epsilon$ for $f \in \mathcal{M}^-$ and u is therefore,

$$N^*_\epsilon(f) \stackrel{\text{def}}{=} \min_t \{N^t_\epsilon(f)\}.$$

The test achieving the minimum is denoted t^*.

Example 1. *For $\delta > 0$, let*

$$f(x) = \begin{cases} 1+\delta & \text{if } x \in [0, \frac{1}{2}), \\ 1-\delta & \text{if } x \in [\frac{1}{2}, 1). \end{cases}$$

Since both f and u are constant over $[0, \frac{1}{2})$ and $[\frac{1}{2}, 1)$, all the information X^n conveys about f and u is contained in $M = |\{i : X_i < \frac{1}{2}\}|$, the number of samples in $[0, \frac{1}{2})$. If the underlying distribution is u, then $M \sim \text{Bin}(\frac{1}{2}, n)$, with expected value $\frac{1}{2}n$ and standard deviation $\frac{1}{2}\sqrt{n}$. If the underlying distribution is f, then $M \sim \text{Bin}(\frac{1}{2}(1+\delta), n)$, with expected value $\frac{1}{2}(1+\delta)n$ and standard deviation $\frac{1}{2}\sqrt{(1-\delta^2)n}$.

*Consider the test t that decides **uni** if $M < \frac{1}{2}(1+\frac{\delta}{2})n$ and **non** otherwise. The test errs only if M strays from its mean by $> n\frac{\delta}{2}$, namely by roughly $\delta\sqrt{n}$ standard deviations. If $n > C\frac{1}{\delta^2}$, then the error probability will be small, showing that for any fixed ϵ, $N^*_\epsilon(f) \le N^t_\epsilon(f) = \mathcal{O}(\frac{1}{\delta^2})$.*

For the lower bound, observe that if $n \ll \frac{1}{\delta^2}$, then the difference between the means of M when sampled under f and u will be significantly smaller than a standard deviation, and no test will distinguish the two alternatives. Hence $N^*_\epsilon(f) = \Omega(\frac{1}{\delta^2})$, and therefore $N^*_\epsilon(f) = \Theta(\frac{1}{\delta^2})$. This argument is of course just intuitive. A rigorous proof is based on Equation (1) and computes the ℓ_1 distance between the two binomial distributions.

Clearly, the closer f gets to u, the larger $N^*(f)$ becomes. Previous approaches to this problem have therefore considered the sub-collection $\mathcal{M}_\delta \subseteq \mathcal{M}$ of all monotone distributions whose ℓ_1 distance from u is at least δ. (Daskalakis et al., 2011) showed that a test similar to t in the last example satisfies $N^t_\epsilon(f) = \mathcal{O}(\frac{1}{\delta^2})$ for all $f \in \mathcal{M}_\delta$, and that for some $f \in \mathcal{M}_\delta$, $N^*_\epsilon(f) = \Omega(\frac{1}{\delta^2})$. Therefore the highest sample complexity of any distribution in $\mathcal{M}_\delta$ is $\Theta(\frac{1}{\delta^2})$. However, the next example shows that some $\mathcal{M}_\delta$ distributions. have a far lower complexity.

Example 2. *Let $f(x) = \frac{\delta}{2}\Delta(x) + (1 - \frac{\delta}{2})$. where $\Delta(x)$ is Dirac Delta function. Clearly $\ell_1(f, u) = \delta$. Consider the test that declares **non** if a sample at $x = 0$ is observed, and **uni** otherwise. For the uniform distribution u, the test errs with probability 0, and to ensure that for f it errs with probability $\le 1/3$, we need $\frac{1}{3} \ge (1 - \frac{\delta}{2})^n$. Namely $n = \lceil -\log 3/\log(1-\frac{\delta}{2})\rceil \le \lceil 2\delta^{-1}\log 3\rceil$ samples suffice. It can be easily shown that roughly this number is also necessary, hence $N^*_\epsilon(f) = \Theta(\frac{1}{\delta})$.*

The two examples show that the ℓ_1 distance of f from u is not an accurate measure for the number of samples needed to test its uniformity, and in particular that some distributions in $\mathcal{M}_\delta$ can be tested for uniformity using much less than $\Theta(\frac{1}{\delta}^2)$ samples. They also show that achieving $N^*_\epsilon(f)$ may require advance knowledge of f as the test may depend on the underlying distribution.

In this paper we derive a test that performs almost as well as possible, not just for the worst distribution in a subclass of $\mathcal{M}$, but for every distribution in $\mathcal{M}$. Its performance for every underlying distribution approaches that of the best test designed with knowledge of all but one (albeit carefully chosen) bit about the underlying distribution.

Specifically, if a genie knew the underlying monotone distribution, it could determine whether the distribution is uniform or not without any samples at all. We show that if the underlying distribution is f, and a very knowledgeable genie, who knew that the underlying distribution was either f or u but did not know which, in a sense missing just a single bit about the distribution, then if that genie needed n samples to determine whether the distribution was uniform or not, our test, designed of course without knowledge of f, would require at most $\mathcal{O}(n\sqrt{\log n})$ samples.

Before formally stating the results, note that for many data-based decision problems, including this one, once an error probability $\epsilon < 1/2$ is achieved, one can repeat the test several times and take the majority of all decisions, resulting in an error diminishing exponentially fast in the number of repetitions. Specifically, if we repeat the test on T independent data sets, it can be easily shown that the error probability of the combined test is at most $\frac{1}{2}(4\epsilon(1-\epsilon))^{T/2}$.

For simplicity therefore we consider the number of samples needed to get error probability $< 1/3$. Any other desired error can be achieved by repeating the experiment a constant number of times. For $f \in \mathcal{M}^-$ and test t, let

$$N^t(f) \stackrel{\text{def}}{=} N^t_{1/3}(f),$$

and

$$N^*(f) \stackrel{\text{def}}{=} N^*_{1/3}(f),$$

be the number of samples that t, and the best test, need to have error probability $< 1/3$.

We design a *single* test t_c that is nearly optimal regardless of the underlying distribution. Namely, for

every $f \in \mathcal{M}^-$

$$N^c(f) \le \mathcal{O}\Big(N^*(f) \cdot \sqrt{\log N^*(f)}\Big).$$

We also show that this extra factor is necessary. More precisely, for every test t, there is a $f \in \mathcal{M}^-$ such that

$$N^t(f) = \Omega\Big(N^*(f) \cdot \sqrt{\log N^*(f)}\Big).$$

Two observations are in order. First, all uniformity tests make some inherent assumption about the underlying distribution. For example, the test (Daskalakis et al., 2011) for distributions in $\mathcal{M}_\delta$ assumes that the tested distribution is either uniform or in $\mathcal{M}_\delta$. If the distribution is not in these two classes, namely $0 < \ell_1(f, u) < \delta$, then the test's performance is not guaranteed. Similarly, for our problem, if $n = o(N^*(f)\sqrt{\log N^*(f)})$, then there is no guarantee on the performance of t_c. In fact, we show that in that case, for sufficiently large n, t_c outputs `uni`with probability at least $\frac{2}{3}$.

Second, with the stated number of samples, t_c has error $\le 1/3$. If a smaller error ϵ is desired, then as above, the number of samples can be multiplied by $17 \ln \frac{1}{\epsilon}$. Note however that in that case, we would be comparing $1/3$ error for the genie with ϵ error for t_c. If we required the genie to achieve ϵ error too, the dependence on ϵ in the constant will decrease and perhaps vanish.

2 RELATED WORKS

The statistics literature offers various uniformity tests, see (Woodroofe and Sun, 1999) and references therein. The computer science community has considered the problem more recently, typically addressing discrete distributions. Without loss of generality, assume the underlying distribution is over $[k] \stackrel{\text{def}}{=} \{1, 2, \dots k\}$. (Paninski, 2008) showed that testing if the distribution is uniform or ϵ far away from uniform in ℓ_1 distance, requires $\Theta\Big(\frac{\sqrt{k}}{\epsilon^2}\Big)$ samples.

(Batu et al., 2004) showed that testing if the distribution is monotone or far from all monotone distributions in ℓ_1 distance requires $\widetilde{\mathcal{O}}(\sqrt{k})$ samples, where the implied constant is an inverse polynomial in the ℓ_1 distance. They also showed that testing if two monotone distributions are close in ℓ_1 distance requires $\mathcal{O}(\text{polylog}(k))$ samples.

(Daskalakis et al., 2011) extended these results to m-modal distributions. In particular, they showed that testing if a monotone distribution equals a prespecified distribution requires $\mathcal{O}(\sqrt{\log k} \log\log k \epsilon^{-2.5})$ samples and the dependence on k is optimal up to a $\mathcal{O}(\log\log k)$ factor.

Although not directly related to this paper, multidimensional distributions were considered as well. (Rubinfeld and Servedio, 2009) showed that testing whether a monotone distribution over the n dimensional boolean hypercube $\{-1, 1\}^n$ is uniform or far from uniform in ℓ_1 distance, requires $\widetilde{\Theta}(n)$ samples, where $\widetilde{\Theta}$ represents Θ with possible poly-logarithmic factors. Similarly, (Adamaszek et al., 2010) showed that testing uniformity of monotone distributions over the continuous $[0, 1]^n$ hypercube requires $\tilde{\Theta}(n/\epsilon^2)$ samples.

All the above results consider the sample complexity of the worst distribution in a class. A different, competitive, approach that compares the performance for each distribution to the best possible, was considered as well.

(Acharya et al., 2011) and (Acharya et al., 2012) considered *closeness testing*, where we would like to determine whether two sequences are drawn from the same or from different distributions, and *classification* where we are given two training sequences generated by unknown distributions and would like to determine which of the two distributions generated a new test sequence. For both problems, they proposed tests that require $\tilde{\mathcal{O}}(n^{3/2})$ samples, where n is the number of samples required by the optimal test, designed with knowledge of the underlying distribution multiset.

3 PRELIMINARIES

3.1 Poisson sampling

Consider any partition of the range $[0, 1)$ into disjoint intervals. When a distribution on $[0, 1)$ is sampled exactly n times, the number of times elements appearing in the intervals are dependent (for example, they sum to exactly n), complicating the analysis of many properties. A standard approach to overcome the dependence, *e.g.*, (Mitzenmacher and Upfal, 2005) is to sample the distribution a random $\text{poi}(n)$ times, the Poisson distribution with parameter n, resulting in sequences of random length close to n.

Poisson tail bounds show that for any $\alpha > 0$, with high probability a random variable $\sim \text{poi}(n(1+\alpha))$ is larger than n. Thus, any test t with error probability $\frac{1}{3} - \epsilon$ for n samples, can be modified to work with error probability $< \frac{1}{3}$ for $\text{poi}(n(1+\alpha))$ samples. Since α can be any positive constant, the test works with a fractionally larger Poisson parameter.

For a distribution f, the probability that $\text{poi}(n)$ samples generated according to f will result in samples $\overline{x}$

is

$$f(\overline{x}) = \frac{e^{-n}n^{|\overline{x}|}}{|\overline{x}|!}.$$

In particular, for the uniform distribution u,

$$u(\overline{x}) = \frac{e^{-n}n^{|\overline{x}|}}{|\overline{x}|!}.$$

We will also use an equivalent formulation of Poisson sampling. Let h be a non-negative function over $[0,1)$, and let $H \stackrel{\text{def}}{=} \int_0^1 h(x)$ denote its integral. Then a distribution over $[0,1)^*$ is $\text{poi}(nH)$ according to h iff

- The number of samples in any two disjoint subsets are independent.
- For any $\mathcal{A} \subset [0,1)$, the number of samples in $\mathcal{A}$ is $\text{poi}(n\int_{x\in\mathcal{A}} h(x)dx)$.

3.2 Simulation

Another useful property of Poisson sampling is that in some cases one can use samples generated by one distribution to simulate samples generated by another, without necessarily knowing the underlying distribution.

Let h and f be any non-negative functions over $[0,1)$ such that h is further from 1 than f is, namely for all x, $(h(x)-1)/(f(x)-1) \geq 1$.

Let $\overline{Y}$ be $\text{poi}(n)$ samples that are generated according to either h or u. We show that without the knowledge of the underlying distribution, one can convert $\overline{Y}$ to $\overline{X}$ such that if the underlying distribution is h then $\overline{X} \sim f$ and if $\overline{Y} \sim u$ then $\overline{X} \sim u$.

Lemma 3. *There exists an algorithm with input $\overline{Y}, n$ and output $\overline{X}$ such that*

- *If $\overline{Y} \sim h$, then $\overline{X} \sim f$,*
- *If $\overline{Y} \sim u$, then $\overline{X} \sim u$.*

Proof. Generate $\overline{Z} \sim u$ of length $\text{poi}(n)$. Construct the set of $\overline{X}$ as follows. Add any sample Y_i to the initially empty set $\overline{X}$ w.p. $(f(X_i)-1)/(h(X_i)-1)$ and add any sample Z_i to the set w.p. $(h(X_i) - f(X_i))/(h(X_i)-1)$. One can show that for any bin $\mathcal{A}$, and the lemma will follow from

- if $\overline{Y} \sim h$, then $E[|\overline{X}_{\mathcal{A}}|] = n\int_{x\in\mathcal{A}} f(x)$,
- if $\overline{Y} \sim u$, then $E[|\overline{X}_{\mathcal{A}}|] = n|\mathcal{A}|$. □

3.3 Error bounds

One can consider two type of error probabilities for any test t. The average error, which is the error of the test t when the samples are either from f or u with equal probability,

$$\bar{P}_e^t(f,n) \stackrel{\text{def}}{=} \frac{1}{2}(P_e^t(f,n) + P_e^t(u,n)).$$

The worst-case error of t which is the larger of the errors for f and u,

$$\hat{P}_e^t(f,n) \stackrel{\text{def}}{=} \max(P_e^t(f,n), P_e^t(u,n)).$$

As with all hypothesis testing problems,

$$\begin{aligned}\bar{P}_e^t(f,n) &\geq \frac{1}{2}\min_t(P_e^t(f,n) + P_e^t(u,n)) \\ &= \frac{1}{2} - \frac{|f-u|_{1,n}}{4}.\end{aligned}$$

Clearly,

$$\begin{aligned}\hat{P}_e^t(f,n) &\geq \bar{P}_e^t(f,n) \\ &\geq \frac{1}{2} - \frac{|f-u|_{1,n}}{4}, \qquad (1)\end{aligned}$$

and

$$\begin{aligned}\min_t \hat{P}_e^t(f,n) &\leq 2\min_t \bar{P}_e^t(f,n) \\ &= \sum_{\overline{x}} \min(f(\overline{x}), u(\overline{x})) \qquad (2) \\ &= 1 - \frac{|f-u|_{1,n}}{2}.\end{aligned}$$

Where the ℓ_1 distance between f and u is

$$|f-u|_{1,n} \stackrel{\text{def}}{=} \sum_{\overline{x}} |f(\overline{x}) - u(\overline{x})|.$$

4 UPPER BOUND

4.1 Test

We propose a test that distinguishes every f from u with Poisson $n' = \mathcal{O}(N^*(f)\sqrt{\log N^*(f)})$ samples with error probability $\leq \frac{1}{3}$, without the knowledge of f. For brevity, let $n = N^*(f)$.

Let $k \stackrel{\text{def}}{=} \lceil \log_2 n \rceil + 5$. Partition the interval $[0,1)$ into $2k$ bins $\mathcal{I}_1, \ldots, \mathcal{I}_{2k}$ as follows,

$$\mathcal{I}_i \stackrel{\text{def}}{=} \begin{cases} [0, 2^{-k}) & i = 1, \\ [2^{i-2-k}, 2^{i-1-k}) & i = 2, \ldots, k, \\ [1-2^{k-i}, 1-2^{k-i-1}) & i = k{+}1, \ldots, 2k{-}1, \\ [1-2^{-k}, 1) & i = 2k. \end{cases}$$

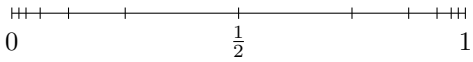

Figure 1: Partition of $[0, 1)$ into $\mathcal{I}_1, \ldots, \mathcal{I}_{2k}$

Note that the bin sizes increase exponentially from $\mathcal{I}_2$ to $\mathcal{I}_k$ and decrease exponentially from $\mathcal{I}_{k+1}$ to $\mathcal{I}_{2k-1}$. This partition is shown in Figure 1.

The test t_c, described below, performs a variation of the χ^2-test on the number of samples appearing in each bin, if either one of them, or their sum, is large, the test declares f to be `non`. Note that since the variation of the χ^2-test we use can be negative, we need to check both the individual values and their sum.

Test $\mathbf{t_c}$

Input: Sequence $\overline{x}$ of length $\text{poi}(n' \stackrel{\text{def}}{=} 1000n\sqrt{k})$
Output: `uni` or `non`
for $1 \le i \le 2k$, let $\nu_i = |\{x : x \in \mathcal{I}_i\}|$ and $\lambda_i = n'|\mathcal{I}_i|$
if $\exists i$ s.t. $\frac{(\nu_i - \lambda_i)^2 - \nu_i}{\lambda_i} \ge 5\sqrt{k}$ or $\sum_{i=1}^{2k} \frac{(\nu_i - \lambda_i)^2 - \nu_i}{\lambda_i} \ge 5\sqrt{k}$
  return `non`
else
  return `uni`

In the remainder of this section, we prove the competitiveness of t_c.

4.2 Bounds on sample complexity

Suppose a monotone distribution f is distinguishable from u using n samples. In this subsection we show that with a constant factor more samples we can distinguish f from u by only using the number of samples within each bin. Also, we show that if f can be distinguished from u, then a variation of χ^2 distance on the number of samples in the bins is large. Conversely, we show that if f cannot be distinguished from u with n samples, then χ^2 distance is small.

Define a staircase distribution g, shown in Figure 2, that assigns the same probability as f to each bin. Namely, for $x \in \mathcal{I}_i$

$$g(x) = \frac{1}{|\mathcal{I}_i|} \int_{y \in \mathcal{I}_i} f(y).$$

The following theorem uses Lemma 3 to relate the error probability of f to that of g.

Theorem 4. $N^*_{\frac{1}{4}}(g) \le 2N^*(f)$.

Proof. Without loss of generality, suppose that f is non-increasing. We show that from samples of g (or u), we can generate samples that are distributed according to f (or u). Then we run the same optimal test for f on the induced samples of g.

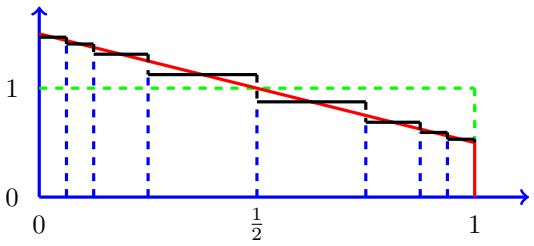

Figure 2: g and f

Let $\overline{Y}$ be $\text{poi}(2n)$ samples from g. Partition $\overline{Y}$ into two sequences $\overline{Y}_1$ and $\overline{Y}_2$, where each sample in $\overline{Y}$ appears in either $\overline{Y}_1$ or $\overline{Y}_2$ with equal probability. The set of elements in $\overline{Y}_1$ and $\overline{Y}_2$ are two independent copies of Poisson n samples from g and their union is $\overline{Y}$.

Since g is not further away from 1 than f, Lemma 3 cannot be applied directly. Instead, observe that the average of f (and hence g) in $\mathcal{I}_i$ is higher than all the values of f in I_{i+1} and it is less than all the values of f in I_{i-1}. Also observe that $|\mathcal{I}_{i-1}|, |\mathcal{I}_{i+1}| \ge \frac{1}{2}|\mathcal{I}_i|$ and we are taking twice the number of samples from the distribution g. By scaling and transforming and using the following steps we generate samples $\overline{X}$ within each bin.

1. For any $i \ge 3$, if $\forall x \in \mathcal{I}_i, f(x) \ge 1$ then samples of $\overline{X}$ belong to $\mathcal{I}_i$ are generated from samples of $\overline{Y}$ within $\mathcal{I}_{i-1}$. Similarly if $\forall x \in \mathcal{I}_2, f(x) \ge 1$, then we use samples of $\overline{Y}_1$ from bin $\mathcal{I}_1$ to generate samples in $\mathcal{I}_2$.

2. For any $i \le 2k-2$, if $\forall x \in \mathcal{I}_i, f(x) \le 1$ then samples of $\overline{X}$ belong to $\mathcal{I}_i$ are generated from samples of $\overline{Y}$ within $\mathcal{I}_{i+1}$. Similarly if $\forall x \in \mathcal{I}_{2k-1}, f(x) \le 1$, then we use samples of $\overline{Y}_1$ from bin $\mathcal{I}_{2k}$ to generate samples in $\mathcal{I}_{2k-1}$.

3. If $f(x) - 1$ changes sign inside I_i where $i \in \{2, \ldots, 2k-1\}$ for the portion of the $\mathcal{I}_i$ where $f(x) \ge 1$ use item 1 or 2 and for the portion where $f(x) \le 1$ use item 3 or 4 to generate the samples of $\overline{X}$ within $\mathcal{I}_i$.

The above procedure is illustrated in Figure 3 for $k = 4$. An arrow from $\mathcal{I}_i$ to $\mathcal{I}_j$ indicates that the samples of $\overline{X}$ in bin $\mathcal{I}_j$ are generated from samples of $\overline{Y}$ within bin $\mathcal{I}_i$. The label of the arrow indicates the step used.

Similar to the proof technique in Lemma 3 it can be shown that the samples $\overline{X}$ are distributed according to f (or u) in bins $\mathcal{I}_2, \ldots, \mathcal{I}_{2k-1}$ and has no samples in $\mathcal{I}_1 \cup \mathcal{I}_{2k}$.

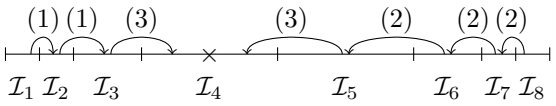

Figure 3: Generation scheme

Let t^* be a test such that $P_e^{t^*}(f,n) < \frac{1}{3}$ and $P_e^{t^*}(u,n) < \frac{1}{3}$. The following test, t has error probability less than 0.4.

1. If $\overline{Y}_2$ has samples in $\mathcal{I}_1$ or $\mathcal{I}_{2k}$ then $t(\overline{Y}) = \texttt{non}$.
2. Run the test t^* on samples $\overline{X}$.

When the underlying distribution is u, after taking Poisson n samples, w.p. $\geq e^{-\frac{1}{16}}$ no sample appears in $\mathcal{I}_1 \cup \mathcal{I}_{2k}$. Since the same test t^* is used, the total error in this case is $< \frac{1}{3}$. Adding the probability of having sample in the $\mathcal{I}_1 \cup \mathcal{I}_{2k}$ we have $P_e^t(u,2n) < \frac{1}{3} + (1 - e^{-\frac{1}{16}}) < 0.4$.

When the underlying distribution is nonuniform, if a sample appears in $\mathcal{I}_1 \cup \mathcal{I}_{2k}$, then the error is zero while in the other case it runs the test t^* and has error probability less than $\frac{1}{3}$. □

So far we have shown that $N^*_{0.4}(g) \leq 2n$. Hence, from Equation (1),

$$0.4 \geq \frac{1}{2} - \frac{|g-u|_{1,2n}}{4}. \tag{3}$$

The following observation helps us evaluate the ℓ_1 distance between g and u.

Observation 5. *Since g and u are constant within each bin, their number of samples within each bin is a sufficient statistic to distinguish them.*

Let ν_i denote the number of samples in $\mathcal{I}_i$. Let $\lambda_i = E[\nu_i]$ when the samples are distributed according to u, and $\lambda'_i = E[\nu_i]$ when the samples are distributed according to f. Note that λ_i and λ'_i are functions of number of samples and for the ease of notation, parameter n is omitted. When the underlying distribution is u, ν_i is Poisson distributed with mean at λ_i and when the underlying distribution is f or g, ν_i is Poisson distributed with mean at λ'_i. Let f_i be the average of f in $\mathcal{I}_i$, and $\overline{\nu} = \nu_1, \ldots, \nu_{2k}$.

Let $u(\overline{\nu})$ be the distribution of $\overline{\nu}$ when $\overline{X} \sim u$ and $f(\overline{\nu})$ be the distribution of $\overline{\nu}$ when $\overline{X} \sim f$. Then,

$$u(\overline{\nu}) \stackrel{\text{def}}{=} \prod_i \frac{e^{-\lambda_i} {\lambda_i}^{\nu_i}}{\nu_i!} \text{ and } f(\overline{\nu}) \stackrel{\text{def}}{=} \prod_i \frac{e^{-\lambda'_i} {\lambda'_i}^{\nu_i}}{\nu_i!}.$$

Let E_h denote the expectation with respect to h. The following lemma, its proof omitted, bounds the ℓ_1 distance between two distributions.

Lemma 6. *For any two distributions h_1 and h_2,*

$$\ell_1(h_1,h_2)^2 \leq E_{h_1}\left[\frac{h_1}{h_2}\right] - 1.$$

The next lemma lower bounds a variant of the χ^2 distance between f and u.

Lemma 7. *For $n = N^*(f)$, with $\text{poi}(2n)$ samples,*

$$\Delta \stackrel{\text{def}}{=} \sum_{i=1}^{2k} \frac{(\lambda'_i - \lambda_i)^2}{\lambda_i} > 0.14.$$

Proof. Using Observation 5 and Equation (3),

$$|f(\overline{\nu}) - u(\overline{\nu})| \geq 0.4.$$

Substituting $f(\overline{\nu})$ and $u(\overline{\nu})$ and using Lemma 6,

$$\begin{aligned} 1.16 &\leq \sum_{\overline{\nu}} \prod_{i=1}^{2k} \frac{e^{-\lambda'_i} {\lambda'_i}^{\nu_i}}{\nu_i!} e^{\lambda_i - \lambda'_i} \left(\frac{\lambda'_i}{\lambda_i}\right)^{\nu_i} \\ &= \exp\left(\sum_{i=1}^{2k} \frac{(\lambda'_i - \lambda_i)^2}{\lambda_i}\right), \end{aligned}$$

where the last inequality follows from generating function of Poisson random variables. Hence,

$$\sum_{i=1}^{2k} \frac{(\lambda'_i - \lambda_i)^2}{\lambda_i} \geq \ln 1.16 > 0.14. \qquad \square$$

Next we upper bound the χ^2 distance between two distributions.

Lemma 8. *For $n = N^*(f)$, with $\text{poi}(n)$ samples,*

$$\sum_{i=1}^{2k} \frac{(\lambda_i - \lambda'_i)^2}{\lambda_i + \lambda'_i} \leq 4 \ln 3.$$

Proof. Since samples from f can be used to generate samples from g, the error of f is at most that of g, hence $N^*(g) \geq N^*(f)$. From Equation (2),

$$\begin{aligned} \frac{1}{3} &\leq \sum_{\overline{\nu}} \min(u(\overline{\nu}), f(\overline{\nu})) \\ &\leq \sum_{\overline{\nu}} \sqrt{u(\overline{\nu}) f(\overline{\nu})} \\ &= \exp(-\frac{1}{2} \sum_{i=1}^{2k} (\sqrt{\lambda_i} - \sqrt{\lambda'_i})^2) \\ &\leq \exp\left(-\frac{1}{4} \sum_{i=1}^{2k} \frac{(\lambda_i - \lambda'_i)^2}{\lambda_i + \lambda'_i}\right). \end{aligned}$$

Hence,

$$\sum_{i=1}^{2k} \frac{(\lambda_i - \lambda'_i)^2}{\lambda_i + \lambda'_i} \leq 4 \ln 3.$$

□

4.3 Proof of competitiveness

Theorem 9. *If $n' > 1000n\sqrt{\log n}$, then $P_e^{t_c}(u, n') \le \frac{1}{3}$ and $P_e^{t_c}(f, n') \le \frac{1}{3}$.*

Proof. Let $\overline{\nu}$ be distributed according to f, and

$$\alpha_i \stackrel{\text{def}}{=} \frac{(\nu_i - \lambda_i)^2 - \nu_i}{\lambda_i},$$
$$\alpha \stackrel{\text{def}}{=} \sum_{i=1}^{2k} \alpha_i.$$

One can show that

$$E[\alpha] = \sum_{i=1}^{2k} \frac{(\lambda_i' - \lambda_i)^2}{\lambda_i} = \frac{n'}{2n}\Delta,$$
$$\begin{aligned}\texttt{var}[\alpha] &= \sum_{i=1}^{2k} \frac{(\lambda_i' - \lambda_i)^4 + 4\lambda_i'(\lambda_i' - \lambda_i)^2 + 2{\lambda_i'}^2}{{\lambda_i}^2} \\ &\le \sum_{i=1}^{2k} \frac{3(\lambda_i' - \lambda_i)^4 + 4{\lambda_i'}^2}{{\lambda_i}^2}.\end{aligned}$$

Uniform case: If $\overline{\nu}$ is distributed according to u, then $E[\alpha] = 0$ and $\texttt{var}[\alpha] = 4k$. Therefore, by Chebyshev's Inequality, $P_e^{t_c}(u, n') \le \frac{1}{3}$.

Non-uniform case: Observe that all of the λ_i's are larger than $1000\sqrt{k}/64$. Hence by Chebyshev's Inequality, if $\exists i$ s.t. $(\lambda_i' - \lambda_i)^2/\lambda_i \ge 50\Delta\sqrt{k}$, then $\frac{(\nu_i - \lambda_i)^2 - \nu_i}{\lambda_i} \ge 5\sqrt{k}$ w.p. at least $2/3$. Otherwise it can be shown that, $\texttt{var}[\alpha] \le \frac{3}{10}(E[\alpha])^2 + 8k$ and by Chebyshev's Inequality the error probability is $< \frac{1}{3}$. □

Next we show that if the number of samples is small, then t_c errs on samples of f with probability $> \frac{2}{3}$.

Theorem 10. *For sufficiently large n, if $n' = o(n\sqrt{\log n})$ then $P_e^{t_c}(f, n') \ge \frac{2}{3}$.*

Proof Sketch. By Lemma 8, the χ^2 distance between f and u is $o(\sqrt{\log n'})$. Suppose the underlying distribution is f. The test t_c compares a variant of χ^2 quantity, α, to the threshold $\Theta(\sqrt{\log n'})$. Similar to the proof of Theorem 9, it can be shown that α concentrates around its mean, which is smaller than the threshold. Hence, the test t_c declares `uni` with probability at least $\frac{2}{3}$. □

5 LOWER BOUND

Similar to the previous section, we define bins whose sizes increase exponentially and construct a family $\mathcal{C}^n$ of monotone decreasing distributions that are flat within each bin. We show that for any $f \in \mathcal{C}^n$, $N^*(f) \le n$, while no single test can distinguish u from all $f \in \mathcal{C}^n$ with $o(n\sqrt{\log n})$ samples.

Let $m = \left\lfloor \sqrt{\log_4 \frac{n}{4}} \right\rfloor$, for simplicity assume m is even, and $k = m^2$. Define,

$$\mathcal{I}_i \stackrel{\text{def}}{=} \begin{cases} [\frac{4^{i-1}-1}{3n}, \frac{4^i-1}{3n}) & i = 1, \ldots, k \\ [\frac{4^k-1}{3n}, 1) & i = k+1. \end{cases}$$

Then, $|\mathcal{I}_i| = \frac{4^{i-1}}{n}$ for $i = 1, \ldots, k$. Let $I(x)$ be the index of the bin containing x.

Let $\mathcal{S} \stackrel{\text{def}}{=} \{1, \ldots, \frac{m}{2}\} \times \{m+1, \ldots, \frac{3m}{2}\} \times \ldots \times \{k - m + 1, \ldots, k - \frac{m}{2}\}$. Then, $|\mathcal{S}| = \left(\frac{m}{2}\right)^m$. For any $\overline{j} = (j_1, \ldots, j_m) \in \mathcal{S}$,

$$f_{\overline{j}}(x) \stackrel{\text{def}}{=} \begin{cases} 1 + \sum_{r: j_r \ge I(x)} \frac{4}{\sqrt{m} 2^{j_r - 1}} & x \notin I_{k+1} \\ 1 - \sum_{r=1}^m \frac{4}{\sqrt{m} 2^{j_r - 1}} \frac{4^{j_r} - 1}{3n - (4^k - 1)} & x \in I_{k+1}. \end{cases}$$

The distribution $f_{\overline{j}}$ is induced from u, by removing some probability mass from I_{k+1} and spreading it across $I_1 \ldots, I_{j_r}$ for each j_r. The amount of probability mass shifted is proportional to the standard deviation of the number of samples appearing in I_{j_r}. This ensures that $f_{\overline{j}}$ is monotone and has exactly $m+1$ jumps. Let

$$\mathcal{C}^n \stackrel{\text{def}}{=} \{f_{\overline{j}} : \overline{j} \in \mathcal{S}\}. \tag{4}$$

Thus, $|\mathcal{C}^n| = |\mathcal{S}| = \left(\frac{m}{2}\right)^m$. First, we show that $N^*(f_{\overline{j}}) \le n$.

Lemma 11. $\forall f_{\overline{j}} \in \mathcal{C}^n$,

$$N^*(f_{\overline{j}}) \le n.$$

Proof. For ν's and λ's defined in the previous section, let

$$\beta \stackrel{\text{def}}{=} \sum_{i=1}^m \frac{\nu_{j_i} - \lambda_{j_i}}{\sqrt{\lambda_{j_i}}}.$$

We show that β concentrates around different values for u and $f_{\overline{j}}$, and use that to prove that they can be distinguished with $\le n$ samples.

When u is sampled $\text{poi}(n)$ times,

$$E[\beta] = 0$$
$$\texttt{var}[\beta] = \sum_{i=1}^m \frac{n|\mathcal{I}_{j_i}|}{n|\mathcal{I}_{j_i}|} = m,$$

whereas if $f_{\overline{j}}$ is sampled,

$$E[\beta] = \sum_{i=1}^m \frac{\lambda_{j_i} \sum_{k=i}^m \frac{4}{\sqrt{m} 2^{j_k - 1}}}{\sqrt{\lambda_{j_i}}} \ge \sum_{i=1}^m \frac{4}{\sqrt{m}} = 4\sqrt{m}$$
$$\texttt{var}[\beta] = \sum_{i=1}^m \left(1 + \sum_{k=i}^m \frac{4}{\sqrt{m} 2^{j_k - 1}}\right) < 1.6m,$$

where the last inequality holds for $m \geq 4$. Using Chebyshev's Inequality,

$$u: \qquad P(\beta \geq \sqrt{3m}) \leq \frac{m}{3m} = \frac{1}{3},$$
$$f_{\overline{j}}: \qquad P(\beta \leq \sqrt{3m}) \leq \frac{1.6m}{(4-\sqrt{3})^2 m} \leq \frac{1}{3},$$

hence, $\forall \overline{j} \in \mathcal{S}, N^*(f_{\overline{j}}) \leq n$. □

The following lemma helps to lower bound the number of samples necessary to distinguish u from all distributions in $\mathcal{C}^n$.

Lemma 12. *For positive $x_1, \ldots, x_n$,*

$$\sum_{i=1}^{n} e^{x_i} \leq n - 1 + \exp(\sum_{i=1}^{n} x_i).$$

The next theorem lower bounds the number of samples needed to distinguish all distributions in $\mathcal{C}^n$ from u.

Theorem 13. *For any test t, $\exists f_{\overline{j}} \in \mathcal{C}^n$ such that,*

$$N^t(f_{\overline{j}}) = \Omega\left(N^*(f_{\overline{j}})\sqrt{\log N^*(f_{\overline{j}})}\right).$$

Proof. We find a lower bound on n' such that $\text{poi}(n')$ samples are necessary to distinguish all $f_{\overline{j}} \in \mathcal{C}^n$ from u with error $\leq \frac{1}{3}$. Consider the average function

$$f_{\text{ave}}(\overline{x}) = \frac{1}{\left(\frac{m}{2}\right)^m} \sum_{\overline{j} \in \mathcal{S}} f_{\overline{j}}(\overline{x}),$$

and let

$$\lambda_{r\overline{j}} = \lambda_r + \lambda_r \sum_{l: j_l \geq r} \frac{4}{\sqrt{m} 2^{j_l - 1}}.$$

Next we relate the maximum error of all distributions contained in $\mathcal{C}^n$, to the error of their mixture.

$$\begin{aligned}\frac{1}{3} &\geq \min_t \max_{f \in \mathcal{C}^n} \max(P_e^t(f, n'), P_e^t(u, n')) \\ &\geq \min_t \max(P_e^t(f_{\text{ave}}, n'), P_e^t(u, n')) \\ &\geq \frac{1}{2} - \frac{|f_{\text{ave}} - u|_{1,n'}}{4},\end{aligned}$$

where the last inequality follows from Equation (1). By Lemma 6,

$$\frac{4}{9} \leq (|f_{\text{ave}} - u|_{1,n'})^2 \leq E_{f_{\text{ave}}}\left[\frac{f_{\text{ave}}(\overline{x})}{u(\overline{x})}\right] - 1.$$

After moving the constants to left hand side we have,

$$\begin{aligned}&\frac{13}{9} \leq E_{f_{\text{ave}}}\left[\frac{f_{\text{ave}}(\overline{x})}{u(\overline{x})}\right] \\ &\overset{(a)}{=} \frac{1}{|\mathcal{S}|^2} \sum_{\overline{\nu}} \prod_{i=1}^{k+1} \frac{e^{-\lambda_i} {\lambda_i}^{\nu_i}}{\nu_i!} \left(\sum_{\overline{j} \in \mathcal{S}} \prod_{r=1}^{k+1} e^{\lambda_r - \lambda_{r\overline{j}}} \left(\frac{\lambda_{r\overline{j}}}{\lambda_r}\right)^{\nu_r}\right)^2 \\ &\overset{(b)}{=} \frac{1}{|\mathcal{S}|^2} \sum_{\overline{j} \in \mathcal{S}} \sum_{\overline{k} \in \mathcal{S}} \sum_{\overline{\nu}} \prod_{i=1}^{k+1} \frac{e^{-\lambda_i} {\lambda_i}^{\nu_i}}{\nu_i!} \prod_{r=1}^{k+1} \left(\frac{\lambda_{r\overline{j}} \lambda_{r\overline{k}}}{\lambda_r^2}\right)^{\nu_r} \\ &\overset{(c)}{\leq} \frac{1}{|\mathcal{S}|^2} \sum_{\overline{j} \in \mathcal{S}} \sum_{\overline{k} \in \mathcal{S}} \exp\left(\frac{16n'}{mn} \sum_{i_1=1}^{m} \sum_{i_2=1}^{m} 2^{-|j_{i_1} - k_{i_2}|}\right) \\ &\overset{(d)}{\leq} \frac{1}{|\mathcal{S}|^2} \sum_{\overline{j} \in \mathcal{S}} \sum_{\overline{k} \in \mathcal{S}} \exp\left(\frac{27n'}{mn} \sum_{i=1}^{m} 2^{-|j_i - k_i|}\right) \\ &\overset{(e)}{=} \frac{1}{|\mathcal{S}|^2} \prod_{i=1}^{m} \left(\sum_{j=im-m+1}^{im-m/2} \sum_{k=im-m+1}^{im-m/2} \exp\left(\frac{27n'}{mn} 2^{-|j-k|}\right)\right) \\ &\overset{(f)}{\leq} \frac{1}{|\mathcal{S}|^2} \left(\frac{m}{2} \sum_{j=1}^{m/2} \exp\left(\frac{27n'}{mn} 2^{-|j - \lfloor \frac{m}{4} \rfloor|}\right)\right)^m \\ &= \frac{1}{|\mathcal{S}|} \left(\sum_{j=1}^{m/2} \exp\left(\frac{27n'}{mn} 2^{-|j - \lfloor \frac{m}{4} \rfloor|}\right)\right)^m \\ &\overset{(g)}{\leq} \frac{1}{|\mathcal{S}|} \left(m/2 - 1 + \exp\left(\frac{81n'}{mn}\right)\right)^m \\ &\leq \exp\left(2\left(\exp(\frac{81n'}{mn}) - 1\right)\right),\end{aligned}$$

where (a) follows since all $f_{\overline{j}}$'s and u are constant within each $\mathcal{I}_i$ and the number of samples within each bin is a sufficient statistic for $f_{\text{ave}}(\overline{x})$ and $u(\overline{x})$, (b) follows from $\prod_{r=1}^{k+1} e^{\lambda_r - \lambda_{r\overline{j}}} = 1$, (c) by substituting $\lambda_{r\overline{j}}$ and $\lambda_{r\overline{k}}$, (d) since $i_2 = i_1$ is the dominant term and others are exponentially decreasing with ratio at most 1/4, (e) from rewriting the sum of products as product of sums, (f) from replacing each term by the maximum value, which occurs for $k = (i-1)m + \lfloor \frac{m}{4} \rfloor$, and (g) from Lemma 12.

Simplifying, we have for any test t, the number of samples necessary is $\Omega(nm)$ where $m = \left\lceil \sqrt{\log_4 \frac{n}{4}} \right\rceil$. □

References

J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, and S. Pan. Competitive closeness testing. *JMLR - Proceedings Track*, 19:47–68, 2011.

J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, S. Pan, and A. T. Suresh. Competitive classification and closeness testing. *JMLR - Proceedings Track*, 23:22.1–22.18, 2012.

M. Adamaszek, A. Czumaj, and C. Sohler. Testing monotone continuous distributions on high-dimensional real cubes. In *SODA*, pages 56–65, 2010.

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 Theory of computing*, STOC '04, pages 381–390, New York, NY, USA, 2004. ACM. ISBN 1-58113-852-0.

C. Daskalakis, I. Diakonikolas, R. A. Servedio, G. Valiant, and P. Valiant. Testing k-modal distributions: Optimal algorithms via reductions. *CoRR*, abs/1112.5659, 2011.

M. Mitzenmacher and E. Upfal. *Probability and computing - randomized algorithms and probabilistic analysis*. Cambridge University Press, 2005. ISBN 978-0-521-83540-4.

L. Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10), 2008.

R. Rubinfeld and R. A. Servedio. Testing monotone high-dimensional distributions. *Random Struct. Algorithms*, 34(1):24–44, January 2009.

M. Woodroofe and J. Sun. Testing uniformity versus a monotone density. *The Annals of Statistics*, 27(1):pp. 338–360, 1999.