Adaptive Binning Coincidence Test for Uniformity Testing

Sudeep Salgia, Xinyi Wang, Qing Zhao, *Fellow, IEEE*, and Lang Tong, *Fellow, IEEE*

***Abstract*—We consider the problem of uniformity testing of Lipschitz continuous distributions with bounded support. The alternative hypothesis is a composite set of Lipschitz continuous distributions whose ℓ_1 distances from the uniform distribution are bounded by ε from below. We propose a sequential test that adapts to the unknown distribution under the alternative hypothesis. Referred to as the Adaptive Binning Coincidence (ABC) test, the proposed strategy adapts in two ways. First, it partitions the set of alternative distributions into layers based on their distances to the uniform distribution. It then sequentially eliminates the alternative distributions layer by layer in decreasing distance to the uniform, allowing it to take advantage of favorable situations of a distant alternative by terminating early. Second, it adapts, across layers of the alternative distributions, the resolution level of the discretization for computing the coincidence statistic. The farther away the layer is from the uniform, the coarser the discretization necessary for eliminating this layer or terminating altogether. It thus terminates the test both *early* (via the layered partition of the alternative set) and *quickly* (via adaptive discretization) to take advantage of favorable alternative distributions. The ABC test builds on an adaptive sequential test for discrete distributions, which is of independent interest.**

***Index Terms*—Uniformity testing, adaptivity, coincidence test.**

I. Introduction

CONSIDER the following composite hypothesis testing problem: Given samples of a random variable with density function f, we aim to determine whether f is the uniform distribution u over $[0, 1]$ (the null hypothesis) or f is at least ε distance away (in ℓ_1) from the uniform distribution. The objective is to minimize the sample complexity subject to the constraint of both the Type I and Type II errors being capped below a prescribed value $\delta \in (0, 1)$.

It turns out that without imposing structural constraints on the set of alternative distributions, the above hypothesis testing problem is not testable: no algorithm can achieve diminishing error probability as the number of samples increases [1], [2]. One class of distributions that are testable is the class of monotone distributions [2], [3], for which it has been shown that the sample complexity is of the order $O(1/\varepsilon^2)$. General conditions on testability remain unknown, and uniformity testing of continuous distributions has not been well explored.

In contrast, there is an extensive literature on uniformity testing of discrete distributions with a support size m. The problem dates back to the so-called empty-box problem first posed by David in [4] and later generalized by Viktorova and Chistyakov in [5]. David cast the problem as throwing balls in m boxes and proposed to use the number of empty boxes or the number of boxes containing exactly one ball (a.k.a the coincidence number) as test statistics for uniformity testing[1]. More in-depth studies of using the coincidence test statistic (i.e., the number of letters in the distribution alphabet that see exactly one sample) for uniformity testing were given by Paninski in [6] and Huang and Meyn [7]. Specifically, Paninski showed that the sample complexity of the coincidence test is of $O(\sqrt{m}/\varepsilon^2)$ under the condition that $\varepsilon = \Omega(m^{-1/4})$. Paninski also established that the sample complexity of the coincidence test is order optimal by providing a matching lower bound. Other test statistics used for uniformity testing include empirical ℓ_2 distance [8], [9], [10], [11], empirical ℓ_1 distance [12], and modified χ^2-statistic [13].

Most existing algorithms for uniformity testing are batch algorithms in the sense that all samples are collected prior to performing the test, which makes it necessary to focus on the most challenging alternative distributions (i.e., those that are at the minimum distance ε away from the uniform). While such approaches are sufficient to obtain minimax-optimal sample complexity, they result in significantly suboptimal sample complexity for almost all instances in the class of alternate distributions. Such suboptimality may have significant implications in practice. Consider, for example, the application of anomaly detection in critical infrastructure such as the power grid (see Sec. IV for an experimental study using real datasets collected from a power system). In such cases, the null hypothesis corresponds to a known, nominal distribution[2] and it is safe for

Manuscript received 17 August 2023; revised 16 February 2024; accepted 20 April 2024. Date of publication 7 May 2024; date of current version 21 May 2024. This work was supported in part by the National Science Foundation under Award 1932501. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Florian Meyer. *(Corresponding author: Sudeep Salgia.)*

The authors are with the School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14850 USA (e-mail: ss3827@cornell.edu; xw555@cornell.edu; qz16@cornell.edu; lt35@cornell.edu).

Digital Object Identifier 10.1109/TSP.2024.3397560

[1]The work by [4] considered the problem of testing whether samples are drawn i.i.d. from a known continuous distribution. The problem was reduced to a discrete problem by quantizing samples into m bins without discussing the choice of m. The heuristic approaches proposed by the author are for solving the discrete problem.

[2]In most applications, the probabilistic model of the normal state is known, which can be transformed to a uniform distribution. Anomaly detection can then be cast as uniformity testing.

1053-587X © 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Cornell University Library. Downloaded on July 29,2024 at 21:49:11 UTC from IEEE Xplore. Restrictions apply.

the system to operate within a prescribed distance from this nominal distribution. The alternative hypothesis corresponds to anomalous operational conditions that have exceeded the prescribed level of deviation. For such applications, it is crucial to detect quickly those severe anomalies far away from the normal state, as more severe anomalies often carry more risk and may lead to cascading failure. It is thus highly desirable to have a test that adapts to the underlying alternative distributions and reacts faster to more severe anomalous conditions. This motivates the sequential and adaptive tests developed in this work.

A. Main Results

We consider uniformity testing of Lipschitz continuous (density) distributions, which is arguably more general than the class of monotone distributions studied in [2] and [3]. Another theme that separates this work from existing literature is the sequential aspect of the proposed tests that adapt to the underlying alternative distribution.

Referred to as the Adaptive Binning Coincidence (ABC) test, the proposed strategy adapts to the unknown alternative distribution in two ways. First, it partitions the set of alternative distributions into layers based on their distances to the uniform distribution. It then sequentially eliminates the alternative distributions layer by layer in decreasing distance to the uniform, and subsequently takes advantage of favorable situations of a distant alternative by exiting early. Second, it adapts, across layers of the alternative distributions, the resolution level of the discretization for computing the coincidence statistic. The farther away the layer is from the uniform, the coarser the discretization is needed for eliminating/exiting this layer. It thus exits both *early* in the detection process and *quickly* by using a lower resolution to take advantage of favorable alternative distributions. We establish the sample complexity of the ABC test, which characterizes its adaptivity.

The ABC test builds on an adaptive sequential coincidence test for discrete distributions, which is of independent interest. Due to the adaptivity, this sequential test improves the sample complexity under the alternative hypothesis from $O(\sqrt{m}/\varepsilon^2)$ of Paninski's batch algorithm to $O(\gamma^{-2}\sqrt{m}\log(1/\gamma))$, where γ is the distance of the underlying alternative distribution to the uniform and is greater than the minimum distance ε. This demonstrates that the sequential coincidence test adapts to the *realized* distance γ in an optimal order (up to a logarithmic factor) in terms of sample complexity.

B. Related Work

The problem of estimating and testing properties of an unknown distribution has an extensive literature (see an excellent survey in [14]). The problem of uniformity testing is among the most widely studied problems among this class. As discussed earlier, most existing work focuses on discrete distributions and batch algorithms (see [6], [7], [13]). A notable exception to these batch-based strategies is work by Batu and Canonne in [15]. The test developed in [15] uses 2-way and 3-way collisions among the set of samples as test statistics, which in expectation correspond to the ℓ_2 and ℓ_3 norms of the distribution. The sample complexity of this algorithm, while having a distribution dependent term of ℓ_3 norm of the distribution, maintains a term determined by the worst distance to the alternative set, ε. In other words, the adaptivity to the underlying hypothesis is only *partial*, and the second term in terms of the worst distance ε may dominate. The test strategies proposed here, however, are *fully* adaptive as their sample complexities depend only on the ℓ_1 distance of the underlying distribution p to the uniform distribution, with no dependence on ε. In a work concurrent to the first posting of this work [16], Fawzi et al. [17] also proposed a sequential extension of the batch-based algorithm for testing closeness of distributions. Their results also establish the benefit of adaptivity obtained by sequential testing as opposed to the batch based approach. However, their work is limited to testing closeness of discrete distributions while our focus is on continuous distributions.

The literature on uniformity testing of continuous distributions is slim. As mentioned previously, it is necessary to impose a certain structure on the class of distributions being considered in order to ensure feasibility of the problem. For the case when the underlying distribution is monotone, Adamaszek et al. [2] and Acharya et al. [3] have proposed algorithms that offer optimal sample complexities using sample mean and modified χ^2 test as test statistics respectively. Diakonikolas et al. [18] also studied identity testing of a distribution under various structural assumptions. Ba et al. [19] approach the problem of closeness testing through the estimation of the Earth Mover distance. In [20], Ingster studied uniformity testing against an alternative class of smooth densities, which includes the Lipschitz continuous distributions studied here. The key differences are that the algorithm in [20] uses the χ^2 test statistic and does not offer adaptivity to the underlying distribution. The focus of this work is to develop fully adaptive test strategies employing the much simpler test statistic of coincidence number.

II. Uniformity Testing of Discrete Distributions

In this section, we consider uniformity testing of discrete distributions and develop a sequential test employing the coincidence statistic. The results obtained for the discrete problem form the foundation for tackling the continuous problem in the next section.

The key property of this sequential coincidence test (SCT) is that it adapts to the unknown alternative distribution in the composite set. More specifically, the sample complexity of SCT under the alternative hypothesis scales optimally with respect to the distance γ of the realized alternative distribution to the uniform. This is in sharp contrast to batch algorithms whose sample complexity is determined by the worst-case alternative distribution seeing the minimum distance ε to the uniform.

A. Problem Formulation

Consider a binary hypothesis testing problem where the null hypothesis H_0 is the uniform distribution u with a support size of m. Without loss of generality, we assume that the support set is $\{1, 2, \ldots, m\}$ denoted by $[m]$. The alternative hypothesis H_1 is composite: it consists of all distributions over $[m]$ whose

Authorized licensed use limited to: Cornell University Library. Downloaded on July 29,2024 at 21:49:11 UTC from IEEE Xplore. Restrictions apply.

ℓ_1 distance to u is no smaller than ε. More specifically, let $\mathcal{C}(\varepsilon)$ denote the composite set of alternative distributions under H_1, we have

$$\mathcal{C}(\varepsilon) = \{q \in \mathcal{P}([m]) : \|q - u\|_1 > \varepsilon\},$$

where $\mathcal{P}([m])$ denotes the set of all probability distributions over $[m]$, $\|q-u\|_1 = \sum_{i=1}^{m} |q_i - 1/m|$ is the ℓ_1 distance between distribution q and the uniform distribution u.

For the hypothesis testing problems, i.i.d. samples are drawn from either u (if H_0 is true) or a specific distribution in $\mathcal{C}(\varepsilon)$ (if H_1 is true), unknown to the decision maker. The goal is to determine, based on the random samples, which hypothesis is true. The probabilities of false alarm and miss detection need to be capped below a given δ ($\delta \in (0,1)$) for all alternative distributions in $\mathcal{C}(\varepsilon)$.

B. Sequential Coincidence Test

Existing work on uniformity testing all focuses on batch methods (a.k.a. the fixed-sample-size tests). Specifically, based on the reliability constraint δ and the minimum separation ε between H_0 and H_1, the number of required samples is predetermined to ensure δ-reliability in the worst case of a closest (i.e., distance ε) alternative.

We propose a sequential test SCT that adapts to the unknown alternative. When the alternative is at a distance greater than ε from u, the sequential test takes advantage of the favorable situation and exits towards H_1 with fewer samples. In particular, the sample complexity of SCT scales optimally with the *realized* distance γ rather than the minimum distance ε.

SCT employs the coincidence statistic. For a given set of samples $\mathcal{S}$, the coincidence $K_1(\mathcal{S})$ is the number of symbols in $[m]$ that appear exactly once in $\mathcal{S}$. Specifically, let n_j denote the number of appearances of symbol j in $\mathcal{S}$. Then

$$K_1(\mathcal{S}) = \sum_{j=1}^{m} \mathbb{1}\{n_j = 1\},$$

where $\mathbb{1}\{\cdot\}$ is the indicator function. Let $K_1(n)$ denote the coincidence number of n i.i.d. samples drawn from a given distribution p. It is a random variable whose distribution is determined by n and p. We then introduce the constant $c_u(n)$, which is the expected value of $K_1(n)$ under the uniform distribution:

$$c_u(n) = \mathbb{E}_u[K_1(\mathcal{S})], \quad \text{where } |\mathcal{S}| = n,\ \mathcal{S} \overset{i.i.d.}{\sim} u.$$

The intuition behind the coincidence test statistic is as follows. Since all the symbols are equally likely under the uniform distribution, a set of n i.i.d. samples from a uniform distribution results in fewer repetitions of the observed symbols resulting in a larger value of $K_1(n)$. On the other hand, for a distribution far away from the uniform distribution, some symbols have higher probability than others resulting in multiple observations of such symbols and smaller values of $K_1(n)$. Thus, by comparing the value of $K_1(n)$ to a carefully designed threshold, one can distinguish the uniform distribution from a distribution in the alternative set.

Algorithm 1 Sequential Coincidence Test (SCT)

Input: $m, \varepsilon, \delta \in (0,1)$
Set $k \leftarrow 1$, $t \leftarrow 0$, $\kappa = 112\varepsilon^{-2}$, $\mathcal{S} = \{\}$
while $k \le \kappa$ **do**
  $n_k \leftarrow \left\lceil k\sqrt{m\log(k+2/\delta)} \right\rceil$
  $\tau_k \leftarrow 7n_k\sqrt{\dfrac{\log(k+2/\delta)}{m}}$
  repeat
    Obtain a sample X_t
    $\mathcal{S} \leftarrow \mathcal{S} \cup X_t$
    $t \leftarrow t+1$
  until $t == n_k$
  if $Z_k := c_u(n_k) - K_1(\mathcal{S}) > \tau_k$ **then**
    Output $\leftarrow H_1$
    break
  end if
  $k \leftarrow k+1$
end while
if $k > \kappa$ **then**
  Output $\leftarrow H_0$
end if
return Output

SCT proceeds in epochs. In each epoch the test takes $\Theta(\sqrt{m})$ additional samples. At the end of each epoch, based on all the samples $\mathcal{S}$ collected so far, the algorithm decides whether there is sufficient evidence to exit towards H_1. This decision is made by comparing the difference between $K_1(\mathcal{S})$ and $c_u(|\mathcal{S}|)$ to a carefully chosen threshold. If the difference exceeds the threshold (indicating a sufficient separation between the coincidence number of the samples $\mathcal{S}$ and the expected coincidence number of the uniform), the algorithm terminates and declares H_1. Otherwise, the algorithm enters the next epoch. In the event that the process reaches the maximum number $\Theta(\varepsilon^{-2})$ of epochs without exiting towards H_1, the algorithm terminates and declares H_0. A pseudo code for the algorithm is given in Algorithm 1.

The sequential detection process of SCT can be visualized as peeling an onion: the core of the onion is the uniform distribution and its ε-neighbors; the layers represent alternative distributions at increasing distance to the uniform distribution[3]. Each epoch peels one layer of the onion, either by exiting towards H_1 (if the realized alternative distribution resides in this layer) or by eliminating this set of alternative distributions and moving to the next layer closer to the core. If all outer layers are eliminated (i.e., all alternative distributions in $\mathcal{C}(\varepsilon)$ are eliminated), SCT terminates and declares H_0. The ability of peeling the onion layer by layer is rooted in the fact that when the samples $\mathcal{S}$ are drawn from a distribution γ-distance away from u, the difference in coincidence numbers $c_u(n_k) - K_1(\mathcal{S})$

[3]The epoch structure of SCT effectively quantizes the distance to u, hence forming a finite partition of the alternative distributions in $\mathcal{C}(\varepsilon)$. More specifically, $\mathcal{C}(\varepsilon)$ is partitioned into κ layers, where $\kappa = 112\varepsilon^{-2}$ is the maximum epoch number defined in Algorithm 1. Each epoch peels off one layer.

Authorized licensed use limited to: Cornell University Library. Downloaded on July 29,2024 at 21:49:11 UTC from IEEE Xplore. Restrictions apply.

scales proportionally with γ^2. This difference hence exceeds the threshold early when γ is large (i.e., when the alternative distribution resides farther from the core of the onion).

C. Sample Complexity

The expected sample complexity of SCT is characterized in the following theorem.

Theorem 1: For $m > m_0$ and $\varepsilon = \Omega(m^{-1/8})$, where $m_0 = \min\{l > 0 : 1123 l^{1/4} e^{-0.25\sqrt{l}} \leq \delta\varepsilon^2\}$, we have

- Under H_1 with an alternative distribution p that is γ away from u, the expected sample complexity of SCT is $O\left((\sqrt{m}/\gamma^2)\sqrt{\log(1/\gamma + 1/\delta)}\right)$
- Under H_0, the expected sample complexity of SCT is $O\left((\sqrt{m}/\varepsilon^2)\sqrt{\log(1/\varepsilon + 1/\delta)}\right)$
- The probabilities of miss detection and false alarm are at most δ.

As evident from the above theorem, the sample complexity of SCT under H_1 adapts to the distance γ of the alternative distribution p to the uniform distribution. Since $p \in \mathcal{C}(\varepsilon)$, we have $\gamma > \varepsilon$, implying that the sample complexity is smaller than the fixed-sample-size approaches which offer a sample complexity of $O(\varepsilon^{-2}\sqrt{m})$. Moreover, the adaptivity of SCT to the realized distance γ is order-optimal (up to a logarithmic factor). This can be shown by noting that even with the knowledge of γ, the lower bound given by Paninski dictates $\Omega(\gamma^{-2}\sqrt{m})$ samples to ensure a constant probability of reliability.

Furthermore, in addition to the near-optimal scaling with γ, SCT offers better scaling of sample complexity with respect to δ, the error probability, as compared to the batch algorithms. In particular, the batch algorithms are designed to guarantee a certain constant probability of error, and the common technique to extend such tests to guarantee an arbitrary probability of error δ is to repeat the test sufficiently many times so that the result declared by the majority vote meets the confidence requirements. Such an approach results in a $\log(1/\delta)$ dependence of sample complexity on δ as opposed to the $\sqrt{\log(1/\delta)}$ offered by SCT. Thus, the refined analysis required to analyze the sequential coincidence test not only demonstrates adaptivity to the underlying distribution but also results in improved dependence on δ.

In addition to the dependence on γ and δ, we would also like to briefly mention the regime of input parameters m and ε for which this result holds. The lower bound m_0 is required to ensure the support size is large enough to ensure a confidence of δ in the sparse regime. Moreover, the regime of ε for which this result is applicable can also in part be attributed to the fact that the coincidence statistic works well only in the sparse regime. We believe that the $\varepsilon = \Omega(m^{-1/8})$ requirement as opposed to the standard requirement of $\varepsilon = \Omega(m^{-1/4})$ for sparse regime is merely an analysis artifact and can be improved using better techniques for bounding the moment generating function of K_1.

Proof of Theorem 1: The central idea of this proof is to establish bounds on the moment generating function of the K_1 statistic. Once we have obtained the bound on MGF, the final result follows directly by an application of Markov's inequality and union bound. Note that K_1 is a sub-Gaussian random variable as it is bounded. In order to obtain tight bounds on the variance proxy, we employ an approach similar to the one used in [7], [21] by appropriately modifying it to obtain bounds for finite sample regime as opposed to for an asymptotic analysis.

Let p denote the underlying distribution and $\mathcal{S}$ be a set of n i.i.d. samples from p. With a slight abuse of notation, we denote $K_1(\mathcal{S})$ as $K_1(n)$. We derive the bounds for a general $n \geq \sqrt{m}$ number of samples, followed by substituting the particular choices later. Lastly, we assume that $n/m \leq \varepsilon^2/1536$[4] and throughout the analysis, θ is a variable lies in the range $[-0.4, 0.4]$.

The following lemma establishes bounds on the moment generating function of K_1.

Lemma 1: Let $p = \{p_1, p_2, \ldots, p_m\} \in \mathcal{P}([m])$ and $K_1(n)$ be as defined above. Then for $\theta \in [-0.4, 0.4]$, we have,

$$\mathbb{E}_p\left[\exp\left(\theta\tilde{K}_1(n)\right)\right] \leq \frac{n!}{2\pi n^n}\left(\frac{ne^{-\theta}}{\lambda_0}\right)^n \times \left(e^{H_p}\sqrt{\frac{\pi}{0.1\lambda_0}} + \pi e^{\lambda_0|(e^\theta - 1)|}\right),$$

where $\tilde{K}_1(n) = K_1(n) - n$, λ_0 is the solution to the equation $\frac{\lambda p_j(e^\theta - 1 + e^{\lambda p_j})}{\lambda p_j(e^\theta - 1) + e^{\lambda p_j}} = n$, and $H_p = \sum_{j=1}^m \log\left(\lambda_0 p_j(e^\theta - 1) + e^{\lambda_0 p_j}\right)$.

The above lemma forms a key step in the proof as it gives the required bound the moment generating function of K_1, which is used to obtain the bounds on the error probabilities. The proof of this lemma is based on a finite sample analysis of the following expression, which is adopted from [7, Eqn. (38)].

$$\mathbb{E}_p\left[\exp\left(\theta\tilde{K}_1(n)\right)\right] = e^{-\theta n}\frac{n!}{2\pi i}\oint g(\lambda)d\lambda, \tag{1}$$

where,

$$g(\lambda) = e^\lambda \prod_{j=1}^m \left(1 - \lambda p_j e^{-\lambda p_j} + \lambda p_j e^{-\lambda p_j} e^\theta\right)\frac{1}{\lambda^{n+1}}.$$

The integral in Eqn. (1) is estimated using the saddle point method [22]. This approach consists of two steps. In the first step, a particular contour around $\lambda = 0$ is chosen to carry out the integration. The choice of this contour is such that $g(\lambda)$ behaves violently along it, i.e., $g(\lambda)$ is very large for a very small interval and significantly smaller on the rest of the contour, similar to a dirac delta function. This violent behaviour allows one to approximate the integral by only evaluating it on the small interval where the function value is very large. Such a contour is usually obtained by identifying a saddle point of $g(\lambda)$, a point where the derivative of g goes to zero and choosing to contour to pass through this point. The second step involves estimating the integral along the contour using the Laplace method. We refer the reader to Appendix D for a detailed proof of Lemma 1.

[4]The additional dependence on ε is a result of the particular technique being employed to bound the MGF. We conjecture that requirement can be relaxed by using better analysis tools and techniques.

Authorized licensed use limited to: Cornell University Library. Downloaded on July 29,2024 at 21:49:11 UTC from IEEE Xplore. Restrictions apply.

Using this bound on the MGF, we obtain the following bounds on the probability of error at the decision instant for any epoch k:

Lemma 2: Let $P_e(n_k)$ denote the probability of error at the decision instant for epoch k of SCT. Then for the choice of parameters described in Alg. 1, we have,

- Under H_0,

$$P_e(n_k) \le e^{1/12n_k}\left(\frac{3}{(k+2/\delta)^3}+\sqrt{\frac{n_k\pi}{2}}e^{-0.25n_k}\right).$$

- Under H_1 *for* $k \ge 112\gamma^{-2}$,

$$P_e(n_k) \le e^{1/12n_k}\left(\frac{3}{(k+2/\delta)^{2.5}}+\sqrt{\frac{n_k\pi}{2}}e^{-0.25n_k}\right).$$

Lemma 2 allows us to obtain the required bounds on probability of error and sample complexity. Note that the probability of miss detection and false alarm can be obtained by summing the corresponding bounds from Lemma 2 over the range of k. In particular, for the probability of false alarm, denoted by $\Pr(\text{err}|H_0)$, we have,

$$\begin{aligned}\Pr(\text{err}|H_0) &= \Pr\left(\bigcup_{k=1}^{\kappa}\{\mathbb{E}_u[K_1(n_k)]-K_1(n_k)>\tau_k\}\right)\\ &\le \sum_{k=1}^{\kappa}\Pr\left(K_1^{(n_k)}-\mathbb{E}_u[K_1^{(n_k)}]<-\tau_k\right)\le\sum_{k=1}^{\kappa}P_e(n_k)\\ &\le \sum_{k=1}^{\kappa}e^{1/12n_k}\left(\frac{3}{(k+2/\delta)^3}+\sqrt{\frac{n_k\pi}{2}}e^{-0.25n_k}\right)\\ &\le \left(1+\frac{1}{6\sqrt{m}}\right)\sum_{k=1}^{\kappa}\left(\frac{3}{(k+2/\delta)^3}+\sqrt{\frac{\sqrt{m}\pi}{2}}e^{-0.25\sqrt{m}}\right)\\ &\le \left(1+\frac{1}{6\sqrt{m}}\right)\left(3\left(\frac{\delta}{2}\right)^2+\sqrt{\frac{\sqrt{m}\pi}{2}}\frac{112e^{-0.25\sqrt{m}}}{\varepsilon^2}\right),\end{aligned}$$

where the fifth line follows by noting that $e^x \le 1+2x$ for $x \le 1$, $\sqrt{x}e^{-x/4}$ is decreasing for $x>2$ and $n_k \ge \sqrt{m}$. On plugging in the lower bound for m, we obtain that the above expression is less than δ, as required. Similarly, the probability of miss detection, denoted by $\Pr(\text{err}|H_1)$, can be bounded as

$$\begin{aligned}\Pr(\text{err}|H_1) &= \Pr\left(\bigcap_{k=1}^{\kappa}\{\mathbb{E}_u[K_1(n_k)]-K_1(n_k)<\tau_k\}\right)\\ &\le P_e(n_\kappa)\le e^{1/12n_\kappa}\left(\frac{3}{(\kappa+2/\delta)^{2.5}}+\sqrt{\frac{n_\kappa\pi}{2}}e^{-0.25n_\kappa}\right)\le\delta.\end{aligned}$$

We now move on to establish the bound on the expected sample complexity of SCT. Let Γ denote the random number of sample taken by the procedure. For the scenario when the underlying distribution is uniform, i.e., H_0, we use a simple upper bound given as

$$\mathbb{E}[\Gamma|H_0] \le n_\kappa \le \frac{112\sqrt{m}}{\varepsilon^2}\sqrt{\log\left(\frac{112}{\varepsilon^2}+\frac{2}{\delta}\right)}+1.$$

For the case when the underlying distribution belongs to $\mathcal{C}(\varepsilon)$ such that $\|p-u\|_1=\gamma>\varepsilon$, the expected sample complexity is given as

$$\mathbb{E}[\Gamma|H_1]=\sum_{k=1}^{\kappa}n_k\Pr(\Gamma=n_k)\le n_{k_0(\gamma)}+\sum_{k=k_0(\gamma)}^{\kappa}n_{k+1}$$

The second term in the expression can be bounded as

$$\begin{aligned}&\sum_{k=k_0(\gamma)}^{\kappa}n_{k+1}P_e(n_k)\\ &\le \sum_{k=k_0(\gamma)}^{\kappa}n_{k+1}e^{1/12n_k}\left(\frac{3}{(k+2/\delta)^{2.5}}+\sqrt{\frac{n_k\pi}{2}}e^{-n_k/4}\right)\\ &\le e^{\frac{1}{12}}\sum_{k=k_0(\gamma)}^{\kappa}\left(\frac{4.5\sqrt{m\log(k+2/\delta)}}{(k+2/\delta)^{1.5}}+\sqrt{\frac{9\pi}{8}}n_k^{1.5}e^{-n_k/4}\right)\\ &\le e^{\frac{1}{12}}\left(27\sqrt{m}\left(k_0(\gamma)-1+\frac{2}{\delta}\right)^{-1/2}\log\left(k_0(\gamma)-1+\frac{2}{\delta}\right)\right.\\ &\qquad\left.+\sqrt{\frac{9\pi}{8}}C_0\delta\sqrt{m}\right)\\ &\le C_1 n_{k_0(\gamma)},\end{aligned}$$

for some universal constants C_0, C_1. Consequently, $\mathbb{E}[\Gamma|H_1]=O\left(\gamma^{-2}\sqrt{m\log\left(\frac{1}{\gamma}+\frac{1}{\delta}\right)}\right)$, as required. □

III. Uniformity Testing of Continuous Distributions

A. Problem Formulation

We now consider uniformity testing of continuous distributions, particularly Lipschitz continuous distributions with bounded support. Specifically, the null hypothesis H_0 is the uniform distribution u over $[0,1]$. The alternative composite hypothesis H_1 is the set of L-Lipschitz distributions whose ℓ_1 distance to u is no smaller than ε. Let $\mathcal{P}([0,1];L)$ denote the set of distributions over $[0,1]$ that are absolutely continuous with the Lebesgue measure on $[0,1]$ and whose density functions are L-Lipschitz. Specifically, for all distributions $q \in \mathcal{P}([0,1];L)$, the density functions $f_q(x)$ satisfies, for all $x, y \in [0,1]$,

$$|f_q(x)-f_q(y)| \le L|x-y|.$$

The composite set of alternative distributions under H_1 is given by

$$\mathcal{C}_L(\varepsilon)=\left\{q\in\mathcal{P}([0,1];L):\int_0^1|f_q(x)-1|\,\mathrm{d}x>\varepsilon\right\}.$$

The objective of the uniformity testing is the same as in the discrete problem: to determine, with a reliability constraint of δ, whether the observed random samples are generated from u (H_0) or from a distribution in $\mathcal{C}_L(\varepsilon)$ (H_1).

B. Adaptive Binning Coincidence Test

We now generalize SCT to the continuous problem specified above. Our goal is to preserve the adaptivity of SCT to the underlying alternative distribution under H_1.

Authorized licensed use limited to: Cornell University Library. Downloaded on July 29,2024 at 21:49:11 UTC from IEEE Xplore. Restrictions apply.

The relation among the set of continuous distributions can still be visualized as an onion: u and its ε-neighbors form the core, and alternative distributions in $\mathcal{C}_L(\varepsilon)$ form layers according to their distances to u. The algorithm still aims to determine, sequentially over epochs, in which layer the underlying distribution of the observed samples resides, starting from the outer-most and moving towards the core. The key question in the continuous problem is how to infer, from a coincidence type of statistic, whether the underlying distribution resides in the current layer. A straightforward answer to this question is discretization: a uniform binning of the support set $[0, 1]$ with coincidence number defined with respect to the bin labels of the random samples. Much less obvious is the choice of the resolution level for the discretization, i.e., how finely to bin the continuum domain. A key rationale behind the proposed ABC (Adaptive Binning Coincidence) test is that the farther away the layer is from the core, the coarser the discretization is needed for inferring whether the underlying distribution resides in this layer. More specifically, not only the test can exit early when the realized distance γ is favorable, but also the number of required samples for making the peeling decision is fewer due to a coarser discretization. In other words, ABC adapts to the unknown realized distance γ by exiting both *early* in the detection process and *quickly* by using a lower resolution to take advantage of favorable alternative distributions.

We can now describe the ABC test, which proceeds in a similar epoch structure as SCT with two key modifications. First, the resolution of the discretization increases at a carefully chosen rate over epochs, and the number of samples taken in each epoch is adjusted accordingly. Specifically, let m_k denote the number of discretization bins in epoch k, where m_k increases at the rate of $k^4 \log k$. The number of samples taken in epoch k is $\Omega(\sqrt{m_k})$, which retains the same squared-root relation to the effective support size m_k as in the discrete case. Second, the coincidence number $K_1(\mathcal{S})$ in each epoch is computed by rebinning all observed samples (including those obtained in previous epochs) based on the refined discretization m_k in the current epoch. A detailed description of the algorithm is given in Algorithm 2, where $K_1(\mathcal{S}; m)$ denotes the coincidence number computed over the set $\mathcal{S}$ of samples when the interval is uniformly divided into m bins. Similarly, $c_u(n; m)$ denotes the expected coincidence number of n samples from the uniform distribution with a support of m bins. The constant[5] $c_0 \geq \max\{28212, m_0, 2L\}$, where m_0 is defined in Theorem 1.

We would like to point out that while we have described ABC for uniform distribution, it can be easily extended to cases where the null hypothesis corresponds to an arbitrary distribution in $\mathcal{P}([0, 1]; L)$. In this case, we can employ a non-uniform binning strategy that bins the null hypothesis into a discrete uniform distribution (See Sec. IV for empirical results). It can be shown that the sample complexity as analyzed next holds for this more general problem provided that the null hypothesis distribution is lower bounded by a constant.

[5]The constant 28212 as a lower bound for c_0 arises from the conditions imposed on n and m during analysis. We would like to point out that these numerical constants are not optimized. The analysis focuses on the order.

Algorithm 2 Adaptive Binning Coincidence (ABC) Test

Input: $\varepsilon, L, \delta \in (0, 1)$
Set $k \leftarrow 1$, $t \leftarrow 0$, $\kappa \leftarrow 576\varepsilon^{-2}$, $\mathcal{S} \leftarrow \{\}$
while $k \leq \kappa$ **do**
  $m_k \leftarrow \lceil c_0 k^4 \log(k + 2/\delta) \rceil$
  $n_k \leftarrow \lceil \sqrt{c_0} k^3 \log(k + 2/\delta) \rceil$
  $\tau_k \leftarrow 9 n_k \sqrt{\dfrac{\log(k + 2/\delta)}{m_k}}$
  repeat
    Obtain a sample X_t
    $\mathcal{S} \leftarrow \mathcal{S} \cup X_t$
    $t \leftarrow t + 1$
  until $t == n_k$
  if $Z_k := c_u(n_k; m_k) - K_1(\mathcal{S}; m_k) > \tau_k$ **then**
    Output $\leftarrow H_1$
    break
  end if
  $k \leftarrow k + 1$
end while
if $k > \kappa$ **then**
  Output $\leftarrow H_0$
end if
return Output

C. Sample Complexity

The theorem below establishes the sample complexity of ABC and its adaptivity to the realized distance γ under H_1.

Theorem 2:

- Under H_1 with an alternative distribution p that is γ away from u, the expected sample complexity of ABC is $O\left(\sqrt{L}\gamma^{-6} \log\left(1/\gamma + 1/\delta\right)\right)$.
- Under H_0, the expected sample complexity of ABC is $O\left(\sqrt{L}\varepsilon^{-6} \log\left(1/\varepsilon + 1/\delta\right)\right)$.
- The probabilities of miss detection and false alarm are at most δ.

The lower bound on the sample complexity for this problem is $\Omega(\sqrt{L}\varepsilon^{-5/2})$ as shown in [20]. Evidently, there is a significant gap between the lower bound on sample complexity and the sample complexity guarantees offered by ABC. This gap, we believe is rooted in that coincidence statistic is informative only in sparse regimes where the number of samples is of sublinear order of the support size. We conjecture that the gap to the lower bound is unavoidable for tests using the coincidence statistic. In other words, we conjecture that while coincidence statistic is sufficient for achieving order-optimal sample complexity in the discrete case, it ceases to remain so in the continuous case. An interesting question is to explore is a lower bound on the sample complexity achievable by the simple test statistic of coincidence number. While the dependence on ε is suboptimal, we would like to point out that ABC achieves optimal scaling with the Lipschitz constant L.

Proof of Theorem 2: The proof of this theorem is heavily built on the proof of Theorem 1. The basic idea is to first show that for a fine enough discretization, the ℓ_1 distance of

Authorized licensed use limited to: Cornell University Library. Downloaded on July 29,2024 at 21:49:11 UTC from IEEE Xplore. Restrictions apply.

resulting discrete distribution is the same as that of the continuous distribution up to a constant factor. Once this relation is established, we can simply invoke the results obtained in the previous theorem to obtain the results. We begin with the following lemma that relates the ℓ_1 distances of the discrete and continuous distributions.

Lemma 3: Let p be a distribution in $\mathcal{C}_L(\varepsilon)$ such that $\|p-u\|_1=\gamma$ and let p^Δ be the discrete distribution obtained by a uniform discretization of the interval $[0,1]$ into m bins. The ℓ_1 distance of p^Δ from the uniform distribution, denoted by $[\gamma]_m$, satisfies the following relation

$$[\gamma]_m=\sum_{i=1}^{m}|p_i^\Delta-1/m|\geq\gamma-L/m,$$

where p_i^Δ denotes the probability mass of p^Δ in the i^{th} bin.

The proof of this lemma can be found in Appendix C.

Once we have obtained a discretization, the proof for the probability of error and sample complexity is almost identical to that in the case of SCT. We have the following lemma for the continuous setting that is the counterpart of Lemma 2 in the discrete setting.

Lemma 4: Let $P_e(n_k)$ denote the probability of error at the decision instant for epoch k of ABC. Then for the choice of parameters described in Alg. 2, we have,

- Under H_0,

$$P_e(n_k)\leq e^{1/12n_k}\left(\frac{3}{(k+2/\delta)^3}+\sqrt{\frac{n_k\pi}{2}}e^{-0.25n_k}\right).$$

- Under H_1 for $k\geq k_0(\gamma)$,

$$P_e(n_k)\leq e^{1/12n_k}\left(\frac{3}{(k+2/\delta)^{4.5}}+\sqrt{\frac{n_k\pi}{2}}e^{-0.25n_k}\right),$$

where $k_0(\gamma):=\min\{k:k\geq 144[\gamma]_{m_k}^{-2}\}$.

Note that the bounds on the probability of error for ABC are smaller than those of SCT for both H_0 and H_1. Consequently, using the same sequence of arguments, we can conclude that $\Pr(\text{err}|H_0)\leq\delta$ and $\Pr(\text{err}|H_1)\leq\delta$.

The expected sample complexity for the uniform case can once again be simply bounded as n_κ implying that $\mathbb{E}[\Gamma|H_0]$ is $O(\sqrt{L}\varepsilon^{-6}\log(\varepsilon^{-1}+\delta^{-1})$, where the $\sqrt{L}$ factor comes from the leading constant in the expression for n_k. The sample complexity for case when the underlying distribution belongs to $\mathcal{C}_L(\varepsilon)$ such that $\|p-u\|_1=\gamma>\varepsilon$ is given as follows:

$$\mathbb{E}[\Gamma|H_1]=\sum_{k=1}^{\kappa}n_k\Pr(\Gamma=n_k)\leq n_{k_0(\gamma)}+\sum_{k=k_0(\gamma)}^{\kappa}n_{k+1}P_e(k).$$

Once again, the second term can be bounded using the sequence of steps similar to that in proof of Theorem 1.

$$\begin{aligned}&\sum_{k=k_0(\gamma)}^{\kappa}n_{k+1}P_e(n_k)\\&\leq\sum_{k=k_0(\gamma)}^{\kappa}n_{k+1}e^{1/12n_k}\left(\frac{3}{(k+2/\delta)^{4.5}}+\sqrt{\frac{n_k\pi}{2}}e^{-n_k/4}\right)\\&\leq e^{1/12}\sum_{k=k_0(\gamma)}^{\kappa}\left(\frac{3\sqrt{c_0}\log(k+2/\delta)}{(k+2/\delta)^{1.5}}+7.5\sqrt{\frac{\pi}{2}}n_k^{1.5}e^{-n_k/4}\right)\\&\leq e^{1/12}\left(18\sqrt{c_0}\left(k_0(\gamma)-1+\frac{2}{\delta}\right)^{-1/2}\right.\\&\qquad\left.\times\log\left(k_0(\gamma)-1+\frac{2}{\delta}\right)+7.5\sqrt{\frac{\pi}{2}}C_0'\delta\sqrt{c_0}\right)\\&\leq C_1'n_{k_0(\gamma)},\end{aligned}$$

for some universal constants C_0', C_1'. As a result, the expected sample complexity is of the order $n_{k_0(\gamma)}$. From the particular choice of m_k and Lemma 3, we can conclude that $k_0(\gamma)\leq 576\gamma^{-2}$. This bound along with the expression for $n_{k_0(\gamma)}$ yields $\mathbb{E}[\Gamma|H_1]=O\left(\sqrt{L}\gamma^{-6}\log\left(\gamma^{-1}+\delta^{-1}\right)\right)$, as required. □

IV. Simulations

We corroborate our theoretical findings by testing the algorithms empirically using both synthetic and real-world datasets. We carry out three different sets of experiments to demonstrate the effectiveness of our proposed approach. In the first set of experiments, we demonstrate the adaptivity of our proposed approach by comparing the performance of our proposed algorithms against several fixed-sample-size algorithms in terms of true positive rates and sample complexities. In the second set of experiments, we compare the performance of ABC against a sequential version of the fixed-sample-size algorithms obtained using the "doubling search technique". Lastly, we conduct experiments where the null hypothesis is a non-uniform distribution to demonstrate the applicability of our process to scenarios where the null distribution is non-uniform.

We compare our proposed algorithms with several representative fixed-sample-size tests, including the fixed-sample-size coincidence test in [6], the fixed-sample-size chi-squared test in [13], the high probability identity test in [12], and the closeness test between two distributions proposed in [11]. These existing algorithms are referred to in the legends as coincidence test, chi-squared test, identity test and closeness test respectively.

A. Adaptivity of ABC

For the first set of experiments, we consider two synthetic and one real-world dataset. We first consider the discrete problem as described in Sec. II with $m=20000$ and $\varepsilon=0.3$. The composite set of distributions $\mathcal{C}(\varepsilon)$ under the alternative hypothesis is parameterized by γ with the probability mass functions given by

$$p_{2i-1}^\gamma=(1+\gamma)/m;\ p_{2i}^\gamma=(1-\gamma)/m,$$

for $i=\{1,2,\ldots,m/2\}$. Note that the ℓ_1 distance of p^γ from the uniform distribution is γ. We consider 7 distributions corresponding to the values of $\gamma\in\{0.3,0.35,0.4,0.45,0.5,0.55,0.6\}$. For all algorithms, the thresholds are set to obtain an false positive rate bounded below 0.2. The constants in the thresholds for all the algorithms are optimized using grid search to give the best empirical sample complexity. Fig. 1 shows the empirical sample complexities and true positive rates obtained under different distributions. For SCT, the expected sample complexity is obtained by taking an average over 1000 Monte Carlo runs. As expected, the sample

Authorized licensed use limited to: Cornell University Library. Downloaded on July 29,2024 at 21:49:11 UTC from IEEE Xplore. Restrictions apply.

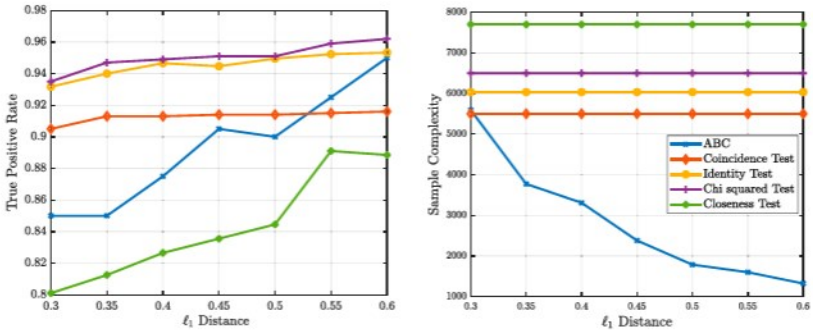

Fig. 1. True positive rate (left) and empirical sample complexity (right) vs ℓ_1 distance for discrete alternative distributions.

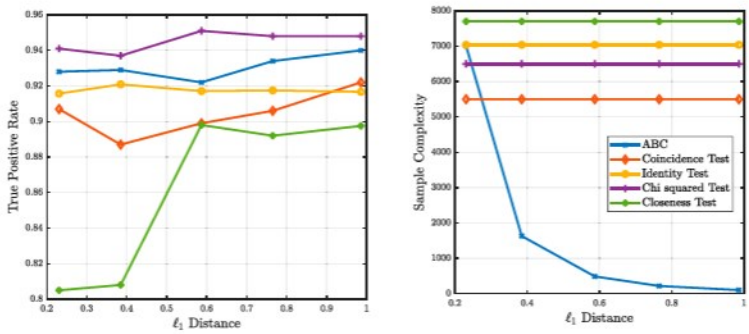

Fig. 2. True positive rate (left) and empirical sample complexity (right) vs ℓ_1 distance for continuous alternative distributions.

complexities for all the fixed-sample-size tests are the same for all the distributions. In contrast, the sample complexity of the ABC adapts to the ℓ_1 distance of the underlying distribution and decreases as the ℓ_1 distance increases, demonstrating the adaptivity of the proposed approach to the unknown alternative distribution. As shown by Fig. 1(left), we also see that the ABC achieves similar level of true positive rates with better empirical sample complexities as compared with the fixed sample-size methods.

In the next example, we consider the continuous uniformity testing problem as described in Sec. III. We take beta distributions as the alternative distributions. The parameters of beta distribution (α, β) were taken in the set $\{(1.5, 1.5), (2, 2), (3, 3), (3, 5), (3, 8)\}$. Fig. 2 shows the empirical sample complexities and false positive rates obtained by all methods. Similar to the discrete case, ABC demonstrates its adaptive sample complexity over fixed-sample-size methods with similar level of accuracy for all continuous alternative distributions.

In the third example, we consider a real-world application. The dataset consists of current samples collected from a power system provided by EPFL[6] [23]. To obtain alternative cases with different ℓ_1 distance, we constructed a synthetic system similar to the EPFL campus grid through MATLAB/Simulink. The power system topology is shown in Fig. 7, where $\mathbf{L_1}, \mathbf{L_2}$, and $\mathbf{I_{DG}}$ are driven by EPFL current measurement collected at bus A, B and C, respectively, to approximate the randomness of consumer load profiles and the power output of the distributed generation. We simulated the fault scenarios located on the line of Z_3, with different electrical distance. We then adopted method in [24] to transform the waveform to samples on $[0, 1]$, and anomaly-free samples are transformed to *i.i.d* samples of continuous uniform distribution. The ABC algorithm and other testing methods were applied on the transformed sequence to detect anomalous and anomaly-free segments.

Fig. 3 shows the true positive rate and sample complexity for the transformed power system data, collected at false positive rate of 0.05. ABC is shown to have the optimal true positive rate and sample complexity over other baselines, demonstrating its effectiveness in real-world applications.

B. Batch Methods With Doubling Search Technique

For the second set of experiments, we compare our proposed sequential algorithm against a sequential version of fixed-sample-size algorithms. Specifically, the sequential version of a fixed-sample-size algorithm $\mathcal{A}$ is defined using the doubling search technique [17] as follows. For each $k = 0, 1, \ldots, \log_2(1/\varepsilon)$, the algorithm $\mathcal{A}$ is run to distinguish distributions that are 2^{-k} away in ℓ_1 distance. If for any value of k, the algorithm returns H_1, the process is terminated and H_1 is returned. Otherwise, after $\log_2(1/\varepsilon)$ iterations, H_0 is returned.

While this doubling search technique offers similar sample complexity to SCT, our proposed algorithm has two advantages over this doubling search technique. First, the sample complexity of the doubling search technique has a suboptimal $\log(1/\delta)$

[6]https://github.com/DESL-EPFL/Point-on-wave-Data-of-EPFL-campus-Distribution-Network

Authorized licensed use limited to: Cornell University Library. Downloaded on July 29,2024 at 21:49:11 UTC from IEEE Xplore. Restrictions apply.

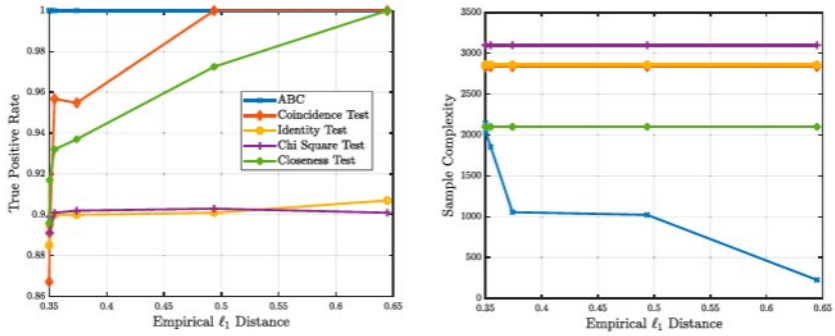

Fig. 3. True positive rate (left) and empirical sample complexity (right) vs ℓ_1 distance for alternative samples collected from simulated power system.

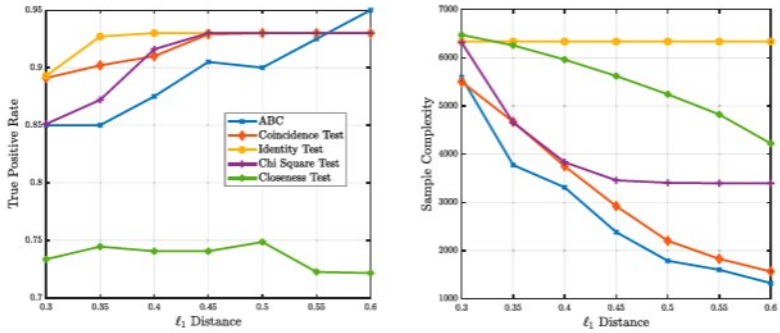

Fig. 4. True positive rate (left) and empirical sample complexity (right) vs ℓ_1 distance for discrete alternative distributions. Batch algorithms are equipped with the doubling search technique.

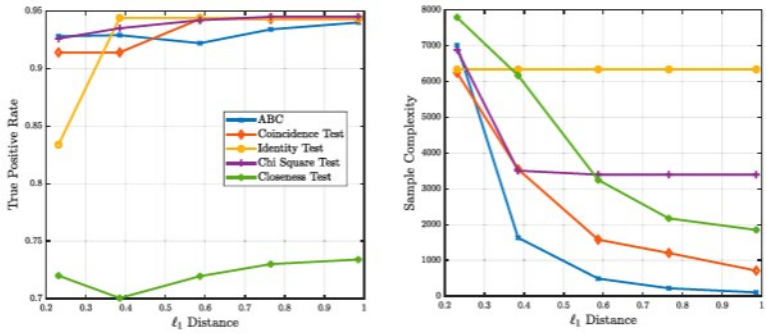

Fig. 5. True positive rate (left) and empirical sample complexity (right) vs ℓ_1 distance for continuous alternative distributions. Batch algorithms are equipped with the doubling search technique.

dependence on the error probability δ which our proposed approach improves it to the optimal order $\sqrt{\log(1/\delta)}$. Second, our algorithm performs a linear discretization of the interval of possible ℓ_1 distances instead of a geometric discretization (as done in the doubling search technique), thereby offering a finer control on the estimated ℓ_1 distance. As a result, our estimate of the ℓ_1 distance is always within a $(1 + o(1))$ factor of the true value as opposed to within a factor of 2 achieved by the doubling search technique. Consequently, the leading constant in the sample complexity of the doubling approach is up to 4 times worse than that of our proposed algorithm, as the sample complexity is inversely proportional to the square of the ℓ_1 distance. Our empirical studies corroborate our theoretical claims. The superior empirical performance of sequential test over the doubling search technique extensions of fixed-sample-sized tests was also noted in [17].

We compare the algorithms on same the synthetic datasets considered in the previous experiment. Figs. 4 and 5 shows the true positive rate and sample complexity for the discrete and continuous settings respectively. As evident from the figures, SCT offers a significant improvement in sample complexity over the doubling search technique based approaches without any compromise in the error rate.

C. Non-Uniformity Testing

In this experiment, we assess the performance of ABC for the case when the null hypothesis is non-uniform. As described in

Authorized licensed use limited to: Cornell University Library. Downloaded on July 29,2024 at 21:49:11 UTC from IEEE Xplore. Restrictions apply.

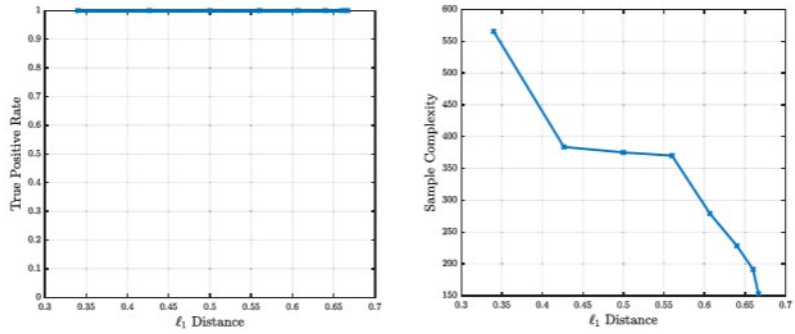

Fig. 6. True positive rate (left) and empirical sample complexity (right) vs ℓ_1 distance for nonuniform null hypothesis for ABC.

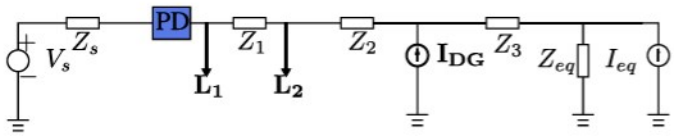

Fig. 7. Power system model.

Sec. III, we can extend the ABC approach to scenarios with non-uniform null hypothesis by using a non-uniform binning. The choice of non-uniform binning transforms the nominal distribution to a uniform one, after which the ABC test can be applied.

For this experiment, we consider the following distribution $f_0(x)$ as the null hypothesis:

$$f_0(x) = \begin{cases} 1 + \frac{2}{3}(1 - 4|x - 0.25|) & x \in [0, 0.5], \\ 1 - \frac{2}{3}(1 - 4|x - 0.75|) & x \in (0.5, 1]. \end{cases}$$

For the alternate hypothesis, we chose a distribution $f_p(x) \equiv f_0(x + p) \mod 1$, given by

$$f_p(x) = \begin{cases} f_0(x - p) & \text{if } x \in [p, 1], \\ f_0(x - p + 1) & \text{if } x \in [0, p), \end{cases}$$

for different values of $p \in [0, 1]$. It is straightforward to show that the ℓ_1 distance between $f_0(x)$ and $f_p(x)$ is given by $8p(1 - p)/3$. The results for true positive rate and sample complexity are plotted in Fig. 6. As shown in Fig. 6, the ABC test can effectively handle the scenario for non-uniform null hypothesis by employing a non-uniform binning of the domain.

V. Conclusion and Future Work

In this work, we considered the problem of uniformity testing of an unknown, Lipschitz continuous distribution on $[0, 1]$. We proposed a *sequential* test called Adaptive Binning Coincidence (ABC) test that adapts to the unknown distribution under the alternative hypothesis. It builds on our proposed adaptive sequential test for discrete distributions with a sample complexity of $\tilde{O}(\sqrt{m}\gamma^{-2})$ improving upon that of $O(\sqrt{m}\varepsilon^{-2})$ offered by offline tests, where $\gamma > \varepsilon$ is the *actual* ℓ_1 distance of the unknown distribution to the uniform distribution. ABC offers a two-fold adaptivity — it terminates the test both *early* and *quickly* using a layered partition of the alternative set and adaptive discretization respectively, resulting in an improved sample complexity for favorable alternative distributions. For ABC, we established a sample complexity of $\tilde{O}(\gamma^{-6}\sqrt{\log(1/\delta)})$ and $\tilde{O}(\varepsilon^{-6}\sqrt{\log(1/\delta)})$ samples under the alternative and null hypotheses respectively. Lastly, we provided an empirical demonstration of the benefit of the proposed approach for both synthetic and real-world settings.

An interesting future direction is to develop sequential tests based on other test statistics that are also valid beyond the sparse regime. We conjecture this might also help address the gap between the upper and lower bounds for continuous distributions.

Appendix

A. Proof of Lemma 1

As mentioned earlier, the proof is based on evaluating the integral in Eqn. (1) using the saddle point method [22]. Recall that, we defined the function g as

$$g(\lambda) = e^{\lambda} \prod_{j=1}^{m} \left(1 - \lambda p_j e^{-\lambda p_j} + \lambda p_j e^{-\lambda p_j} e^{\theta}\right) \frac{1}{\lambda^{n+1}}.$$

To choose the contour, we first differentiate g to find the saddle point. On differentiating, we obtain,

$$g'(\lambda) = g(\lambda) \left(\sum_{j=1}^{m} \left(\frac{p_j(e^{\theta} - 1 + e^{\lambda p_j})}{\lambda p_j(e^{\theta} - 1) + e^{\lambda p_j}} \right) - \frac{n+1}{\lambda} \right).$$

Instead of exactly choosing the minimizer, we choose a point very close to it, defined as the solution to the following equation:

$$\frac{\lambda p_j(e^{\theta} - 1 + e^{\lambda p_j})}{\lambda p_j(e^{\theta} - 1) + e^{\lambda p_j}} = n. \tag{2}$$

It is established in [7] that the above equation has a unique non-negative solution, which we denote using λ_0. Furthermore, it satisfies the following inequalities:

$$\begin{aligned} &\text{For } \theta \geq 0: \quad ne^{-\theta} \leq \lambda_0 \leq n(1 + e^{-1}(e^{\theta} - 1)) \\ &\text{For } \theta < 0: \quad n(1 + e^{-1}(e^{\theta} - 1)) \leq \lambda_0 \leq ne^{-\theta}. \end{aligned} \tag{3}$$

From the above inequalities, it is clear that $\lambda_0 = O(n)$. We can also obtain a more refined approximation of the relation

Authorized licensed use limited to: Cornell University Library. Downloaded on July 29,2024 at 21:49:11 UTC from IEEE Xplore. Restrictions apply.

between λ_0 and n as follows. We first split the set of symbols into two sets based on the magnitude of their probabilities and obtain the approximation separately for these sets. In particular, we define $\mathcal{W} = \{j : p_j \geq 8/m\}$ and $\beta(p) = \sum_{j\in\mathcal{W}} p_j$. We first focus on the symbols that not do not belong in $\mathcal{W}$. For these symbols, the following set of inequalities hold.

$$\lambda_0 p_j e^{\theta} + \lambda_0^2 p_j^2(1-e^{2\theta}) + l(\theta)\lambda_0^3 p_j^3 \leq \frac{\lambda_0 p_j(e^{\theta}-1+e^{\lambda p_j})}{\lambda_0 p_j(e^{\theta}-1)+e^{\lambda_0 p_j}}$$
$$\lambda_0 p_j e^{\theta} + \lambda_0^2 p_j^2(1-e^{2\theta}) + u(\theta)\lambda_0^3 p_j^3 \geq \frac{\lambda_0 p_j(e^{\theta}\ 1+e^{\lambda p_j})}{\lambda_0 p_j(e^{\theta}-1)+e^{\lambda_0 p_j}}, \tag{4}$$

where $l(\theta) := \frac{1}{6}\big(6e^{3\theta} - 9e^{\theta} + 3\big)\mathbb{1}\{\theta < 0\}$ and $u(\theta) := \frac{1}{6}\big(6e^{3\theta} - 9e^{\theta} + 3\big)\mathbb{1}\{u \geq 0\}$. For the symbols in $\mathcal{W}$, we have,

$$\frac{\lambda_0 p_j(e^{\theta}-1+e^{\lambda p_j})}{\lambda_0 p_j(e^{\theta}-1)+e^{\lambda_0 p_j}} = \underbrace{\frac{e^{-\theta}+e^{-\lambda_0 p_j}(1-e^{-\theta})}{1+\lambda_0 p_j e^{-\lambda_0 p_j}(e^{\theta}-1)}}_{:=D_j}\lambda_0 p_j e^{\theta},$$

On plugging these approximations for different regimes in (2), we obtain that λ_0 is of the form, $\lambda_0 = \dfrac{ne^{-\theta}(1+w)}{1+\sum_{j\in\mathcal{W}} p_j(D_j-1)}$, where w is the small approximating factor, on which we will obtain bounds. For brevity, we define $D_* := 1+\sum_{j\in\mathcal{W}} p_j (D_j-1)$. Lastly, we can obtain the range on the ratio $(1+w)/D_*$ from the relations in (3).

On summing the lower bound in (4) over j, we obtain,

$$n \geq \sum_{j\in\mathcal{W}} D_j \lambda_0 p_j e^{\theta} + \sum_{j\notin\mathcal{W}} \left[\lambda_0 p_j e^{\theta} + \lambda_0^2 p_j^2(1-e^{2\theta}) + l(\theta)\lambda_0^3 p_j^3\right]$$
$$n \geq \lambda_0 D_* e^{\theta} + \lambda_0^2 \sum_{j\notin\mathcal{W}} p_j^2(1-e^{2\theta}) + l(\theta)\lambda_0^3 \sum_{j\notin\mathcal{W}} p_j^3$$

On plugging in the value of λ_0 and rearranging the equation, we obtain,

$$w \leq \frac{n}{D_*^2}(1+w)^2 \sum_{j\notin\mathcal{W}} p_j^2(1-e^{-2\theta}) - l(\theta)n^2\left(\frac{1+w}{D_*}\right)^3 \sum_{j=1}^{m} p_j^3. \tag{5}$$

A similar series of steps using the upper bound in (4) yields

$$w \geq \frac{n(1+w)^2}{D_*^2}\left(\sum_{j\notin\mathcal{W}} p_j^2\right)(1-e^{-2\theta}) - u(\theta)n^2\left(\frac{1+w}{D_*}\right)^3 \sum_{j\notin\mathcal{W}} p_j^3. \tag{6}$$

With these relations at our disposal, we now move on to evaluating the integral in (1) along the closed contour $\lambda = \lambda_0 e^{i\psi}$ for $\psi \in [-\pi, \pi]$. The integral reads as

$$\mathbb{E}_p\left[\exp\left(\theta\tilde{S}_n^*\right)\right] = e^{-\theta n}\frac{n!}{2\pi}\int_{-\pi}^{\pi} g(\lambda_0 e^{i\psi})\lambda_0 e^{i\psi}\,\mathrm{d}\psi = \frac{n!}{2\pi}\lambda_0^{-n}e^{-\theta n}\mathrm{Re}\left[\int_{-\pi}^{\pi} h(\psi)\,\mathrm{d}\psi\right],$$

where

$$h(\psi) = e^{-in\psi}\prod_{j=1}^{m}\left(\lambda_0 p_j(e^{\theta}-1)e^{i\psi} + e^{\lambda_0 p_j e^{i\psi}}\right).$$

Instead of $h(\psi)$, we focus on bounding $H(\psi) = \log(h(\psi))$ since it is easier to deal with sums than products. We have,

$$H(\psi) = -in\psi + \sum_{j=1}^{m}\log\left(\lambda_0 p_j(e^{\theta}-1)e^{i\psi} + e^{\lambda_0 p_j e^{i\psi}}\right).$$

We split the integral into three parts by evaluating it over three different ranges, i.e., $[-\pi, -\pi/2)$, $[-\pi/2, \pi/2]$ and $(\pi/2, \pi]$. We first consider the integral over $[-\pi/2, \pi/2]$. This is the region where the integrand behaviour behaves violently and thus will correspond to the dominant term. We have,

$$\begin{aligned}\mathrm{Re}(H(\psi)) &= \sum_{j=1}^{m}\mathrm{Re}\left[\log\left(\lambda_0 p_j(e^{\theta}-1)e^{i\psi} + e^{\lambda_0 p_j e^{i\psi}}\right)\right] \\ &= \sum_{j=1}^{m}\mathrm{Re}\Big[\log(e^{\lambda_0 p_j e^{i\psi}}) \\ &\qquad + \log\left(1+\lambda_0 p_j(e^{\theta}-1)e^{i\psi}e^{-\lambda_0 p_j e^{i\psi}}\right)\Big] \\ &\leq \sum_{j=1}^{m}\Big[\lambda_0 p_j\cos(\psi) \\ &\qquad + \log\left(1+\lambda_0 p_j e^{-\lambda_0 p_j\cos(\psi)}(e^{\theta}-1)\right)\Big].\end{aligned}$$

We define a function $G(\psi; u)$ that corresponds to the general form of the RHS in the above expression. That is for $u \geq 0$,

$$G(\psi; u) = u\cos(\psi) + \log\left(1+ue^{-u\cos(\psi)}(e^{\theta}-1)\right).$$

Note that $G'(0;u) = 0$ and $G''(0;u) \leq -0.4u$ for $\psi \in [-\pi/2, \pi/2]$. Thus, by using mean value theorem it can be shown that $G(\psi;u)$ satisfies $G(\psi;u) \leq G(0;u) - 0.2u\psi^2$ for $\psi \in [-\pi/2, \pi/2]$ for all $u \geq 0$. On plugging this inequality in previous one, we have,

$$\begin{aligned}\mathrm{Re}(H(\psi)) &\leq \sum_{j=1}^{m} G(\psi; \lambda_0 p_j) \leq \sum_{j=1}^{m}\left[G(0;\lambda_0 p_j) - 0.2\lambda_0 p_j\psi^2\right] \\ &\leq H(0) - 0.2\lambda_0\psi^2.\end{aligned}$$

Using this relation, we can establish the following bound.

$$\begin{aligned}\mathrm{Re}\left[\int_{-\pi/2}^{\pi/2} h(\psi)\,\mathrm{d}\psi\right] &= \mathrm{Re}\left[\int_{-\pi/2}^{\pi/2}\exp(H(\psi))\,\mathrm{d}\psi\right] \\ &\leq \int_{-\pi/2}^{\pi/2}\exp(\mathrm{Re}(H(\psi)))\,\mathrm{d}\psi \\ &\leq e^{H(0)}\int_{-\pi/2}^{\pi/2} e^{-0.2\lambda_0\psi^2}\,\mathrm{d}\psi \\ &\leq e^{H(0)}\sqrt{\frac{\pi}{0.1\lambda_0}}.\end{aligned} \tag{7}$$

Authorized licensed use limited to: Cornell University Library. Downloaded on July 29,2024 at 21:49:11 UTC from IEEE Xplore. Restrictions apply.

For $\psi \in [-\pi, -\pi/2] \cup [\pi/2, \pi]$, we have $|e^{\lambda_0 p_j e^{i\psi}}| \leq 1$. Consequently, $|\lambda_0 p_j (e^\theta - 1) e^{i\psi} + e^{\lambda_0 p_j e^{i\psi}}| \leq 1 + \lambda_0 p_j (e^\theta - 1)$. Thus, we have,

$$\begin{aligned}\mathrm{Re}(H(\psi)) &= \sum_{j=1}^{m} \mathrm{Re}\left[\log\left(\lambda_0 p_j (e^\theta - 1) e^{i\psi} + e^{\lambda_0 p_j e^{i\psi}}\right)\right] \\ &\leq \sum_{j=1}^{m} \log\left(1 + \lambda_0 p_j |(e^\theta - 1)|\right) \leq \lambda_0 |(e^\theta - 1)|.\end{aligned}$$

Using this bound, we can bound the following integral.

$$\begin{aligned}\mathrm{Re}\left[\int_{-\pi}^{-\pi/2} h(\psi)\, \mathrm{d}\psi\right] &= \mathrm{Re}\left[\int_{-\pi}^{-\pi/2} \exp(H(\psi))\, \mathrm{d}\psi\right] \\ &\leq \int_{-\pi}^{-\pi/2} \exp(\mathrm{Re}(H(\psi)))\, \mathrm{d}\psi \leq \frac{\pi}{2} e^{\lambda_0 |e^\theta - 1|}.\end{aligned}$$

We can similarly bound the integral for $[\pi/2, \pi]$. On combining all of them, we have,

$$\begin{aligned}\mathbb{E}_p\left[\exp\left(\theta \tilde{K}_1(n)\right)\right] &= \frac{n!}{2\pi} \lambda_0^{-n} e^{-\theta n} \mathrm{Re}\left[\int_{-\pi}^{\pi} h(\psi)\, \mathrm{d}\psi\right] \\ &\leq \frac{n!}{2\pi n^n} \left(\frac{D_*}{1+w}\right)^n \left(e^{H(0)} \sqrt{\frac{\pi}{0.1\lambda_0}} + \pi e^{\lambda_0 |(e^\theta - 1)|}\right).\end{aligned} \tag{8}$$

The statement of the lemma follows by noting $H_p := H(0)$ and $\frac{D_*}{1+w} = \frac{ne^{-\theta}}{\lambda_0}$.

B. Proof of Lemma 2

We begin with bounding the probability of false alarm, that is, the underlying distribution is uniform (H_0) and we fail to detect it. For this scenario, the following relation holds for $\theta < 0$, where τ_n denotes the threshold when n samples have been taken and $P_e(n)$ denotes the error probability at that time instant.

$$\begin{aligned}P_e(n) &= \Pr(K_1(n) - \mathbb{E}_u[K_1(n)] < -\tau_n) \\ &\leq \mathbb{E}\left[\exp(\theta(K_1(n) - \mathbb{E}_u[K_1(n)]))\right] e^{\theta\tau_n} \\ &\leq \mathbb{E}\left[\exp(\theta(K_1(n) - n))\right] \exp\left(\theta(n - \mathbb{E}_u[K_1(n)] + \tau_n)\right) \\ &\leq \mathbb{E}\left[\exp(\theta \tilde{K}_1(n)))\right] \\ &\quad \times \exp\left[\theta\left(\tau_n + \frac{n(n-1)}{m} - \frac{n}{2}\left(\frac{n-1}{m}\right)^2\right)\right] \\ &\leq \frac{n!}{2\pi n^n} \left(\frac{D_*}{1+w}\right)^n \left(e^{H(0)} \sqrt{\frac{\pi}{0.1\lambda_0}} + \pi e^{0.7n}\right) \\ &\quad \times \exp\left(\theta\tau_n + \theta\left(\frac{n(n-1)}{m} - \frac{n}{2}\left(\frac{n-1}{m}\right)^2\right)\right) \\ &\leq \frac{n! e^n}{2\pi n^n} \left(\frac{D_*}{1+w}\right)^n \left(e^{H(0)-n} \sqrt{\frac{\pi}{0.1\lambda_0}} + \pi e^{-0.3n}\right) \\ &\quad \times \exp\left(\theta\tau_n + \theta\left(\frac{n(n-1)}{m} - \frac{n}{2}\left(\frac{n-1}{m}\right)^2\right)\right)\end{aligned}$$

Since the underlying distribution is uniform, $\mathcal{W} = \emptyset$ and consequently $D_* = 1$. Moreover, for $\theta < 0$, $H(0) \leq \sum_{j=1}^{m} \left[\lambda_0 p_j e^\theta + \frac{1}{2}\lambda_0^2 p_j^2 (1 - e^{2\theta})\right]$. Using these relations along with polynomial expansions of e^x and $\log(1+x)$ near $x = 0$, we can bound the first term in the expression as

$$\begin{aligned}T_1 &:= \left(\frac{D_*}{1+w}\right)^n \exp\left[H(0) - n \right. \\ &\qquad \left. + \theta\left[\tau_n + \frac{n(n-1)}{m} - \frac{n}{2}\left(\frac{n-1}{m}\right)^2\right]\right] \qquad (9) \\ &\leq \exp\left(\frac{2n^2\theta^2}{m} + \theta\left(\tau_n - \frac{n}{m} - \frac{n}{2}\left(\frac{n}{m}\right)^2\right) + 2nw^2\right). \qquad (10)\end{aligned}$$

Using the condition on the ratio of n and m, we have $\frac{n}{m} + \frac{n^3}{2m^2} \leq \frac{n^3}{m^2} \leq \frac{n^2\varepsilon^2}{1536m}$. Note that for the particular choice of τ and corresponding decision instant in the Sequential Coincidence Test satisfies $n^2\varepsilon^2/1536m \leq \tau_n/75$ for all epochs. Lastly, using (5), we have, $2nw^2 \leq 2n\left(3.76\frac{n}{m}(1 - e^{-2\theta})\right)^2 \leq \frac{114n^2\varepsilon^2\theta^2}{1536m}$. On combining these two bounds with the bound on T_1, we obtain,

$$T_1 \leq \exp\left(\frac{n^2\theta^2}{m}\left(2 + \frac{114\varepsilon^2}{1536}\right) + \frac{74\theta\tau_n}{75}\right).$$

On plugging $\theta = -\frac{74m\tau_n}{375n^2}$, we obtain $T_1 \leq \exp\left(-\frac{5476}{56250} \cdot \frac{m\tau_n^2}{n^2}\right)$. When evaluated at the decision for epoch k, the above relation simplifies to

$$T_1 \leq \exp\left(-3\log(k + 2/\delta)\right) \leq \frac{1}{(k+2/\delta)^3}.$$

With the above choice of θ and the bound on w obtained from (6), the second term in the expression of $P_e(n)$ can be simplified as,

$$\begin{aligned}T_2 &:= \left(\frac{D_*}{1+w}\right)^n e^{-0.3n} \\ &\quad \times \exp\left[\theta\left[\tau_n + \frac{n(n-1)}{m} - \frac{n}{2}\left(\frac{n-1}{m}\right)^2\right]\right] \\ &\leq \exp\left[-0.3n - \frac{n\theta}{2}\left(\frac{n-1}{m}\right)^2 - n\log(1+w)\right] \leq e^{-n/4}.\end{aligned}$$

On combining the bounds on T_1 and T_2 and plugging them back in the expression of $P_e(n)$, we obtain the following bound on the error probability at the decision instant during epoch k:

$$\begin{aligned}P_e(n_k) &\leq \frac{n_k! e^{n_k}}{2\pi n_k^{n_k}} \left(\frac{1}{(k+2/\delta)^3} \cdot \sqrt{\frac{18\pi}{n_k}} + \pi e^{-0.25n_k}\right) \\ &\leq e^{1/12n_k} \left(\frac{3}{(k+2/\delta)^3} + \sqrt{\frac{n_k\pi}{2}} e^{-0.25n_k}\right).\end{aligned}$$

The second step follows by invoking Stirling's approximation for factorials.

We bound the probability of miss detection using a similar process. In this case, the underlying distribution p belongs to

Authorized licensed use limited to: Cornell University Library. Downloaded on July 29,2024 at 21:49:11 UTC from IEEE Xplore. Restrictions apply.

$\mathcal{C}(\varepsilon)$ such that $\|p-u\|_1=\gamma>\varepsilon$. For this scenario, the following relation holds for $\theta\geq 0$, with τ_n and $P_e(n)$ defined as before.

$$\begin{aligned}\Pr(\text{err}) &= \Pr(\mathbb{E}_u[K_1(n)]-K_1(n)<\tau_n)\\ &\leq \mathbb{E}\left[\exp(\theta(K_1(n)-n))\right]\\ &\quad\times\exp\left(\theta(n-\mathbb{E}_u[K_1(n)]+\tau_n)\right)\\ &\leq \mathbb{E}\left[\exp(\theta(K_1(n)-n))\right]\exp\left(\theta\left(\tau_n+\frac{n^2}{m}\right)\right)\\ &\leq \frac{n!e^n}{2\pi n^n}\left(\frac{D_*}{1+w}\right)^n\left(e^{H(0)-n}\sqrt{\frac{\pi}{0.1\lambda_0}}+\pi e^{-0.3n}\right)\\ &\quad\times\exp\left(\theta\left(\tau_n+\frac{n^2}{m}\right)\right).\end{aligned}$$

For $\theta\geq 0$, $H(0)$ can be bounded as

$$\begin{aligned}H(0)\leq\sum_{j\notin\mathcal{W}}&\left[\lambda_0 p_j e^{\theta}+\frac{1}{2}\lambda_0^2p_j^2(1-e^{2\theta})+\frac{\theta e^{3\theta}}{2}\lambda_0^3p_j^3\right]\\ &+\sum_{j\in\mathcal{W}}\left(\lambda_0p_je^{\theta}+\lambda_0p_j(1-e^{\lambda_0p_j})(1-e^{\theta})\right)\end{aligned}$$

Similar to the previous case, we express $P_e(n)$ as a sum of two terms, T_1' and T_2' and bound them separately. Using the above bound on $H(0)$, we can simplify the first term as,

$$\begin{aligned}T_1' &:= \left(\frac{D_*}{1+w}\right)^n\exp\left(H(0)-n+\theta\left(\tau_n+\frac{n^2}{m}\right)\right)\\ &\leq\exp\left[\frac{n^2}{2}\sum_{j\notin\mathcal{W}}p_j^2(e^{-2\theta}-1)+\theta\left(\tau_n+\frac{n^2}{m}\right)\right.\\ &\quad\left.+\frac{32\theta n^3}{m^2}\left(\frac{1+w}{D_*}\right)^3\right]\\ &\quad\times\exp\left(n\frac{(1+w)}{D_*}\left(1+\sum_{j\in\mathcal{W}}p_j(1-e^{\lambda_0p_j})(e^{-\theta}-1)\right)\right)\\ &\quad\times\exp\left(n\left\{-1-\log\left(\frac{1+w}{D_*}\right)\right\}\right).\end{aligned}\tag{11}$$

To further simplify T_1', we define

$$\begin{aligned}J &:= \frac{(1+w)}{1+\sum_{j\in\mathcal{W}}p_j(D_j-1)}\\ &\quad\times\left(1+\sum_{j\in\mathcal{W}}p_j(1-e^{\lambda_0p_j})(e^{-\theta}-1)\right)\\ &\quad-1-\log\left(\frac{1+w}{1+\sum_{j\in\mathcal{W}}p_j(D_j-1)}\right).\end{aligned}$$

This corresponds to the term in the exponential in the third and fourth lines of (11). We also define two functions $D(x)$ and $E(x)$ as follows:

$$\begin{aligned}D(x) &:= \frac{e^{-\theta}+e^{-x}(1-e^{-\theta})}{1+xe^{-x}(e^{\theta}-1)}\\ E(x) &:= (e^{-\theta}-1)(1-e^{-x}).\end{aligned}$$

Note that $D_j=D(\lambda_0p_j)$ and $E(\lambda_0p_j)=(1-e^{\lambda_0p_j})(e^{-\theta}-1)$. Moreover, the function $F(x)=\frac{D(x)-1}{E(x)}$ is decreasing for all $x\geq 0$ and satisfies the relation $1\leq F(x)\leq 1+e^{\theta}$. Consequently, $(1+e^{\theta})E(x)\leq D(x)-1\leq E(x)$ holds for all $x\geq 0$ since $E(x)\leq 0$. Thus, we have,

$$\begin{aligned}\sum_{j\in\mathcal{W}}p_j(D_j-1)&\geq(1+e^{\theta})\sum_{j\in\mathcal{W}}p_j(1-e^{\lambda_0p_j})(e^{-\theta}-1)\\ \sum_{j\in\mathcal{W}}p_j(D_j-1)&\leq\sum_{j\in\mathcal{W}}p_j(1-e^{\lambda_0p_j})(e^{-\theta}-1).\end{aligned}$$

If we let $x=\sum_{j\in\mathcal{W}}p_j(D_j-1)$, then we can write J as

$$J(x)=\frac{(1+w)(1+\rho x)}{1+x}-1-\log\left(\frac{(1+w)}{1+x}\right),$$

where $\rho\in[(1+e^{\theta})^{-1},1]$. Since $D(y)\geq e^{-2\theta}$ for all $y\geq 0$, the domain of x is given as $x\in[-\beta(p)(1-e^{-2\theta}),0]$. As $\beta(p)<1$ and $\theta<0.4$, the function $J(x)$ is increasing throughout the domain of x. Consequently, $J(x)\leq J(0)\leq w^2$. Furthermore, over this domain, we can upper bound $J(x)$ as $J(x)=w^2+0.2x$. Lastly, since $D(x)$ is a decreasing function $D_j\leq D\left(\frac{8\lambda_0}{m}\right)$. Once again, using a local linear approximation of $D(x)$, we have the upper bound $D(x)\leq 1+0.7(e^{-\theta}-e^{\theta})x$. On combining everything, we can bound J as,

$$\begin{aligned}J &\leq w^2+0.2\sum_{j\in\mathcal{W}}p_j(D_j-1)\\ &\leq w^2+0.2\sum_{j\in\mathcal{W}}p_j\left(D\left(\frac{8\lambda_0}{m}\right)-1\right)\\ &\leq w^2+0.14\sum_{j\in\mathcal{W}}p_j\frac{8\lambda_0}{m}(e^{-\theta}-e^{\theta})\\ &\leq w^2+1.12\beta(p)\frac{n(1+w)}{mD_*}(e^{-2\theta}-1).\end{aligned}$$

On plugging this back into the bound for T_1', we obtain,

$$\begin{aligned}T_1'\leq&\exp\left(\frac{n^2}{2m}\left(m\sum_{j\notin\mathcal{W}}p_j^2+2\beta(p)\right)(e^{-2\theta}-1)\right)\\ &\times\exp\left(\theta\left(\tau_n+\frac{n^2}{m}\right)+\frac{32\theta n^3}{m^2}\left(\frac{1+w}{D_*}\right)^3+nw^2\right).\end{aligned}$$

Let $p_j=\frac{1}{m}+\Delta_j$. So $\sum_{j=1}^m\Delta_j=0$ and $\sum_{j=1}^m|\Delta_j|=\gamma$, the actual ℓ_1 distance to the uniform distribution. Using this, the first term can be written as,

$$\begin{aligned}&\frac{n^2}{2m}\left(m\sum_{j\notin\mathcal{W}}p_j^2+2\sum_{j\in\mathcal{W}}p_j\right)\\ &\geq\frac{n^2}{2m}\left(m\sum_{j\notin\mathcal{W}}\left(\frac{1}{m}+\Delta_j\right)^2+2\sum_{j\notin\mathcal{W}}\left(\frac{1}{m}+\Delta_j\right)\right)\\ &\geq\frac{n^2}{2m}\left(1+m\sum_{j\notin\mathcal{W}}\Delta_j^2+\frac{|\mathcal{W}|}{m}\right)\geq\frac{n^2}{2m}\left(1+\frac{\gamma^2}{4}\right),\end{aligned}$$

Authorized licensed use limited to: Cornell University Library. Downloaded on July 29,2024 at 21:49:11 UTC from IEEE Xplore. Restrictions apply.

where the last step follows from $\sum_{j\notin\mathcal{W}}|\Delta_j| \geq \gamma/2$. The bound on n/m and (3) together imply $\frac{32\theta n^3}{m^2}\left(\frac{1+w}{D_*}\right)^3 \leq \frac{\theta n^2\varepsilon^2}{8m}$. Lastly using (5) and (3), we have, $nw^2 \leq \frac{n^3}{D_*^4}(1+w)^4\left(\sum_{j\notin\mathcal{W}} p_j^2\right)^2(1-e^{-2\theta})^2 \leq \theta^2\cdot\frac{7n^2\varepsilon^2}{4m}$. On plugging on these bounds in the bound for T_1', we obtain,

$$T_1' \leq \exp\left(\frac{2n^2\theta^2}{m}\left(1+\frac{\gamma^2}{4}+\frac{7\varepsilon^2}{8}\right)-\theta\left(\frac{n^2\gamma^2}{8m}-\tau_n\right)\right).$$

Note that the particular choice of τ and corresponding decision instant as used in Alg. 1 satisfy $\frac{n^2\gamma^2}{8m} \geq 2\tau_n$ for all $k \geq 112/\gamma^2$. Thus, if we define $k_0(\gamma) = 112/\gamma^2$, then for all $k \geq k_0(\gamma)$, we have

$$T_1' \leq \exp\left(\frac{4.25n^2\theta^2}{m}-\theta\tau_n\right).$$

On plugging $\theta = \frac{m\tau_n}{9.5n^2}$, we obtain, $T_1' \leq \exp\left(-\frac{m\tau_n^2}{18n^2}\right)$. On evaluating the above expression at decision instant for epoch $k \geq k_0(\gamma)$, we obtain

$$T_1' \leq \exp\left(-2.5\log(k+2/\delta)\right) \leq \frac{1}{(k+2/\delta)^{2.5}}.$$

Similar to the previous case, the above choice of θ along with bound on w obtained from (5), yields the bound

$$T_2 := \left(\frac{D_*}{1+w}\right)^n \exp\left(-0.3n+\theta\left(\tau_n+\frac{n^2}{m}\right)\right) \leq e^{-n/4},$$

On combining the bounds on T_1' and T_2', we obtain the following result for the probability of error at the decision instant of epoch k.

$$\begin{aligned}P_e(n_k) &\leq \frac{n_k! e^{n_k}}{2\pi {n_k}^{n_k}}\left(\frac{1}{(k+2/\delta)^{2.5}}\cdot\sqrt{\frac{18\pi}{n_k}}+\pi e^{-0.25n_k}\right)\\ &\leq e^{1/12n_k}\left(\frac{3}{(k+2/\delta)^{2.5}}+\sqrt{\frac{n_k\pi}{2}}e^{-0.25n_k}\right),\end{aligned}$$

where the second line again employs Stirling's Approximation.

C. Proof of Lemma 3

The proof of Lemma 3 uses the Lipschitz continuity of the PDF of p to bound the error between the ℓ_1 distances of the continuous and the discrete distributions. We denote the continuous uniform distribution on $[0,1]$ using $u(x)$. From the definition of ℓ_1 distance, we have,

$$\begin{aligned}\gamma &= \int_0^1 |p(x)-u(x)|\,\mathrm{d}x\\ &= \sum_{i=0}^{m-1}\int_{i/m}^{(i+1)/m}|p(x)-u(x)|\,\mathrm{d}x\\ &= \sum_{i=0}^{m-1}\int_{i/m}^{(i+1)/m}\left|p(x)-mp_i^\Delta+mp_i^\Delta-u(x)\right|\,\mathrm{d}x\\ &\leq \sum_{i=0}^{m-1}\int_{i/m}^{(i+1)/m}\left(|p(x)-mp_i^\Delta|+|mp_i^\Delta-1|\right)\mathrm{d}x\\ &\leq \sum_{i=0}^{m-1}\int_{i/m}^{(i+1)/m}\left(\frac{L}{m}+|mp_i^\Delta-1|\right)\mathrm{d}x\\ &\leq \sum_{i=0}^{m-1}\left[\frac{L}{m^2}+\left|\int_{(i-1)/m}^{i/m}p(y)\,\mathrm{d}y-\frac{1}{m}\right|\right.\\ &\qquad\left.\times\left(\int_{i/m}^{(i+1)/m}m\,\mathrm{d}x\right)\right]\\ &\leq \sum_{i=0}^{m-1}\frac{L}{m^2}+\sum_{i=0}^{m-1}\left|p_i^\Delta-\frac{1}{m}\right|\\ &\leq \frac{L}{m}+[\gamma]_m.\end{aligned}$$

In the fifth step, we have used the mean value theorem along with the Lipschitz condition on the PDF. From the mean value theorem, we can conclude that there exists an $x_i \in [i/m,(i+1)/m]$ such that $p(x_i) = p_i^\Delta/(1/m) = mp_i$ for all $i = 0,1,2,\ldots,m-1$ and since PDF is L-Lipschitz, we have $|p(x)-p(x_i)| \leq L/m$ for all $x \in [i/m,(i+1)/m]$. Consequently, we can relate the ℓ_1 distances between the continuous and discrete distributions. In particular, if $m \geq L/2\gamma$, then $[\gamma]_m \geq \gamma/2$.

D. Proof of Lemma 4

The proof of this lemma is almost the same as that of Lemma 2 with very minor modifications. Firstly, we modify the set $\mathcal{W}$ to $\mathcal{W}_k$ defined for each epoch as $\mathcal{W}_k = \{j : p_j^\Delta \geq 8/m_k\}$, where once again p_j^Δ is the mass in the j^{th} bin in the discretization. Also for this case, instead of computing the error for any sample n, we just compute it for the decision instant n_k and the corresponding number of bins m_k.

We begin with the probability of false alarm. It can be verified that for the choice of n_k and m_k, all the conditions in the proof of Lemma 2 (Appendix B) are satisfied yielding us same bounds on T_1 and T_2 and consequently, the following result holds for the probability of error at the decision instant corresponding to epoch k.

$$P_e(n_k) \leq e^{1/12n_k}\left(\frac{3}{(k+2/\delta)^3}+\sqrt{\frac{n_k\pi}{2}}e^{-0.25n_k}\right).$$

Similarly, we consider the probability of miss detection. In this scenario, the ℓ_1 distance to the uniform distribution of the underlying distribution p is γ and for simplicity, we denote the discretized ℓ_1 distance as $[\gamma]_k$ instead of $[\gamma]_{m_k}$. If we define $k_0(\gamma) = \min\{k : k \geq 144[\gamma]_k^{-2}\}$, then using this definition of $k_0(\gamma)$, we can obtain all the results from Appendix C. Specifically, for $k \geq k_0(\gamma)$, we have, $\frac{32\theta n_k^3}{m_k^2}\left(\frac{1+w}{D_*}\right)^3 \leq \frac{\theta n_k^2[\gamma]_k^2}{8m_k}$, $n_k w^2 \leq \theta^2\cdot\frac{7n_k^2[\gamma]_k^2}{4m_k}$ and $2\tau_k \leq \frac{n_k^2[\gamma]_k^2}{8m_k}$. Consequently, for $k \geq k_0(\gamma)$, we have,

$$T_1' \leq \exp\left(\frac{4.25n_k^2\theta^2}{m_k}-\theta\tau_k\right).$$

Authorized licensed use limited to: Cornell University Library. Downloaded on July 29,2024 at 21:49:11 UTC from IEEE Xplore. Restrictions apply.

On plugging $\theta = \frac{m_k \tau_k}{9.5 n_k^2}$, we obtain, $T_1' \leq \exp\left(-\frac{m_k \tau_k^2}{18 n_k^2}\right) \leq (k + 2/\delta)^{-9/2}$. It is not difficult to see that $T_2' \leq \exp(-n_k/4)$ holds for $k \geq k_0(\gamma)$, yielding the required expression for $P_e(n_k)$.

REFERENCES

[1] L. LeCam, "Convergence of estimates under dimensionality restrictions," *Ann. Statist.*, vol. 1, no. 1, pp. 38–53, Jan. 1973.

[2] M. Adamaszek, A. Czumaj, and C. Sohler, "Testing monotone continuous distributions on high-dimensional real cubes," in *Proc. Annu. ACM-SIAM Symp. Discrete Algorithms*, New York, NY, USA: ACM, 2010, pp. 56–65.

[3] J. Acharya, A. Jafarpour, A. Orlitsky, and A. T. Suresh, "A competitive test for uniformity of monotone distributions," in *J. Mach. Learn. Res.*, vol. 31, PMLR, 2013, pp. 57–65.

[4] F. N. David, "Two combinatorial tests of whether a sample has come from a given population," *Biometrika*, vol. 37, nos. 1–2, pp. 97–110, 1950.

[5] I. I. Viktorova and V. P. Chistyakov, "Some gneralizations of the test of eEmpty boxes," *Theory Probability Appl.*, vol. 11, no. 2, pp. 270–276, 1966.

[6] L. Paninski, "A coincidence-based test for uniformity given very sparsely sampled discrete data," *IEEE Trans. Inf. Theory*, vol. 54, no. 10, pp. 4750–4755, Oct. 2008.

[7] D. Huang and S. Meyn, "Generalized error exponents for small sample universal hypothesis testing," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8157–8181, Dec. 2013.

[8] T. Batu, E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White, "Testing random variables for independence and identity," in *Proc. Annu. Symp. Found. Comput. Sci.*, 2001, pp. 442–451.

[9] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White, "Testing closeness of discrete distributions," *J. ACM*, vol. 60, no. 1, 2013.

[10] O. Goldreich and D. Ron, *On Testing Expansion in Bounded-Degree Graphs*, Berlin, Germany: Springer-Verlag, 2011, pp. 68–75.

[11] S. O. Chan, I. Diakonikolas, P. Valiant, and G. Valiant, "Optimal algorithms for testing closeness of discrete distributions," in *Proc. Annu. ACM-SIAM Symp. Discrete Algorithms*, 2014, pp. 1193–1203.

[12] I. Diakonikolas, T. Gouleakis, J. Peebles, and E. Price, "Sample-optimal identity testing with high probability," in *Leibniz Int. Proc. Inform. (LIPIcs)*, vol. 107, 2018, pp. 41:1–41:14.

[13] J. Acharya, C. Daskalakis, and G. Kamath, "Optimal testing for properties of distributions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 3591–3599.

[14] C. Canonne, "A survey on distribution testing: Your data is big. But is it blue?" *Theory Comput.*, vol. 1, no. 1, pp. 1–100, 2020.

[15] T. Batu and C. L. Canonne, "Generalized uniformity testing," in *Proc. Annu. Symp. Found. Comput. Sci.*, Los Alamitos, CA, USA: IEEE Comput. Soc. Press 2017, pp. 880–889.

[16] S. Salgia, Q. Zhao, and L. Tong, "As easy as ABC: Adaptive binning coincidence test for uniformity testing," 2021.

[17] O. Fawzi, N. Flammarion, A. Garivier, and A. Oufkir, "Sequential algorithms for testing identity and closeness of distributions," in *Proc. 35th Annu. Conf. Neural Inf. Process. Syst.*, Curran Associates, Inc., vol. 34, 2021, pp. 11655–11664.

[18] I. Diakonikolas, D. M. Kane, and V. Nikishkin, "Testing identity of structured distributions," *Proc. Annu. ACM-SIAM Symp. Discrete Algorithms*, 2015, pp. 1841–1854.

[19] K. D. Ba, H. L. Nguyen, H. N. Nguyen, and R. Rubinfeld, "Sublinear time algorithms for earth mover's distance," *Theory Comput. Syst.*, vol. 48, no. 2, pp. 428–442, 2011.

[20] Y. I. Ingster, "Adaptive chi-square tests," *Zapiski Nauchnykh Seminarov POMI*, vol. 244, no. 2, pp. 150–166, 1997.

[21] D. Huang and S. Meyn, "Error exponents for composite hypothesis testing with small samples," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2012, pp. 3261–3264.

[22] N.G. de Bruijn, *Asymptotic Methods in Analysis* (Bibliotheca Mathematica). New York, NY, USA: Dover, 1981.

[23] F. Sossan, E. Namor, R. Cherkaoui, and M. Paolone, "Achieving the dispatchability of distribution feeders through prosumers data driven forecasting and model predictive control of electrochemical storage," *IEEE Trans. Sustain. Energy*, vol. 7, no. 4, pp. 1762–1777, Oct. 2016.

[24] X. Wang and L. Tong, "Innovations autoencoder and its application in one-class anomalous sequence detection," *J. Mach. Learn. Res.*, vol. 23, no. 49, pp. 1–27, 2022.

Authorized licensed use limited to: Cornell University Library. Downloaded on July 29,2024 at 21:49:11 UTC from IEEE Xplore. Restrictions apply.