Proceedings of the 2024 Winter Simulation Conference
H. Lam, E. Azar, D. Batur, S. Gao, W. Xie, S. R. Hunter, and M. D. Rossetti, eds.

ADVANCED TUTORIAL: LABEL-EFFICIENT TWO-SAMPLE TESTS

Weizhi Li[1,2], Visar Berisha[2], and Gautam Dasarathy[2]

[1]Los Alamos National Laboratory, Los Alamos, NM, USA
[2]Arizona State University, Tempe, AZ, USA

ABSTRACT

Hypothesis testing is a statistical inference approach used to determine whether data supports a specific hypothesis. An important type is the two-sample test, which evaluates whether two sets of data points are from identical distributions. This test is widely used, such as by clinical researchers comparing treatment effectiveness. This tutorial explores two-sample testing in a context where an analyst has many features from two samples, but determining the sample membership (or labels) of these features is costly. In machine learning, a similar scenario is studied in active learning. This tutorial extends active learning concepts to two-sample testing within this *label-costly* setting while maintaining statistical validity and high testing power. Additionally, the tutorial discusses practical applications of these label-efficient two-sample tests.

1 INTRODUCTION

Much scientific research involves conducting experiments and drawing conclusions from the results. This process is prevalent in fields such as physics, chemistry, biology, and other natural sciences, where scientists perform extensive experiments to verify hypotheses. To address the common pitfall of scientists making false discoveries or incorrectly claiming a hypothesis is true, statisticians have developed hypothesis testing, a statistical inference approach that evaluates whether a hypothesis is supported by experimental data.

Conducting experiments is often resource-intensive. For example, clinical trials in drug discovery can be very costly, running an emulator for physics experiments can demand significant computational power, and chemical experiments may involve lengthy reaction times. To address these challenges, the research community has proposed "data-efficient" hypothesis testing (Kartik et al. 2019; Naghshvar et al. 2013; Cohen and Zhao 2015; Naghshvar and Javidi 2010; Dasarathy et al. 2017; Cecchi and Hegde 2017), which aims to perform a limited number of experiments to make effective decisions. In this approach, an action space that characterizes the data sources (e.g., various experiments) is typically defined. The hypothesis testing procedure then adaptively selects an informative data source to facilitate decision-making regarding the hypotheses. This tutorial is focused on an important type of hypothesis testing, *two-sample hypothesis tests*, in a *label-costly* setting.

Two-sample testing: Two-sample tests are applied to data samples (or measurements) from two distributions to determine if the data support the hypothesis that the distributions are different. If we consider each data point as a feature and label (indicating which distribution the data is from) pair, then two-sample testing is equivalent to testing the dependence between the features and the labels. Viewing the problem through this lens, the null hypothesis for two-sample testing states that the feature and label variables are independent, while the alternative hypothesis states the opposite. An analyst performing a two-sample test is to decide between the null and alternative hypotheses based on the data from the two distributions. Depending on the flexibility in examining the two samples, she can choose to perform a *batch* two-sample test, where hypothesis testing is conducted only after all data collection is complete, with no continuous data collection and testing allowed. Alternatively, she can opt for *sequential* two-sample testing, where both data collection and hypothesis testing proceed sequentially and concurrently.

Label-costly setting: In traditional two-sample testing, the underlying assumption is that both the features

and their corresponding labels are simultaneously available. In this tutorial, we discuss two-sample testing in a rarely considered setting, where the measurements (or features) are readily accessible, but their groups (or labels) are unknown and difficult/costly to obtain. This scenario is analogous to active learning (Cohn et al. 1994) in machine learning community, leading to many label-efficient algorithms (Gal et al. 2017; Balcan et al. 2006; Dasarathy et al. 2015; Hanneke et al. 2014) for classification purposes. In contrast, this tutorial focuses on making decisions on hypotheses, i.e., hypothesis testing, which requires a different label query strategy to achieve label efficiency while effectively determining whether the null hypothesis is supported by the data.

Motivating application: Computer scientists and clinicians have collaborated to develop digital twins for simulating patient data in clinical trials. Identifying potential biomarkers indicative of disease severity is crucial before constructing these twins. Without this step, significant resources could be wasted on non-indicative biomarkers. Clinicians must validate these biomarkers through experiments, which can be costly. Efficient validation against expensive and labor-intensive labels is essential. Label-efficient two-sample tests address this challenge by enabling validation with a small subset of labeled data. A motivating example is the validation of digital biomarkers in Alzheimer's disease relative to imaging markers. Say we want to determine whether a series of digital biomarkers that can be collected at scale (e.g. gait, speech, typing speed measured using a patient's smartphone) is related to amyloid buildup in the brain (measured via neuroimaging, and an indication of increased risk of Alzheimer's disease). In this scenario, we can obtain the digital biomarkers on a large scale by distributing the tests via the internet. However, actually determining if a particular patient is amyloid positive (higher risk of Alzheimer's disease) or negative (lower risk) involves bringing participants in expensive and burdensome neuroimaging tests. It is of considerable interest to validate the digital biomarkers relative to the biomarker by only collecting neuroimaging data from a small number of participants. Notice that this scenario is in stark contrast to traditional formulations of two sample testing, where the class label (amyloid positivity) is assumed to be available for participants in the study.

Questions to address: Developing a legitimate label-efficient two-sample test requires addressing the following questions,

- What is an effective query scheme for the label querying?
- Does the developed test maintain statistical validity, i.e., can the error of rejecting the null hypothesis be controlled when the null is true?
- Does the developed test have high testing power, i.e., can it reject the null hypothesis with high probability when the alternative hypothesis is true?

In this tutorial, we discuss label-efficient two-sample testing and demonstrate that it is possible to achieve performance similar to traditional two-sample testing while minimizing the number of label queries. We will present label-efficient testing algorithms that can answer these questions with supporting theorems. The organization of this tutorial is as follows. In Section 2, we review the two-sample test. In section 3, we present the label-efficient two-sample testing problem. In Section 4, we introduce an effective label query scheme called the bimodal query. In Section 5 and 6, we present batch and sequential label-efficient two-sample tests, respectively.

2 REVIEW OF THE TWO-SAMPLE TESTING

2.1 A Traditional Two-Sample Testing Problem

Two-sample hypothesis testing evaluates whether two samples (or sets of data points) are generated from the same distribution (null hypothesis) or different distributions (alternative hypothesis). Let $\mathscr{X} = \{\mathbf{x}_1, \ldots \mathbf{x}_{n_0}\}$ and $\mathscr{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_{n_1}\}$ denote two samples of the realizations of *i.i.d.* random variables $\{\mathbf{X}_i\}_{i=1}^{n_0}$ and $\{\mathbf{Y}_i\}_{i=1}^{n_0}$ for $\mathbf{X} \sim p_{\mathbf{X}}(\mathbf{x})$ and $\mathbf{Y} \sim p_{\mathbf{Y}}(\mathbf{y})$. The null and alternative hypotheses are formulated as H_0 and H_1 in the

following:

$$H_0 : p_{\mathbf{X}}(\mathbf{x}) = p_{\mathbf{Y}}(\mathbf{y}),$$
$$H_1 : p_{\mathbf{X}}(\mathbf{x}) \neq p_{\mathbf{Y}}(\mathbf{y}). \tag{1}$$

Then, **the *traditional* two-sample testing problem is to decide whether to reject or retain H_0 after examining $\mathscr{X}$ and $\mathscr{Y}$.**

2.2 Desired Properties for the Traditional Two-Sample Testing

A conventional two-sample test is formulated as follows Johnson and Kuby 2011: (a) the analyst obtains two sets of data points $\mathscr{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{n_0}\}$ and $\mathscr{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_{n_1}\}$; (b) she computes a test statistic t from $\mathscr{X}$ and $\mathscr{Y}$; (c) she then computes the p-value for the statistic t under H_0 (both $\mathscr{X}$ and $\mathscr{Y}$ come from the same distribution). Formally, the p-value of a *valid* two-sample test represents the probability of obtaining test results at least as extreme as the test statistic t, assuming the null H_0 is true. In other words, a low p-value indicates that the test statistic t is unlikely to have been generated under H_0, suggesting that H_0 should be rejected. Typically, the analyst compares the p-value to a predefined significance-level α and reject H_0 if $p \leq \alpha$. Throughout this tutorial, we will consider the following desired properties for a two-sample test:
Valid p-value: Given $p_{\mathbf{X}}(\mathbf{x})$, $p_{\mathbf{Y}}(\mathbf{y})$, the resulting realization sets $\mathscr{X}$ and $\mathscr{Y}$ and a significance level α, a two-sample test generates a *valid* p-value computed from $\mathscr{X} \bigcup \mathscr{Y}$ if $P(p \leq \alpha) \leq \alpha$ when H_0 is true.
Consistency: Given $p_{\mathbf{X}}(\mathbf{x})$, $p_{\mathbf{Y}}(\mathbf{y})$, sets $\mathscr{X}$ and $\mathscr{Y}$ containing n total samples drawn from these distributions, and a significance level α, a two-sample test is *consistent* if $\lim_{n\to\infty} P(p \leq \alpha) = 1$ when H_1 is true.
High testing power: Given $p_{\mathbf{X}}(\mathbf{x})$, $p_{\mathbf{Y}}(\mathbf{y})$, $n = |\mathscr{X}| + |\mathscr{Y}|$, the resulting realization sets $\mathscr{X}$ and $\mathscr{Y}$ and a significance level α, a two-sample test has high testing power if $P(p \leq \alpha)$ is high for the finite n when H_1 is true. In other words, a two-sample test is more powerful than another if its $P(p \leq \alpha)$ is higher given a specific n.

Here, we write $P_0(p \leq \alpha)$ and $P_1(p \leq \alpha)$ to denote $P(p \leq \alpha)$ under H_0 and H_1, respectively. $P_0(p \leq \alpha)$ represents the *Type I error*, which is the probability of making a wrong decision when H_0 is true, while $1 - P_1(p \leq \alpha)$ represents the *Type II error*, which is the probability of making a wrong decision when H_1 is true. A valid two-sample test (i.e., one that produces a valid p-value) can upper-bound its Type I error P_0 by a predefined significance level α. On the other hand, the properties of consistency and high testing power serve to reduce the Type II error $1 - P_1$.

2.3 Classical Two-Sample Tests

The development of two-sample tests has a rich history (Friedman and Rafsky 1979; Chen and Friedman 2017; Hotelling 1992; Friedman 2004; Vayatis et al. 2009; Lhéritier and Cazals 2018; Hajnal 1961). In this section, we introduce representative examples of the non-parametric two-sample test and the sequential non-parametric two-sample test. In section 3, we will see how label-efficient two-sample tests are extended from these classical two-sample tests.

2.3.1 Nonparametric Two-Sample Testing

This section introduces the Friedman-Rafsky (FR) test (Friedman and Rafsky 1979), a non-parametric, graph-based test to determine whether two multivariate samples are realization of the same distribution. The null and alternative hypotheses and the notations used are stated in (1).
The FR test statistic: The FR test statistic is computed as follows. First, one constructs a Euclidean minimum spanning tree (MST) over the samples $\mathscr{X} \bigcup \mathscr{Y}$, i.e., the MST of a complete graph whose vertices are the samples, and edge weights are the Euclidean distance between the samples. Then, one counts the edges connecting samples from opposite classes (i.e., cut edges). We use r_n to denote the cut-edge number for the MST constructed over two samples $\mathscr{X} \bigcup \mathscr{Y}$ with a total size of n, and use R_n to denote the corresponding random variable. Under the alternative H_1, r_n is expected to be small, and under the null

H_0, r_n is expected to be large. The FR test statistic w_n is a normalized version of r_n,

$$w_n = \frac{r_n - \mathbb{E}[R_n \mid H_0]}{\sqrt{\mathrm{Var}[R_n \mid H_0]}}, \tag{2}$$

where $\mathbb{E}[R_n \mid H_0]$ and $\mathrm{Var}[R_n \mid H_0]$ are the expectation and the variance of R_n under the null H_0. We use W_n to denote a random variable of which w_n is a realization. The FR test rejects H_0 if a small W_n is observed.
A permutation test: In practice, as stated in Friedman and Rafsky 1979, the FR test is carried out as a permutation test where the null distribution (distribution of a statistic under the null H_0) of W_n is obtained by calculating all possible values of w_n (2) under all possible rearrangements of the observations of $\mathscr{X} \bigcup \mathscr{Y}$. Then a p-value is obtained using the permutation null distribution and the w_n computed from $\mathscr{X}$ and $\mathscr{Y}$. The p-value is compared to a significance level α to reject H_0 for $p \leq \alpha$. We refer readers to Welch 1990 for the procedure of the permutation test.
Approximation of the permutation test: Both Theorem 4.1.2 in Bloemena 1964 and Section 4 in Friedman and Rafsky 1979 demonstrate that, if W_n is generated under H_0, then the *permutation distribution* of W_n approaches a standard normal distribution for large sample size $n \to \infty$: $W_n \xrightarrow{\mathscr{D}} \mathscr{N}(0,1)$, where $\xrightarrow{\mathscr{D}}$ stands for distributional convergence. Therefore, a valid p-value for the FR test is given by

$$p = \Phi[W_n], \tag{3}$$

An analyst can construct the FR statistic w_n to generate the p-value and compare the p-value with α to decide H_0 or H_1. The Type I error is upper-bounded by α.
Consistency: The consistency of the FR test is proved in Henze and Penrose 1999 based on the following.
Theorem 1 (Henze and Penrose 1999) Given $\mathscr{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_{n_0}\}$ and $\mathscr{Y} = \{\mathbf{y}_1, \cdots, \mathbf{y}_{n_1}\}$ which are *i.i.d.* realizations of $\mathbf{X} \sim p_{\mathbf{X}}(\mathbf{x})$ and $\mathbf{Y} \sim p_{\mathbf{Y}}(\mathbf{y})$, and $n = n_0 + n_1$, suppose $\lim_{n\to\infty} \frac{n_0}{n} = u$, $\lim_{n\to\infty} \frac{n_1}{n} = v$ where $u + v = 1$, then

$$\frac{R_n}{n} \overset{a.s.}{\to} 2uv \int \frac{p_{\mathbf{X}}(\mathbf{s}) p_{\mathbf{Y}}(\mathbf{s})}{u p_{\mathbf{X}}(\mathbf{s}) + v p_{\mathbf{Y}}(\mathbf{s})} d\mathbf{s}. \tag{4}$$

Under H_0, where $p_{\mathbf{X}}(\mathbf{s}) = p_{\mathbf{Y}}(\mathbf{s})$, the R.H.S of (4) is strictly larger than that under H_1, where $p_{\mathbf{X}}(\mathbf{s}) \neq p_{\mathbf{Y}}(\mathbf{s})$, resulting in the consistency of the FR test.

2.3.2 Sequential Nonparametric Two-Sample Testing

FR (Friedman and Rafsky 1979) is a *batch* two-sample test, where two samples, $\mathscr{X}$ and $\mathscr{Y}$ are collected, and then a two-sample hypothesis test is performed on $\mathscr{X} \bigcup \mathscr{Y}$ to decide on H_0 or H_1. In this method, *continuous data collection and hypothesis testing are not allowed*. The research community has proposed sequential two-sample tests (Wald 1992; Lhéritier and Cazals 2018; Hajnal 1961; Shekhar and Ramdas 2021; Balsubramani and Ramdas 2015) that allow the analyst to sequentially collect data and monitor statistical evidence, i.e., a statistic computed from the data. The test can stop anytime when sufficient evidence has been accumulated to make a decision.

In this section, we present a *sequential nonparametric two-sample test* constructed by Lhéritier and Cazals 2018. We consider the observed data a *measurement sequence* $\mathscr{S} = \mathscr{X} \bigcup \mathscr{Y} = \{\mathbf{s}_i\}_{i=1}^n$ and a corresponding *label sequence* $\mathscr{Z} = \{z_1, \ldots, z_n\}$, where $z_i = 0$ if $\mathbf{s}_i \in \mathscr{X}$ and 1 otherwise. Accordingly, our observation model is n *i.i.d.* draws from the joint distribution $p_{\mathbf{S}Z}(\mathbf{s}, z)$. The two sample testing problem under this formulation is equivalent to testing if $p_{\mathbf{S}|Z}(\cdot \mid 0) = p_{\mathbf{S}|Z}(\cdot \mid 1)$ (i.e., $\mathbf{S}$ and Z are independent).
(Lhéritier and Cazals 2018) test statistic: Lhéritier and Cazals 2018 assumes the class prior $P_Z(z)$ is known in the problem setting. Additionally, they build a *nonparametric* class-probability predictor $Q(z \mid \mathbf{s})$ to model $P_{Z|\mathbf{S}}(z \mid \mathbf{s})$. Specifically, $Q_n(z \mid \mathbf{s})$ represents a class-probability predictor built with the past observed

sample sequence $\{\mathbf{s}_i, z_i\}_{i=1}^{n-1}$ to model $P_{Z|\mathbf{S}}(z_n \mid \mathbf{s}_n)$–the posterior probability of z_n given newly observed $\mathbf{s}_n$. Any nonparametric probabilistic classifier, such as a decision tree and KNN classifier, can be used to build $Q_n(z \mid \mathbf{s})$. The following statistic is then constructed,

$$t_n = \frac{P_Z(z^n)}{Q(z^n \mid \mathbf{s}^n)} = \prod_{i=1}^{n} \frac{P_Z(z_i)}{Q_i(z_i \mid \mathbf{s}_i)}. \tag{5}$$

The statistic (5) stems from the well-known likelihood ratio test (Wilks 1938). Here, the likelihood under H_0 where $\mathbf{S} \perp\!\!\!\perp Z$ is $\prod_{i=1}^{n} p_{\mathbf{S}}(\mathbf{s}_i) P_Z(z_i)$ and the likelihood under H_1 where $\mathbf{S} \not\perp\!\!\!\perp Z$ is $\prod_{i=1}^{n} P_{Z|\mathbf{S}}(z_i \mid \mathbf{s}_i) p_{\mathbf{S}}(s_i)$. Replacing $P_{Z|\mathbf{S}}$ with an approximation Q in the likelihood ratio $\prod_{i=1}^{n} \frac{p_{\mathbf{S}}(\mathbf{s}_i) P_Z(z_i)}{P_{Z|\mathbf{S}}(z_i|\mathbf{s}_i) p_{\mathbf{S}}(s_i)}$ leads to the (Lhéritier and Cazals 2018) statistic in (5).

Anytime-valid p-value: Typically, a sequential test iteratively collects $(\mathbf{s}, z)$, adds $(\mathbf{s}, z)$ to the measurement and label sequences $\mathscr{S}$ and $\mathscr{Z}$, and monitors the p-values computed from the sequences for every n. Implementing the FR in a sequential way generates *invalid* p-values, meaning that $P(p \leq \alpha)$ is not necessarily upper-bounded by α for every n. This occurs when multiple p-values are generated without *adjustment* for deciding between H_0 and H_1, the probability of observing a rare event increases, thereby raising the likelihood of incorrectly rejecting H_0 (i.e., making a Type I error); alternatively, one may make the test extremely weak by using a very conservative Bonferroni correction (union bound) (Dunn 1961) across all time steps. In contrast, the (Lhéritier and Cazals 2018) statistic in (5) produces the *anytime-valid* p-value stated in the following theorem,

Theorem 2 (Lhéritier and Cazals 2018) An analyst sequentially collects $(\mathbf{s}, z)$ resulting in $\{\mathbf{s}_i, z_i\}_{i=1}^{n}$ which are *i.i.d.* realizations of $(\mathbf{S}, Z) \sim p_{\mathbf{S}Z}(\mathbf{s}, z)$, and computes the statistic t_n as in (5) with $\{\mathbf{s}_i, z_i\}_{i=1}^{n}$. Under H_0, the following holds for the random variable T_n from which t_n is realized,

$$P_0\left(\exists n \geq 1, T_n = \prod_{i=1}^{n} \frac{P(Z_i)}{Q_i(Z_i \mid \mathbf{S}_i)} \leq \alpha\right) \leq \alpha. \tag{6}$$

Theorem 2 implies that t_n is an anytime-valid p-value and rejects H_0 whenever $t_n \leq \alpha$ leading to a Type I error upper-bounded by α. Briefly, it is proved by observing that the sequence $\left(\frac{1}{T_1}, \cdots, \frac{1}{T_n}\right)$ is a non-negative martingale, and hence Ville's maximal inequality (Durrett 2019; Doob 1939) can be used to bound the tail probability of the infimum of an infinite $\{T_i\}_{i=1}^{n}$ to develop Theorem 2.

Consistency: The authors of Györfi et al. 2002 prove that when $Q(z \mid \mathbf{s})$ is a kernel, KNN or partition estimates with proper smoothing parameters (e.g., bandwidth for the kernel), then $Q_n(z \mid \mathbf{s})$ converges to $P(z \mid \mathbf{s})$. Then, the following theorem holds:

Theorem 3 When $Q(z \mid \mathbf{s})$ is a kernel, KNN or partition estimate with proper smoothing parameters (e.g., bandwidth for the kernel), then

$$\lim_{n\to\infty} \frac{\log T_n}{n} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \log \frac{P_Z(Z_i)}{Q_i(Z_i \mid \mathbf{S}_i)} = -\left(H(Z) - H(Z \mid \mathbf{S})\right) = -I(\mathbf{S}, Z) \tag{7}$$

where $H(Z)$, $H(Z \mid \mathbf{S})$, and $I(\mathbf{S}; Z)$ are the entropy, conditional entropy and mutual information for $(\mathbf{S}, Z) \sim p_{\mathbf{S}Z}(\mathbf{s}, z)$.

Theorem 3 implies that when Q is built with a proper nonparametric method, $\frac{\log T_n}{n}$ converges to the negation of mutual information, which is smaller than zero under H_1 where $\mathbf{S} \not\perp\!\!\!\perp Z$, leading to $\lim_{n\to\infty} P_1(T_n \leq \alpha) = 1$.

3 THE LABEL-EFFICIENT TWO-SAMPLE TESTING PROBLEM

To motivate our *label-efficient* two-sample testing problem, we use a pair of random variables $(\mathbf{S}, Z)$ to denote a feature and its label variables whose realization is $(\mathbf{s}, z) \in \mathbb{R}^d \times \{0,1\}$. The variable pair $(\mathbf{S}, Z)$ admits a joint distribution $p_{\mathbf{S}Z}(\mathbf{s}, z)$. Furthermore, we write $\mathbb{S}$ to denote the support of $p_{\mathbf{S}}(\mathbf{s})$.

In the traditional two-sample testing problem, discussed in section 2.1, it is assumed that both features and their corresponding labels are available simultaneously. This tutorial introduces a novel problem formulation where an analyst has free access to an unlabeled feature set $\mathscr{S}_u = \{\mathbf{s}_i\}_{i=1}^n$, but obtaining the corresponding labels z_i in the label set $\mathscr{Z}$ is costly. The analyst is provided with a label budget $N_q \leq n$, allowing her to select a subset $\overline{\mathscr{S}} \subseteq \mathscr{S}_u$ with at most N_q features, for which an oracle returns the corresponding label set $\overline{\mathscr{Z}} \subseteq \mathscr{Z}$. The original unlabeled feature set $\mathscr{S}_u$ consists of realizations of *i.i.d.* variables $\mathbf{S} \sim p_{\mathbf{S}}(\mathbf{s})$, but the selected $\overline{\mathscr{S}}$ *does not* necessarily contain *i.i.d.* variables.

The label-efficient two sample testing problem under this formulation aims to test the following null and alternative hypotheses:

$$H_0 : p_{S|Z}(\cdot \mid 0) = p_{S|Z}(\cdot \mid 1) \qquad H_1 : p_{S|Z}(\cdot \mid 0) \neq p_{S|Z}(\cdot \mid 1),$$

or equivalently, to test the independence between $\mathbf{S}$ and Z:

$$H_0 : p_{\mathbf{S}Z}(\mathbf{s}, z) = p_{\mathbf{S}}(\mathbf{s}) P_Z(z), \forall \mathbf{s} \in \mathbb{S} \qquad H_1 : p_{\mathbf{S}Z}(\mathbf{s}, z) \neq p_{\mathbf{S}}(\mathbf{s}) P_Z(z), \exists \mathbf{s} \in \mathbb{S}$$

by examining $\overline{\mathscr{S}}$ and $\overline{\mathscr{Z}}$, whose sizes do not exceed the label budget N_q. Moving forward, we omit the subscripts in $p_{\mathbf{S}Z}(\mathbf{s}, z)$, $P_Z(z)$ and $p_{\mathbf{S}}(\mathbf{s})$ and write them as $p(\mathbf{s}, z)$, $P(z)$ and $p(\mathbf{s})$. In addition, we use $\mathbf{s}^N$, z^N and $(\mathbf{s}, z)^N$ to denote sequences of samples $\{\mathbf{s}_i\}_{i=1}^N$, $\{z_i\}_{i=1}^N$ and $\{(\mathbf{s}, z)_i\}_{i=1}^N$ respectively.

4 BIMODAL QUERY

Our label-efficient tests employ a probabilistic classification-based query scheme called the *bimodal query* to determine which features to select from an unlabeled feature set. This query scheme was first proposed, to the best of our knowledge, in Li et al. 2022. Let $\overline{\mathscr{S}}$ and $\overline{\mathscr{Z}}$ be the features and their revealed labels. The bimodal query uses a class-probability predictor $Q(z \mid \mathbf{s})$ and the unlabeled set $\mathscr{S}_u$ as input. This classifier $Q(z \mid \mathbf{s})$ is trained using $\overline{\mathscr{S}}$ and $\overline{\mathscr{Z}}$. The bimodal query then selects two features, $s_{q_0} \in \mathscr{S}_u$ and $s_{q_1} \in \mathscr{S}_u$ to query for their labels, where $s_{q_0} = \arg\max_{\mathbf{s}} [Q(Z=0|\mathbf{s})]$ and $s_{q_1} = \arg\max_{\mathbf{s}} [Q(Z=1|\mathbf{s})], \forall \mathbf{s} \in \mathscr{S}_u$. In other words, the features predicted to have the highest class one or zero posterior probabilities are labeled, as these features lie in the most "informative" region.

The authors of Györfi et al. 2002 prove that when $Q(z \mid \mathbf{s})$ is a kernel, KNN or partition estimates with proper smoothing parameters (e.g., bandwidth for the kernel) and labels are sufficiently revealed in the proximity of $\mathbf{s}, \forall \mathbf{s} \in \mathbb{S}$, then $Q(z \mid \mathbf{s})$ converges to $P(z \mid \mathbf{s})$. A bimodal query using Q with such a convergence property leads to the following,

Definition 1 (*Consistent bimodal query*) Let $\mathbb{S}$ be the support of $p(\mathbf{s})$ and let $\mathscr{S}_u$ be an unlabeled set where sample features are sampled *i.i.d.* from $p(\mathbf{s})$, and let $P(z \mid \mathbf{s})$ denote the posterior probability of z given $\mathbf{s} \in \mathbb{S}$. An analyst adopts a label query scheme, for every $n > 0$, to query the label Z_n of $\mathbf{S}_n \in \mathscr{S}_u$ such that $\mathbf{S}_n$ admits a probability density function (pdf) $p_n(\mathbf{s})$. The label query scheme is a *consistent bimodal query* if $\lim_{n\to\infty} p_n(\mathbf{s}) = p^*(\mathbf{s})$ where

$$\begin{aligned} p^*(\mathbf{s}) &= 0, \forall \mathbf{s} \in \mathbb{S} \setminus \left(\mathbb{S}_{q_0} \bigcup \mathbb{S}_{q_1}\right), \text{ and } p^*(\mathbf{s}) > 0, \forall \mathbf{s} \in \mathbb{S}_{q_0} \bigcup \mathbb{S}_{q_1}, \\ \mathbb{S}_{q_0} &= \left\{ \mathbf{s}_{q_0} \middle| P(Z=0 \mid \mathbf{s}_{q_0}) = \max_{\mathbf{s}\in\mathbb{S}} P(Z=0 \mid \mathbf{s}) \right\}, \\ \mathbb{S}_{q_1} &= \left\{ \mathbf{s}_{q_1} \middle| P(Z=1 \mid \mathbf{s}_{q_1}) = \max_{\mathbf{s}\in\mathbb{S}} P(Z=1 \mid \mathbf{s}) \right\}. \end{aligned} \tag{8}$$

5 A BATCH LABEL-EFFICIENT TWO-SAMPLE TEST

In this section, we propose a three-stage framework for a batch label-efficient two-sample test and outline its statistical properties. We refer interested readers to Li et al. 2022 for a more comprehensive presentation.

5.1 A Three-Stage Two-Sample Testing Framework

The algorithmic description of the three-stage testing framework is listed in Algorithm 1. The inputs of the algorithm 1 are as follows: an unlabeled feature set $\mathscr{S}_u$, a classification algorithm $\mathscr{A}$ that takes a training set as input and outputs a classifier, the number N_0 of labels used to construct a training set, the label budget N_q and a pre-defined significance level α. The output of algorithm 1 is a single bit of information: was the null hypothesis H_0 rejected? During the **first stage**, a classification algorithm $\mathscr{A}$ takes N_0 uniformly labeled samples (and corresponding labels provided by the oracle) as a training set input, and outputs a class-probability estimation function $Q(z \mid \mathbf{s})$ used to model $P(z|\mathbf{s})$ subsequently. During the **second stage**, we propose a bimodal query algorithm that queries the labels of samples with highest class one probability $Q(1 \mid \mathbf{s})$ and highest class zero probability $Q(0 \mid \mathbf{s})$ until the label query budget, N_q, is exhausted. During the **third stage**, we split a labeled feature set $\overline{\mathscr{S}}$ to two samples $\overline{\mathscr{X}}$ and $\overline{\mathscr{Y}}$, where each set only contains features from one class. Then the FR two-sample test is performed with the following steps: (1) compute the FR statistic (see Section 2.3.1) from $\overline{\mathscr{X}}$ and $\overline{\mathscr{Y}}$; (2) compute p-value; (3) rejects the null hypothesis if the p-value is smaller than the pre-defined significance level α.

Algorithm 1 A three-stage framework for the batch label-efficient two-sample testing

input $\mathscr{S}, N_0, N_q, \alpha, \mathscr{A}$
output Reject or accept H_0

First stage: model $P(Z=1|\mathbf{s})$
Uniformly sample N_0 features $\overline{\mathscr{S}} \subset \mathscr{S}_u$ and query their labels $\overline{\mathscr{Z}}$; $\mathscr{S}_u = \mathscr{S}_u / \overline{\mathscr{S}}$;
$\mathscr{A}$ takes input $\overline{\mathscr{S}}$ and $\overline{\mathscr{Z}}$, and outputs class-probability predictor $Q(z \mid \mathbf{s})$ used to model $P(z \mid \mathbf{s})$;

Second stage: bimodal query
Select $\lfloor (N_q - N_0)/2 \rfloor$ features $\overline{\mathscr{S}}_0 \subseteq \mathscr{S}_u$ which corresponds to $\lfloor (N_q - N_0)/2 \rfloor$ highest $Q(0 \mid \mathbf{s})$, and query their labels $\overline{\mathscr{Z}}_0$;
Select $N_q - N_0 - \lfloor (N_q - N_0)/2 \rfloor$ features $\overline{\mathscr{S}}_1 \subseteq \mathscr{S}_u$ which corresponds to $N_q - N_t - \lfloor (N_q - N_0)/2 \rfloor$ highest $Q(1 \mid \mathbf{s})$, and query their labels $\overline{\mathscr{Z}}_1$;
$\overline{\mathscr{S}} = \overline{\mathscr{S}} \bigcup \overline{\mathscr{S}}_0 \bigcup \overline{\mathscr{S}}_1$; $\overline{\mathscr{Z}} = \overline{\mathscr{Z}} \bigcup \overline{\mathscr{Z}}_0 \bigcup \overline{\mathscr{Z}}_1$

Third stage: FR two-sample test
Split $\overline{\mathscr{S}}$ to two samples $\overline{\mathscr{X}}$ and $\overline{\mathscr{Y}}$ based on the label set $\overline{\mathscr{Z}}$; compute FR statistic using $\overline{\mathscr{X}}$ and $\overline{\mathscr{Y}}$; compute p-value;
If $p < \alpha$ **Then** Reject H_0 **Else** Accept H_0.

5.2 Consistent Bimodal Query Minimizes the FR Statistic W_n

Our problem statement in Section 3 assumes that the original unlabeled feature set $\mathscr{S}_u$ includes i.i.d realizations of $\mathbf{S} \sim p(\mathbf{s})$, and that the access to every $\mathbf{s}_i \in \mathscr{S}$ is free; but it is costly to obtain the corresponding label $z_i \in \mathscr{Z}$. However, we are assigned a label budget N_q such that we can select a set $\overline{\mathscr{S}} \subseteq \mathscr{S}$ to query labels from an oracle, and each random variable Z_i corresponding to the returned label z_i admits $P(z|\mathbf{s}_i)$. We then divide $\overline{\mathscr{S}}$ to $\overline{\mathscr{X}}$ from class zero and $\overline{\mathscr{Y}}$ from class one and perform a two-sample test on $\overline{\mathscr{X}}$ and $\overline{\mathscr{Y}}$. We write $|\overline{\mathscr{X}}| = \overline{n}_0$ and $|\overline{\mathscr{Y}}| = \overline{n}_1$ and we have $N_q = \overline{n}_0 + \overline{n}_1$.

Our aim is to find a query scheme that increases the testing power of a test performed on the selected samples $\overline{\mathscr{X}}$ and $\overline{\mathscr{Y}}$. For a uniform sampling query scheme, then we will have $\overline{\mathscr{S}}$ as a set of N_q *i.i.d* realizations generated from the original marginal distribution $p(\mathbf{s})$, and we can rewrite p-value in (3) as

$p = \Phi[W_{N_q}]$ where W_{N_q} is a FR statistic random variable obtained from N_q *i.i.d.* pairs of $(\mathbf{S}_i, Z_i) \sim p(\mathbf{s}, z)$. Instead of directly tackling the query scheme, we consider to find an optimal marginal distribution $p_q(\mathbf{s})$ such that, under H_1, performing the FR test on a set of *i.i.d.* $\mathbf{S} \sim p_q(\mathbf{s})$ generates large testing power than performing on the uniformly sampled data points with the same number of labels N_q.

Under H_0, the feature variable $\mathbf{S}$ is always independent of label variable $\mathbf{Z}$ and the Type I error can be controlled by running a permutation test or approximation of that regardless of $p_q(\mathbf{s})$. Here, we focus on $p_q(\mathbf{s})$ under H_1. Specifically, given N_q i.i.d. realizations generated from $p_q(\mathbf{s})$, we seek a $p_q(\mathbf{s})$ to minimize the FR statistic W_{N_q} in (2), thereby creating a more powerful FR test under H_1. Before presenting our theoretical analysis, we need the following proposition, which follows from Theorem 1.

Proposition 1 Suppose $(\mathbf{S}, Z)^n$ comprises *i.i.d.* pairs $\{\mathbf{S}, Z\} \sim p(\mathbf{s}, z)$, and let $u = P(Z=0)$ and $v = P(Z=1)$. Then, the following holds for a normalized FR statistic $\frac{W_n}{n}$ generated from $(\mathbf{S}, Z)^n$,

$$\frac{W_n}{n} \xrightarrow{a.s.} \frac{[\int 2P(Z=0 \mid \mathbf{s})p(Z=1 \mid \mathbf{s})p(\mathbf{s})d\mathbf{s} - 2uv]}{\sqrt{2uv[2uv + (A_d - 1)(1-4uv)]}} \tag{9}$$

where A_d is a constant dependent on the dimension d of $\mathbf{S}$.

We direct readers to Section 7 in Li et al. 2022 for the proof. From Proposition 1 we know that the convergence of $\frac{W_{N_q}}{N_q}$ only depends on $p(\mathbf{s})$. Therefore, we construct the following optimization problem:

$$\begin{aligned} &\min_{p_q(\mathbf{s})} \int P(Z=0 \mid \mathbf{s})P(Z=1 \mid \mathbf{s})p_q(\mathbf{s})d\mathbf{s} \\ &\text{subject to} \int P(Z=0 \mid \mathbf{s})p_q(\mathbf{s})d\mathbf{s} = u, \quad \int p_q(\mathbf{s})d\mathbf{s} = 1, \quad p_q(\mathbf{s}) \geq 0. \end{aligned} \tag{10}$$

Under the null hypothesis H_0, Z and S are independent and thus $p(\mathbf{s}, z) = p(\mathbf{s})p(z)$, and $\int P(Z=0|\mathbf{s})P(Z=1|\mathbf{s})p_q(\mathbf{s})ds = uv$ for any $p_q(\mathbf{s})$. Therefore, minimizing (10) with $p_q(\mathbf{s})$ does not alter the Type I error. On the other hand, under the alternate H_1, solving the optimization problem (10) leads to a solution that minimizes W_{N_q} in (3) for large sample sizes $N_q \to \infty$, leading to a decreasing Type II error of the FR test.

Theorem 4 The optimal solution to (10) is $p^*(\mathbf{s})$ defined in (8).

We direct readers to Section 8 in Li et al. 2022 for the proof. Briefly, the proof of Theorem 4 comes about when we combine the linear constraints in Eq. (10) with the fact that the optimum value is always achieved on the boundary of the constraint set for Linear Programming (LP) (Korte et al. 2011). The optimal solution $p^*(\mathbf{s})$ of (10) is a pdf that samples the highest posterior probabilities of $P(Z=0|\mathbf{s})$ and $P(Z=1|\mathbf{s})$.

Remark 1 Let $\mathbb{S}$ be the support of $p(\mathbf{s})$ and let $\mathscr{S}_u$ be an unlabeled set where sample features are *i.i.d.* sampled from $p(\mathbf{s})$. Suppose $p_n(\mathbf{s})$ denotes the pdf of $\mathbf{S}_n$, whose label Z_n is queried by the *consistent bimodal query* (see Definition 1) for every $n > 0$. Then $p_n(\mathbf{s})$ converges to the optimal solution of (10). In the finite-sample regime, we use a bimodal query where the class-probability predictor $Q(z \mid \mathbf{s})$ is an approximation of $P(z \mid \mathbf{s})$ to query the labels of features in the regions with high posterior probability for class one and zero.

5.3 Type I Error of the Three-Stage Framework

One important observation for the proposed framework is that the features labeled in the second stage are dependent on the uniform sampled features in the first stage. For every n i.i.d. realizations $(\mathbf{s}, z)^n$ of $(\mathbf{S}_i, Z_i)^n \sim p(\mathbf{s}, z)$ under the null hypothesis H_0, we write $\overline{\mathscr{S}} = \{\bar{\mathbf{s}}_1, \ldots, \bar{\mathbf{s}}_{N_q}\}$ to denote a set that our query scheme (comprised of uniform sampling and bimodal query) selects from the original unlabelled set $\mathscr{S}_u$, and write $\overline{\mathscr{Z}} = \{\bar{z}_1, \ldots, \bar{z}_{N_q}\}$ to denote a set of label observations corresponding to $\overline{\mathscr{S}}$. We use $\bar{\mathbf{S}}_i$ and $\bar{Z}_i$ to denote the random variables corresponding to $\bar{\mathbf{s}}_i$ and $\bar{z}_i$. In the following, we present our theorem regarding the Type I error control:

Theorem 5 Suppose $\left(\bar{\mathbf{S}},\bar{Z}\right)^{N_q}$ are pairs of random feature variables and label variables acquired in the end of the second stage of the framework, using a permutation test in the third stage of the framework to obtain p-value from $\left(\bar{\mathbf{S}},\bar{Z}\right)^{N_q}$ for any two-sample test have Type I error $P(p \leq \alpha) \leq \alpha, \forall \alpha$ for the framework.

We direct readers to Section 10 in Li et al. 2022 for a detailed proof. Theorem 5 states that the Type I error of our framework is upper-bounded by α for any two-sample test carried out as a permutation test in the third stage. A permutation test rearranges labels of features, obtains permutation distribution of a statistic computed from the rearrangements, and rejects H_0 if a true observed statistic is contained in α probability range of the permutation distribution. This process does not need features to be *i.i.d.* sampled to control the Type I error at exact α, and it is applicable to any two-sample tests testing independency between $\bar{\mathbf{S}}_i$ and $\bar{Z}_i$. However, we need to make sure our query procedure maintains $\bar{\mathbf{S}}_i \perp\!\!\!\perp \bar{Z}_i$ under the H_0. Our framework only trains a classifier one time with uniformly sampled data points in the first stage, and then the bimodal query selects a subset of features from $\mathscr{S}$ to label based on the trained classifier. For a set of feature and label variables $\mathscr{Q} = \{\bar{\mathbf{S}}_i, \bar{Z}_i\}_{i=1}^{N_q}$ obtained in the end of the second stage, we write $\mathscr{Q}_u \subseteq \mathscr{Q}$ to denote the set obtained from uniform sampling, and write $\mathscr{Q}_b \subseteq \mathscr{Q}$ to denote the set obtained from bimodal query. Considering that a uniform sampling scheme does not change the original distributional properties ($\mathbf{S} \perp\!\!\!\perp Z$ under the null) to generate $(\bar{\mathbf{S}}_i, \bar{Z}_i) \in \mathscr{Q}_u$, we have $\bar{\mathbf{S}}_i \perp\!\!\!\perp \bar{Z}_i, \forall (\bar{\mathbf{S}}_i, \bar{Z}_i) \in \mathscr{Q}_u$. $\mathscr{Q}_b$ is not used to train the classifier, so we also have $\bar{\mathbf{S}}_i \perp\!\!\!\perp \bar{Z}_i, \forall (\bar{\mathbf{S}}_i, \bar{Z}_i) \in \mathscr{Q}_b$.

6 A SEQUENTIAL LABEL-EFFICIENT TWO-SAMPLE TEST

The batch design discussed in Section 5 invariably exhausts the label budget. Fixing the number of labels can lead to inefficiencies: if the two-sample testing problem is difficult, additional label query for more evidence may be required for a final decision, whereas if the problem is simple, the test may collect excessive labels. Therefore, this section introduces a sequential label-efficient two-sample testing framework that adapts the number of label queries based on the problem's complexity while maintaining statistical validity. For a more comprehensive presentation, we refer interested readers to Li et al. 2024.

6.1 A Sequential Label-Efficient Framework

Suppose N_q is a label budget and α is a significance level. An analyst uses the proposed framework to sequentially and actively query the label z_n of $\mathbf{s}_n$ from an unlabelled feature set $\mathscr{S}_u$ based on the predictions of $Q_n(z \mid \mathbf{s})$. As a new z_n of $\mathbf{s}_n$ is queried, the analyst constructs the following statistic,

$$u_n = \prod_{i=1}^{n} \frac{\hat{P}(z_i)}{Q_i(z_i \mid \mathbf{s}_i)}. \tag{11}$$

The analyst evaluates w_n and makes one of the following decisions: (1) rejects H_0 if $u_n \leq \alpha$; (2) retains H_0 if the label budget N_q is exhausted and (1) is not satisfied; and (3) continues the test and updates Q_n to Q_{n+1} if (1) and (2) are not satisfied.

Figure 1: The sequential label-efficient framework

In Li et al. 2024, we introduced a novel framework for the sequential label-efficient two-sample test. We use a statistic u_n slightly modified from the sequential testing statistic v_n (5). Given that labels are costly to access in the problem setting (See Section 3), the class prior $P(z)$ required to compute (5) is also unknown. Therefore, *we will use a likelihood estimate $\hat{P}(z^n)$, maximized over all class priors, to replace $P(z^n)$—the product of the class prior in* (5). This results in $u_n = \frac{\hat{P}(z^n)}{Q(z^n|\mathbf{s}^n)} = \prod_{i=1}^{n} \frac{\hat{P}(z_i)}{Q_i(z_i|\mathbf{s}_i)}$ where $\hat{P}(Z=1) = \frac{\sum_{i=1}^{n} z_i}{n}$. Our

framework begins by labeling sample features randomly selected from from $\mathscr{S}_u$, the unlabeled feature set, and then initializing the class-probability predictor $Q_n(z \mid \mathbf{s})$ at $n = 1$ with the labeled features. Then, the framework enters the sequential testing stage to sequentially query label, update and monitor the statistic u_n. We formally introduce our framework in Fig 1.

We provide a *framework instantiation* called *bimodal query based active sequential two-sample testing (BQ-AST)* described in Algorithm 2. The algorithm takes the following input: an unlabelled feature set $\mathscr{S}_u$, a probabilistic classification algorithm $\mathscr{A}$, the size N_0 of an initialization set used for $\mathscr{A}$, a label budget N_q and a significance level α. Then, the algorithm initializes a class-probability predictor Q using $\mathscr{A}$ with a small set of randomly labeled samples. In the sequential testing stage, the algorithm uses bimodal query discussed in Section 4 to sample $\mathbf{s}_n$ with samples having the highest posteriors from either class (e.g. a fair chance to select the highest $Q_n(Z = 0 \mid \mathbf{s})$ or $Q_n(Z = 1 \mid \mathbf{s})$) from $\mathscr{S}_u$, queries its label z_n and updates the statistic w_n. Next, the algorithm compares w_n with α and make a decision, e.g., reject H_0 or continue testing. *In what follows in this section, we simply use N_q to denote the "label budget" allowed to be used after the initialization.*

Algorithm 2 Bimodal Query Based Active Sequential Two-Sample Testing (BQ-AST)

Input: $\mathscr{S}_u, \mathscr{A}, N_0, N_q, \alpha$
Output: Reject or fail to reject H_0
Initialization: Initialize $Q_1(z \mid \mathbf{s})$ using $\mathscr{A}$ with N_0 features uniformly sampled from $\mathscr{S}_u$ without replacement and then labeled.
Active Sequential testing:
for $n = 1$ to $N_q - N_0$ **do**
    Sample a feature $\mathbf{s}_n = \mathbf{s}_{q_0}$ or $\mathbf{s}_{q_1}$ with **fair chance** where $\mathbf{s}_{q_0} = \arg\max_{\mathbf{s}} [Q_n(Z = 0|\mathbf{s})], \forall \mathbf{s} \in \mathscr{S}_u$ and $\mathbf{s}_{q_1} = \arg\max_{\mathbf{s}} [Q_n(Z = 1|\mathbf{s})], \forall \mathbf{s} \in \mathscr{S}_u$
    Query the label z_n of $\mathbf{s}_n$
    Update u_n in (11) with $(\mathbf{s}_n, z_n)$ and $Q_n(z_n \mid \mathbf{s}_n)$
    if $u_n \leq \alpha$ **then**
        Return Reject H_0
    else
        Update $Q_n(z \mid \mathbf{s})$ with newly queried $(\mathbf{s}_n, z_n)$ and past training examples.
    end if
end for
Return Retain H_0

6.2 The Proposed Framework Results in an Anytime-Valid p-value

Our framework rejects H_0 if the statistic $u_n \leq \alpha$. The following theorem states that under H_0, u_n is an *anytime-valid* p-value.

Theorem 6 If an analyst uses the proposed framework to sequentially query the oracle for Z with $\mathbf{S} \in \mathscr{S}_u$ resulting in $(\mathbf{S}, Z)^n$, then we have the following under H_0,

$$P_0\left(\exists n \in [N_q], U_n = \prod_{i=1}^{n} \frac{\hat{P}(Z_i)}{Q_i(Z_i \mid \mathbf{S}_i)} \leq \alpha\right) \leq \alpha$$

where N_q is a label budget and α is the pre-specified significance level.

We direct readers to Section A in Li et al. 2024 for the detailed proof. Theorem 6 implies the probability P_0 (or Type I error) that our framework mistakenly rejects H_0 is upper-bounded by α. The proof of Theorem 6 builds upon Theorem 2, and we can demonstrate that $P(U_n \geq T_n) = 1, \forall n$, assuming t_n (5) is constructed from the same $(s, z)^n$ with a known class prior $P(z)$.

6.3 Asymptotic Properties of the Proposed Framework

We consider normalizing the test statistic in (11) as follows,

$$\overline{U}_n = \frac{1}{n}\sum_{i=1}^{n} \log \frac{\hat{P}(Z_i)}{Q_i\left(Z_i \mid \mathbf{S}_i\right)}, (\mathbf{S}_i, Z_i) \sim p_i\left(\mathbf{s}, z\right) = p(z \mid \mathbf{s})\, p_i\left(\mathbf{s}\right)$$

where $(\mathbf{S}_i, Z_i)$ denotes a feature-label pair returned by a label query scheme when querying the i-th label. We then state the following theorem.

Theorem 7 Let $\mathbb{S}$ be the support of $p\left(\mathbf{s}\right)$ that sample features are collected from and added to an unlabeled set $\mathscr{S}_u$, and let $P\left(z \mid \mathbf{s}\right)$ denote the posterior probability of z given $\mathbf{s} \in \mathscr{S}$. There exists a consistent bimodal query scheme (see Definition 1); when an analyst uses such a scheme in the proposed active sequential framework, then, under H_1, $\overline{U}_n$ converges to the negation of mutual information (MI), and the converged negated MI lower-bounds the negated MI generated by any $p\left(\mathbf{s}\right)$ subject to $P\left(z \mid \mathbf{s}\right), \forall \mathbf{s} \in \mathbb{S}$. Precisely, there exists a consistent bimodal query leading to the following

$$\lim_{n\to\infty} \overline{U}_n = -\left(H^*(Z) - H^*\left(Z \mid \mathbf{S}\right)\right) = -I^*\left(\mathbf{S}; Z\right) \leq -I\left(\mathbf{S}; Z\right).$$

$I^*\left(\mathbf{S}; Z\right)$ is the MI constructed with $(\mathbf{S}, Z) \sim p^*\left(\mathbf{s}, z\right) = P\left(z \mid \mathbf{s}\right) p^*\left(\mathbf{s}\right)$ (See (8) for $p^*\left(\mathbf{s}\right)$); $I\left(\mathbf{S}; Z\right)$ is MI constructed with $(\mathbf{S}, Z) \sim p\left(\mathbf{s}, z\right) = P\left(z \mid \mathbf{s}\right) p\left(\mathbf{s}\right)$.

We direct readers to Section B in Li et al. 2024 for the detailed proof. Recalling the null H_0 is rejected when the test statistic u_n in (11) is smaller than α; hence, the proposed framework, when used with a consistent bimodal query to asymptotically minimize the normalized u_n in (11), favorably increases the testing power when $|\mathscr{S}_u|$ is large and $Q(z \mid \mathbf{s})$ is close to $P(z \mid \mathbf{s})$. Theorem 7 alludes that the proposed framework asymptotically turns the original hard two-sample testing problem to a simply by increasing the dependency between $\mathbf{S}$ and Z.

Remark 2 Our testing framework is also *consistent* under H_1 and the same conditions of Theorem 7 as $\lim_{n\to\infty} P_1\left(\prod_{i=1}^{n} \frac{\hat{P}(Z_i)}{Q_i(Z_i|\mathbf{S}_i)} \leq \alpha\right) = \lim_{n\to\infty} P\left(\overline{U}_n \leq \frac{1}{n}\log(\alpha)\right) = P_1(-I^*\left(\mathbf{S}, Z\right) \leq 0) = 1$. The last equality holds due to $I^*\left(\mathbf{S}, Z\right) > 0$ under H_1.

6.4 Finite-Sample Analysis for the Proposed Framework

This section analyzes the testing power of the proposed framework in the finite-sample case. Section 6.4.1 and 6.4.2 offer metrics that assess the approximation error of $Q(z \mid \mathbf{s})$ and an irreducible Type II error. They together determine the finite-sample testing power. Furthermore, Section 6.4.3 presents an illustrative example of using our framework. In Section 6.4.4, we conduct a finite-sample analysis for the example, incorporating both the metrics that characterize the approximation error and the irreducible Type II error.

6.4.1 Characterizing the Approximation Error of $Q(z \mid \mathbf{s})$

As our framework constructs the test statistic in (5) with the approximation $Q(z \mid \mathbf{s})$, there arises a need to establish a metric for assessing the approximation error of $Q(z \mid \mathbf{s})$ for our finite-sample analysis. To this end, we introduce KL^2-divergence,

Definition 2 (**KL^2-divergence**) Let p_0 and q_0 be two probability density functions on the same support $\mathscr{X}$. Let $f(t) = \log^2(t)$. Then, the KL^2-divergence between p_0 and q_0 is

$$D_{\mathrm{KL}^2}\left(q_0 \| p_0\right) = \mathbb{E}_{\mathbf{X}\sim p_0(\mathbf{x})}\left[f\left(\frac{q_0(\mathbf{X})}{p_0(\mathbf{X})}\right)\right] = \mathbb{E}_{\mathbf{X}\sim p_0(\mathbf{x})}\left[\log^2\left(\frac{q_0(\mathbf{X})}{p_0(\mathbf{X})}\right)\right].$$

$D_{\mathrm{KL}^2}\left(q_0 \| p_0\right)$ is the second moment of the log-likelihood ratio and has been used (see, e.g., (3.1.14) in Koga et al. 2002) to understand the behavior of the distribution of $\log\left(\frac{q_0(\mathbf{x})}{p_0(\mathbf{x})}\right)$. We use $D_{\mathrm{KL}^2}\left(q_0 || p_0\right)$ to

evaluate the distance between $p(\mathbf{s},z) = P(z \mid \mathbf{s})\, p(\mathbf{s})$ and $q(\mathbf{s},z) = Q(z \mid \mathbf{s})\, p(\mathbf{s})$, which yields the following

$$D_{\mathrm{KL}^2}(q(\mathbf{s},z) \| p(\mathbf{s},z)) = \mathbb{E}_{(\mathbf{S},Z)\sim p(\mathbf{s},z)}\left[\log^2\left(\frac{q(\mathbf{S},Z)}{p(\mathbf{S},Z)}\right)\right] = \mathbb{E}_{(\mathbf{S},Z)\sim p(\mathbf{s},z)}\left[\log^2\left(\frac{Q(Z \mid \mathbf{S})}{P(Z \mid \mathbf{S})}\right)\right]. \tag{12}$$

Remarkably, $D_{\mathrm{KL}^2}(q(\mathbf{s},z) \| p(\mathbf{s},z))$ in (12) also characterizes the discrepancy between $P(z \mid \mathbf{s})$ and $Q(z \mid \mathbf{s})$ by averaging their log square distance over $\mathscr{S}$. *$D_{KL^2}(q(\mathbf{s},z) \| p(\mathbf{s},z))$ and $D_{KL^2}(p(\mathbf{s},z) \| q(\mathbf{s},z))$ both characterize the approximation error of $Q(z \mid \mathbf{s})$, and in our main results, we will also see they jointly determine the testing power of the proposed framework in Section 6.4.4.*

6.4.2 Characterizing the Factor that Leads to the Irreducible Type II Error in Finite-Sample Case

We also introduce another factor influencing testing power, which persists even in the absence of approximation error, i.e., $Q(z \mid \mathbf{s}) = P(z \mid \mathbf{s})$. To see this, we recall the following random variable,

$$\bar{I}(\mathbf{S}^n; Z^n) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{p(\mathbf{S}_i, Z_i)}{p_{\mathbf{S}}(Z_i) p_Z(Z_i)}, \quad (\mathbf{S},Z) \sim p(\mathbf{s},z)$$

The distribution of $\bar{I}$ is formally defined as the *information spectrum* in Han and Verdú 1993. Han 2000 leverages the dispersion of $\bar{I}(\mathbf{S}^n; Z^n)$ to quantify the rate that Type II error goes to zero with increasing n. Their underlying rationale is that, for a larger variance of $\bar{I}$, the probability of $\bar{I}$ falling outside the acceptance region for an alternative hypothesis also increases, thereby resulting in a slower convergence rate for the Type II error. In our work, we will make use of the variance of the log-likelihood ratio between $p(\mathbf{s},z)$ and $p(\mathbf{s})p(z)$

$$\mathrm{Var}_{(\mathbf{S};Z)\sim p(\mathbf{s},z)} \bar{I}(\mathbf{S},Z) = n\mathrm{Var}_{(\mathbf{S};Z)^n \sim p((\mathbf{s},z)^n)} \bar{I}(\mathbf{S}^n, Z^n) = \mathrm{Var}_{(\mathbf{S},Z)\sim p(\mathbf{s},z)}\left[-\log \frac{P(Z)}{P(Z \mid \mathbf{S})}\right].$$

It remains present even in the absence of approximation error (i.e., $Q(z \mid \mathbf{s}) = P(z \mid \mathbf{s})$). *As we will see in Section 6.4.4, the persistent $Var_{(\mathbf{S},Z)\sim p(\mathbf{s},z)} \bar{I}(\mathbf{S};Z)$ leads to a non-zero Type II error in the finite-sample case.*

(***An example of the proposed framework***) Given a label budget N_q, α, an unlabeled set $\mathscr{S}_u$, a partition $\mathscr{P} = \{A_1, \cdots, A_m\}$, and class priors $\{P(Z=0 \mid A_1), \cdots, P(Z=0 \mid A_m)\}$, an analyst initializes $Q(z \mid \mathbf{s})$ with a set of labeled features randomly sampled from $\mathscr{S}_u$, then, she estimates $I(\mathbf{S}; Z \mid A_i)$ by

$$\begin{aligned}\hat{I}(\mathbf{S}; Z \mid A_i) &= H(Z \mid A_i) - \hat{H}(Z \mid \mathbf{S}, A_i) \\ &= -\sum_{z=0}^{1} P(Z=z \mid A_i) \log P(Z=z \mid A_i) + \frac{\sum_{\mathbf{s}\in A_i \bigcap \mathscr{S}_u} \sum_{z=0}^{1} Q(Z=z \mid \mathbf{s}) \log Q(Z=z \mid \mathbf{s})}{|A_i \bigcap \mathscr{S}_u|},\end{aligned} \tag{13}$$

selects $A^* = \arg\max_{A\in\mathbb{P}} \hat{I}(\mathbf{S}; Z \mid A)$, and sequentially constructs the statistic $u_n = \prod_{i=1}^{n} \frac{P(z_i)}{Q(z_i|\mathbf{s}_i)}$ by labelling features randomly sampled from $A^* \bigcap \mathscr{S}_u$. The analyst rejects H_0 whenever $u_n \le \alpha$ or retains H_0 if the label budget runs out.

(***The baseline***) Given a label budget N_q, α, an unlabeled set $\mathscr{S}_u$ and the class prior $P(Z=0)$, an analyst initializes $Q(z \mid \mathbf{s})$ with a set of labeled features randomly sampled from $\mathscr{S}_u$, then, she sequentially constructs the statistic $u_n = \prod_{i=1}^{n} \frac{P(z_i)}{Q(z_i|\mathbf{s}_i)}$ by labelling features randomly sampled from $\mathscr{S}_u$. The analyst rejects H_0 whenever $u_n \le \alpha$ or retains H_0 if the label budget runs out.

Figure 2: An example of the proposed framework and the baseline.

6.4.3 An Example of Using the Proposed Framework

We write $\mathscr{P} = \{A_1, \cdots, A_m\}$ to denote a partition of the support $\mathbb{S}$ of $p(\mathbf{s})$ from which unlabeled sample features in $\mathscr{S}_u$ are generated; in other words, $\bigcup_{i=1}^{m} A_i = \mathbb{S}$. In the following, we quantitatively compare an example of the framework to a baseline where features are randomly sampled from $\mathscr{S}_u$ and labeled. The example of the framework and the baseline are as in Fig 2. The class priors $\{P(z \mid A_i)\}$ are given to simplify our analytical results. In addition, the label-efficient test chooses the partition A^* predicted by $Q(z \mid \mathbf{s})$ to have the highest dependency between $\mathbf{S}$ and Z and only conducts sequential testing with the labeled points in A^*. In contrast, the baseline case conducts the sequential test entirely the same, except that the analyst queries the labels of features that are randomly generated from $\mathscr{S}_u$. Both the proposed framework and the baseline cases assert the use of a stable $Q(z \mid \mathbf{s})$ with no updates in the sequential testing, which is sufficient and convenient for our analysis.

6.4.4 Finite-Sample Analysis for the Example

We use $\varepsilon_1 = \max_{A \in \mathscr{P}} D_{\mathrm{KL}^2}(q(\mathbf{s}, z) \| p(\mathbf{s}, z) \mid A)$ and $\varepsilon_2 = \max_{A \in \mathscr{P}} D_{\mathrm{KL}^2}(p(\mathbf{s}, z) \| q(\mathbf{s}, z) \mid A)$ to capture the maximum approximation error of $Q(z \mid \mathbf{s})$ over the partition $\mathscr{P} = \{A_1, \cdots, A_m\}$, and use $\sigma^2 = \max\left\{\max_{A \in \mathscr{P}} \mathrm{Var}_{(\mathbf{S},Z) \sim p(\mathbf{s},z \mid A)} \bar{I}(\mathbf{S};Z), \mathrm{Var}_{(\mathbf{S},Z) \sim p(\mathbf{s},z)} \bar{I}(\mathbf{S};Z)\right\}$ to capture the maximum irreducible Type II error over the same partition $\mathscr{P}$. We will need to make the following assumptions.

Assumption 1 (**Maximum mutual information gain**) $\max_{A \in \mathscr{P}} I(\mathbf{S}; Z \mid A) - I(\mathbf{S}; Z) = \Delta \geq 0$.

Assumption 1 characterizes the largest MI gain of the proposed framework in the case study over the baseline; that is the direct reason for the increased testing power of the proposed framework.

Assumption 2 (**Sufficient size of unlabeled samples**)

$$\frac{\sum_{\mathbf{s} \in A \cap \mathscr{S}_u} \mathbb{E}_{Z \sim Q(z \mid \mathbf{s})} \left[\log\left(\frac{Q(Z \mid \mathbf{s})}{P(Z \mid \mathbf{s})}\right)\right]}{|A \cap \mathscr{S}_u|} \approx D_{\mathrm{KL}}(Q(z \mid \mathbf{s}) \| P(z \mid \mathbf{s}) \mid A), \forall A \in \mathscr{P}.$$

Assumption 2 assumes a sufficient supply of unlabeled samples to simplify the analysis and concentrate solely on the number of labels needed for the case study of the proposed framework. Now, we present our theorem to address the testing power of the case study in the finite-sample case.

Theorem 8 Under Assumption 1 and 2, the example of the proposed framework with a label budget N_q and α has a testing power of approximately at least

$$\Phi\left(\frac{\frac{\log \alpha}{\sqrt{N_q}} + \sqrt{N_q}\left(I(\mathbf{S};Z) + \Delta - 2\sqrt{\varepsilon_1} - \sqrt{\varepsilon_2}\right)}{\left(\varepsilon_1 + \sigma^2 + 2\sigma\sqrt{\varepsilon_1}\right)^{1/2}}\right); \tag{14}$$

and the baseline test has a testing power of approximately at least

$$\Phi\left(\frac{\frac{\log \alpha}{\sqrt{N_q}} + \sqrt{N_q}\left(I(\mathbf{S};Z) - \sqrt{\varepsilon_1}\right)}{\left(\varepsilon_1 + \sigma^2 + 2\sigma\sqrt{\varepsilon_1}\right)^{1/2}}\right). \tag{15}$$

We direct readers to Section C in Li et al. 2024 for the detailed proof. We observe that

- Given α, the lower bounds of the testing powers for both the proposed framework and the baseline, increase with a larger budget N_q and smaller approximation errors characterized by ε_1.
- Comparing (14) for the proposed framework to the (15) for the baseline, the extra Δ is ascribed to the maximum power gain, and $\sqrt{\varepsilon_1} + \sqrt{\varepsilon_2}$ accounts for the diminishing of the maximum power gain in selecting a $A^* \in \mathscr{P}$ that does not have the highest MI over $A \in \mathscr{P}$.

- When the approximation errors $\varepsilon_1 = 0$ and/or $\varepsilon_2 = 0$, both testing power's lower-bounds are decreased by a factor of σ, resulting in the irreducible Type II error.
- When the maximum MI gain Δ can compensate the approximation error of $Q(z \mid \mathbf{s})$ being larger than $\sqrt{\varepsilon_1} + \sqrt{\varepsilon_2}$, our framework in the example has higher testing power's lower bound than the baseline test given the same label budget N_q and α.

7 FUTURE DEVELOPMENT

Beyond the digital health example for Alzheimer's Disease, there are many other applications where label-efficient two-sample tests improve validation efficiency. In cancer research, genetic biomarkers can be collected non-invasively from patients, but confirming the presence of a specific type of cancer may require invasive biopsies or expensive imaging tests. Using a limited number of biopsy-confirmed cases can help efficiently validate the association between genetic biomarkers and cancer presence. In wildlife studies, data from non-invasive methods like camera traps or acoustic sensors can be collected over large areas, while confirming species presence or health status may necessitate capturing and tagging animals, a costly and labor-intensive process. Similarly, in financial transactions, features such as transaction amount, frequency, and location can be collected automatically, but identifying fraudulent transactions often demands human investigation and verification. Finally, for assessing pollution levels, inexpensive sensor data (e.g., air quality indices) can be collected widely, while confirming the exact pollutant composition and levels may require costly laboratory analyses. In all these examples, label-efficient two sample testing allows researchers to validate sensor data against lab results using a small sample of lab-verified data.

ACKNOWLEDGEMENT

This work was funded in part by Office of Naval Research grant N00014-21-1-2615 and by the National Science Foundation (NSF) under grants CNS-2003111, and CCF-2048223.

REFERENCES

Balcan, M.-F., A. Beygelzimer, and J. Langford. 2006. "Agnostic active learning". In *Proceedings of the 23rd international conference on Machine learning*, 65–72.

Balsubramani, A. and A. Ramdas. 2015. "Sequential nonparametric testing with the law of the iterated logarithm". *arXiv preprint arXiv:1506.03486*.

Bloemena, A. 1964. "Sampling from a Graph". *MC Tracts*.

Cecchi, F. and N. Hegde. 2017. "Adaptive active hypothesis testing under limited information". *Advances in Neural Information Processing Systems* 30.

Chen, H. and J. H. Friedman. 2017. "A new graph-based two-sample test for multivariate and object data". *Journal of the American statistical association* 112(517):397–409.

Cohen, K. and Q. Zhao. 2015. "Active hypothesis testing for anomaly detection". *IEEE Transactions on Information Theory* 61(3):1432–1450.

Cohn, D., L. Atlas, and R. Ladner. 1994. "Improving generalization with active learning". *Machine learning* 15:201–221.

Dasarathy, G., R. Nowak, and X. Zhu. 2015. "S2: An efficient graph based active learning algorithm with application to nonparametric classification". In *Conference on Learning Theory*, 503–522. PMLR.

Dasarathy, G., P. Shah, and R. G. Baraniuk. 2017. "Sketched covariance testing: A compression-statistics tradeoff". In *2017 51st Asilomar Conference on Signals, Systems, and Computers*, 676–680. IEEE.

Doob, J. 1939. "Jean Ville, Étude Critique de la Notion de Collectif". *Bulletin of the American mathematical society* 45(11):824–824.

Dunn, O. J. 1961. "Multiple comparisons among means". *Journal of the American statistical association* 56(293):52–64.

Durrett, R. 2019. *Probability: theory and examples*, Volume 49. Cambridge university press.

Friedman, J. 2004. "On multivariate goodness-of-fit and two-sample testing". Technical report, Citeseer.

Friedman, J. H. and L. C. Rafsky. 1979. "Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests". *The Annals of Statistics*:697–717.

Gal, Y., R. Islam, and Z. Ghahramani. 2017. "Deep bayesian active learning with image data". In *International conference on machine learning*, 1183–1192. PMLR.

Györfi, L., M. Köhler, A. Krzyżak, and H. Walk. 2002. *A distribution-free theory of nonparametric regression*, Volume 1. Springer.

Hajnal, J. 1961. "A two-sample sequential t-test". *Biometrika* 48(1/2):65–75.

Han, T. S. 2000. "Hypothesis testing with the general source". *arXiv preprint math/0004121*.

Han, T. S. and S. Verdú. 1993. "Approximation theory of output statistics". *IEEE Transactions on Information Theory* 39(3):752–772.

Hanneke, S. *et al*. 2014. "Theory of disagreement-based active learning". *Foundations and Trends® in Machine Learning* 7(2-3):131–309.

Henze, N. and M. D. Penrose. 1999. "On the multivariate runs test". *Annals of statistics*:290–298.

Hotelling, H. 1992. "The generalization of Student's ratio". In *Breakthroughs in statistics*, 54–65. Springer.

Johnson, R. R. and P. J. Kuby. 2011. *Elementary statistics*. Cengage Learning.

Kartik, D., A. Nayyar, and U. Mitra. 2019. "Active hypothesis testing: Beyond chernoff-stein". In *2019 IEEE International Symposium on Information Theory (ISIT)*, 897–901. IEEE.

Koga, H. *et al*. 2002. *Information-spectrum methods in information theory*, Volume 50. Springer Science & Business Media.

Korte, B. H., J. Vygen, B. Korte, and J. Vygen. 2011. *Combinatorial optimization*, Volume 1. Springer.

Lhéritier, A. and F. Cazals. 2018. "A sequential non-parametric multivariate two-sample test". *IEEE Transactions on Information Theory* 64(5):3361–3370.

Li, W., G. Dasarathy, K. N. Ramamurthy, and V. Berisha. 2022. "A label efficient two-sample test". In *Uncertainty in Artificial Intelligence*, 1168–1177. PMLR.

Li, W., K. N. Ramamurthy, P. Kadambi, P. Saidi, G. Dasarathy and V. Berisha. 2024. "Active Sequential Two-Sample Testing". *Accepted to Transactions on Machine Learning Research*.

Naghshvar, M. and T. Javidi. 2010. "Active M-ary sequential hypothesis testing". In *2010 IEEE International Symposium on Information Theory*, 1623–1627. IEEE.

Naghshvar, M., T. Javidi, *et al*. 2013. "Active sequential hypothesis testing". *Annals of Statistics* 41(6):2703–2738.

Shekhar, S. and A. Ramdas. 2021. "Game-theoretic Formulations of Sequential Nonparametric One-and Two-Sample Tests". *arXiv preprint arXiv:2112.09162*.

Vayatis, N., M. Depecker, and S. Clémençcon. 2009. "AUC optimization and the two-sample problem". *Advances in Neural Information Processing Systems* 22.

Wald, A. 1992. "Sequential tests of statistical hypotheses". In *Breakthroughs in Statistics*, 256–298. Springer.

Welch, W. J. 1990. "Construction of permutation tests". *Journal of the American Statistical Association* 85(411):693–698.

Wilks, S. S. 1938. "The large-sample distribution of the likelihood ratio for testing composite hypotheses". *The annals of mathematical statistics* 9(1):60–62.

AUTHOR BIOGRAPHIES

WEIZHI LI is a postdoctoral associate in the Statistical Science Group at Los Alamos National Laboratory. His research primarily focuses on developing data-efficient machine learning algorithms with statistical guarantees. His email address is weizhili@lanl.gov and his website is https://wayne0908.github.io/.

VISAR BERISHA is a Professor in the College of Engineering and the College of Health Solutions at Arizona State University. At its core, his work is interdisciplinary and use-inspired; it lies at the intersection of engineering and human health and is driven by a desire to use technology to improve the human condition. His email address is visar@asu.edu and his website is http://vees.ar/.

GAUTAM DASARATHY is an Associate Professor in the School of Electrical, Computer and Energy Engineering at Arizona State University. His research interests include various topics in machine learning, high-dimensional statistics, information processing, and networked systems. His email address is gautamd@asu.edu and his website is http://gautamdasarathy.com.