Proceedings of Machine Learning Research vol 195:1–38, 2023 36th Annual Conference on Learning Theory

Minimax optimal testing via classification

Patrik Róbert Gerber PRGERBER@MIT.EDU
Yanjun Han YJHAN@MIT.EDU
Yury Polyanskiy YP@MIT.EDU
Massachusetts Institute of Technology

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

This paper considers an ML inspired approach to hypothesis testing known as classifier/classification-accuracy testing (CAT). In CAT, one first trains a classifier by feeding it labeled synthetic samples generated by the null and alternative distributions, which is then used to predict labels of the actual data samples. This method is widely used in practice when the null and alternative are only specified via simulators (as in many scientific experiments).

We study goodness-of-fit, two-sample (TS) and likelihood-free hypothesis testing (LFHT), and show that CAT achieves (near-)minimax optimal sample complexity in both the dependence on the total-variation (TV) separation ϵ and the probability of error δ in a variety of non-parametric settings, including discrete distributions, d-dimensional distributions with a smooth density, and the Gaussian sequence model. In particular, we close the high probability sample complexity of LFHT for each class. As another highlight, we recover the minimax optimal complexity of TS over discrete distributions, which was recently established by Diakonikolas et al. (2021). The corresponding CAT simply compares empirical frequencies in the first half of the data, and rejects the null when the classification accuracy on the second half is better than random.

Keywords: Goodness-of-fit testing, Identity testing, Two-sample testing, Closeness testing, Likelihood-free hypothesis testing, Scheffé's test, Classifier-accuracy testing, Likelihood-free inference

1. Introduction

The rapid development of machine learning over the past three decades has had a profound impact on many areas of science and technology. It has replaced or enhanced traditional statistical procedures and automated feature extraction and prediction where in the past human experts had to intervene manually. One example is the technique that has become known as 'classification accuracy testing' (CAT). The idea, first explicitly described in Friedman (2004), is extremely simple. Consider the setting of two-sample testing: suppose the statistician has samples X and Y of size n from two distributions $\mathbb{P}_{\mathsf{X}}$ and $\mathbb{P}_{\mathsf{Y}}$ respectively on some space $\mathcal{X}$, and wishes to test the hypotheses

$$H_0 : \mathbb{P}_{\mathsf{X}} = \mathbb{P}_{\mathsf{Y}} \qquad \text{versus} \qquad H_1 : \mathbb{P}_{\mathsf{X}} \neq \mathbb{P}_{\mathsf{Y}}. \tag{TS}$$

The statistician has many classical methods at their disposal such as the Kolmogorov-Smirnov or the Wilcoxon – Mann – Whitney test. Friedman's idea was to use machine learning as a powerful tool to summarize the data and subsequently apply a classical two-sample test to the transformed data. More concretely, the proposal is to train a binary classifier $\mathcal{C} : \mathcal{X} \to \{0, 1\}$ on the labeled data $\cup_{i=1}^{n}\{(X_i, 0), (Y_i, 1)\}$ and compare the samples $\mathcal{C}(X_1), \ldots, \mathcal{C}(X_n)$ and $\mathcal{C}(Y_1), \ldots, \mathcal{C}(Y_n)$.

Friedman's idea to use classifiers to summarize data before applying classical statistical analysis downstream can be generalized beyond two-sample testing (TS). Likelihood-free inference (LFI),

also known as simulation-based inference (SBI), has seen a flurry of interest recently. In LFI, the scientist has a dataset $Z_1, \dots, Z_m \overset{\text{iid}}{\sim} \mathbb{P}_{\theta^\star}$ and is given access to a black box simulator which given a parameter θ produces a random variable with distribution $\mathbb{P}_\theta$. The goal is to do inference on $\theta^\star$. The key aspect of the problem, lending the name 'likelihood-free', is that the scientist doesn't know the inner workings of the simulator. In particular its output is not necessarily differentiable with respect to θ and the density of $\mathbb{P}_\theta$ cannot be evaluated even up to normalization. This setting arises in numerous areas of science where highly complex, mechanistic, stochastic simulators are used such as climate modeling, particle physics, phylogenetics and epidemiology to name a few, and its importance was realized as early as Diggle and Gratton (1984). In this paper we study the problem of likelihood-free hypothesis testing (LFHT) proposed recently in Gerber and Polyanskiy (2022) as a simplified model of likelihood-free inference. Compared to two-sample testing, here in addition to the dataset Z of size m, we have two 'simulated' samples X, Y of size n each from $\mathbb{P}_{\mathsf{X}}$ and $\mathbb{P}_{\mathsf{Y}}$ respectively. The goal is to test the hypotheses

$$H_0 : Z_i \sim \mathbb{P}_{\mathsf{X}} \qquad \text{versus} \qquad H_1 : Z_i \sim \mathbb{P}_{\mathsf{Y}}. \tag{LFHT}$$

It is important that apriori $\mathbb{P}_{\mathsf{X}}$ and $\mathbb{P}_{\mathsf{Y}}$ are only known to belong to a certain ambient (usually non-parametric) class. This stands in contrast with the earliest appearances of (LFHT) in Ziv (1988); Gutman (1989), where authors studied the rate of decay of the type-I and type-II error probabilities for fixed $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}$.

In the context of (LFHT) the idea of Friedman materializes as follows. First, train a classifier $\mathcal{C} : \mathcal{X} \to \{0, 1\}$ to distinguish between $\mathbb{P}_{\mathsf{X}}$ and $\mathbb{P}_{\mathsf{Y}}$ and second, compare the transformed dataset $\{\mathcal{C}(Z_j))\}_{j=1}^m$ to $\{\mathcal{C}(X_i)\}_{i=1}^n$ and $\{\mathcal{C}(Y_i)\}_{i=1}^n$. The second step compares iid samples of Bernoulli random variables (provided $\mathcal{C}$ is trained on held out data), thus any reasonable test simply thresholds the number of Z_j classified as 1, namely the test is of the form

$$\frac{1}{m}\sum_{j=1}^{m} \mathcal{C}(Z_j) \geq \gamma \tag{1}$$

for some $\gamma \in [0, 1]$. The idea to classify Z as coming from either $\mathbb{P}_{\mathsf{X}}$ or $\mathbb{P}_{\mathsf{Y}}$ based on the empirical mass on some separating set $S = \mathcal{C}^{-1}(\{1\}) \approx \{\mathrm{d}\mathbb{P}_{\mathsf{Y}}/\mathrm{d}\mathbb{P}_{\mathsf{X}} \geq 1\}$ has been attributed to Scheffé in folklore (Devroye and Lugosi, 2001, Section 6). To illustrate the genuine importance of these ideas, we draw on the famous Higgs boson discovery. In 2012 Chatrchyan et al. (2012); Adam-Bourdarios et al. (2015) at the Large Hadron Collider (LHC) a team of physicists announced that they observed the Higgs boson, an elementary particle theorized to exist in 1964. It is regarded as the crowning achievement of the LHC, the most expensive instrument ever built. They achieved this feat via likelihood-free inference, using the ideas of classification accuracy testing/Scheffé's test in particular. As part of their analysis pipeline they trained a boosted decision tree classifier on simulated data and thresholded counts of observations falling in the classification region.

This work was initiated as an attempt to understand the theoretical properties of classifier-accuracy testing, motivated by the clear practical interest in these questions. Our intuition told us that restricting the classifier to have binary output might throw away too much statistical power. In regions with large (small) density ratio, the binary output ought to loose useful information about the (un)certainty of the classifier output. The Neyman-Pearson Lemma phrases this succinctly: the optimal classifier aggregates the log density ratio, while heuristically Scheffé's test aggregates indicators that the log density ratio exceeds some threshold. The operational implication of this would

be to train probabilistic classifiers $\mathcal{C} : \mathcal{X} \to \mathbb{R}$ approximating the log density ratio, and to aggregate this $\mathbb{R}$-valued output instead of the binary output. However, our results show that this is not necessary for optimality, at least in the minimax sense.

1.1. Informal description of the results

We study the problems of goodness-of-fit testing, two-sample testing and likelihood-free hypothesis testing in a minimax framework (see Section 2.1.1 for precise definitions). Namely, given a family of probability distributions $\mathcal{P}$, we study the minimum number of observations n (and m for LFHT) that are required to perform the test with error probability less than $\delta \in (0, 1/2)$ in the worst case over the distributions $\mathbb{P}_{\mathsf{X}}$ and $\mathbb{P}_{\mathsf{Y}}$. We show for multiple natural classes $\mathcal{P}$ that there exist minimax optimal (with some restrictions) classification accuracy tests.

Let us clarify what we mean by 'classification-accuracy' tests for goodness-of-fit testing (GoF) and the problems TS and LFHT. Suppose we have a sample X of size $2n$ from the unknown distribution $\mathbb{P}_{\mathsf{X}}$. We also have a second sample Y of size $2n$ from $\mathbb{P}_{\mathsf{Y}} \in \mathcal{P}$ which corresponds to the *known* null distribution in the case of GoF and is *unknown* in the case of TS, LFHT. Finally, for LFHT we have an additional sample Z of size $2m$ from $\mathbb{P}_{\mathsf{Z}} \in \{\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}\}$. Write $\mathcal{D}_{\mathsf{tr}} \triangleq \{X^{\mathsf{tr}}, Y^{\mathsf{tr}}, Z^{\mathsf{tr}}\}$ for the first halves of each sample and $\mathcal{D}_{\mathsf{te}} \triangleq \{X^{\mathsf{te}}, Y^{\mathsf{te}}, Z^{\mathsf{te}}\}$ for the rest. We train a classifier $\mathcal{C} : \mathcal{X} \to \{0, 1\}$ on the input $\mathcal{D}_{\mathsf{tr}}$ that aims to assign 1 to $\mathbb{P}_{\mathsf{X}}$ and 0 to $\mathbb{P}_{\mathsf{Y}}$. Going forward, it will be easier to think of $\mathcal{C}$ in terms of the 'separating set' $S \triangleq \mathcal{C}^{-1}(\{1\})$. Thus, S is a random subset of $\mathcal{X}$ whose randomness comes from $\mathcal{D}_{\mathsf{tr}}$ and potentially an external seed. Given two datasets $\{A_i\}_{i=1}^a, \{B_j\}_{j=1}^b$, we define the classifier-accuracy statistic

$$T_S(A, B) \triangleq \frac{1}{a} \sum_{i=1}^{a} \mathbb{1}\{A_i \in S\} - \frac{1}{b} \sum_{j=1}^{b} \mathbb{1}\{B_j \in S\}. \tag{2}$$

The name 'classifier-accuracy' is given due to the fact that $T_S(X^{\mathsf{te}}, Y^{\mathsf{te}}) + 1$ is equal to the sum of the fraction of correctly classified test instances under the two classes. Finally, we say a test is a classifier-accuracy test if its output is obtained by thresholding $|T_S|$ for some classifier $\mathcal{C} = \mathbb{1}_S$ on the test data $\mathcal{D}_{\mathsf{te}}$.

Theorem 1 (informal) *There exist classifier-accuracy tests with minimax (near-)optimal sample complexity for all problems* GoF, TS, LFHT *and multiple classes of distributions* $\mathcal{P}$.

1.2. Proof sketch

The bulk of the technical difficulty lies in finding a good separating set $S \subseteq \mathcal{X}$. But how do we measure the quality of S? Define the "separation" $\mathrm{sep}(S) \triangleq \mathbb{P}_{\mathsf{X}}(S) - \mathbb{P}_{\mathsf{Y}}(S)$, and the "size" $\tau(S) \triangleq \min\{\mathbb{P}_{\mathsf{X}}(S)\mathbb{P}_{\mathsf{X}}(S^{\mathsf{c}}), \mathbb{P}_{\mathsf{Y}}(S)\mathbb{P}_{\mathsf{Y}}(S^{\mathsf{c}})\}$. The following lemma describes the performance of classifier-accuracy tests (2) in terms of sep and τ.

Lemma 2 *Consider the hypothesis testing problem $H_0 : p = q$ versus an arbitrary alternative H_1. Suppose that the learner has constructed a separating set S such that $|\mathrm{sep}(S)| = |p(S) - q(S)| \geq \overline{\mathrm{sep}}$ for every $(p, q) \in H_1$, and $\tau(S) = (p(S)(1 - p(S)) \wedge (q(S)(1 - q(S))) \leq \overline{\tau}$ for every $(p, q) \in H_0 \cup H_1$. Then using only the knowledge of $\overline{\tau}$, the classifier-accuracy test (2) with n test*

samples from both p and q and an appropriate threshold achieves type-I and type-II errors at most δ, provided that

$$n \geq c \frac{\log(1/\delta)}{\underline{\text{sep}}} \left(1 + \frac{\overline{\tau}}{\underline{\text{sep}}}\right)$$

for a large enough universal constant $c > 0$.

With Lemma 2 in hand it is clear how we need to design S. It should satisfy

$$\big|\operatorname{sep}(S)\big| \text{ is big under } H_1, \text{ and } \tau(S) \text{ is small under both } H_0 \text{ and } H_1 \tag{3}$$

with probability $1 - \delta$. The latter condition, namely that τ is small i.e. $\mathcal{C} = \mathbb{1}_S$ is imbalanced, may seem unintuitive as given any two (sufficiently regular) probability distributions there always exists a balanced classifier whose separation is optimal up to constant.

Proposition 3 *Let $\mathbb{P}, \mathbb{Q}$ be two distributions on a generic probability space $(\mathcal{X}, \mathcal{F})$. Then*

$$\mathsf{TV}(\mathbb{P}, \mathbb{Q}) \leq 2 \sup\{\mathbb{P}(\mathcal{C}(X) = 0) - \mathbb{Q}(\mathcal{C}(X) = 0) : \mathbb{P}(\mathcal{C}(X) = 0) = \mathbb{Q}(\mathcal{C}(X) = 1)\},$$

where $\mathcal{C} : \mathcal{X} \to \{0, 1\}$ is a possibly randomized classifier. Here the constant 2 is tight.

Despite Proposition 3, we find that choosing a highly imbalanced classifier $\mathcal{C}$ is crucial in obtaining the minimax sample complexity in some classes. This has interesting implications for practical classifier-accuracy testing. Indeed, classifiers are commonly trained to minimize some proxy of misclassification error; however, the above heuristics show that this is not necessarily optimal, instead one should seek *imbalanced* classifiers with large separation. Another way to phrase it is that when training a classifier for testing one should have the downstream task in mind, namely, maximizing the power of the resulting test, and not classification accuracy.

1.3. Prior work and contribution

The problem of two-sample (TS) testing (aka closeness testing) and the related problem of goodness-of-fit (GoF) testing (aka identity testing) has a long history in both statistics and computer science. We only mention a small subset of the literature, directly relevant to our work. In seminal works Ingster studied (GoF) for the Gaussian sequence model Ingster (1982); Ingster and Suslina (2003) and for smooth densities Ingster (1987) in one dimension. Extensions to multiple dimensions and (TS) can be found in works such as Li and Yuan (2019); Arias-Castro et al. (2018). For discrete distributions on a large alphabet the two problems appeared first in Goldreich and Ron (2000); Batu et al. (2000), see also Chan et al. (2014); Valiant and Valiant (2017) and the survey Canonne (2020). Recent work Diakonikolas et al. (2018, 2021) has focused on GoF and TS with vanishing error probability.

The problem of likelihood-free hypothesis testing appeared first in the works Ziv (1988); Gutman (1989), who studied the asymptotic setting. Minimax likelihood-free hypothesis testing (LFHT) was first studied by the information theory community in Kelly et al. (2010, 2012) for a restricted class of discrete distributions on a large alphabet, with a strengthening by Huang and Meyn (2012) to vanishing error probability (in some regimes). More recently, the problem was proposed in Gerber and Polyanskiy (2022) as a simplified model of likelihood-free inference, and authors derived minimax optimal sample complexities for constant error in the settings studied in the present paper.

The idea of using classifiers for two-sample testing was proposed in Friedman (2004) and has seen a flurry of interest Golland and Fischl (2003); Lopez-Paz and Oquab (2016); Kim et al. (2021); Hediger et al. (2022). In likelihood-free inference the output of classifiers can be used as summary statistics for Approximate Bayesian Computation Jiang et al. (2017); Gutmann et al. (2018) or to approximate density ratios Cranmer et al. (2020) via the 'likelihood-ratio trick'. A classifier with binary $\{0,1\}$ output was used in the discovery of the Higgs boson Chatrchyan et al. (2012); Adam-Bourdarios et al. (2015) to determine the detection region.

Our work is the first to study the non-asymptotic properties of classifier-based tests in any setting and we find that classifier-accuracy tests are minimax optimal for a wide range of problems. As a consequence of our results we resolve the minimax high probability sample complexity of LFHT over all classes studied, and also obtain new, tight results on high probability GoF and TS.

1.4. Structure

In Sections 2.1.1 and 2.1.2 we define the statistical problems and distribution classes we study. In Tables 1 and 2 we present all sample complexity results, and in Section 2.2 we indicate how to derive them. Sections 3.1, 3.2 and 3.3 study the problem of learning good separating sets for discrete and smooth distributions and the Gaussian sequence model respectively. The appendix contains all proofs omitted from the main text, including all lower bounds in Appendix D.

2. Results

2.1. Technical preliminaries

2.1.1. TWO-SAMPLE, GOODNESS-OF-FIT AND LIKELIHOOD-FREE HYPOTHESIS TESTING

Formally, we define a hypothesis as a set of probability measures. Given two hypotheses H_0 and H_1 consisting of distributions on some measurable space $\mathcal{X}$, we say that a function $\psi : \mathcal{X} \to \{0,1\}$ tests the two hypotheses against each other with error at most $\delta \in (0, 1/2)$ if

$$\max_{i=0,1} \max_{P \in H_i} \mathbb{P}_{S\sim P}(\psi(S) \neq i) \leq \delta. \tag{4}$$

Throughout the remainder of this section let $\mathcal{P}$ be a class of probability distributions on $\mathcal{X}$. Suppose we observe independent samples $X \sim \mathbb{P}_{\mathsf{X}}^{\otimes n}$, $Y \sim \mathbb{P}_{\mathsf{Y}}^{\otimes n}$ and $Z \sim \mathbb{P}_{\mathsf{Z}}^{\otimes m}$ whose distributions $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}, \mathbb{P}_{\mathsf{Z}} \in \mathcal{P}$ are *unknown* to us. We now define the problems at the center of our work.

Definition 4 *Given a known $\mathbb{P}_0 \in \mathcal{P}$, **goodness-of-fit testing** is the comparison of*

$$H_0 : \mathbb{P}_{\mathsf{X}} = \mathbb{P}_0 \qquad \textit{versus} \qquad H_1 : \mathsf{TV}(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_0) \geq \epsilon \tag{GoF}$$

based on the sample X. Write $n_{\mathsf{GoF}}(\epsilon, \delta, \mathcal{P})$ for the smallest number such that for all $n \geq n_{\mathsf{TS}}$ there exists a function $\psi : \mathcal{X}^n \to \{0,1\}$ which given X as input tests between H_0 and H_1 with error probability at most δ, for arbitrary $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_0 \in \mathcal{P}$.

Definition 5 ***Two-sample testing** is the comparison of*

$$H_0 : \mathbb{P}_{\mathsf{X}} = \mathbb{P}_{\mathsf{Y}} \qquad \textit{versus} \qquad H_1 : \mathsf{TV}(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}) \geq \epsilon \tag{TS}$$

based on the samples X, Y. Write $n_{\mathsf{TS}}(\epsilon, \delta, \mathcal{P})$ for the smallest number such that for all $n \geq n_{\mathsf{TS}}$ there exists a function $\psi : \mathcal{X}^n \times \mathcal{X}^n \to \{0,1\}$ which given X, Y as input tests between H_0 and H_1 with error probability at most δ, for arbitrary $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}} \in \mathcal{P}$.

Definition 6 ***Likelihood-free hypothesis testing*** *is the comparison of*

$$H_0 : \mathbb{P}_{\mathsf{Z}} = \mathbb{P}_{\mathsf{X}} \qquad \textit{versus} \qquad H_1 : \mathbb{P}_{\mathsf{Z}} = \mathbb{P}_{\mathsf{Y}} \tag{LF}$$

based on the samples X, Y, Z. Write $\mathcal{R}_{\mathsf{LF}}(\epsilon, \delta, \mathcal{P}) \subseteq \mathbb{R}^2$ for the maximal set such that for all $(n, m) \in \mathbb{N}^2$ with $n \geq x, m \geq y$ for some $(x, y) \in \mathcal{R}_{\mathsf{LF}}$, there exists a function $\psi : \mathcal{X}^n \times \mathcal{X}^n \times \mathcal{X}^m \to \{0, 1\}$ which given X, Y, Z as input, successfully tests H_0 against H_1 with error probability at most δ, provided $\mathsf{TV}(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}) \geq \epsilon$ and $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}} \in \mathcal{P}$.

2.1.2. CLASSES OF DISTRIBUTIONS

We consider the following nonparametric families of distributions.

Smooth density. Let $\mathcal{C}(\beta, d, C)$ denote the set of functions $f : [0, 1]^d \to \mathbb{R}$ that are $\lceil \beta - 1 \rceil$-times differentiable and satisfy

$$\|f\|_{\mathcal{C}_\beta} \triangleq \max\left(\max_{0 \leq |\alpha| \leq \lceil \beta - 1 \rceil} \|f^{(\alpha)}\|_\infty, \sup_{x \neq y \in [0,1]^d, |\alpha| = \lceil \beta - 1 \rceil} \frac{|f^{(\alpha)}(x) - f^{(\alpha)}(y)|}{\|x - y\|_2^{\beta - \lceil \beta - 1 \rceil}} \right) \leq C,$$

where $\lceil \beta - 1 \rceil$ denotes the largest integer strictly smaller than β and $|\alpha| = \sum_{i=1}^d \alpha_i$ for the multiindex $\alpha \in \mathbb{N}^d$. We write $\mathcal{P}_{\mathsf{H}}(\beta, d, C_{\mathsf{H}})$ for the class of distributions with Lebesgue-densities in $\mathcal{C}(\beta, d, C_{\mathsf{H}})$.

Distributions on a finite alphabet. For $k \in \mathbb{N}$, let

$$\mathcal{P}_{\mathsf{D}}(k) \triangleq \{\text{all distributions on the finite alphabet } [k]\},$$
$$\mathcal{P}_{\mathsf{Db}}(k, C_{\mathsf{Db}}) \triangleq \{p \in \mathcal{P}_{\mathsf{D}}(k) : \|p\|_\infty \leq C_{\mathsf{Db}}/k\},$$

where $C_{\mathsf{Db}} > 1$ is a constant. In other words, $\mathcal{P}_{\mathsf{Db}}$ are those discrete distributions that are bounded by a constant multiple of the uniform distribution.

Gaussian sequence model on the Sobolev ellipsoid. Define the Sobolev ellipsoid $\mathcal{E}(s, C)$ of smoothness $s > 0$ and size $C > 0$ as $\{\theta \in \mathbb{R}^{\mathbb{N}} : \sum_{j=1}^\infty j^{2s}\theta_j^2 \leq C\}$. For $\theta \in \mathbb{R}^\infty$ let $\mu_\theta = \otimes_{i=1}^\infty \mathcal{N}(\theta_i, 1)$, and define our second class as

$$\mathcal{P}_{\mathsf{G}}(s, C_{\mathsf{G}}) \triangleq \{\mu_\theta : \theta \in \mathcal{E}(s, C_{\mathsf{G}})\}.$$

To briefly motivate the study of $\mathcal{P}_{\mathsf{G}}$, consider the classical Gaussian white noise model. Here we have iid observations of the stochastic process

$$\mathrm{d}Y_t = f(t)\mathrm{d}t + \mathrm{d}W_t, \ t \in [0, 1],$$

where $(W_t)_{t \geq 0}$ denotes Brownian motion and $f \in L^2[0, 1]$ is unknown. Suppose now that $\{\phi_i\}_{i \geq 1}$ forms an orthonormal basis for $L^2[0, 1]$ and given an observation Y define the values

$$y_i \triangleq \langle Y, \phi_i \rangle = \int_0^1 f(t)\phi_i(t)\mathrm{d}t + \int_0^1 \phi_i(t)\mathrm{d}W_t \triangleq \theta_i + \epsilon_i.$$

Notice that $\epsilon_i \sim \mathcal{N}(0, 1)$ and that $\mathbb{E}[\epsilon_i \epsilon_j] = \mathbb{1}_{i=j}$. In other words, the sequence $\{y_i\}_{i \geq 1}$ is an observation from the distribution μ_θ. Consider the particular case of $\phi_1 \equiv 1$ and $\phi_{2k} = \sqrt{2}\cos(2\pi k x), \phi_{2k+1} = \sqrt{2}\sin(2\pi k x)$ for $k \geq 1$ and assume that f satisfies periodic boundary conditions. Then θ denotes the Fourier coefficients of f and the condition that $\sum_{j=1}^\infty j^{2s}\theta_j^2 \leq C$ is equivalent to an upper bound on the order $(s, 2)$-Sobolev norm of f, see e.g. Proposition 1.14 of Tsybakov (2008). In other words, by studying the class $\mathcal{P}_{\mathsf{G}}$ we can deduce results for signal detection in Gaussian white noise, where the signal has bounded Sobolev norm.

Table 1: Minimax sample complexity of testing (up to constant factors) over $\mathcal{P}_{\mathsf{H}}, \mathcal{P}_{\mathsf{G}}, \mathcal{P}_{\mathsf{Db}}$.

	n_{GoF}	n_{TS}	$\mathcal{R}_{\mathsf{LF}}$
$\mathcal{P}_{\mathsf{Db}}(k)$	$\frac{\sqrt{k\log(1/\delta)}}{\epsilon^2} + \frac{\log(1/\delta)}{\epsilon^2}$	n_{GoF}	$m \geq \frac{\log(1/\delta)}{\epsilon^2}$ and $n \geq n_{\mathsf{GoF}}$ and $nm \geq n_{\mathsf{GoF}}^2$
$\mathcal{P}_{\mathsf{H}}(\beta, d)$	$\frac{\sqrt{\log(1/\delta)}}{\epsilon^{(2\beta+d/2)/\beta}} + \frac{\log(1/\delta)}{\epsilon^2}$	n_{GoF}	$m \geq \frac{\log(1/\delta)}{\epsilon^2}$ and $n \geq n_{\mathsf{GoF}}$ and $nm \geq n_{\mathsf{GoF}}^2$
$\mathcal{P}_{\mathsf{G}}(s)$	$\frac{\sqrt{\log(1/\delta)}}{\epsilon^{(2s+1/2)/s}} + \frac{\log(1/\delta)}{\epsilon^2}$	n_{GoF}	$m \geq \frac{\log(1/\delta)}{\epsilon^2}$ and $n \geq n_{\mathsf{GoF}}$ and $nm \geq n_{\mathsf{GoF}}^2$

Table 2: Minimax sample complexity of testing (up to constant factors) over $\mathcal{P}_{\mathsf{D}}$.

	$n_{\mathsf{GoF}}(\mathcal{P}_{\mathsf{D}})$	$n_{\mathsf{TS}}(\mathcal{P}_{\mathsf{D}})$	$\mathcal{R}_{\mathsf{LF}}(\mathcal{P}_{\mathsf{D}})$	
$k \geq \frac{\log(\frac{1}{\delta})}{\epsilon^4}$	(OPT) $n_{\mathsf{GoF}}(\mathcal{P}_{\mathsf{Db}})$	$\left(\frac{k^2\log(\frac{1}{\delta})}{\epsilon^4}\right)^{\frac{1}{3}}$	$n \geq m$	$m \geq \frac{\log(1/\delta)}{\epsilon^2}$ and $mn^2 \geq kn_{\mathsf{GoF}}^2$
			$m > n$	(OPT) $mn^2 \geq kn_{\mathsf{GoF}}^2$ and $n \geq n_{\mathsf{GoF}}$
	(CAT) $n_{\mathsf{GoF}}\left(\frac{\epsilon}{\log(k)}, \frac{\delta}{k}, \mathcal{P}_{\mathsf{Db}}\right)$			(CAT) $\frac{mn^2}{\log(\frac{k}{\delta})} \geq kn_{\mathsf{GoF}}^2\left(\frac{\epsilon}{\log(k)}, \frac{\delta}{k}\right)$ and $n \geq n_{\mathsf{GoF}}(\frac{\epsilon}{\log(k)}, \frac{\delta}{k})$
$k < \frac{\log(\frac{1}{\delta})}{\epsilon^4}$	$n_{\mathsf{GoF}}(\mathcal{P}_{\mathsf{Db}})$	$n_{\mathsf{GoF}}(\mathcal{P}_{\mathsf{Db}})$	$m \geq \frac{\log(1/\delta)}{\epsilon^2}$ and $n \geq n_{\mathsf{GoF}}$ and $nm \geq n_{\mathsf{GoF}}^2$	

2.2. Minimax sample complexity of classifier-accuracy tests

In Tables 1 and 2 we present our and prior results on the minimax sample complexity of GoF, TS and LFHT; here

- unmarked entries denote minimax optimal results achievable by a classifier-accuracy test;
- entries marked with (OPT) denote minimax optimal results that are not known to be achievable by any classifier-accuracy test;
- entries marked with (CAT) denote the best known result using a classifier-accuracy test.

In the constant error regime ($\delta = \Theta(1)$) the results of Tables 1 and 2 are well known; for instance, the sample complexities of GoF, TS, and LFHT under $\mathcal{P}_{\mathsf{D}}$ were characterized in Paninski (2008); Bhattacharya and Valiant (2015); Gerber and Polyanskiy (2022), respectively[1]. Less is known under the high-probability regime ($\delta = o(1)$): for $\mathcal{P}_{\mathsf{D}}$, n_{GoF} was characterized in Huang and Meyn (2013); Diakonikolas et al. (2018) for uniformity testing, with the general case following from the flattening reduction Diakonikolas and Kane (2016); n_{TS} was characterized in Diakonikolas et al. (2021). For $\mathcal{R}_{\mathsf{LF}}$, the $k > n$ case for $\mathcal{P}_{\mathsf{Db}}$ is resolved by Huang and Meyn (2012), and the achievability direction of the case $m > n$ of $\mathcal{R}_{\mathsf{LF}}$ for $\mathcal{P}_{\mathsf{D}}$ can be deduced from Diakonikolas et al.

[1] Gerber and Polyanskiy (2022) only resolved the minimax sample complexity of LFHT for $\mathcal{P}_{\mathsf{D}}$ up to $\log(k)$-factors in some regimes. However, by combining the classifier accuracy tests of this paper for $m \leq n$ and the reduction to two-sample testing with unequal sample size Bhattacharya and Valiant (2015); Diakonikolas et al. (2021) for $m > n$ these gaps are filled.

(2021) via the natural reduction between TS and LFHT (see Gerber and Polyanskiy (2022)). The remaining upper bounds are achievable by the classifier-accuracy tests below, and the proofs of all lower bounds are deferred to Appendix D.

As for the efficacy of classifier-accuracy tests, the upper bounds in Tables 1 and 2 follow from the combination of Lemma 2 and the following results:

- $\mathcal{P}_{\mathsf{Db}}$: see Corollary 10;
- $\mathcal{P}_{\mathsf{H}}$: see Section 3.2 and Corollary 10;
- $\mathcal{P}_{\mathsf{G}}$: see Proposition 14;
- $\mathcal{P}_{\mathsf{D}}$: for GoF, see Proposition 7 if $k < \log(1/\delta)/\epsilon^4$, and Proposition 12 otherwise; for TS, see Proposition 7; for LFHT, see Proposition 7 if $n \geq k \wedge m$, and Section 3.1.3 and Proposition 12 otherwise.

3. Learning separating sets

In this section, we construct the separating sets S used in the classifier-accuracy test (2). Section 3.1 is devoted to discrete distribution models $\mathcal{P}_{\mathsf{Db}}$ and $\mathcal{P}_{\mathsf{D}}$, where we need a delicate tradeoff between the expected separation and the size of S. A similar construction in the Gaussian sequence model $\mathcal{P}_{\mathsf{G}}$ is presented in Section 3.3.

3.1. The discrete case

Given two iid samples X, Y of sizes $N_X, N_Y \overset{iid}{\sim} \mathrm{Poi}(n)$ from unknown discrete distributions $p = (p_1, \ldots, p_k), q = (q_1, \ldots, q_k)$ over a finite alphabet $[k] = \{1, 2, \ldots, k\}$, can we learn a set $\hat{S} \subseteq [k]$ using X, Y that separates p from q? To measure the quality of a given separating set $A \subseteq [k]$, we define two quantities $\mathrm{sep}(A) \triangleq p(A) - q(A)$ and $\tau(A) \triangleq \min\{p(A)p(A^{\mathsf{c}}), q(A)q(A^{\mathsf{c}})\}$. Intuitively, the first quantity $\mathrm{sep}(A)$ measures the separation of A, and the second quantity $\tau(A)$ measures the size of A. Recall that by Lemma 2, in order to perform the classifier-accuracy test (2), we aim to find a separating set $\hat{S}$ such that

$$|\mathrm{sep}(\hat{S})| \text{ is large and } \tau(\hat{S}) \text{ is small.} \tag{5}$$

The rest of this section is devoted to the construction of $\hat{S}$ satisfying (5), and we will present our results on learning separating sets in order of increasing complexity.

Notation: for a random variable X we write $\sigma^2(X)$ for the optimal sub-Gaussian variance proxy of X. In other words, $\sigma^2(X)$ is the smallest value such that $\mathbb{E}\exp(\lambda(X - \mathbb{E}X)) \leq \exp(\lambda^2\sigma^2(X)/2)$ holds for all $\lambda \in \mathbb{R}$.

3.1.1. A NATURAL SEPARATING SET

Let $\{X_i, Y_i\}_{i \in [k]}$ be the empirical frequencies of each bin $i \in [k]$ in our samples X, Y, i.e. $nX_i \sim \mathrm{Poi}(np_i)$ and $nY_i \sim \mathrm{Poi}(nq_i)$. A natural separating set is the following:

$$\hat{S}_{1/2} \triangleq \{i : X_i > Y_i \text{ or } X_i = Y_i \text{ and } C_i = 1\},$$

where $C_1, C_2 \ldots C_k$ are iid $\mathrm{Ber}(1/2)$ random variables. We use the subscript "$1/2$" to illustrate our tie-breaking rule: when $X_i = Y_i$, the symbol i is added to the set with probability $1/2$.

Our first result concerns the separating power of the above set.

Proposition 7 *Suppose $p, q \in \mathcal{P}_{\mathsf{D}}(k)$ with $\mathsf{TV}(p, q) \geq \epsilon$. There exists a universal constant $c > 0$ such that*

$$\mathbb{P}\left(\mathrm{sep}(\hat{S}_{1/2}) \geq c\epsilon^2\left(\frac{n}{k} \wedge \sqrt{\frac{n}{k}} \wedge \frac{1}{\epsilon}\right)\right) \geq 1 - \delta,$$

provided $n \geq \frac{1}{c} n_{\mathsf{TS}}(\epsilon, \delta, \mathcal{P}_{\mathsf{D}}(k))$.

Together with the trivial upper bound $\tau(\hat{S}_{1/2}) \leq 1/4$, Proposition 7 and Lemma 2 imply that using $\hat{S}_{1/2}$ achieves the minimax sample complexity for the following problems:

- GoF in $\mathcal{P}_{\mathsf{Db}}$ and $\mathcal{P}_{\mathsf{D}}$ as long as $k = \mathcal{O}(\log(1/\delta)/\epsilon^4)$;
- TS in $\mathcal{P}_{\mathsf{Db}}$ as long as $k = \mathcal{O}(\log(1/\delta)/\epsilon^4)$, and in $\mathcal{P}_{\mathsf{D}}$ for all (k, ϵ, δ);
- LFHT in $\mathcal{P}_{\mathsf{Db}}$ as long as $k = \mathcal{O}(\log(1/\delta)/\epsilon^4)$, and in $\mathcal{P}_{\mathsf{D}}$ as long as $n \geq m$.

However, in the remaining regimes the above test could be strictly sub-optimal. This failure comes down to two issues. First, Proposition 7 requires $n \gtrsim n_{\mathsf{TS}}(\epsilon, \delta, \mathcal{P}_{\mathsf{D}}(k))$ in order to find a good separating set, which can be sub-optimal when the optimal sample complexity for the original testing problem is only $n \gtrsim n_{\mathsf{GoF}}(\epsilon, \delta, \mathcal{P}_{\mathsf{D}}(k))$. Second, the quantity $\tau(\hat{S}_{1/2})$ is $\Omega(1)$ in the general case because the tie-breaking rule adds too many symbols to the set. These issues will be addressed separately in the next two sections.

3.1.2. THE "BETTER OF TWO" SEPARATING SETS

This section aims to find a separating set $\hat{S}$ with essentially the same separation as $\hat{S}_{1/2}$ in Proposition 7, but with a smaller $\tau(\hat{S})$. The central idea is to use a different tie-breaking rule from $\hat{S}_{1/2}$. Given a subset $D \subseteq [k]$, we define the imbalanced separating sets

$$\hat{S}_{>}(D) = \{i \in D : X_i > Y_i\},$$
$$\hat{S}_{<}(D) = \{i \in D : X_i < Y_i\}.$$

In other words, in both $\hat{S}_{>}$ and $\hat{S}_{<}$, we do not include the symbols with $X_i = Y_i$ in the separating set. Consequently, $|\hat{S}_{>}(D)| \vee |\hat{S}_{<}(D)|$ is upper bounded by the sample size; if in addition q_i is bounded from above uniformly over $i \in D$, this will yield good control of τ for both separating sets $\hat{S}_{>}(D)$ and $\hat{S}_{<}(D)$. In particular, $\tau(\hat{S}_{>}(D)) \vee \tau(\hat{S}_{<}(D)) = \mathcal{O}(1 \wedge (n \max_{i \in D} q_i))$.

Next we aim to show that the above sets achieve good separation. However, there is a subtlety here: removing the ties from $\hat{S}_{1/2}$ may no longer guarantee the desired separation, as illustrated in the following proposition.

Proposition 8 *Consider the distributions p, q on $[3k]$ with $p_i = \mathbb{1}\{i \leq k\}/(2k) + \mathbb{1}\{i > k\}/(4k)$ and $q_i = \mathbb{1}\{i \leq k\}/k$. Then, for $n \leq 0.6k$,*

$$\mathbb{E}\,\mathrm{sep}(\hat{S}_{>}([3k])) < 0.$$

Proposition 8 shows that sticking to only one set $\hat{S}_{>}$ or $\hat{S}_{<}$ fails to give the same separation guarantees as Proposition 7. A priori it may seem that $\hat{S}_{>}$ is designed to capture elements of the

support where p is greater than q, but it fails to do so spectacularly. An intuitive explanation of this phenomenon is as follows. Since the probability of each bin is small ($\lesssim 1/k$) under both p and q, in the small n regime[2] can expect that (a) each bin appears either once or not at all and (b) there is no overlap between the observed bins in sample X and Y. In this heuristic picture, the set $\hat{S}_>$ is simply the set of observed bins in the X-sample. Each X-sample falling in the first k bins contributes $-\frac{1}{2k}$ to the separation, while each X-sample in the last $2k$ bins contributes only $+\frac{1}{4k}$ to the separation. Since p puts mass $1/2$ on both the first k and last $2k$ bins, there is an equal number of $n/2$ observations in each part and the overall separation is $\asymp -\frac{n}{8k}$. Similar results can be proved for $\hat{S}_<$ with p, q as above but swapped, and also for modified p, q separated by smaller ϵ in TV for any $\epsilon \in (0, 1)$.

Motivated by the above discussion, in the sequel we consider the sets $\hat{S}_>, \hat{S}_<$ jointly. Specifically, the next proposition shows that *at least one of* the sets $\hat{S}_>$ and $\hat{S}_<$ have a good separation.

Proposition 9 *There exists a universal constant $c > 0$ such that for any $D \subseteq [k]$ and probability mass functions p, q, it holds that*

$$\mathbb{E}\left[\mathrm{sep}(\hat{S}_>(D)) - \mathrm{sep}(\hat{S}_<(D))\right] \geq c \sum_{i \in D} \frac{n(p_i - q_i)^2}{\sqrt{n(p_i \wedge q_i) + 1}} \wedge |p_i - q_i|,$$
$$\sigma^2(\mathrm{sep}(\hat{S}_>(D))) + \sigma^2(\mathrm{sep}(\hat{S}_<(D))) \leq \frac{1}{c} \sum_{i \in D} \frac{p_i + q_i}{n} \wedge |p_i - q_i|^2.$$

Based on Proposition 9, our final separating set is chosen from these two options, based on evaluation on held out data. As for the choice of D, in this section we choose $D = [k]$. The following corollary summarizes the performance of this choice under $\mathcal{P}_{\mathsf{Db}}$.

Corollary 10 *Suppose $p, q \in \mathcal{P}_{\mathsf{Db}}(k, \mathcal{O}(1))$ with $\mathsf{TV}(p, q) \geq \epsilon$. There exists a universal constant $c > 0$ such that using the samples X, Y we can find a set $\hat{S} \subseteq [k]$ which, with probability $1 - \delta$, satisfies*

$$\left|\mathrm{sep}(\hat{S})\right| \geq c\epsilon^2 \left(\frac{1}{\epsilon} \wedge \sqrt{\frac{n}{k}} \wedge \frac{n}{k}\right) \qquad \text{and} \qquad \tau(\hat{S}) \leq \frac{1}{c}\left(1 \wedge \frac{n}{k}\right), \tag{6}$$

provided $n \geq \frac{1}{c} n_{\mathsf{GoF}}(\epsilon, \delta, \mathcal{P}_{\mathsf{Db}}(k, \mathcal{O}(1)))$.

By Corollary 10 and Lemma 2, using the above set $\hat{S}$ achieves the minimax sample complexity for all problems GoF, TS, and LFHT and all parameters (k, ϵ, δ) under $\mathcal{P}_{\mathsf{Db}}$. However, under $\mathcal{P}_{\mathsf{D}}$, the performance of $\hat{S}$ is no better than that of $\hat{S}_{1/2}$. This is because a good control of $\tau(\hat{S}_>([k]))$ requires a bounded probability mass function; in other words, choosing $D = [k]$ is not optimal for finding the best separating set under $\mathcal{P}_{\mathsf{D}}$. In the next section, we address this issue by choosing D to be one of $\mathcal{O}(\log k)$ subsets of $[k]$.

3.1.3. THE "BEST OF $\mathcal{O}(\log k)$" SEPARATING SETS

This section is devoted to the two missing regimes $m \geq n$ for LFHT over $\mathcal{P}_{\mathsf{D}}$ and $k \gtrsim \log(1/\delta)/\epsilon^4$ for GoF over $\mathcal{P}_{\mathsf{D}}$ (cf. discussion after Proposition 3 and Corollary 10). For the former, recall that

[2] Technically, to satisfy the stated conditions we would require $n \lesssim \sqrt{k}$, but the described event captures dominant effects even for larger $\sqrt{k} \ll n \ll k$.

the classifier-accuracy test based on $\hat{S}_{1/2}$ achieves the sample complexity

$$n \gtrsim n_{\mathsf{GoF}}(\epsilon, \delta, \mathcal{P}_{\mathsf{D}}) + \frac{k\sqrt{\log(1/\delta)}}{\sqrt{n}\epsilon^2}. \tag{7}$$

If $n \gtrsim k$ then (7) is the same as $n \gtrsim n_{\mathsf{GoF}}$; if $m/\log(1/\delta) \lesssim n$ then (7) is implied by $n \gtrsim n_{\mathsf{GoF}} + \frac{k\log(1/\delta)}{\sqrt{m}\epsilon^2}$, which is optimal within an $O(\log^{1/2}(1/\delta))$ factor (cf. Table 2). In our application to GoF we take $m = \infty$, and the missing regime $k \gtrsim \log(1/\delta)/\epsilon^4$ corresponds precisely to $n_{\mathsf{GoF}} \lesssim k$. Summarizing, in the remainder of this section we may assume that $k \wedge (m/\log(1/\delta)) \gtrsim n$.

Let $t = k \wedge (c_0 m/\log(1/\delta))$, where $c_0 > 0$ is a small absolute constant. By the previous paragraph, we assume without loss of generality that $t > n$. For $\ell = \lceil \log_2(t/n) \rceil \geq 1$, define the following $\ell + 2$ subsets of $[k]$:

$$D_0 = \left\{ i : \hat{q}_i^0 \leq \frac{1}{t} \right\}, \quad D_j = \left\{ i : \hat{q}_i^0 \in \left(\frac{2^{j-1}}{t}, \frac{2^j}{t} \right] \right\} \text{ for } j \in [\ell], \quad D_{\ell+1} = \left\{ i : \hat{q}_i^0 > \frac{2^\ell}{t} \right\}.$$

Here $\hat{q}_i^0$ denotes the empirical pmf of $m/2$ held out samples drawn from q (for GoF, one can understand $\hat{q}_i^0 = q_i$ for the distribution q is known). The motivation behind the above choices is the "localization" of each $\hat{q}_i^0$, as shown in the following lemma.

Lemma 11 *For a small enough universal constant $c_0 > 0$, with probability at least $1 - k\delta$ it holds that for each $i \in [k]$:*

1. *if $\hat{q}_i^0 \in D_0$, then $q_i < 2/t$;*
2. *if $\hat{q}_i^0 \in D_j$ for some $j \in [\ell]$, then $q_i \in (2^{j-2}/t, 2^{j+1}/t]$;*
3. *if $\hat{q}_i^0 \in D_{\ell+1}$, then $q_i > 2^{\ell-1}/t$.*

Lemma 11 ensures that with high probability, the distribution q restricted to each set D_j is near-uniform. This is similar in spirit to the idea of flattening used in distribution testing Diakonikolas and Kane (2016). The proof of Lemma 11 directly follows from the Poisson concentration in Lemma 17 and is thus omitted.

Our main result of this section is the next proposition, which shows that there exist some $j \in \{0, 1, \cdots, \ell+1\}$ and $\hat{S} \subseteq D_j$ such that $\hat{S}$ is a near-optimal separating set within logarithmic factors.

Proposition 12 *Suppose $p, q \in \mathcal{P}_{\mathsf{D}}(k)$ with $\mathsf{TV}(p, q) \geq \epsilon$, and X, Y are n iid samples drawn from p, q respectively. There exists a universal constant $c > 0$ such that using the samples X, Y, we can find some $j \in \{0, 1, \cdots, \ell + 1\}$ and a set $\hat{S} \subseteq D_j$ which, with probability $1 - \mathcal{O}(k\delta)$, satisfies*

$$\left|\mathrm{sep}(\hat{S})\right| \geq c \left(\frac{\epsilon}{\ell}\right)^2 \left\{ \begin{array}{ll} n/k & \text{if } j = 0 \\ n/\sqrt{kt/2^j} & \text{if } j \in [\ell + 1] \end{array} \right\} \qquad \text{and} \quad \tau(\hat{S}) \leq \frac{n2^j}{ct}$$

provided that

$$n\sqrt{1 \wedge \frac{m}{\log(1/\delta)k}} \geq \frac{1}{c} n_{\mathsf{GoF}}(\epsilon/\ell, \delta, \mathcal{P}_{\mathsf{D}}).$$

By Proposition 12 and Lemma 2, using the above set $\hat{S}$ leads to the following sample complexity guarantee for the problems GoF and LFHT:

- for GoF under $\mathcal{P}_{\mathsf{D}}$, it succeeds with $n = \Theta(n_{\mathsf{GoF}}(\epsilon/\ell, \delta/k, \mathcal{P}_{\mathsf{D}}))$ observations, which is within a multiplicative $\mathcal{O}(\log^{\Theta(1)}(k))$ factor of the minimax optimal sample complexity in the missing $k \geq \log(1/\delta)/\epsilon^4$ regime;
- for LFHT under $\mathcal{P}_{\mathsf{D}}$ and $m \geq n$, it succeeds with $n = \Theta(n_{\mathsf{GoF}}(\epsilon/\ell, \delta/k, \mathcal{P}_{\mathsf{D}})\sqrt{k\log(k/\delta)/m})$ observations, which is within a multiplicative $\mathcal{O}(\log^{\Theta(1)}(k)\log(k/\delta))$ factor of the minimax optimal sample complexity in the missing $n \leq m \wedge k$.

Therefore, classifier-accuracy tests always lead to near-optimal sample complexities for all GoF, TS, and LFHT problems under both $\mathcal{P}_{\mathsf{Db}}$ and $\mathcal{P}_{\mathsf{D}}$, within polylogarithmic factors in $(k, 1/\delta)$. We leave the removal of extra logarithmic factors for classifier-accuracy tests as an open problem.

3.2. The smooth density case

We briefly explain how Corollary 10 can be used to learn separating sets between distributions in the class $\mathcal{P}_{\mathsf{H}}$ of β-Hölder smooth distributions on $[0,1]^d$. The reduction relies on an approximation result due to Ingster Ingster (1987); Ingster and Suslina (2003), see also (Arias-Castro et al., 2018, Lemma 7.2). Let P_r be the L^2-projection onto piecewise constant functions on the regular grid on $[0,1]^d$ with r^d cells.

Lemma 13 *There exist constants c_1, c_2 independent of r such that for any $f \in \mathcal{P}_{\mathsf{H}}(\beta, d, C_{\mathsf{H}})$,*

$$\|P_r f\|_2 \geq c_1\|f\|_2 - c_2 r^{-\beta}.$$

For simplicity write f, g for the Lebesgue densities of $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}} \in \mathcal{P}_{\mathsf{H}}$. Suppose $\mathsf{TV}(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}) = \frac{1}{2}\|f-g\|_1 \geq \epsilon$. By Jensen's inequality and Lemma 13, $\epsilon \lesssim \|P_r(f-g)\|_2$ for $r \asymp \epsilon^{-1/\beta}$. The key observation is that $P_r f$ is essentially the probability mass function of the distribution $\mathbb{P}_{\mathsf{X}}$ when binned on the regular grid with r^d cells. We can now directly apply the results for $\mathcal{P}_{\mathsf{Db}}$ (Corollary 10) with alphabet size $k \asymp \epsilon^{-d/\beta}$, which combined with Lemma 2 leads to the sample complexity guarantees in Table 1 for the smooth density class $\mathcal{P}_{\mathsf{H}}$ in all three problems GoF, TS and LFHT.

3.3. The Gaussian case

Suppose we have two samples X, Y of size n from $\otimes_{j=1}^{\infty}\mathcal{N}(\theta_j^X, 1) =: \mu_{\theta^X}$ and μ_{θ^Y} respectively, where θ^X, θ^Y have Sobolev norm $\|\theta\|_s^2 \triangleq \sum_j \theta_j^2 j^{2s}$ bounded by a constant. In addition, $\mathsf{TV}(\mu_{\theta^X}, \mu_{\theta^Y}) \geq \epsilon > 0$. We use $\hat{\theta}^X$ and $\hat{\theta}^Y$ to denote the empirical mean vector from samples X and Y, respectively.

The separating set is constructed as follows:

$$\hat{S} = \{Z \in \mathbb{R}^{\mathbb{N}} : T(Z) \geq 0\},$$

where $T(Z) = 2\sum_{j=1}^{J}(\hat{\theta}_j^X - \hat{\theta}_j^Y)(Z_j - (\hat{\theta}_j^X + \hat{\theta}_j^Y)/2)$ for some $J \in \mathbb{N}$ to be specified. This is simply a truncated version of the likelihood-ratio test between $\mu_{\hat{\theta}^X}$ and $\mu_{\hat{\theta}^Y}$, where we set all but the first J coordinates of $\hat{\theta}^X$ and $\hat{\theta}^Y$ to zero. The performance of the separating set is summarized in the next proposition.

Proposition 14 *There exists universal constants c, c' such that when $J = \lfloor c\epsilon^{-1/s} \rfloor$ the inequality*

$$\mathbb{P}\left(\mu_{\theta^X}(\hat{S}) - \mu_{\theta^Y}(\hat{S}) \geq c'\left(\sqrt{n\epsilon^{1/s}} \wedge \frac{1}{\epsilon}\right)\epsilon^2\right) \geq 1 - \delta$$

holds, provided $n \gtrsim \frac{1}{c'} n_{\mathsf{TS}}(\epsilon, \delta, \mathcal{P}_{\mathsf{G}})$.

Applying Proposition 14 and Lemma 2 with the trivial bound $\tau(\hat{S}) \leq 1/4$ leads to the sample complexity guarantees in Table 1 for the Gaussian sequence model class $\mathcal{P}_{\mathsf{G}}$ in all three problems GoF, TS and LFHT.

Acknowledgments

YH was generously supported by the Norbert Wiener postdoctoral fellowship in statistics at MIT IDSS. YP was supported in part by the National Science Foundation under Grant No CCF-2131115. Research was sponsored by the United States Air Force Research Laboratory and the Department of the Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department of the Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Claire Adam-Bourdarios, Glen Cowan, Cécile Germain, Isabelle Guyon, Balázs Kégl, and David Rousseau. The higgs boson machine learning challenge. In *NIPS 2014 workshop on high-energy physics and machine learning*, pages 19–55. PMLR, 2015.

Ery Arias-Castro, Bruno Pelletier, and Venkatesh Saligrama. Remember the curse of dimensionality: The case of goodness-of-fit testing in arbitrary dimension. *Journal of Nonparametric Statistics*, 30(2):448–471, 2018.

Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D Smith, and Patrick White. Testing that distributions are close. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 259–269. IEEE, 2000.

Bhaswar Bhattacharya and Gregory Valiant. Testing closeness with unequal sized samples. *Advances in Neural Information Processing Systems*, 28, 2015.

V Buldygin and K Moskvichova. The sub-gaussian norm of a binary random variable. *Theory of probability and mathematical statistics*, 86:33–49, 2013.

Clément L Canonne. A survey on distribution testing: Your data is big. but is it blue? *Theory of Computing*, pages 1–100, 2020.

Siu-On Chan, Ilias Diakonikolas, Paul Valiant, and Gregory Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 1193–1203. SIAM, 2014.

Serguei Chatrchyan, Vardan Khachatryan, Albert M Sirunyan, Armen Tumasyan, Wolfgang Adam, Ernest Aguilo, Thomas Bergauer, M Dragicevic, J Erö, C Fabjan, et al. Observation of a new boson at a mass of 125 gev with the cms experiment at the lhc. *Physics Letters B*, 716(1):30–61, 2012.

Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2001.

Ilias Diakonikolas and Daniel M Kane. A new approach for testing properties of discrete distributions. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 685–694. IEEE, 2016.

Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. Sample-optimal identity testing with high probability. In *45th International Colloquium on Automata, Languages, and Programming (ICALP 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

Ilias Diakonikolas, Themis Gouleakis, Daniel M Kane, John Peebles, and Eric Price. Optimal testing of discrete distributions with high probability. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 542–555, 2021.

Peter J Diggle and Richard J Gratton. Monte carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(2):193–212, 1984.

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

Patrik R Gerber and Yury Polyanskiy. Likelihood-free hypothesis testing. *arXiv preprint arXiv:2211.01126*, 2022.

Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. 2000.

Polina Golland and Bruce Fischl. Permutation tests for classification: towards statistical significance in image-based studies. In *Biennial international conference on information processing in medical imaging*, pages 330–341. Springer, 2003.

Michael Gutman. Asymptotically optimal classification for multiple tests with empirically observed statistics. *IEEE Transactions on Information Theory*, 35(2):401–408, 1989.

Michael U Gutmann, Ritabrata Dutta, Samuel Kaski, and Jukka Corander. Likelihood-free inference via classification. *Statistics and Computing*, 28(2):411–425, 2018.

Simon Hediger, Loris Michel, and Jeffrey Näf. On the use of random forest for two-sample testing. *Computational Statistics & Data Analysis*, 170:107435, 2022.

Dayu Huang and Sean Meyn. Classification with high-dimensional sparse samples. In *2012 IEEE International Symposium on Information Theory Proceedings*, pages 2586–2590. IEEE, 2012.

Dayu Huang and Sean Meyn. Generalized error exponents for small sample universal hypothesis testing. *IEEE transactions on information theory*, 59(12):8157–8181, 2013.

Yuri I Ingster. On the minimax nonparametric detection of signals in white gaussian noise. *Problemy Peredachi Informatsii*, 18(2):61–73, 1982.

Yuri I Ingster. Minimax testing of nonparametric hypotheses on a distribution density in the l_p metrics. *Theory of Probability & Its Applications*, 31(2):333–337, 1987.

Yuri I Ingster and Irina A Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*, volume 169. Springer Science & Business Media, 2003.

Bai Jiang, Tung-yu Wu, Charles Zheng, and Wing H Wong. Learning summary statistic for approximate bayesian computation via deep neural network. *Statistica Sinica*, pages 1595–1618, 2017.

Benjamin G Kelly, Thitidej Tularak, Aaron B Wagner, and Pramod Viswanath. Universal hypothesis testing in the learning-limited regime. In *2010 IEEE International Symposium on Information Theory*, pages 1478–1482. IEEE, 2010.

Benjamin G Kelly, Aaron B Wagner, Thitidej Tularak, and Pramod Viswanath. Classification of homogeneous data with large alphabets. *IEEE transactions on information theory*, 59(2):782–795, 2012.

Ilmun Kim, Aaditya Ramdas, Aarti Singh, and Larry Wasserman. Classification accuracy as a proxy for two-sample testing. *The Annals of Statistics*, 49(1):411–434, 2021.

Tong Li and Ming Yuan. On the optimality of gaussian kernel based nonparametric tests against smooth alternatives. *arXiv preprint arXiv:1909.03302*, 2019.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. *arXiv preprint arXiv:1610.06545*, 2016.

Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge university press, 2017.

Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008.

Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2023+.

Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519.

Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46(1):429–455, 2017.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Jacob Ziv. On classification with empirically observed statistics and universal data compression. *IEEE Trans. Inf. Theory*, 34:278–286, 1988.

Appendix A. Auxiliary Lemmas

We state some auxiliary lemmas which will be used for the proof. We begin with a simple identity for standard normal distributions.

Lemma 15 *Take $a, b \in \mathbb{R}$ and let Z be standard normal. Then*

$$\mathbb{E}\Phi(aZ+b) = \Phi\left(\frac{b}{\sqrt{1+a^2}}\right).$$

Proof Let Z' be a standard Gaussian independent of Z. Then

$$\mathbb{E}\Phi(aZ+b) = \mathbb{P}(aZ+b \geq Z') = \mathbb{P}\left(\frac{Z'-aZ}{\sqrt{1+a^2}} \leq \frac{b}{\sqrt{1+a^2}}\right) = \Phi\left(\frac{b}{\sqrt{1+a^2}}\right).$$

■

The following lemma is the celebrated result of Gaussian Lipschitz concentration.

Lemma 16 (Lipschitz concentration for Gaussians (Vershynin, 2018, Theorem 5.2.1)) *Let Q be a d-dimensional standard Gaussian and let $f : \mathbb{R}^d \to \mathbb{R}$ be σ-Lipschitz. Then $f(Q)$ is sub-Gaussian with variance proxy σ^2.*

The next lemma states the Chernoff bound for Poisson random variables.

Lemma 17 ((Mitzenmacher and Upfal, 2017, Theorem 5.4)) *For all $\lambda > 0$ and $x \geq 0$ we have*

$$\mathbb{P}(\mathrm{Poi}(\lambda) - \lambda \geq x) \leq \exp\left(-\frac{x^2}{2(\lambda+x)}\right),$$
$$\mathbb{P}(\mathrm{Poi}(\lambda) - \lambda \leq -x) \leq \exp\left(-\frac{x^2}{2\lambda}\right).$$

The following technical lemma is helpful in establishing the Bernstein concentration in Lemma 19.

Lemma 18 *Let $a \geq 0, p, q \in [0,1]$ and define $\tau = p(1-p) \wedge q(1-q), \nu = p(1-p) \vee q(1-q)$. Then it always holds that*

$$a\sqrt{\frac{\nu}{2}} \leq a\sqrt{\tau} + a^2 + |p-q|.$$

In particular, if $|p-q| \geq a\sqrt{\tau} + a^2$, then

$$4|p-q| \geq a\sqrt{\tau} + a\sqrt{\nu} + a^2.$$

Proof After rearranging and noting that $1 + 2\sqrt{2} < 4$, it is clear that the first inequality implies the second. Below we prove the first inequality.

Since the claim is invariant under the transformations $(p, q) \mapsto (q, p)$ and $(p, q) \mapsto (1-p, 1-q)$, it suffices to consider the case where $p \le 1/2$ and $p(1-p) \le q(1-q)$. It further suffices to consider the case where $p \le q \le 1/2$: if not, then $p \le 1-q \le 1/2$, and the transformation $(p, q) \mapsto (p, 1-q)$ keeps (τ, ν) invariant while makes $|p - q|$ smaller. The proof is then completed by considering the following two scenarios:

- if $p \ge q/2$, then $\nu = q(1 - q) \le 2p(1 - p) = 2\tau$, so $a\sqrt{\nu/2} \le a\sqrt{\tau}$;
- if $p \le q/2$, then $2a\sqrt{\nu} \le a^2 + \nu \le a^2 + q \le a^2 + 2(q - p)$.

■

Appendix B. Omitted Proofs from Section 1

B.1. Proof of Lemma 2

Before we prove Lemma 2, we begin with a technical lemma on the Bernstein concentration of the classifier-accuracy test (2).

Lemma 19 *Suppose* $A_1, \dots, A_n \overset{iid}{\sim} \mathrm{Ber}(p)$ *and* $B_1, \dots, B_m \overset{iid}{\sim} \mathrm{Ber}(q)$. *Let* $\tau = p(1-p) \wedge q(1-q)$ *and define the averages* $\bar{A} = \frac{1}{n}\sum_{i=1}^n A_i$ *and* $\bar{B} = \frac{1}{m}\sum_{j=1}^m B_j$. *There exists a universal constant* $c > 0$ *such that*

$$\mathbb{P}\left(|\bar{A} - \bar{B}| \le \frac{1}{2}|p - q| - \frac{1}{2}\sqrt{\frac{c\log(1/\delta)\tau}{n \wedge m}} - \frac{1}{2}\frac{c\log(1/\delta)}{n \wedge m}\right) \le \delta,$$

$$\mathbb{P}\left(|\bar{A} - \bar{B}| \ge 2|p - q| + 2\sqrt{\frac{c\log(1/\delta)\tau}{n \wedge m}} + 2\frac{c\log(1/\delta)}{n \wedge m}\right) \le \delta.$$

Proof Let $\nu = p(1 - p) \vee q(1 - q)$. Note that the first inequality is trivially true if

$$|p - q| \le \sqrt{\frac{c\log(1/\delta)\tau}{n \wedge m}} + \frac{c\log(1/\delta)}{n \wedge m}.$$

Assuming otherwise, by the second statement of Lemma 18, the first probability is upper bounded by

$$\mathbb{P}\left(|\bar{A} - \bar{B}| \le |p - q| - \frac{5}{8}\sqrt{\frac{c\log(1/\delta)\tau}{n \wedge m}} - \frac{1}{8}\sqrt{\frac{c\log(1/\delta)\nu}{n \wedge m}} - \frac{5}{8}\frac{c\log(1/\delta)}{n \wedge m}\right).$$

By choosing c sufficiently large (independently of p, q, n, m, δ), and applying Bernstein's inequality separately to both $\bar{A}$ and $\bar{B}$, the above probability can be made smaller than δ.

For the second inequality, using the first statement of Lemma 18, it is upper bounded by

$$\mathbb{P}\left(|\bar{A} - \bar{B}| \ge |p - q| + \sqrt{\frac{c\log(1/\delta)\tau}{n \wedge m}} + \frac{1}{\sqrt{2}}\sqrt{\frac{c\log(1/\delta)\nu}{n \wedge m}} + \frac{c\log(1/\delta)}{n \wedge m}\right).$$

Again, taking c sufficiently large (independently of p, q, n, m, δ) and applying Bernstein's inequality separately to both $\bar{A}$ and $\bar{B}$, the above probability can be made smaller than δ. ∎

Now we proceed to prove Lemma 2. Using n test samples (X, Y) from both p and q, consider the following classifier-accuracy test: we accept H_0 if

$$\left|\frac{1}{n}\sum_{i=1}^{n}(\mathbb{1}(X_i \in S) - \mathbb{1}(Y_i \in S))\right| \le \sqrt{\frac{c\overline{\tau}\log(1/\delta)}{n}} + \frac{c\log(1/\delta)}{n},$$

and reject H_0 otherwise. Here $c > 0$ is a large absolute constant, and we note that the threshold only relies on the knowledge of $\overline{\tau}$ in addition to (n, δ).

To analyze the type-I and type-II errors, first assume that H_0 holds. Since $\mathrm{sep}(S) = 0$ under H_0, the second statement of Lemma 19 implies that we accept H_0 with probability at least $1 - \delta/2$ if $c > 0$ is large enough. If H_1 holds, with probability at least $1 - \delta/2$, by the first statement of Lemma 19 we have

$$\left|\frac{1}{n}\sum_{i=1}^{n}(\mathbb{1}(X_i \in S) - \mathbb{1}(Y_i \in S))\right| \ge |\underline{\mathrm{sep}}| - \left(\sqrt{\frac{c\overline{\tau}\log(1/\delta)}{n}} + \frac{c\log(1/\delta)}{n}\right).$$

By the lower bound of n assumed in Lemma 2, in this case we will reject H_0, as desired.

B.2. Proof of Proposition 3

Lemma 20 *Let μ be a non-negative measure on some space $\mathcal{X}$ and let $a, b : \mathcal{X} \to \mathbb{R}_+$ such that $\int a(x)\mathrm{d}\mu(x) > 0$ and $b(x) = 0$ only if $a(x) = 0$. Then*

$$\inf_{x\in\mathrm{spt}(\mu)}\left(\frac{a(x)}{b(x)}\right) \le \frac{\int a(x)\mathrm{d}\mu(x)}{\int b(x)\mathrm{d}\mu(x)} \le \sup_{x\in\mathrm{spt}(\mu)}\left(\frac{a(x)}{b(x)}\right).$$

Proof Defining $0/0 = 1$, we have

$$\begin{aligned}\int a(x)\mathrm{d}\mu(x) &= \int \frac{a(x)}{b(x)}b(x)\mathrm{d}\mu(x)\\ &\le \sup_{x\in\mathrm{spt}(\mu)}\left(\frac{a(x)}{b(x)}\right)\int b(x)\mathrm{d}\mu(x).\end{aligned}$$

The other direction follows analogously. ∎

Proof [Proof of Proposition 3] Let p, q be the densities of $\mathbb{P}, \mathbb{Q}$ with respect to a common dominating measure, and let $E \triangleq \{x : p(x) > q(x)\}$ so that $\mathsf{TV}(\mathbb{P}, \mathbb{Q}) = \mathbb{P}(E) - \mathbb{Q}(E) > 0$. Assume without loss of generality that $\mathbb{P}(E) + \mathbb{Q}(E) \ge 1$. Given $t \in [0, 1]$ define $E_t \triangleq \{x : \frac{p(x)-q(x)}{p(x)+q(x)} \ge t\}$, so that the map $t \mapsto \mathbb{P}(E_t) + \mathbb{Q}(E_t)$ is non-increasing and left-continuous. Note that $E_0 = E$ while $E_1 = \varnothing$, so that $t^\star = \max\{t \in [0, 1] : \mathbb{P}(E_t) + \mathbb{Q}(E_t) \ge 1\}$ exists. Now choose the randomized classifier $\mathcal{C}$ as follows:

$$\mathcal{C}(x) = \begin{cases} 0 & \text{if } x \in E_{(t^\star)^+},\\ 1 & \text{if } x \notin E_{t^\star},\\ \mathrm{Ber}(r) & \text{if } x \in E_{t^\star} - E_{(t^\star)^+},\end{cases}$$

where $E_{(t^\star)^+} = \cap_{t>t^\star} E_t \subseteq E_{t^\star}$, and

$$r := \frac{1 - \mathbb{P}(E_{(t^\star)^+}) - \mathbb{Q}(E_{(t^\star)^+})}{\mathbb{P}(E_{t^\star}) + \mathbb{Q}(E_{t^\star}) - \mathbb{P}(E_{(t^\star)^+}) - \mathbb{Q}(E_{(t^\star)^+})} \in [0,1].$$

This classifier is balanced, as

$$\begin{aligned}
&\mathbb{P}(\mathcal{C}(X)=0) + \mathbb{Q}(\mathcal{C}(X)=0) \\
&= \mathbb{P}(E_{(t^\star)^+}) + \mathbb{Q}(E_{(t^\star)^+}) + r(\mathbb{P}(E_{t^\star}) + \mathbb{Q}(E_{t^\star}) - \mathbb{P}(E_{(t^\star)^+}) - \mathbb{Q}(E_{(t^\star)^+})) \\
&= 1.
\end{aligned}$$

For $t \in [0,1]$ define

$$f(t) \triangleq \begin{cases} (\mathbb{P}(E_t) - \mathbb{Q}(E_t))/(\mathbb{P}(E_t) + \mathbb{Q}(E_t)) & \text{if } \mathbb{P}(E_t) + \mathbb{Q}(E_t) > 0, \\ 1 & \text{otherwise.} \end{cases}$$

Let $0 \le t \le s \le 1$, we show that $f(t) \le f(s)$. Without loss of generality assume that $f(s) < 1$ and that $\mathbb{P}(E_s \backslash E_t) + \mathbb{Q}(E_s \backslash E_t) > 0$. Notice that $f(t) \le f(s)$ if and only if

$$\frac{\int_{E_t \backslash E_s} (p(x) - q(x)) \mathrm{d}x}{\int_{E_t \backslash E_s} (p(x) + q(x)) \mathrm{d}x} \overset{!}{\le} \frac{\int_{E_s} (p(x) - q(x)) \mathrm{d}x}{\int_{E_s} (p(x) + q(x)) \mathrm{d}x}.$$

However, the above inequality follows from Lemma 20. Thus, it holds that

$$\frac{\mathbb{P}(\mathcal{C}(X)=0) - \mathbb{Q}(\mathcal{C}(X)=0)}{\mathbb{P}(\mathcal{C}(X)=0) + \mathbb{Q}(\mathcal{C}(X)=0)} \ge f(t^\star) \ge f(0) = \frac{\mathbb{P}(E) - \mathbb{Q}(E)}{\mathbb{P}(E) + \mathbb{Q}(E)}.$$

Plugging in $\mathbb{P}(\mathcal{C}(X)=0) + \mathbb{Q}(\mathcal{C}(X)=0) = 1$ and $\mathbb{P}(E) + \mathbb{Q}(E) \le 2$ yields the result.

To show tightness, one can consider $p(x) = \mathbb{1}_{[0,1]}$, $q(x) = (1+\epsilon)\mathbb{1}_{[0,1/(1+\epsilon)]}$, $\mathcal{C}(x) = \mathbb{1}_{x \in (1/(2+\epsilon),1]}$, and let $\epsilon \to 0^+$. ■

Appendix C. Omitted Proofs from Section 3

C.1. Useful Lemmas

Before we present the formal proofs, this section summarizes some useful lemmas on the expected value and sub-Gaussian concentration of the separation.

Lemma 21 *Let $\mu \ge \lambda \ge 0$ and $X \sim \mathrm{Poi}(\mu), Y \sim \mathrm{Poi}(\lambda)$. Then*

$$\mathbb{P}(X > Y) + \frac{1}{2}\mathbb{P}(X = Y) - \frac{1}{2} \ge c\left(\frac{\mu - \lambda}{\sqrt{\lambda + 1}} \wedge 1\right)$$

holds, where $c > 0$ is a universal constant.

Proof For $t \in [\lambda, \mu]$ define the function

$$f(t) = \mathbb{P}(\mathrm{Poi}(t) > Y) + \frac{1}{2}\mathbb{P}(\mathrm{Poi}(t) = Y).$$

Clearly $f(\lambda) = \frac{1}{2}$. We have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{P}(\mathrm{Poi}(t) > Y) = -\mathbb{P}(\mathrm{Poi}(t) > Y) + \mathbb{P}(\mathrm{Poi}(t) > Y - 1) = \mathbb{P}(\mathrm{Poi}(t) = Y).$$

Similarly we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{P}(\mathrm{Poi}(t) = Y) = -\mathbb{P}(\mathrm{Poi}(t) = Y) + \mathbb{P}(\mathrm{Poi}(t) = Y - 1).$$

Thus, we obtain

$$f'(t) = \frac{1}{2}\mathbb{E}\left[\mathbb{P}(\mathrm{Poi}(t) \in \{Y - 1, Y\})\right].$$

Next we prove the following inequality: if y is a non-negative integer with $|y - t| \le 8\sqrt{t}$, then

$$\mathbb{P}(\mathrm{Poi}(t) = y) = \Omega\left(\frac{1}{\sqrt{t+1}}\right). \tag{8}$$

To prove (8), we distinguish three scenarios:

1. If $t < 1/100$, then the only non-negative integer y with $|y - t| \le 8\sqrt{t}$ is $y = 0$. Therefore $\mathbb{P}(\mathrm{Poi}(t) = y) = e^{-t} = \Omega(1)$.

2. If $1/100 \le t \le 100$, then $0 \le y \le 180$. In this case,

$$\mathbb{P}(\mathrm{Poi}(t) = y) \ge \min_{1/100 \le t \le 100} \min_{0 \le y \le 180} \mathbb{P}(\mathrm{Poi}(t) = y) = \Omega(1).$$

3. If $t > 100$, then for $t - 8\sqrt{t} \le y_1 \le y_2 \le t + 8\sqrt{t}$, we have

$$\frac{\mathbb{P}(\mathrm{Poi}(t) = y_1)}{\mathbb{P}(\mathrm{Poi}(t) = y_2)} = t^{y_2 - y_1}\frac{y_2!}{y_1!} = \prod_{y=y_1+1}^{y_2}\frac{t}{y} = \left(1 \pm \mathcal{O}(t^{-1/2})\right)^{\mathcal{O}(16\sqrt{t})} = \Theta(1).$$

In the above we have used that $|t/y - 1| = \mathcal{O}(t^{-1/2})$ for all $y \in [y_1, y_2]$, and $y_2 - y_1 \le 16\sqrt{t}$. Consequently,

$$\mathbb{P}(\mathrm{Poi}(t) = y) = \Omega\left(\frac{\mathbb{P}(|\mathrm{Poi}(t) - t| \le 8\sqrt{t})}{16\sqrt{t}}\right) = \Omega\left(\frac{1}{\sqrt{t}}\right),$$

where the last step is due to Chebyshev's inequality.

Now we apply (8) to prove Lemma 21. We first show that for non-negative integer y,

$$\{|y-\lambda| \le 2\sqrt{\lambda}\} \wedge \{\sqrt{\lambda} \le \sqrt{t} \le \sqrt{\lambda}+1\} \Longrightarrow \{|y-t| \le 8\sqrt{t}\}. \tag{9}$$

In fact, if $\sqrt{\lambda} < \sqrt{2}-1$, then the LHS of (9) implies that $y = 0$ and $t < 2$, thus (9) holds. If $\sqrt{\lambda} \ge \sqrt{2}-1$, then the LHS of (9) implies that

$$|y-t| \le |y-\lambda| + (t-\lambda) \le 2\sqrt{\lambda} + (2\sqrt{\lambda}+1) < 8\sqrt{\lambda} \le 8\sqrt{t},$$

and (9) holds as well. Next, by (8) and (9), as well as Chebyshev's inequality $\mathbb{P}(|Y-\lambda| \le 2\sqrt{\lambda}) \ge \frac{3}{4}$, we have

$$\begin{aligned} f'(t) &\ge \frac{3}{8} \min_{y\ge 0:|y-\lambda|\le 2\sqrt{\lambda}} \mathbb{P}(\mathrm{Poi}(t)=y) \\ &\ge \frac{3}{8}\mathbb{1}\{\sqrt{\lambda} \le \sqrt{t} \le \sqrt{\lambda}+1\} \cdot \min_{|y-t|\le 8\sqrt{t}} \mathbb{P}(\mathrm{Poi}(t)=y) \\ &= \Omega\left(\frac{\mathbb{1}\{\sqrt{\lambda} \le \sqrt{t} \le \sqrt{\lambda}+1\}}{\sqrt{t+1}}\right) = \Omega\left(\frac{\mathbb{1}\{\sqrt{\lambda} \le \sqrt{t} \le \sqrt{\lambda}+1\}}{\sqrt{\lambda+1}}\right). \end{aligned}$$

Finally, for some absolute constant $c > 0$ it holds that

$$f(\mu) - f(\lambda) = \int_\lambda^\mu f'(t)\mathrm{d}t \ge c\int_\lambda^\mu \frac{\mathbb{1}\{\sqrt{\lambda} \le \sqrt{t} \le \sqrt{\lambda}+1\}}{\sqrt{\lambda+1}}\mathrm{d}t \ge c\left(\frac{\mu-\lambda}{\sqrt{\lambda+1}} \wedge 1\right),$$

which is the statement of the lemma. ■

Lemma 22 *For any $D \subseteq [k]$, each of $\mathrm{sep}(\hat{S}_s(D)), s \in \{>, <, 1/2\}$ is sub-Gaussian with variance proxy σ^2 which can be bounded as*

$$\sigma^2 \lesssim \sum_{i\in D}(p_i-q_i)^2 \wedge \frac{p_i+q_i}{n} = \mathcal{O}\left(\frac{1}{n}\right),$$

with universal hidden constants.

Proof Using standard tail bounds of the Poisson distribution (Lemma 17) we have for any $i \in D$ with $p_i > q_i$,

$$\begin{aligned} &\mathbb{P}(i \in \hat{S}_<(D)) \le \mathbb{P}(i \notin \hat{S}_{1/2}(D)) \le \mathbb{P}(i \notin \hat{S}_>(D)) \\ &= \mathbb{P}(\mathrm{Poi}(np_i) \le \mathrm{Poi}(nq_i)) \\ &\le \mathbb{P}\left(\mathrm{Poi}(np_i) - np_i \le -\frac{1}{2}n(p_i-q_i)\right) + \mathbb{P}\left(\mathrm{Poi}(nq_i) - nq_i > \frac{1}{2}n(p_i-q_i)\right) \\ &\le 2\exp\left(-c\frac{n(p_i-q_i)^2}{p_i+q_i}\right) \end{aligned}$$

for some universal $c > 0$. Similarly, if $i \in D$ with $p_i \le q_i$ we get

$$\mathbb{P}(i \in \hat{S}_>(D)) \le \mathbb{P}(i \in \hat{S}_{1/2}(D)) \le \mathbb{P}(i \notin \hat{S}_<(D)) = \mathbb{P}(\mathrm{Poi}(np_i) \ge \mathrm{Poi}(nq_i)) \le 2\exp\left(-c\frac{n(p_i-q_i)^2}{p_i+q_i}\right).$$

Using these estimates we turn to bounding the moment generating function of $\mathrm{sep}(\hat{S}_s)$ for $s \in \{>, <, 1/2\}$. Before doing so, recall (Buldygin and Moskvichova, 2013, Theorem 2.1) that the best-possible sub-Gaussian variance proxy $\sigma^2_{\mathsf{opt}}(\mu)$ of the $\mathrm{Ber}(\mu)$ distribution satisfies

$$\sigma^2_{\mathsf{opt}}(\mu) = \frac{\frac{1}{2} - \mu}{\log\left(\frac{1}{\mu} - 1\right)},$$

where the values for $\mu \in \{0, \frac{1}{2}, 1\}$ should be understood as the limit of the above expression (resulting in $\sigma^2_{\mathsf{opt}} = 0, \frac{1}{4}, 0$ respectively). Notice also that $\mu \mapsto \sigma^2_{\mathsf{opt}}(\mu)$ is increasing on $[0, \frac{1}{2}]$ and decreasing on $[\frac{1}{2}, 1]$, and

$$\sigma^2_{\mathsf{opt}}(\mu) \le \begin{cases} \frac{2}{\log(2/\mu)} & \text{if } 0 < \mu < 1/4, \\ 1/4 & \text{if } 1/4 \le \mu \le 3/4, \\ \frac{2}{\log(2/(1-\mu))} & \text{if } 3/4 < \mu < 1. \end{cases}$$

Let $T \subseteq D$ denote the subset of indices given by

$$T = \left\{ i \in D : 2\exp\left(-c\frac{n(p_i - q_i)^2}{p_i + q_i}\right) \ge \frac{1}{4} \right\} = \left\{ i \in D : (p_i - q_i)^2 \le \frac{p_i + q_i}{n}\frac{\log(8)}{c} \right\}.$$

Now, for any $s \in \{>, <, 1/2\}$, the sub-Gaussian variance proxy σ_s^2 of $\mathrm{sep}(\hat{S}_s) - \mathbb{E}\,\mathrm{sep}(\hat{S}_s) = \sum_{i \in D}(p_i - q_i)(\mathbb{1}\{i \in \hat{S}_s\} - \mathbb{P}(i \in \hat{S}_s)))$ is at most

$$\sigma_s^2 \le \sum_{i \in T} \frac{(p_i - q_i)^2}{4} + \sum_{i \in D \setminus T} (p_i - q_i)^2 \cdot \frac{2(p_i + q_i)}{cn(p_i - q_i)^2} \lesssim \sum_{i \in D} (p_i - q_i)^2 \wedge \frac{p_i + q_i}{n},$$

where the second step used the definition of T. In particular, since $\sum_{i \in D}(p_i + q_i)/n \le 2/n$, the above expression is always upper bounded by $\mathcal{O}(1/n)$. ∎

C.2. Proof of Proposition 7

By Lemma 21, we have

$$\begin{aligned}
\mathbb{E}\,\mathrm{sep}(\hat{S}_{1/2}) &= \sum_{i \in [k]} \mathbb{P}(i \in \hat{S}_{1/2})(p_i - q_i) \\
&= \sum_{i \in [k]} (\mathbb{P}(i \in \hat{S}_{1/2}) - \frac{1}{2})(p_i - q_i) \\
&\gtrsim \sum_{i \in [k]} \left(\frac{n|p_i - q_i|}{\sqrt{n(p_i \wedge q_i) + 1}} \wedge 1 \right) |p_i - q_i| \\
&\ge \min_{G \subseteq [k]} \left\{ \sum_{i \in G} \frac{n(p_i - q_i)^2}{\sqrt{n(q_i \wedge p_i) + 1}} + \sum_{i \notin G} |p_i - q_i| \right\}.
\end{aligned}$$

Applying the Cauchy-Schwarz inequality twice, we can bound the first term above by

$$\sum_{i\in G}\frac{n(p_i-q_i)^2}{\sqrt{n(q_i+p_i)+1}} \geq \frac{n\left(\sum_{i\in G}|p_i-q_i|\right)^2}{\sum_{i\in G}\sqrt{n(q_i+p_i)+1}} \geq \frac{n\left(\sum_{i\in G}|p_i-q_i|\right)^2}{\sqrt{2nk+k^2}}.$$

Therefore, we get the lower bound

$$\mathbb{E}\,\mathrm{sep}(\hat{S}_{1/2}) \gtrsim \min_{0\leq\epsilon_1\leq\epsilon}\left\{\frac{n\epsilon_1^2}{\sqrt{k(n+k)}}+\epsilon-\epsilon_1\right\} = \begin{cases}\frac{\epsilon^2}{\lambda} & \text{if } \epsilon<\frac{\lambda}{2}\\ \epsilon-\frac{\lambda}{4}\geq\frac{\epsilon}{2} & \text{if } \epsilon\geq\frac{\lambda}{2}\end{cases} \gtrsim \epsilon^2\left(\frac{1}{\epsilon}\wedge\sqrt{\frac{n}{k}}\wedge\frac{n}{k}\right)$$

where $\lambda=\frac{\sqrt{k(n+k)}}{n}\asymp\sqrt{\frac{k}{n}}\vee\frac{k}{n}$.

By Lemma 22 we know that $\mathrm{sep}(\hat{S}_{1/2})$ is sub-Gaussian with variance proxy $\mathcal{O}(1/n)$, which implies that $|\,\mathrm{sep}(\hat{S}_{1/2})| \gtrsim \epsilon^2(\frac{1}{\epsilon}\wedge\sqrt{\frac{n}{k}}\wedge\frac{n}{k})$ with probability at least $1-\delta$, provided that

$$\epsilon^2\left(\frac{1}{\epsilon}\wedge\sqrt{\frac{n}{k}}\wedge\frac{n}{k}\right)\gtrsim\sqrt{\frac{\log(1/\delta)}{n}}.$$

The above rearranges to $n\gtrsim n_{\mathsf{TS}}(\epsilon,\delta,\mathcal{P}_{\mathsf{D}})$.

C.3. Proof of Proposition 8

A direct computation gives

$$\begin{aligned}2\mathbb{E}\,\mathrm{sep}(\hat{S}_{>}) &= 2\sum_{i=1}^{3k}(p_i-q_i)\mathbb{P}(i\in\hat{S}_{>})\\ &= -\mathbb{P}\left(\mathrm{Poi}\left(\frac{n}{2k}\right)>\mathrm{Poi}\left(\frac{n}{k}\right)\right)+1-e^{-n/(4k)}\\ &\leq -(1-e^{-n/(2k)})e^{-n/k}+1-e^{-n/(4k)}\\ &= -e^{-n/k}+e^{-3n/(2k)}+1-e^{-n/(4k)}\leq 0,\end{aligned}$$

for $\exp(-n/(4k))\gtrapprox 0.86$. Rearranging, this gives the sufficient condition $n/k\leq 0.6$.

C.4. Proof of Proposition 9

Similar to the proof of Proposition 7, we have by Lemma 21 that

$$\begin{aligned}\mathbb{E}\,\mathrm{sep}(\hat{S}_{1/2}(D)) &= \sum_{i\in D}(p_i-q_i)\mathbb{P}(i\in\hat{S}_{1/2}(D))\geq c\mathcal{E}(D)+\frac{1}{2}\{p(D)-q(D)\}\\ -\mathbb{E}\,\mathrm{sep}(D\setminus\hat{S}_{1/2}(D)) &= \sum_{i\in D}(q_i-p_i)\mathbb{P}(i\notin\hat{S}_{1/2}(D))\geq c\mathcal{E}(D)+\frac{1}{2}\{q(D)-p(D)\}\end{aligned}$$

where $c>0$ is universal and $\mathcal{E}(D)=\sum_{i\in D}\frac{n|p_i-q_i|^2}{\sqrt{n(p_i\wedge q_i)+1}}\wedge|p_i-q_i|$. Therefore,

$$\mathbb{E}\left[\mathrm{sep}(\hat{S}_{>}(D))-\mathrm{sep}(\hat{S}_{<}(D))\right]=\mathbb{E}\left[\mathrm{sep}(\hat{S}_{1/2}(D))-\mathrm{sep}(D\setminus\hat{S}_{1/2}(D))\right]\geq 2c\mathcal{E}(D). \tag{10}$$

The bound on the sub-Gaussian variance proxy follows directly from Lemma 22.

C.5. Proof of Corollary 10

By a two-fold sample splitting, suppose that we have independent held out samples $(\tilde{X}, \tilde{Y})$ identical in distribution to (X, Y). In the sequel we will use samples (X, Y) to construct two separating sets, and use samples $(\tilde{X}, \tilde{Y})$ to make a choice between them.

Let the sets $\hat{S}_> \triangleq \hat{S}_>([k]), \hat{S}_< \triangleq \hat{S}_<([k])$ be constructed using X, Y. By Proposition 7 and 9, we have

$$|\mathbb{E}\,\mathrm{sep}(\hat{S}_>)| \vee |\mathbb{E}\,\mathrm{sep}(\hat{S}_<)| \gtrsim \epsilon^2 \left(\frac{1}{\epsilon} \wedge \sqrt{\frac{n}{k}} \wedge \frac{n}{k}\right),$$
$$\sigma^2(\hat{S}_>) + \sigma^2(\hat{S}_<) \lesssim \sum_{i\in[k]} \frac{p_i + q_i}{n} \lesssim \frac{1}{k \vee n},$$

where the last step have used that $p_i + q_i \lesssim 1/k$ in $\mathcal{P}_{\mathsf{Db}}$. Going forward, we assume that

$$\epsilon^2 \left(\frac{1}{\epsilon} \wedge \sqrt{\frac{n}{k}} \wedge \frac{n}{k}\right) \gtrsim \sqrt{\frac{\log(1/\delta)}{k \vee n}},$$

which rearranges to $n \gtrsim n_{\mathsf{GoF}}(\epsilon, \delta, \mathcal{P}_{\mathsf{D}})$. Consequently, this ensures that $|\,\mathrm{sep}(\hat{S}_>)| \vee |\,\mathrm{sep}(\hat{S}_<)| \gtrsim \epsilon^2 \left(\frac{1}{\epsilon} \wedge \sqrt{\frac{n}{k}} \wedge \frac{n}{k}\right)$ with probability $1 - \mathcal{O}(\delta)$. Moreover, as $n \gtrsim \log(1/\delta)$, with probability at least $1 - \delta$ we have $\mathrm{Poi}(n) \le 2n$ (cf. Lemma 17). Under this event, one has $|\hat{S}_>| \vee |\hat{S}_<| \le 2n$, and

$$\tau(\hat{S}_>) \vee \tau(\hat{S}_<) \lesssim \frac{|\hat{S}_>| \vee |\hat{S}_<|}{k} \wedge 1 \le \frac{2n}{k} \wedge 1.$$

Next we make a choice between $\hat{S}_>$ and $\hat{S}_<$ based on held out samples $(\tilde{X}, \tilde{Y})$. Let $\hat{p}, \hat{q}$ denote the empirical pmfs constructed using $\tilde{X}, \tilde{Y}$ respectively. For any set $A \subseteq [k]$ write $\widehat{\mathrm{sep}}(A) = \hat{p}(A) - \hat{q}(A)$. We define our final estimator to be

$$\hat{S} = \begin{cases} \hat{S}_> & \text{if } |\widehat{\mathrm{sep}}(\hat{S}_>)| \ge |\widehat{\mathrm{sep}}(\hat{S}_<)|, \\ \hat{S}_< & \text{otherwise.} \end{cases}$$

Clearly $\tau(\hat{S}) \le \tau(\hat{S}_>) \vee \tau(\hat{S}_<) \lesssim 1 \wedge (n/k)$. To show the high-probability separation of $\hat{S}$, note that by Lemma 19, it holds with probability at least $1 - \mathcal{O}(\delta)$ that

$$\begin{aligned}
|\,\mathrm{sep}(\hat{S})| &\ge \frac{1}{2}|\widehat{\mathrm{sep}}(\hat{S})| - \mathcal{O}\left(\sqrt{\frac{\tau(\hat{S})\log(1/\delta)}{n}} + \frac{\log(1/\delta)}{n}\right) \\
&= \frac{1}{2}|\widehat{\mathrm{sep}}(\hat{S}_>)| \vee |\widehat{\mathrm{sep}}(\hat{S}_<)| - \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n \vee k}} + \frac{\log(1/\delta)}{n}\right) \\
&\ge \frac{1}{4}|\,\mathrm{sep}(\hat{S}_>)| \vee |\,\mathrm{sep}(\hat{S}_<)| - \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n \vee k}} + \frac{\log(1/\delta)}{n}\right) \\
&= \Omega\left(\epsilon^2 \left(\frac{1}{\epsilon} \wedge \sqrt{\frac{n}{k}} \wedge \frac{n}{k}\right)\right) - \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n \vee k}} + \frac{\log(1/\delta)}{n}\right).
\end{aligned}$$

Here the first term always dominates the second as long as $n \gtrsim n_{\mathsf{GoF}}(\epsilon, \delta, \mathcal{P}_{\mathsf{D}})$.

C.6. Proof of Proposition 12

Similar to the proof of Corollary 10, we apply a two-fold sample splitting to obtain n independent held out samples $(\tilde{X}, \tilde{Y})$. In the sequel we construct $2(\ell+2)$ candidate separating sets from (X, Y), and make a choice among them using held out samples $(\tilde{X}, \tilde{Y})$.

The construction of the $2(\ell+2)$ separating sets is simple: for each $j \in \{0, 1, \cdots, \ell+1\}$, we construct two sets $\hat{S}_>(D_j)$ and $\hat{S}_<(D_j)$. The following lemma summarizes some properties of these separating sets. Recall that we assume that $t = k \wedge (c_0 m/\log(1/\delta)) > n$ so that $\ell = \lceil \log_2(t/n) \rceil \geq 1$.

Lemma 23 *Fix any $j \in \{0, 1, \cdots, \ell+1\}$, and let $\epsilon_j = \sum_{i \in D_j} |p_i - q_i|$. With probability at least $1-\delta$, the following statements hold:*

1. *if $j = 0$, then*

$$\left|\mathrm{sep}(\hat{S}_>(D_0))\right| \vee \left|\mathrm{sep}(\hat{S}_<(D_0))\right| \gtrsim E_0 - \mathcal{O}\left(\sqrt{\frac{E_0 \log(1/\delta)}{n}}\right),$$

   where

$$E_0 = \sum_{i \in D_0} n|p_i - q_i|^2 \wedge |p_i - q_i| \gtrsim \frac{n\epsilon_0^2}{k} =: \tilde{E}_0(\epsilon_0).$$

2. *if $j \in [\ell]$, then*

$$\left|\mathrm{sep}(\hat{S}_>(D_j))\right| \vee \left|\mathrm{sep}(\hat{S}_<(D_j))\right| \gtrsim E_j - \mathcal{O}\left(\sqrt{\frac{E_j \log(1/\delta)}{n}}\right),$$

   where

$$E_j = \sum_{i \in D_j} n|p_i - q_i|^2 \wedge |p_i - q_i| \gtrsim \frac{n\epsilon_j^2}{\sqrt{kt/2^j}} =: \tilde{E}_j(\epsilon_j).$$

3. *if $j = \ell+1$, then*

$$\left|\mathrm{sep}(\hat{S}_>(D_{\ell+1}))\right| \vee \left|\mathrm{sep}(\hat{S}_<(D_{\ell+1}))\right| \gtrsim E_{\ell+1} - \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right),$$

   where

$$E_{\ell+1} = \sum_{i \in D_{\ell+1}} \frac{n|p_i - q_i|^2}{\sqrt{nq_i}} \wedge |p_i - q_i| \gtrsim \sqrt{\frac{n}{k}}\epsilon_{\ell+1}^2 =: \tilde{E}_{\ell+1}(\epsilon_{\ell+1}).$$

Proof We prove the above statements separately.

1. Case I: $j = 0$. By Proposition 9, it holds that

$$\mathbb{E}[\mathrm{sep}(\hat{S}_>(D_0)) - \mathrm{sep}(\hat{S}_<(D_0))] \gtrsim \sum_{i \in D_0} n|p_i - q_i|^2 \wedge |p_i - q_i| = E_0,$$

where we have used Lemma 11 that $q_i \le 2/t \le 2/n$ for all $i \in D_0$. Moreover,

$$\begin{aligned}&\sigma^2(\mathrm{sep}(\hat{S}_>(D_0))) \vee \sigma^2(\mathrm{sep}(\hat{S}_<(D_0)))\\ &\lesssim \sum_{i \in D_0} |p_i - q_i|^2 \wedge \frac{p_i + q_i}{n} \lesssim \sum_{i \in D_0} \frac{1}{n}\left(n|p_i - q_i|^2 \wedge |p_i - q_i|\right) = \frac{E_0}{n},\end{aligned}$$

where the last inequality is due to the following deterministic inequality: if $q \le 2/n$, then

$$|p - q|^2 \wedge \frac{p + q}{n} \lesssim \frac{1}{n}\left(n|p - q|^2 \wedge |p - q|\right).$$

The proof of the above deterministic inequality is based on two cases:

- if $p \le 3/n$, then $|p - q|^2 \lesssim |p - q|^2 \wedge (|p - q|/n)$;
- if $p > 3/n$, then $p + q \lesssim n|p - q|^2 \wedge |p - q|$.

Consequently, we have the first statement. For the second statement, similar to the proof of Proposition 7 we have

$$E_0 \ge \min_{\epsilon_0' \in [0, \epsilon_0]} \left(\frac{n(\epsilon_0')^2}{k} + \epsilon_0 - \epsilon_0'\right) \gtrsim \epsilon_0^2 \left(\frac{1}{\epsilon_0} \wedge \frac{n}{k}\right) \asymp \frac{n\epsilon_0^2}{k}.$$

2. Case II: $j \in [\ell]$. By Proposition 9 and Lemma 11 we have

$$\mathbb{E}[\mathrm{sep}(\hat{S}_>(D_j)) - \mathrm{sep}(\hat{S}_<(D_j))] \gtrsim \sum_{i \in D_j} n(p_i - q_i)^2 \wedge |p_i - q_i| = E_j.$$

Similar to Case I, we have

$$\sigma^2(\mathrm{sep}(\hat{S}_>(D_j))) \vee \sigma^2(\mathrm{sep}(\hat{S}_<(D_j))) \lesssim \sum_{i \in D_j} |p_i - q_i|^2 \wedge \frac{p_i + q_i}{n} \lesssim \frac{E_j}{n},$$

and the first statement follows.

For the second statement, note that $|D_j| \le t/2^{j-1} = \mathcal{O}(\sqrt{kt/2^j})$ by Lemma 11. Therefore,

$$E_j \ge \min_{\epsilon_j' \in [0, \epsilon_j]} \left(\frac{n(\epsilon_j')^2}{|D_j|} + \epsilon_j - \epsilon_j'\right) \gtrsim \epsilon_j^2 \left(\frac{1}{\epsilon_j} \wedge \frac{n}{\sqrt{kt/2^j}}\right) \asymp \frac{n\epsilon_j^2}{\sqrt{kt/2^j}}.$$

3. Case III: $j = \ell + 1$. By Proposition 9 and Lemma 11, we have

$$\mathbb{E}[\mathrm{sep}(\hat{S}_>(D_{\ell+1})) - \mathrm{sep}(\hat{S}_<(D_{\ell+1}))] \gtrsim \sum_{i \in D_{\ell+1}} \frac{n(p_i - q_i)^2}{\sqrt{nq_i}} \wedge |p_i - q_i| = E_{\ell+1}.$$

The first statement then follows from Lemma 22. The second statement then follows from

$$E_{\ell+1} \ge \min_{\epsilon_{\ell+1}' \in [0, \epsilon_{\ell+1}]} \left(\frac{n(\epsilon_{\ell+1}')^2}{\sqrt{nk}} + \epsilon_{\ell+1} - \epsilon_{\ell+1}'\right) \gtrsim \epsilon_{\ell+1}^2 \left(\frac{1}{\epsilon_{\ell+1}} \wedge \sqrt{\frac{n}{k}}\right) \asymp \sqrt{\frac{n}{k}}\epsilon_{\ell+1}^2.$$

The proof is complete. ■

Based on Lemma 23, we are about to describe how we choose from the sets $\{\hat{S}_>(D_j), \hat{S}_<(D_j)\}_{j=0}^{\ell+1}$. Similar to the proof of Corollary 10, using the held out samples $(\tilde{X}, \tilde{Y})$, we can obtain the empirical estimates $\widehat{\mathrm{sep}}(\hat{S}_s(D_j))$ for all $s \in \{>, <\}$ and $j \in \{0, 1, \cdots, \ell+1\}$. With a small absolute constant $c_1 > 0$ and $\tilde{E}_j$ as defined in Lemma 23, the selection rule is as follows: if there is some $s \in \{>, <\}$ and $j \in \{0, 1, \cdots, \ell+1\}$ such that

$$|\widehat{\mathrm{sep}}(\hat{S}_s(D_j))| \geq c_1 \tilde{E}_j(\epsilon/(\ell+2)),$$

then choose $\hat{S} = \hat{S}_s(D_j)$; if there is no such pair (s, j), choose an arbitrary $\hat{S}$.

We first show that with probability at least $1 - \mathcal{O}(k\delta)$, such a pair (s, j) exists. Since $\|p - q\|_1 \geq \epsilon$, there must exist some $j \in \{0, 1, \cdots, \ell+1\}$ such that $\epsilon_j \geq \epsilon/(\ell+2)$. As long as

$$n \geq c_2 n_{\mathsf{GoF}}(\epsilon/\ell, \delta, \mathcal{P}_{\mathsf{D}})$$

for a large constant $c_2 > 0$, one can check via Lemma 23 that $|\mathrm{sep}(\hat{S}_>(D_j))| \vee |\mathrm{sep}(\hat{S}_<(D_j))| \geq 4c_1\tilde{E}_j(\epsilon/(\ell+2))$ for a small enough universal constant $c_1 > 0$. Assuming that $n \gtrsim \log(1/\delta)$, we have $\tau(\hat{S}_>(D_j)) \vee \tau(\hat{S}_<(D_j)) = \mathcal{O}(n2^j/t)$ with probability $1 - \mathcal{O}(\delta)$ due to Poisson concentration (Lemma 17). On this event, it holds with probability at least $1 - \delta$ that (cf. Lemma 19)

$$|\widehat{\mathrm{sep}}(\hat{S}_>(D_j))| \vee |\widehat{\mathrm{sep}}(\hat{S}_<(D_j))| \geq 2c_1\tilde{E}_j(\epsilon/(\ell+2)) - \mathcal{O}\left(\sqrt{\frac{2^j \log(1/\delta)}{t}} + \frac{\log(1/\delta)}{n}\right),$$

which is at least $c_1\tilde{E}_j(\epsilon/(\ell+2))$ as long as

$$n\sqrt{\frac{t}{k}} \asymp n\sqrt{1 \wedge \frac{m}{\log(1/\delta)k}} \geq c_3 n_{\mathsf{GoF}}(\epsilon/\ell, \delta, \mathcal{P}_{\mathsf{D}}) \tag{11}$$

for some large $c_3 > 0$. Therefore, provided (11) holds, the desired pair (j, s) exists with probability $1 - \mathcal{O}(k\delta)$ due to a union bound.

Conversely, if $|\widehat{\mathrm{sep}}(\hat{S}_s(D_j))| \geq c_1\tilde{E}_j(\epsilon/(\ell+2))$ holds for some (s, j), the true separation $|\mathrm{sep}(\hat{S}_s(D_j))|$ is at least of the same order as well. Indeed, Lemma 19 shows that

$$|\mathrm{sep}(\hat{S}_s(D_j))| \geq \frac{1}{2}|\widehat{\mathrm{sep}}(\hat{S}_s(D_j))| - \mathcal{O}\left(\sqrt{\frac{2^j \log(1/\delta)}{t}} + \frac{\log(1/\delta)}{n}\right),$$

which is at least $c_1 E_j(\epsilon/(\ell+2))/4$ as long as (11) holds. This completes the proof.

C.7. Proof of Proposition 14

The statement of Proposition 14 follows immediately from the following lemma.

Lemma 24 *Let* $\mathrm{sep}(\hat{S}) \triangleq \mu_{\theta^X}(\hat{S}) - \mu_{\theta^Y}(\hat{S})$. *There exist universal constants* $c_i > 0, i \in [5]$ *such that for* $J = \lfloor c_1 \epsilon^{-1/s} \rfloor$ *we have*

$$\mathbb{E}[\mathrm{sep}(\hat{S})] + \frac{c_2}{\sqrt{n}} \geq \frac{c_3\epsilon^2}{\epsilon + \sqrt{J/n}}$$

$$\mathbb{P}\left(\left|\mathrm{sep}(\hat{S}) - \mathbb{E}\,\mathrm{sep}(\hat{S})\right| \geq t + \frac{c_4}{\sqrt{n}}\right) \leq 2\exp(-c_5 n t^2)$$

for all $t \geq 0$.

Proof Write $\|\cdot\|, \langle\cdot,\cdot\rangle$ for the ℓ^2 norm/inner product restricted to the first J coordinates. Notice that given $\hat{\theta}^X$ and $\hat{\theta}^Y$, $T(\theta)$ is simply a Gaussian random variable with $\mathbb{E}T(\theta) = \|\hat{\theta}^Y - \theta\|^2 - \|\hat{\theta}^X - \theta\|^2$ and $\mathrm{var}(T) = 4\|\hat{\theta}^X - \hat{\theta}^Y\|^2$. Define the vectors

$$U = \{\hat{\theta}^X_j - \hat{\theta}^Y_j\}_{j=1}^J$$
$$V = \{\hat{\theta}^X_j + \hat{\theta}^Y_j\}_{j=1}^J.$$

Note that they are independent, jointly Gaussain with variance $2I_J/n$ and means equal to the first J coordinates of $\theta^X \mp \theta^Y$ respectively. Let Φ be the cdf of the standard Gaussian and $\phi = \Phi'$ be its density. The separation can be written as

$$\mathrm{sep}(\hat{S}) = f(\theta^X) - f(\theta^Y),$$

where

$$f(\theta) = \Phi\left(\frac{\|\hat{\theta}^Y - \theta\|^2 - \|\hat{\theta}^X - \theta\|^2}{2\|\hat{\theta}^X - \hat{\theta}^Y\|}\right) = \Phi\left(-\frac{1}{2}\left\langle V, \frac{U}{\|U\|}\right\rangle + \left\langle \theta, \frac{U}{\|U\|}\right\rangle\right). \tag{12}$$

We focus on proving the desired tail bound first. To make the dependence on the variables explicit, write $g(U,V) = f(\theta^X) - f(\theta^Y)$ for the separation. Given U, V is a $\mathcal{N}(\theta^X + \theta^Y, 2I_j/n)$ random variable. Differentiating g and using that ϕ is $1/\sqrt{2\pi e}$-Lipschitz we have

$$\begin{aligned}
\|\nabla_V g(U,V)\| &= \Big\| -\frac{1}{2}\frac{U}{\|U\|}\left(\phi\left(-\frac{1}{2}\left\langle V, \frac{U}{\|U\|}\right\rangle + \left\langle \theta^X, \frac{U}{\|U\|}\right\rangle\right)\right. \\
&\qquad \left. - \phi\left(-\frac{1}{2}\left\langle V, \frac{U}{\|U\|}\right\rangle + \left\langle \theta^Y, \frac{U}{\|U\|}\right\rangle\right)\right)\Big\| \\
&\le \frac{1}{\sqrt{8\pi e}}\left|\left\langle \theta^X - \theta^Y, \frac{U}{\|U\|}\right\rangle\right| \\
&\le \frac{C_{\mathsf{G}}}{\sqrt{8\pi e}}.
\end{aligned}$$

By Lipschitz concentration of the Gaussian distribution (Lemma 16) we conclude that $g - \mathbb{E}[g|U]$ is sub-Gaussian with variance proxy $C_{\mathsf{G}}^2/(4\pi e n)$. Next we study the concentration of $\mathbb{E}[g|U]$. To this end, note that

$$-\frac{1}{2}\left\langle V, \frac{U}{\|U\|}\right\rangle + \left\langle \theta, \frac{U}{\|U\|}\right\rangle \Big| U \sim \mathcal{N}\left(\left\langle \theta - \frac{1}{2}(\theta^X + \theta^Y), \frac{U}{\|U\|}\right\rangle, \frac{1}{2n}\right).$$

Thus, using the independence of U and V and Lemma 15 we obtain

$$\begin{aligned}
\mathbb{E}\left[g(U,V)|U\right] &= \mathbb{E}\left[f(\theta^X) - f(\theta^Y)|U\right] \\
&= \Phi\left(\frac{W}{\sqrt{4 + 2/n}}\right) - \Phi\left(-\frac{W}{\sqrt{4+2/n}}\right),
\end{aligned}$$

where we write $W \triangleq \left\langle \theta^X - \theta^Y, \frac{U}{\|U\|} \right\rangle$. Let $\tilde{\Phi} = \Phi(\cdot/\sqrt{4+2/n})$ to ease notation. Once again by Lipschitzness of Φ, we obtain for every $t \geq 0$ that

$$\begin{aligned}\mathbb{P}\left(\left|\tilde{\Phi}(W) - \mathbb{E}\tilde{\Phi}(W)\right| \geq t\right) &\leq \mathbb{P}\left(\left|\tilde{\Phi}(W) - \tilde{\Phi}(\mathbb{E}W)\right| \geq t - \|\tilde{\Phi}\|_{\mathsf{Lip}}\sqrt{\mathrm{var}(W)}\right) \\ &\leq \mathbb{P}\left(|W - \mathbb{E}W| \geq \frac{t}{\|\tilde{\Phi}\|_{\mathsf{Lip}}} - \sqrt{\mathrm{var}(W)}\right),\end{aligned}$$

and an analogous inequality can be obtained for $-W$. The last ingredient is showing that W concentrates well.

Lemma 25 *W is sub-Gaussian with variance proxy $1/(2n)$.*

Proof [Proof of Lemma 25] To simplify notation, let $\tau = \theta^X - \theta^Y$, $\sigma^2 = 1/(2n)$ and let Q be a zero-mean identity-covariance Gaussian random vector so that

$$W \overset{d}{=} \left\langle \tau, \frac{\tau + \sigma Q}{\|\tau + \sigma Q\|} \right\rangle.$$

We have

$$\left\langle \tau, \frac{\tau + \sigma Q}{\|\tau + \sigma Q\|} \right\rangle = \underbrace{\left\langle \frac{\tau}{\mathbb{E}\|\tau + \sigma Q\|}, \frac{\tau + \sigma Q}{\|\tau + \sigma Q\|} \right\rangle}_{|\cdot| \leq 1 \text{ almost surely}} \underbrace{(\|\tau + \sigma Q\| - \mathbb{E}\|\tau + \sigma Q\|)}_{\sigma^2 \text{ sub-Gaussian}} + \underbrace{\sigma \left\langle \frac{\tau}{\mathbb{E}\|\tau + \sigma Q\|}, Q \right\rangle}_{\sigma^2 \text{ sub-Gaussian}},$$

where we use that $\mathbb{E}\|\tau + \sigma Q\| \geq \|\tau\|$ by Jensen's inequality, and apply Lemma 16 twice. Overall, this implies that W is sub-Gaussian with variance proxy $\sigma^2 = 1/(2n)$ as required. ■

Recall that we have decomposed the separation as follows:

$$\mathrm{sep}(\hat{S}) - \mathbb{E}\,\mathrm{sep}(\hat{S}) = \underbrace{g - \mathbb{E}[g|U]}_{\mathcal{O}(1/n) \text{ sub-Gaussian}} + \underbrace{\tilde{\Phi}(W) - \tilde{\Phi}(-W) - \mathbb{E}[\tilde{\Phi}(W) - \tilde{\Phi}(-W)]}_{\mathcal{O}(1/n) \text{ sub-Gaussian tails beyond } \mathcal{O}(1/\sqrt{n})},$$

which completes the proof.

Let us turn to calculating the expected separation. We have already seen that

$$\mathbb{E}\,\mathrm{sep}(\hat{S}) = \mathbb{E}\left[\tilde{\Phi}(W) - \tilde{\Phi}(-W)\right].$$

Again by Lipschitzness we have $|\mathbb{E}\tilde{\Phi}(W) - \tilde{\Phi}(\mathbb{E}W)| \leq \|\tilde{\Phi}\|_{\mathsf{Lip}}\mathbb{E}|W - \mathbb{E}W| \lesssim 1/\sqrt{n}$ by Lemma 25. Thus, we see that

$$\mathbb{E}\,\mathrm{sep}(\hat{S}) + \Omega\left(\frac{1}{\sqrt{n}}\right) \geq \tilde{\Phi}(\mathbb{E}W) - \tilde{\Phi}(-\mathbb{E}W),$$

where the implied constant is universal. To simplify notation, let $\tau = \theta^X - \theta^Y$, $\sigma^2 = 1/(2n)$ and let Q be a standard normal random variable. Looking at $\mathbb{E}W$ we have

$$\mathbb{E}W = \mathbb{E}\left\langle \tau, \frac{\tau + \sigma Q}{\|\tau + \sigma Q\|} \right\rangle = \frac{1}{\sigma}\mathbb{E}\left\langle \tau, \nabla_Q \|\tau + \sigma Q\| \right\rangle = \frac{1}{\sigma}\mathbb{E}\left[\langle \tau, Q \rangle \|\tau + \sigma Q\|\right]$$

by Stein's identity. By the rotational invariance of the Gaussian distribution, the above is equal to

$$\begin{aligned}\mathbb{E}W &= \frac{\|\tau\|}{\sigma}\mathbb{E}\left[Q_1\sqrt{(\|\tau\|+\sigma Q_1)^2+\cdots+\sigma^2Q_J^2}\right]\\ &= \frac{\|\tau\|}{\sigma}\mathbb{E}\left[Q_1\sqrt{(\|\tau\|+\sigma Q_1)^2+\cdots+\sigma^2Q_J^2}-Q_1\sqrt{\|\tau\|^2+\sigma^2Q_1^2+\cdots+\sigma^2Q_J^2}\right]\\ &= 2\|\tau\|^2\mathbb{E}\left[\frac{Q_1^2}{\sqrt{(\|\tau\|+\sigma Q_1)^2+\cdots+\sigma^2Q_J^2}+\sqrt{\|\tau\|^2+\sigma^2Q_1^2+\cdots+\sigma^2Q_J^2}}\right].\end{aligned}$$

By the Cauchy-Schwarz inequality we have

$$(\mathbb{E}|Q_1|)^2 \lesssim \mathbb{E}\left[\frac{Q_1^2}{\sqrt{(\|\tau\|+\sigma Q_1)^2+\cdots+\sigma^2Q_J^2}+\sqrt{\|\tau\|^2+\sigma^2Q_1^2+\cdots+\sigma^2Q_J^2}}\right]\times(\|\tau\|+\sigma\sqrt{J}).$$

Plugging into our expression for $\mathbb{E}W$ this yields

$$\mathbb{E}W \gtrsim \frac{\|\tau\|^2}{\|\tau\|+\sigma\sqrt{J}}.$$

To clarify notation, let us now write $\|\cdot\|_J$ for the ℓ^2-norm restricted to the first J coordinates. Taking $J = c\epsilon^{-1/s}$ it holds that

$$\|\tau\|_J^2 = \|\tau\|^2 - \sum_{j>J}\tau_j^2 \geq \|\tau\|^2 - J^{-2s}\sum_{j>J}\tau_j^2 j^{2s} = \|\tau\|^2 - c^{-2s}\epsilon^2\|\tau\|_s^2.$$

Since $\|\tau\|_s \lesssim 1$ and $\|\tau\| \geq \epsilon$ by assumption, we see that for large enough universal constant c we have $\|\tau\|_J \geq \epsilon/2$. Since the map $x \mapsto x^2/(x+c)$ is increasing for $x, c > 0$ it follows that

$$\mathbb{E}W \gtrsim \frac{\epsilon^2}{\epsilon+\sqrt{J/n}}$$

for a universal implied constant. By the inequality $\Phi(x) - \Phi(-x) \geq x/2$ for $x \in [0,1]$ we obtain

$$\tilde{\Phi}(\mathbb{E}W) - \tilde{\Phi}(-\mathbb{E}W) \geq 1 \wedge \mathbb{E}W/2,$$

which completes the proof. ∎

Appendix D. Lower bounds

Recall the notation of Section 2.1.1. Given two hypotheses H_0, H_1, our aim is to lower bound the minimum achievable worst-case error. To this end, we use the following standard fact:

$$\min_{\psi}\max_{i=0,1}\sup_{P\in H_i}\mathbb{P}_{S\sim P}(\psi(S)\neq i) \geq \frac{1}{2}\big(1-\mathsf{TV}(\mathbb{E}P_0,\mathbb{E}P_1)\big), \tag{13}$$

where P_0, P_1 are any random probability distributions with $\mathbb{P}(P_i \in H_i) = 1$ and $\mathbb{E}P_i$ denote the corresponding mixtures and TV denotes the total variation distance. Hence, deriving a lower bound of order δ on the minimax error reduces to the problem of finding mixtures $\mathbb{E}P_i$ such that $1 - \mathsf{TV}(\mathbb{E}P_0, \mathbb{E}P_1) = \Omega(\delta)$. To this end we utilize standard inequalities between divergences.

Lemma 26 (Polyanskiy and Wu (2023+)) *For any probability measures $\mathbb{P}, \mathbb{Q}$ the inequalities*

$$1 - \mathsf{TV}(\mathbb{P}, \mathbb{Q}) \geq \frac{1}{2} e^{-\mathsf{KL}(\mathbb{P}\|\mathbb{Q})} \geq \frac{1}{2(1+\chi^2(\mathbb{P}\|\mathbb{Q}))}$$

hold, where KL *and χ^2 denote the Kullback-Leibler and χ^2 divergence respectively.*

Many of our lower bounds will follow from reduction to prior work.

D.1. Lower bounds for $\mathcal{P}_{\mathsf{Db}}$

In Gerber and Polyanskiy (2022) the authors gave the construction of distributions $p_{\eta,\epsilon}, p_0 \in \mathcal{P}_{\mathsf{Db}}(k, 2)$ (originally due to Paninski) for a mixing parameter η such that $\mathsf{TV}(p_{\eta,\epsilon}, p_0) = \epsilon \asymp \sqrt{\mathsf{KL}(p_{\eta,\epsilon}, p_0)}$ for all η, where the implied constant is universal. They further showed that

$$\chi^2(\mathbb{E}_\eta p_{\eta,\epsilon}^{\otimes n}, p_0^{\otimes n}) \leq \exp\left(c\frac{n^2\epsilon^4}{k}\right) - 1 \tag{14}$$

and

$$\chi^2\Big(\mathbb{E}_\eta\big[p_0^{\otimes n} \otimes p_{\epsilon,\eta}^{\otimes(n+m)}\big] \Big\| \mathbb{E}_\eta\big[p_0^{\otimes n} \otimes p_{\epsilon,\eta}^{\otimes n} \otimes p_0^{\otimes m}\big]\Big) \leq \exp\left(c\frac{m(n+m)\epsilon^4}{k}\right) - 1 \tag{15}$$

for a universal $c > 0$.

Remark 27 *More precisely, (15) can be extracted from Gerber and Polyanskiy (2022) using the chain rule for χ^2 (as opposed to* KL*).*

D.1.1. Lower bound for TS and GoF

Take $P_0 = p_0^{\otimes 2n}$ and $P_1 = p_{\epsilon,\eta_0}^{\otimes n} \otimes p_0^{\otimes n}$ in (13) for a fixed η_0. Then, by Lemma 26 and the data-processing inequality we have

$$1 - \mathsf{TV}(\mathbb{E}P_0, \mathbb{E}P_1) \geq \frac{1}{2}\exp(-n\mathsf{KL}(p_{\epsilon,\eta}\|p_0)) \geq \frac{1}{2}\exp(-cn\epsilon^2) \overset{!}{=} \Omega(\delta)$$

for a universal $c > 0$. This shows that GoF, TS are impossible at total error δ unless $n \gtrsim \log(1/\delta)/\epsilon^2$, which gives the first term of our lower bound.

For the second term, consider the random measures $P_0 = p_0^{\otimes 2n}$ and $P_1 = p_0^{\otimes n} \otimes p_{\epsilon,\eta}^{\otimes n}$ in (13). Then using (14) and Lemma 26 we have

$$\begin{aligned} 1 - \mathsf{TV}(\mathbb{E}P_1, \mathbb{E}P_0) &\geq \frac{1}{2}\frac{1}{1+\chi^2(\mathbb{E}P_1\|\mathbb{E}P_0)} \\ &\geq \frac{1}{2}\exp\left(-c\frac{n^2\epsilon^4}{k}\right) \overset{!}{=} \Omega(\delta). \end{aligned}$$

Therefore, TS is impossible unless $n \gtrsim \sqrt{k\log(1/\delta)}/\epsilon^2$, which yields the second term of our lower bound.

D.1.2. LOWER BOUND FOR LFHT

The necessity of $m \gtrsim \log(1/\delta)/\epsilon^2$ and $n \gtrsim \sqrt{k\log(1/\delta)}/\epsilon^2$ follows as for TS above. Taking $P_0 = p_0^{\otimes n} \otimes p_{\epsilon,\eta}^{\otimes n} \otimes p_0^{\otimes m}$ and $P_1 = p_0^{\otimes n} \otimes p_{\epsilon,\eta}^{\otimes(n+m)}$ in (13), using (15) and Lemma 26 we obtain the inequality

$$\begin{aligned} 1 - \mathsf{TV}(\mathbb{E}P_0, \mathbb{E}P_1) &\geq \frac{1}{2}\frac{1}{1+\chi^2(\mathbb{E}P_1\|\mathbb{E}P_0)} \\ &\geq \frac{1}{2}\exp\left(-c\frac{m(m+n)\epsilon^4}{k}\right) \stackrel{!}{=} \Omega(\delta). \end{aligned}$$

Therefore, LFHT is impossible with error $\mathcal{O}(\delta)$ unless $mn \gtrsim k\log(1/\delta)/\epsilon^4$ (note that the m^2-term is never active), which completes the lower bound proof.

D.2. Lower bounds for $\mathcal{P}_{\mathsf{H}}$

We don't provide the details because they are entirely analogous to Section D.1 and rely on classical constructions that can be found in Gerber and Polyanskiy (2022).

D.3. Lower bounds for $\mathcal{P}_{\mathsf{G}}$

Given a vector $\eta \in \{\pm 1\}^{\mathbb{N}}$ define the measure

$$\mathbb{P}_\eta = \bigotimes_{j=1}^{\infty} \left\{ \begin{array}{ll} \mathcal{N}(\eta_j c_1 \epsilon^{\frac{2s+1}{2s}}, 1) & \text{if } 1 \leq j \leq c_2\epsilon^{-1/s}, \\ \mathcal{N}(0,1) & \text{otherwise.} \end{array} \right\}$$

Let $\eta_1, \eta_2, \ldots$ be iid uniform signs in $\{\pm 1\}$, and γ_η be the mean vector of $\mathbb{P}_\eta$. Writing $\|\cdot\|_s$ for the Sobolev-norm of smoothness s and $\|\cdot\|$ for the Euclidean norm, we see that for any η

$$\begin{aligned} \|\gamma_\eta\|_s^2 &= \sum_{j=1}^{\infty} j^{2s}\gamma_{\eta j}^2 = \sum_{j=1}^{c_2\epsilon^{-1/s}} j^{2s} c_1^2 \epsilon^{\frac{2s+1}{s}} \leq c_1^2\epsilon^{\frac{2s+1}{s}}\left(2c_2\epsilon^{-1/s}\right)^{2s+1} \asymp c_1^2 c_2^{2s+1}, \\ \|\gamma_\eta\|^2 &= \sum_{j=1}^{\infty}\gamma_{\eta j}^2 = c_1^2\epsilon^{\frac{2s+1}{s}} c_2\epsilon^{-1/s} \asymp c_1^2 c_2 \epsilon^2. \end{aligned}$$

Then for any $C_{\mathsf{G}} > 0$ we can choose c_1, c_2 independently of ϵ such that $\mathbb{P}_0, \mathbb{P}_\eta \in \mathcal{P}_{\mathsf{G}}(s, C_{\mathsf{G}})$ almost surely and $\|\gamma_\eta\| = 10\epsilon$. Then for $\epsilon \leq 1/10$ we know that

$$\mathsf{TV}(\mathbb{P}_0, \mathbb{P}_\eta) = 2\Phi\left(\frac{\|\gamma_\eta\|}{2}\right) - 1 \geq \epsilon.$$

D.3.1. LOWER BOUNDS FOR GOF AND TS

Take $P_0 = \mathbb{P}_0^{\otimes 2n}$ and $P_1 = \mathbb{P}_{\mathbb{1}}^{\otimes n} \otimes \mathbb{P}_0^{\otimes n}$. Then

$$\mathsf{KL}(P_0\|P_1) = n\mathsf{KL}(\mathbb{P}_0\|\mathbb{P}_{\mathbb{1}}) = nc_2\epsilon^{-1/s}\frac{(c_1\epsilon^{\frac{2s+1}{2s}} - 0)^2}{2} \asymp n\epsilon^2.$$

Using Lemma 26 this gives us

$$1 - \mathsf{TV}(P_0, P_1) \gtrsim \exp(-\mathsf{KL}(P_0 \| P_1)) = \exp(-\Theta(n\epsilon^2)) \overset{!}{=} \Omega(\delta).$$

By (13) we know then that $n \gtrsim \log(1/\delta)/\epsilon^2$ is necessary for both GoF and TS over $\mathcal{P}_{\mathsf{G}}$.

To get the second term in the minimax sample complexity consider the construction $P_0 = \mathbb{P}_0^{\otimes 2n}$ and $P_1 = \mathbb{P}_\eta^{\otimes n} \otimes \mathbb{P}_0^{\otimes n}$ where η is a uniformly random vector of signs. Writing $\omega = c_1 \epsilon^{\frac{2s+1}{2s}}$ note that

$$\mathbb{E}\mathbb{P}_\eta^{\otimes n} = \bigotimes_{j=1}^{c_2\epsilon^{-1/s}} \left(\frac{1}{2}\mathcal{N}(\omega, 1)^{\otimes n} + \frac{1}{2}\mathcal{N}(-\omega, 1)^{\otimes n} \right).$$

From here we can compute

$$\begin{aligned}
\mathsf{KL}(P_0 \| \mathbb{E}P_1) &\asymp \epsilon^{-1/s} \mathsf{KL}\left(\mathcal{N}(0,1)^{\otimes n} \middle\| \frac{1}{2}\mathcal{N}(\omega, 1)^{\otimes n} + \frac{1}{2}\mathcal{N}(-\omega, 1)^{\otimes n} \right) \\
&\asymp \epsilon^{-1/s} \left(\frac{n}{2}\omega^2 - \mathbb{E}_{X \sim \mathcal{N}(0, I_n)} \log\cosh\left(\omega \sum_{i=1}^n X_i \right) \right) \\
&\leq \frac{\epsilon^{-1/s}}{4} n^2 \omega^4 \asymp n^2 \epsilon^{\frac{4s+1}{s}},
\end{aligned}$$

where we used the inequality $\log\cosh(x) \geq \frac{x^2}{2} - \frac{x^4}{12}$ for all $x \in \mathbb{R}$. Thus, using Lemma 26,

$$1 - \mathsf{TV}(P_0 \| \mathbb{E}P_1) \gtrsim \exp(-\mathsf{KL}(P_0 \| \mathbb{E}\mathbb{P}_1)) \geq \exp(-\Theta(n^2 \epsilon^{\frac{4s+1}{s}})) \overset{!}{=} \Omega(\delta).$$

By (13) we know then that $n \gtrsim \sqrt{\log(1/\delta)}/\epsilon^{\frac{2s+1/2}{s}}$ is necessary for both GoF and TS over $\mathcal{P}_{\mathsf{G}}$.

D.3.2. LOWER BOUNDS FOR LFHT

If $m \geq n$, from the GoF lower bound $n \gtrsim n_{\mathsf{GoF}}$ we conclude that $mn \gtrsim n_{\mathsf{GoF}}^2$, as desired. Therefore, throughout this section we assume that $m < n$.

Let $P_0 = \mathbb{P}_\eta^{\otimes n} \otimes \mathbb{P}_0^{\otimes n} \otimes \mathbb{P}_\eta^{m}$ and $P_1 = \mathbb{P}_\eta^{\otimes n} \otimes \mathbb{P}_0^{\otimes n} \otimes \mathbb{P}_0^{\otimes m}$, where η is a uniformly random vector of signs. Once again, we define $\omega = c_1 \epsilon^{\frac{2s+1}{2s}}$. We follow a proof similar to the cases $\mathcal{P}_{\mathsf{Db}}, \mathcal{P}_{\mathsf{H}}$ in Gerber and Polyanskiy (2022). We use the data processing inequality, the chain rule and tensorization of χ^2:

$$\begin{aligned}
\chi^2(\mathbb{E}P_0 \| \mathbb{E}P_1) &= \chi^2(\mathbb{E}\mathbb{P}_\eta^{\otimes(n+m)} \| \mathbb{E}\mathbb{P}_\eta^{\otimes n} \otimes \mathbb{P}_0^{\otimes m}) \\
&= \left(\mathbb{E}_{X_1} \mathbb{E}_{\eta_1|X} \mathbb{E}_{\eta_1'|X_1} \int_{\mathbb{R}^m} \frac{\exp\left(-\frac{1}{2}\sum_{j=1}^m \left\{ (z_j - \eta_1\omega)^2 + (z_j - \eta_1'\omega)^2 \right\} \right)}{(2\pi)^{m/2} \exp(-\frac{1}{2}\sum_{j=1}^m z_j^2)} \mathrm{d}z \right)^{c_2\epsilon^{-1/s}} - 1,
\end{aligned}$$

where $X_1 \sim (\frac{1}{2}\mathcal{N}(\omega, 1/n) + \frac{1}{2}\mathcal{N}(-\omega, 1/n))$ and $\eta_1, \eta_1'|X_1$ are iid scalar signs from the posterior $p(\cdot|X_1)$, with joint distribution $p(\eta_1, X_1) = \phi(\sqrt{n}(X_1 - \eta_1\omega))/2$.

The Gaussian integral above can be evaluated exactly and we obtain

$$\chi^2(\mathbb{E}P_0 \| \mathbb{E}P_1) = (\mathbb{E}_{X_1, \eta_1, \eta_1'} \exp(\omega^2 m \eta_1 \eta_1'))^{c_2\epsilon^{-1/s}} - 1.$$

Now, we can calculate

$$\begin{aligned}\mathbb{P}(\eta_1 = \eta_1') &= \mathbb{E}_{X_1}\frac{p(X_1|\eta_1=1)^2 + p(X_1|\eta_1=-1)^2}{(p(X_1|\eta_1=1)+p(X_1|\eta_1=-1))^2} \\ &= \frac{1}{2} + \frac{1}{4}\int \frac{(p(x_1|\eta_1=1)-p(x_1|\eta_1=-1))^2}{p(x_1|\eta_1=1)+p(x_1|\eta_1=-1)}\mathrm{d}x_1 \\ &\le \frac{1}{2} + \frac{1}{16}\sum_{b\in\{\pm 1\}}\chi^2(\mathcal{N}(b\omega,1/n)\|\mathcal{N}(-b\omega,1/n)) \\ &= \frac{1}{2} + \frac{\exp(4\omega^2 n)-1}{8}.\end{aligned}$$

Together with $\mathbb{P}(\eta_1 = \eta_1') \le 1$, we have

$$\begin{aligned}\mathbb{E}_{X_1,\eta_1,\eta_1'}\exp(\omega^2 m\eta_1\eta_1') &\le e^{-\omega^2 m} + \left(\frac{1}{2}+\frac{1}{2}\wedge\frac{e^{4\omega^2 n}-1}{8}\right)(e^{\omega^2 m}-e^{-\omega^2 m}) \\ &= \cosh(\omega^2 m) + t\sinh(\omega^2 m),\end{aligned}$$

with $t = 1 \wedge ((e^{4\omega^2 n}-1)/4)$. Distinguish into two scenarios:

- if $t = 1$, then $4\omega^2 n \ge 1$, and the above expression is $e^{\omega^2 m} \le e^{4\omega^4 nm}$;
- if $t < 1$, then $\omega^2 n \le 1/2$ and $t \le 8\omega^2 n$. Since $m < n$, and $\cosh(x) \le 1 + x^2, \sinh(x) \le 2x$ for all $x \in [0,1]$, the above expression is at most

$$1 + (\omega^2 m)^2 + 2t\omega^2 m \le \exp(17\omega^4 mn).$$

Combining the above scenarios, we have

$$\chi^2(\mathbb{E}P_0\|\mathbb{E}P_1) \le \exp(17\omega^4 nm \cdot c_2\epsilon^{-1/s}) - 1.$$

Thus, we obtain

$$1 - \mathsf{TV}(\mathbb{E}P_0, \mathbb{E}P_1) \gtrsim \frac{1}{1+\chi^2(\mathbb{E}P_0\|\mathbb{E}P_1)} \ge \exp(-17\omega^4 nm \cdot c_2\epsilon^{-1/s}) \overset{!}{=} \Omega(\delta).$$

This gives the desired lower bound

$$nm \gtrsim \frac{\log(1/\delta)}{\epsilon^{\frac{4s+1}{s}}}.$$

D.4. Lower bounds for $\mathcal{P}_{\mathsf{D}}$

Clearly all lower bounds that apply to $\mathcal{P}_{\mathsf{Db}}$ also apply to $\mathcal{P}_{\mathsf{D}}$; in particular this gives the sample complexity lower bound for GoF. In addition, lower bounds on the minimax high-probability sample complexity of TS were derived in Diakonikolas et al. (2021). Hence, inspecting the claimed minimax rates, we only need to consider the problem LFHT in the cases $m \le n \le k$ and $n \le m \le k$. We give two separate constructions for the two cases, both inspired by classical constructions in the literature. As opposed to the i.i.d. sampling models, we will use the Poissonized models and rely

on the formalism of pseudo-distributions as described in Diakonikolas et al. (2021). Specifically, suppose we can construct a random vector $(p,q) \in [0,1]^2$ such that 1) $\mathbb{E}p = \mathbb{E}q = \Theta(1/k)$ and $|\mathbb{E}[p-q]| = \Theta(\epsilon/k)$; and 2) the following χ^2 upper bounds hold for the Poisson mixture:

$$\begin{aligned}\chi^2(\mathbb{E}[\mathrm{Poi}(np)\otimes\mathrm{Poi}(nq)\otimes\mathrm{Poi}(mp)]\|\mathbb{E}[\mathrm{Poi}(np)\otimes\mathrm{Poi}(nq)\otimes\mathrm{Poi}(mq)]) &\le B(n,m,\epsilon,k),\\ \chi^2(\mathbb{E}[\mathrm{Poi}(nq)\otimes\mathrm{Poi}(np)\otimes\mathrm{Poi}(mp)]\|\mathbb{E}[\mathrm{Poi}(np)\otimes\mathrm{Poi}(nq)\otimes\mathrm{Poi}(mp)]) &\le B(n,m,\epsilon,k);\end{aligned} \tag{16}$$

then $(n,m) \in \mathcal{R}_{\mathsf{LF}}(\epsilon,\delta,\mathcal{P}_{\mathsf{D}})$ requires $kB(n,m,\epsilon,k) \gtrsim \log(1/\delta)$ (essentially via Lemma 26).

D.4.1. CASE $m \le n \le k$

Suppose that $m \le n \le k/2$, and let p, q be two random variables defined as

$$(p,q) = \begin{cases} (\frac{1}{n},\frac{1}{n}) & \text{with probability } \frac{n}{k},\\ (\frac{\epsilon}{k},\frac{2\epsilon}{k}) & \text{with probability } \frac{1}{2}(1-\frac{n}{k}),\\ (\frac{\epsilon}{k},0) & \text{with probability } \frac{1}{2}(1-\frac{n}{k}).\end{cases}$$

Note that $\mathbb{E}[p] = \mathbb{E}[q] = \Theta(1/k)$ and $|\mathbb{E}[p-q]| = \Theta(\epsilon/k)$. Let $X, Y \in \mathbb{R}^3$ be random, whose distribution is given by

$$\begin{aligned}X|(p,q) &\sim \mathrm{Poi}(np)\otimes\mathrm{Poi}(nq)\otimes\mathrm{Poi}(mp),\\ Y|(p,q) &\sim \mathrm{Poi}(np)\otimes\mathrm{Poi}(nq)\otimes\mathrm{Poi}(mq).\end{aligned}$$

Now, for any $(a,b,c) \in \mathbb{N}^3$ we have

$$\begin{aligned}\mathbb{P}(X=(a,b,c)) = \frac{1}{a!b!c!}\Big(&\frac{n}{k}e^{-2-\frac{m}{n}}\left(\frac{m}{n}\right)^c + \frac{1}{2}(1-\frac{n}{k})e^{-(3n+m)\epsilon/k}\left(\frac{\epsilon n}{k}\right)^a\left(\frac{2\epsilon n}{k}\right)^b\left(\frac{\epsilon m}{k}\right)^c\\ &+\frac{1}{2}(1-\frac{n}{k})e^{-(n+m)\epsilon/k}\left(\frac{\epsilon n}{k}\right)^a \mathbb{1}_{b=0}\left(\frac{\epsilon m}{k}\right)^c\Big).\end{aligned}$$

Similarly, for Y we get

$$\begin{aligned}\mathbb{P}(Y=(a,b,c)) = \frac{1}{a!b!c!}\Big(&\frac{n}{k}e^{-2-\frac{m}{n}}\left(\frac{m}{n}\right)^c + \frac{1}{2}(1-\frac{n}{k})e^{-(3n+2m)\epsilon/k}\left(\frac{\epsilon n}{k}\right)^a\left(\frac{2\epsilon n}{k}\right)^b\left(\frac{2\epsilon m}{k}\right)^c\\ &+\frac{1}{2}(1-\frac{n}{k})e^{-n\epsilon/k}\left(\frac{\epsilon n}{k}\right)^a \mathbb{1}_{b=c=0}\Big).\end{aligned}$$

In particular, we have

$$\mathbb{P}(Y=(a,b,c)) = \Omega\left(\frac{1}{a!b!c!}\right)\begin{cases}1 & \text{if } (a,b,c)=(0,0,0),\\ \frac{n}{k}\left(\frac{m}{n}\right)^c & \text{otherwise.}\end{cases}$$

Notice also that

$$\begin{aligned}&\mathbb{P}(X=(a,b,c)) - \mathbb{P}(Y=(a,b,c))\\ &= \frac{1}{a!b!c!}\underbrace{\frac{1}{2}(1-\frac{n}{k})e^{-n\epsilon/k}}_{=\Theta(1)}\left(\frac{\epsilon}{k}\right)^{a+b+c} n^{a+b}m^c \underbrace{\left[2^b e^{-(2n+m)\epsilon/k}\left(1-2^c e^{-m\epsilon/k}\right) + \mathbb{1}_{b=0}\left(e^{-m\epsilon/k} - \mathbb{1}_{c=0}\right)\right]}_{\triangleq I_{bc}}\\ &= \frac{\Theta(1)}{a!b!c!}\left(\frac{\epsilon}{k}\right)^{a+b+c} n^{a+b}m^c I_{bc},\end{aligned}$$

where

$$|I_{bc}| \lesssim \begin{cases} \frac{nm\epsilon^2}{k^2} & \text{if } b=c=0, \\ \frac{2^b m\epsilon}{k} & \text{if } b \geq 1, c=0, \\ \frac{n\epsilon}{k} & \text{if } b=0, c=1, \\ 2^{b+c} & \text{otherwise.} \end{cases} \tag{17}$$

We now turn to bounding the χ^2-divergence between X and Y. Using the estimates (17), we obtain

$$\begin{aligned}
\chi^2(X\|Y) &= \sum_{(a,b,c)\in\mathbb{N}^3} \frac{(\mathbb{P}(X=(a,b,c)) - \mathbb{P}(Y=(a,b,c)))^2}{\mathbb{P}(Y=(a,b,c))} \\
&\lesssim I_{00}^2 + \left(\sum_{b=c=0, a\geq 1} + \sum_{a\geq 0, b+c\geq 1} \right) \frac{\frac{1}{a!b!c!}\left(\frac{\epsilon}{k}\right)^{2a+2b+2c} n^{2a+2b} m^{2c} I_{bc}^2}{\frac{n}{k}\left(\frac{m}{n}\right)^c} \\
&= I_{00}^2 \left(1 + \sum_{a\geq 1} \frac{1}{a!} \frac{\epsilon^{2a} n^{2a-1}}{k^{2a-1}}\right) + \left(\sum_{a\geq 0} \frac{1}{a!} \frac{\epsilon^{2a} n^{2a}}{k^{2a}}\right) \sum_{b+c\geq 1} \frac{1}{b!c!} \frac{\epsilon^{2b+2c} n^{2b+c-1} m^c}{k^{2b+2c-1}} I_{bc}^2 \\
&\lesssim \frac{n^2 m^2 \epsilon^4}{k^4} \underbrace{\left(1 + \frac{n\epsilon^2}{k} e^{\epsilon^2 n^2/k^2}\right)}_{=\Theta(1)} + \underbrace{e^{\epsilon^2 n^2/k^2}}_{=\Theta(1)} \sum_{b+c\geq 1} \frac{1}{b!c!} \frac{\epsilon^{2b+2c} n^{2b+c-1} m^c}{k^{2b+2c-1}} I_{bc}^2.
\end{aligned}$$

Focusing on the sum and decomposing it as $\sum_{b+c\geq 1} = \sum_{c=0,b\geq 1} + \sum_{b=0,c=1} + \sum_{b=0,c\geq 2} + \sum_{b,c\geq 1}$ we have the estimates

$$\begin{aligned}
&\sum_{b+c\geq 1} \frac{1}{b!c!} \frac{\epsilon^{2b+2c} n^{2b+c-1} m^c}{k^{2b+2c-1}} I_{bc}^2 \\
&\lesssim \sum_{c=0,b\geq 1} \frac{1}{b!} \frac{\epsilon^{2b+2} n^{2b-1} 4^b m^2}{k^{2b+1}} + \frac{\epsilon^4 mn^2}{k^3} + \sum_{b=0,c\geq 2} \frac{1}{c!} \frac{\epsilon^{2c} n^{c-1} m^c 4^c}{k^{2c-1}} + \sum_{b,c\geq 1} \frac{1}{b!c!} \frac{\epsilon^{2b+2c} n^{2b+c-1} m^c 4^{b+c}}{k^{2b+2c-1}} \\
&\lesssim \frac{\epsilon^4 m^2 n}{k^3} + \frac{\epsilon^4 mn^2}{k^3} + \frac{\epsilon^4 m^2 n}{k^3} + \frac{\epsilon^4 mn^2}{k^3} \lesssim \frac{\epsilon^4 mn^2}{k^3}.
\end{aligned}$$

As $m \leq k$, we obtain

$$\chi^2(X\|Y) \lesssim \frac{\epsilon^4 mn^2}{k^3}.$$

By (16) we conclude that in the regime $m \leq n \leq k$, $(n,m) \in \mathcal{R}_{\mathsf{LF}}(\epsilon, \delta, \mathcal{P}_{\mathsf{D}})$ requires $n^2 m \gtrsim k^2 \log(1/\delta)/\epsilon^4$, as desired.

D.4.2. CASE $n \leq m \leq k$

This case is entirely analogous to the previous case with minor modifications. Suppose that $n \leq m \leq k/2$, and let p, q be two random variables defined as

$$(p,q) = \begin{cases} (\frac{1}{m}, \frac{1}{m}) & \text{with probability } \frac{m}{k}, \\ (\frac{\epsilon}{k}, \frac{2\epsilon}{k}) & \text{with probability } \frac{1}{2}(1-\frac{m}{k}), \\ (\frac{\epsilon}{k}, 0) & \text{with probability } \frac{1}{2}(1-\frac{m}{k}). \end{cases}$$

Let $X, Y \in \mathbb{R}^3$ be random, whose distribution is given by

$$X|(p,q) \sim \mathrm{Poi}(np) \otimes \mathrm{Poi}(nq) \otimes \mathrm{Poi}(mp),$$
$$Y|(p,q) \sim \mathrm{Poi}(nq) \otimes \mathrm{Poi}(np) \otimes \mathrm{Poi}(mp).$$

Now, for any $(a,b,c) \in \mathbb{N}^3$ we have

$$\mathbb{P}(X=(a,b,c)) = \frac{1}{a!b!c!}\Big(\frac{m}{k}e^{-\frac{2n}{m}-1}\left(\frac{n}{m}\right)^{a+b} + \frac{1}{2}(1-\frac{m}{k})e^{-(3n+m)\epsilon/k}\left(\frac{\epsilon n}{k}\right)^{a}\left(\frac{2\epsilon n}{k}\right)^{b}\left(\frac{\epsilon m}{k}\right)^{c} + \frac{1}{2}(1-\frac{m}{k})e^{-(n+m)\epsilon/k}\left(\frac{\epsilon n}{k}\right)^{a}\mathbb{1}_{b=0}\left(\frac{\epsilon m}{k}\right)^{c}\Big).$$

Similarly, for Y we get

$$\mathbb{P}(Y=(a,b,c)) = \frac{1}{a!b!c!}\Big(\frac{m}{k}e^{-\frac{2n}{m}-1}\left(\frac{n}{m}\right)^{a+b} + \frac{1}{2}(1-\frac{m}{k})e^{-(3n+m)\epsilon/k}\left(\frac{2\epsilon n}{k}\right)^{a}\left(\frac{\epsilon n}{k}\right)^{b}\left(\frac{\epsilon m}{k}\right)^{c} + \frac{1}{2}(1-\frac{m}{k})e^{-(n+m)\epsilon/k}\mathbb{1}_{a=0}\left(\frac{\epsilon n}{k}\right)^{b}\left(\frac{\epsilon m}{k}\right)^{c}\Big).$$

In particular, we have

$$\mathbb{P}(Y=(a,b,c)) = \Omega\left(\frac{1}{a!b!c!}\right)\begin{cases}1 & \text{if } (a,b,c)=(0,0,0),\\ \frac{m}{k}\left(\frac{n}{m}\right)^{a+b} & \text{otherwise.}\end{cases}$$

Notice that

$$\begin{aligned}\mathbb{P}(X=(a,b,c)) &- \mathbb{P}(Y=(a,b,c))\\ &= \frac{1}{a!b!c!}\frac{1}{2}(1-\frac{m}{k})e^{-(n+m)\epsilon/k}\left(\frac{\epsilon}{k}\right)^{a+b+c}n^{a+b}m^{c}\underbrace{\left(e^{-2n\epsilon/k}(2^b-2^a)+\mathbb{1}_{b=0}-\mathbb{1}_{a=0}\right)}_{\triangleq J_{ab}}\\ &= \frac{\Theta(1)}{a!b!c!}\left(\frac{\epsilon}{k}\right)^{a+b+c}n^{a+b}m^{c}J_{ab},\end{aligned}$$

where

$$|J_{ab}| \lesssim \begin{cases}0 & \text{if } a+b=0,\\ \frac{n\epsilon}{k} & \text{if } a+b=1,\\ 2^{a+b} & \text{if } a+b\geq 2.\end{cases} \tag{18}$$

We now turn to bounding the χ^2-divergence between X and Y. We have

$$\begin{aligned}\chi^2(X\|Y) &= \sum_{(a,b,c)\in\mathbb{N}^3}\frac{\big(\mathbb{P}(X=(a,b,c))-\mathbb{P}(Y=(a,b,c))\big)^2}{\mathbb{P}(Y=(a,b,c))}\\ &\asymp \sum_{a+b+c\geq 1}\frac{\frac{1}{a!^2b!^2c!^2}\left(\frac{\epsilon}{k}\right)^{2a+2b+2c}n^{2a+2b}m^{2c}J_{ab}^2}{\frac{1}{a!b!c!}\frac{m}{k}\left(\frac{n}{m}\right)^{a+b}}\\ &\asymp \sum_{a+b+c\geq 1}\frac{1}{a!b!c!}\frac{\epsilon^{2a+2b+2c}n^{a+b}m^{2c+a+b-1}J_{ab}^2}{k^{2a+2b+2c-1}}\\ &= \underbrace{e^{\epsilon^2m^2/k^2}}_{\Theta(1)}\sum_{a+b\geq 1}\frac{1}{a!b!}\frac{\epsilon^{2a+2b}n^{a+b}m^{a+b-1}J_{ab}^2}{k^{2a+2b-1}},\end{aligned}$$

where the last step follows from $J_{ab} = 0$ if $a = b = 0$. Now writing $t = a + b$ and distinguishing into cases $t = 1$ and $t \geq 2$, by (18) we have

$$\chi^2(X\|Y) \lesssim \frac{\epsilon^4 n^3}{k^3} + \sum_{t\geq 2} \frac{2^t}{t!} \frac{\epsilon^{2t} n^t m^{t-1} 4^t}{k^{2t-1}} \lesssim \frac{\epsilon^4 n^3}{k^3} + \frac{\epsilon^4 n^2 m}{k^3} \lesssim \frac{\epsilon^4 n^2 m}{k^3},$$

where the last line uses that $n \leq m$. Once again, we can conclude by (16) that $n^2 m \gtrsim \log(1/\delta) k^2/\epsilon^4$ is a lower bound for the sample complexity of LFHT.