Likelihood-Free Hypothesis Testing

Patrik Róbert Gerber, Yury Polyanskiy, *Fellow, IEEE,*

***Abstract*—Consider the problem of binary hypothesis testing. Given Z coming from either $\mathbb{P}^{\otimes m}$ or $\mathbb{Q}^{\otimes m}$, to decide between the two with small probability of error it is sufficient, and in many cases necessary, to have $m \asymp 1/\varepsilon^2$, where ε measures the separation between $\mathbb{P}$ and $\mathbb{Q}$ in total variation (TV). Achieving this, however, requires complete knowledge of the distributions and can be done, for example, using the Neyman-Pearson test. In this paper we consider a variation of the problem which we call likelihood-free hypothesis testing, where access to $\mathbb{P}$ and $\mathbb{Q}$ is given through n i.i.d. observations from each. In the case when $\mathbb{P}$ and $\mathbb{Q}$ are assumed to belong to a non-parametric family, we demonstrate the existence of a fundamental trade-off between n and m given by $nm \asymp n^2_{\mathsf{GoF}}(\varepsilon)$, where $n_{\mathsf{GoF}}(\varepsilon)$ is the minimax sample complexity of testing between the hypotheses $H_0 : \mathbb{P} = \mathbb{Q}$ vs $H_1 : \mathsf{TV}(\mathbb{P}, \mathbb{Q}) \geq \varepsilon$. We show this for three families of distributions, in addition to the family of all discrete distributions for which we obtain a more complicated trade-off exhibiting an additional phase-transition. Our results demonstrate the possibility of testing without fully estimating $\mathbb{P}$ and $\mathbb{Q}$, provided $m \gg 1/\varepsilon^2$.**

***Index Terms*—Hypothesis testing, likelihood-free inference, minimax sample complexity, nonparametric statistics, goodness-of-fit testing, density estimation, total variation**

I. Introduction

THE setting called *likelihood-free inference (LFI)*, also known as simulation based inference (SBI), has independently emerged in many areas of science over the past decades. Given an expensive to collect "experimental" dataset and the ability to simulate from a high fidelity, often mechanistic, stochastic model, whose output distribution and likelihood is intractable and inapproximable, how does one perform model selection, parameter estimation or construct confidence sets? The list of disciplines where such highly complex black-box simulators are used is long, and include particle physics, astrophysics, climate science, epidemiology, neuroscience and ecology to just name a few. For some of the above fields, such as climate modeling, the bottleneck resource is in fact the simulated data as opposed to the experimental data. In either case, understanding the trade-off between the number of simulations and experiments necessary to do valid inference is crucial. Our aim in this paper is to introduce a theoretical framework under which LFI can be studied using the tools of non-parametric statistics and information theory.

To illustrate we draw an example from high energy physics, where LFI methods are used and developed extensively. The discovery of the Higgs boson in 2012 [1], [2] is regarded as the crowning achievement of the Large Hadron collider (LHC) - the most expensive instrument ever built. Using a composition of complex simulators [3]–[7] modeling the standard model and the detection process, physicists are able to simulate the results of LHC experiments. Given actual data $Z_1, \ldots, Z_m$ from the collider, to verify existence of the Higgs boson one tests whether the null hypothesis (physics without the Higgs boson, or $Z_i \overset{iid}{\sim} \mathbb{P}_0$) or the alternative hypothesis (physics with the Higgs boson, or $Z_i \overset{iid}{\sim} \mathbb{P}_1$) describes the experimental data more accurately. The standard Neyman-Pearson likelihood ratio test is not implementable since $\mathbb{P}_0$ and $\mathbb{P}_1$ are only available via simulators. How was this statistical test actually performed? First, a probabilistic classifier C was trained on simulated data to distinguish the two hypotheses (a boosted decision tree to be more specific). Then, the proportion of real data points falling in the set $S = \{x \in \mathbb{R}^d : C(x) \leq t\}$ was computed, where t is chosen to maximize an asymptotic approximation of the power. Finally, p-values are reported based on the asymptotic distribution under a Poisson sampling model [8], [9]. Summarizing, the "Higgs boson" test was performing the simple comparison

$$\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{Z_i \in S\} \lessgtr \gamma, \tag{1}$$

where $Z_1, \ldots, Z_m$ are the real data and γ is some threshold. Such count-based tests, named after Scheffé in folklore [10, Section 6], are quite intuitive.

P. R. Gerber was supported in part by the NSF award IIS-1838071. Y. Polyanskiy was supported in part by the NSF under grant No CCF-2131115 and by the MIT-IBM Watson AI Lab.

0000–0000/00$00.00 ©2024 IEEE

Notice that Scheffé's test converts each observation Z_i into a binary $0/1$ value. This extreme quantization certainly helps robustness, but should raise the suspicion of potential loss of power. Indeed, when the distributions under both hypotheses are completely known, the optimal Neyman-Pearson test thresholds the sum of *real-valued* logarithms of the likelihood-ratio. Thus, it is natural to expect that a good test should aggregate non-binary values. This is what motivated this work originally, although follow-up work [11] has shown that Scheffé's test with a properly trained classifier can also be optimal.

Let us describe the test that we study for most of this paper. Given estimates $\widehat{p}_0, \widehat{p}_1$ of the density of the null and alternative distributions based on simulated samples, our test proceeds via the comparison

$$\frac{2}{m}\sum_{i=1}^{m}(\widehat{p}_0(Z_i) - \widehat{p}_1(Z_i)) \lessgtr \gamma \tag{2}$$

where $Z_1, \ldots, Z_m$ are the real data. Tests of this kind originate from the famous goodness-of-fit work of Ingster [12], which corresponds to taking $\widehat{p}_0 = p_0$, as the null-density is known exactly.[1] The surprising observation of Ingster was that such a test is able to reject the null hypothesis that $Z_i \overset{iid}{\sim} p_0$ even when the true distribution of Z is much closer to p_0 than described by the optimal density-estimation rate; in other words *goodness-of-fit testing is significantly easier than estimation*. In fact we will use $\gamma = \|\widehat{p}_0\|_2^2 - \|\widehat{p}_1\|_2^2$ in which case (2) boils down to the comparison of two squared L^2-distances.

Our overall goal is to understand the trade-off between the number n of simulated observations and the size of the actual data set m. The characterization of this tradeoff is reminiscent of the rate-regions in multi-user information theory, but there is an important difference that we wanted to emphasize for the reader. In information theory, the problem is most often stated in the form "given a distribution $P_{X,Y,Z}$, or a channel $P_{Y,Z|X}$, find the rate region", with the distribution being completely specified ahead of time. In minimax statistics, however, distributions are a priori only known to belong to a certain class. In *estimation problems* the fundamental limits are thus defined by minimizing the estimation error over this class, and the theoretical goal is to characterize the worst-case rate at which this error converges to zero as the sample size grows to infinity. The definition of the fundamental limit in *testing problems*, however, is more subtle. If the total variation separation ε between the null and alternative distribution is fixed, and the number of samples is taken to infinity, then the rate of convergence trivializes and becomes exponentially decreasing in n. By now a standard definition of fundamental limit, as suggested by Ingster following ideas of Pittman efficiency, is to vary ε with n and to find the fastest possible decrease of ε so as to still have an acceptable probability of error. This is the approach taken in the literature on goodness-of-fit and two-sample testing, and also the one we adopt here. This perspective is also widely used in TCS where the optimal value of n, as a function of ε, is referred to as the "sample complexity" of the problem.

Specifically, we assume that it is known a priori that the two distributions $\mathbb{P}_0, \mathbb{P}_1$ belong to a known class $\mathcal{P}$ and are ε-separated under total variation. Given a large number n of samples simulated from $\mathbb{P}_0$ and $\mathbb{P}_1$ and m samples $Z_1, \ldots, Z_m$ from the experiment, our goal is to test which of the $\mathbb{P}_i$ generated the data. If n is sufficiently large to estimate $\mathbb{P}_i$ in total variation to precision $\varepsilon/10$, then one can perform the hypothesis test with $m \asymp 1/\varepsilon^2$ experimental samples, which is information-theoretically optimal even under oracle knowledge of $\mathbb{P}_i$'s. However, looking at the test (1) one may wonder if the full estimation of the distributions $\mathbb{P}_i$ is needed, or whether perhaps a suitable decision boundary could be found with a lot fewer simulated samples n. Unfortunately, our *first main result* disproves this intuition: *any test using the minimal $m \asymp 1/\varepsilon^2$ dataset size will require n so large as to be enough to estimate the distributions of $\mathbb{P}_0$ and $\mathbb{P}_1$ to within accuracy $\asymp \varepsilon$*, which is the distance separating the two hypotheses. In particular, any method minimizing m performs no different in the worst case, than pairing off-the-shelf density estimators $\widehat{p}_0, \widehat{p}_1$ and applying (1) with $S = \{\widehat{p}_1 \geq \widehat{p}_0\}$.

This result appears rather pessimistic and seems to invalidate the whole attraction of LFI, which after all hopes to circumvent the exorbitant number of simulation samples required for fully learning high-dimensional distributions. Fortunately, *our second result* offers a resolution: if more data samples $m \gg 1/\varepsilon^2$ are collected, then testing is possible with n much smaller than required for density estimation. More precisely, when neither p_0 nor p_1 are known except through n i.i.d. samples from each, the test (2) is able to detect which of the two distributions generated the Z-sample, *even when the number of samples n* is insufficient for any estimate $\widehat{p}_i$ to be within a distance $\asymp \varepsilon = \mathsf{TV}(p_0, p_1)$ from the

[1] In the case of discrete distributions on a finite (but large) alphabet, the idea was rediscovered by the computer science community startin with [13]. Moreover, the difference of L^2-norms statistic was first studied in [14]. See Section I-B for more on the latter.

true values. In other words, the test is able to reliably detect the true hypotheses even though the estimates $\widehat{p}_i$ themselves have accuracy that is orders of magnitude larger than the separation ε between the hypotheses.

In summary, this paper shows that likelihood-free hypothesis testing (LFHT) is possible without learning the densities when $m \gg 1/\varepsilon^2$, but not otherwise. It turns out that (appropriate analogues of) the simple test (2) has minimax optimal sample complexity up to constants in both n and m in all "regular" settings, see also the discussion at the end of Section II-B.

A. Informal Statement of the Main Result

Let us formulate the problem using the notation used throughout the rest of the paper. Suppose that we observe true data $Z \sim \mathbb{P}_{\mathsf{Z}}^{\otimes m}$ and that we have two candidate parameter settings for our simulator, from which we generate two artificial datasets $X \sim \mathbb{P}_{\mathsf{X}}^{\otimes n}$ and $Y \sim \mathbb{P}_{\mathsf{Y}}^{\otimes n}$. If we are convinced that one of the settings accurately reflects reality, we are faced with the problem of testing the hypothesis

$$H_0 : \mathbb{P}_{\mathsf{X}} = \mathbb{P}_{\mathsf{Z}} \qquad \text{versus} \qquad H_1 : \mathbb{P}_{\mathsf{Y}} = \mathbb{P}_{\mathsf{Z}}. \tag{3}$$

Remark 1. We emphasize that $\mathbb{P}_{\mathsf{X}}$ and $\mathbb{P}_{\mathsf{Y}}$ are known only through the n simulated samples. Thus, (3) can be interpreted as binary hypothesis testing with approximately specified hypotheses. Alternatively, using the language of machine learning, we may think of this problem as having n labeled samples from both classes, and m unlabeled samples. The twist is that the unlabeled samples are guaranteed to have the same common label, that is, they all come from a single class. One can think of many examples of this setting occurring in genetic, medical and other studies.

To put (3) in a minimax framework, suppose that $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}} \in \mathcal{P}$ for a known class $\mathcal{P}$, and that $\mathsf{TV}(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}) \geq \varepsilon$. Clearly (3) becomes "easier" if we have a lot of data (large sample sizes n and m) or if the hypotheses are well-separated (large ε). We are interested in characterizing the pairs of values (n, m) as functions of ε and $\mathcal{P}$, for which the hypothesis test (3) can be performed with constant type-I and type-II error. Letting $n_{\mathsf{GoF}}(\varepsilon, \mathcal{P})$ denote the minimax sample complexity of goodness-of-fit testing (Definition 2), we show for *several different classes* of $\mathcal{P}$, that (3) is possible with total error, say, 5% if and only if

$$m \gtrsim 1/\varepsilon^2 \quad \text{and} \quad n \gtrsim n_{\mathsf{GoF}} \quad \text{and} \quad mn \gtrsim n_{\mathsf{GoF}}^2. \tag{4}$$

We also make the observation that $n_{\mathsf{GoF}}^2 \varepsilon^2 \asymp n_{\mathsf{Est}}$ for these classes, where $n_{\mathsf{Est}}(\varepsilon, \mathcal{P})$ denotes the minimax complexity of density estimation to ε-accuracy (Definition 4) with respect to total variation. This provides additional meaning to the mysterious formula of Ingster [12] for the sample complexity of goodness-of-fit testing over the class of β-smooth densities over $[0,1]^d$, see Table I.[2] More importantly, it allows us to interpret (3) as an "interpolation" between different fundamental statistical procedures, namely

$\mathsf{A} \leftrightarrow$ Binary hypothesis testing (BHT),
$\mathsf{B} \leftrightarrow$ Estimation followed by robust BHT,
$\mathsf{C} \leftrightarrow$ Two-sample testing,
$\mathsf{D} \leftrightarrow$ Goodness-of-fit testing,

corresponding to the extreme points $\mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D}$ on Fig. 1.

B. Related Work

LHFT as defined in (3) initially appeared in Gutman's paper [15], building on Ziv's work [16], where the problem is studied for distributions on a fixed, finite alphabet. Ziv called the problem *classification with empirically observed statistics*, to emphasize the fact that hypotheses are specified only in terms of samples and the underlying true distributions are unknown. In [17] it is shown that the error exponent of Gutman's test is second order optimal. Recent work [18]–[21] extends this problem to distributed and sequential testing. However, the setting of these papers is fundamentally different from ours, a point which we expand on below.

Given two arbitrary, unknown $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}$ over a finite alphabet of fixed size, Gutman's test (see [17, Equation (4)]) rejects the null hypothesis $H_0 : \mathbb{P}_{\mathsf{Z}} = \mathbb{P}_{\mathsf{X}}$ in favor of the alternative $H_1 : \mathbb{P}_{\mathsf{Z}} = \mathbb{P}_{\mathsf{Y}}$ if the statistic $\mathrm{GJS}(\widehat{\mathbb{P}}_{\mathsf{X}}, \widehat{\mathbb{P}}_{\mathsf{Z}}, \alpha)$ is large, where $\widehat{\mathbb{P}}$ denotes empirical measures, GJS denotes the generalized Jensen-Shannon divergence defined in [17, Equation (3)] and $\alpha = n/m$. In other words, it simply performs a two-sample test using the samples from $\mathbb{P}_{\mathsf{X}}$ and $\mathbb{P}_{\mathsf{Z}}$ of size n and m respectively, and completely discards the sample from $\mathbb{P}_{\mathsf{Y}}$. In light of our sample complexity results this is strictly sub-optimal due to minimax lower bounds on two-sample testing, see the difference of light gray and striped regions in Fig. 1.

[2] A possible reason for this observation having been missed previously is that fundamental limits in statistics are usually presented in the form of *rates* of loss decrease with n, for example $r_{\mathsf{Est}}(n) \triangleq n_{\mathsf{Est}}^{-1}(n) = 1/n^{\beta/(2\beta+d)}$ and $r_{\mathsf{GoF}}(n) \triangleq n_{\mathsf{GoF}}^{-1}(n) = 1/n^{\beta/(2\beta+d/2)}$ for β-smooth densities. Unlike $n_{\mathsf{Est}} \asymp n_{\mathsf{GoF}}^2 \varepsilon^2$ there seems to be no simple relation between r_{Est} and r_{GoF}.

More generally, the method of types, which is a crucial tool for the works cited above, cannot be used to derive our results, because in the regime where the alphabet size k scales with the sample size n, the usual $\binom{n}{k} = e^{o(n)}$ approximation no longer holds, i.e. these factors affect estimation rates and do not lead to tight minimax results. As a consequence, one cannot deduce results about the minimax sample complexity of LFHT from works on the classical regime because the latter do not quantify the speed of convergence of the error terms as a function of the alphabet size. Specifically, let us examine [17, Theorem 1], which is a strengthening of the results of [15]. Paraphrasing, it states that for any fixed ratio $\alpha = n/m$ and pair of distributions $(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}})$, Gutman's test has type-II error bounded by $1/3$ when given samples from $\mathbb{P}_{\mathsf{X}}$ and $\mathbb{P}_{\mathsf{Y}}$ as input, and type-I error bounded by $\exp(-\lambda n)$ given arbitrary input, where

$$\lambda = \mathrm{GJS}(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}, \alpha) + \sqrt{\frac{V(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}, \alpha)}{n}}\Phi^{-1}(1/3) + \mathcal{O}\left(\frac{\log(n)}{n}\right) \quad (5)$$

as $n \to \infty$. Here V denotes the dispersion function defined in [17, Equation (9)] and Φ is the standard normal cdf. The crucial point we make here is that in (5) the dependence of the $\mathcal{O}(\log(n)/n)$ term on $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}$, and in particular their support size k and the ratio $\alpha = n/m$ is unspecified. Because of this, (5) and similar results cannot be used to derive minimax sample complexities as $\min\{n, m, k\} \to \infty$ jointly at possibly different rates.

This distinction between the fixed alphabet size setting studied in [15]–[17] and similar works, and our large alphabet setting was recognized by [14], [22]–[24] whose results are much closer to those of this paper. In [23] Huang and Meyn introduce the concept of "generalized error exponent" to deal with support sizes that grow superlinearly with sample size (referred to as the "sparse sample regime" by them) in the setting of uniformity testing.[3] In [22] they extend this idea to LFHT and say, quote,

> "In the classification problem, the classical error exponent analysis has been applied to the case of fixed alphabet in [16] and [15].... However, in the sparse sample problem, the classical error exponent concept is again not applicable, and thus a different scaling is needed."

[3]Uniformity testing is the problem of goodness-of-fit testing where the null is given by a uniform distribution.

Moving on to [14], [24], their authors study (3) with $n = m$ over the class of discrete distributions p with $\min_i p_i \asymp \max_i p_i \asymp 1/n^\alpha$, which they call α-*large sources*. Disregarding the dependence on the TV-separation ε, effectively setting ε to a constant, they find that achieving non-trivial minimax error is possible if and only if $\alpha \leq 2$, using in fact the same *difference of squared L^2-distances* test (2) that we study in this paper. Follow-up work [22] extends to the case $m \neq n$ and the class of distributions on alphabet $[k]$ with $\max_i p_i \lesssim 1/k$, we also cover this class under the name $\mathcal{P}_{\mathsf{Db}}$. In the regime of constant separation $\varepsilon = \Theta(1)$ and $n, m \to \infty$ they show that LFHT with vanishing error is possible if and only if $k = o(\min(n^2, mn))$, thus discovering for the first time the *trade-off* between m and n.[4] Contrasting with our work, we are the first to characterize the full m, n, ε trade-off in the regime of constant probability of error, and we also consider three other classes of distributions, in addition to $\mathcal{P}_{\mathsf{Db}}$.

Another related problem is that of two-sample testing with unequal sample sizes, studied in [25], [26] for the class of discrete distributions $\mathcal{P}_{\mathsf{D}}$. In Section III-A we present reductions that show that our problem's sample complexity equals, up to constant factors, to that of two-sample testing in the case $m \geq n$. We emphasize that the distinction between $m \geq n$ and $m \leq n$ is necessary for this equivalence: in the latter case the sample complexities of the two problems are not the same. Moreover, our reduction doesn't help us solve classes other than $\mathcal{P}_{\mathsf{D}}$, as two-sample testing with unequal sample size exhibits a trade-off between n and m only in classes for which $n_{\mathsf{TS}} \neq n_{\mathsf{GoF}}$, see also the discussion at the end of Section II-B.

The test (1) has been considered previously [27]–[33] and is also known as a "classification accuracy" test (CAT). Follow-up work [11] to the present paper shows that CATs are able to attain a (near-)minimax optimality in all settings studied here, and also achieve optimal dependence on the probability of error (in this paper we only consider a fixed error probability).

[4]The paper [22] contains implicitly other interesting results. For example, it appears that the constructive (upper bound) part of their proof if done carefully can also handle the case of variable $\varepsilon \to 0$ in the regime $m, n \lesssim k$. Specifically, we believe they also show that for the minimax error $\delta \in (0, 1)$ LFHT is possible if $k \log(1/\delta)/\varepsilon^4 \lesssim \min(n^2, nm)$. The lower bound appears to show LFHT is possible only if $k \log(1/\delta) \lesssim \min(n^2, nm)$. In addition they also apply the flattening technique, later re-discovered in [25].

C. Contributions

Though the likelihood-free hypothesis testing problem (3) has previously appeared under various disguises and was studied in different regimes for the class of bounded discrete distributions, it omitted the key question of understanding the dependence of the sample complexity on the separation ε. Our work fully characterizes the dependence on the separation ε (Theorems 1 and 2). We discover the existence of a rather non-trivial trade-off between the m and n showing that in the likelihood-free setting statistical performance (m) can be traded for computational resources (n). Our results are shown for not just one but multiple distribution classes. In addition, we also demonstrate that LFHT naturally interpolates between its special cases corresponding to goodness-of-fit testing, two-sample testing and density-estimation. As a by-product we observe the relation $n_{\mathsf{GoF}}^2 \varepsilon^2 \asymp n_{\mathsf{Est}}$ that holds over several classes of distributions and measures of separation, hinting at some universality property. On the technical side we provide a unified upper bound analysis for all regular classes we consider, and prove matching lower bounds using techniques of Tsybakov, Ingster and Valiant. Our upper bound analysis is inspired by Ingster [12], [34] whose L^2-distance testing approach, originally designed for goodness-of-fit in smooth-density classes, has been rediscovered in the discrete-alphabet world [13], [14], [24]. Compared to Ingster's work, the new ingredient needed in the discrete case is a "flattening" reduction [22], [25], [35], which we also utilize. Several minor extensions are also shown along the way, namely, robustness with respect to L^2-misspecification (Theorem 3) and characterization of n_{GoF} for the class of β-smooth densities with $\beta \leq 1$ under Hellinger separation (Theorem 4).

D. Structure

Section II defines the statistical problems and the classes of distributions that are studied throughout the paper, and discusses multiple tests for likelihood-free hypothesis testing. Section III contains our main results and the discussion linking to goodness-of-fit and two-sample testing, estimation and robustness. In Section IV we provide sketch proofs for our results. Finally, in Section V we discuss possible future directions of research. The detailed proofs of Theorems 1 to 4 and all auxiliary results are included in the Appendix.

E. Notation

For $k \in \mathbb{N}$ we write $[k] \triangleq \{1, 2, \ldots, k\}$. For $x, y \in \mathbb{R}$ we write $x \wedge y \triangleq \min\{x, y\}$, $x \vee y \triangleq \max\{x, y\}$. We use the Bachmann–Landau notation $\Omega, \Theta, \mathcal{O}, o$ as usual and write $f \lesssim g$ for $f = \mathcal{O}(g)$ and $f \asymp g$ for $f = \Theta(g)$. For $c \in \mathbb{R}$ and $A \subseteq \mathbb{R}^2$ we write $cA \triangleq \{(ca_1, ca_2) \in \mathbb{R}^2 : (a_1, a_2) \in A\}$. For two sets $A, B \subseteq \mathbb{R}^2$ we write $A \asymp B$ if there exists $c \in [1, \infty)$ with $\frac{1}{c}A \subseteq B \subseteq cA$. For two probability measures μ, ν dominated by η with densities p, q we define the following divergences: $\mathsf{TV}(\mu, \nu) \triangleq \frac{1}{2}\int |p - q| \mathrm{d}\eta$, $\mathsf{H}(\mu, \nu) \triangleq (\int (\sqrt{p} - \sqrt{q})^2 \mathrm{d}\eta)^{1/2}$, $\mathsf{KL}(\mu \| \nu) \triangleq \int p \log(p/q) \mathrm{d}\eta$, $\chi^2(\mu \| \nu) \triangleq \int ((p - q)^2/q) \mathrm{d}\eta$. Abusing notation, we sometimes write (p, q) as arguments instead of (μ, ν). We write $\|\cdot\|_p$ for the L^p and ℓ^p norms, where the base measure shall be clear from the context.

II. Sample Complexity, Non-Parametric Classes, and Tests

In the first two parts of this section we go over the technical background and definitions that are required to understand the rest of the paper, after which we give an exposition of multiple alternative approaches for our problem in Section II-C.

A. Five Fundamental Problems in Statistics

Formally, we define a hypothesis as a set of probability measures. Given two hypotheses H_0 and H_1 on some space $\mathcal{X}$, we say that a function $\psi : \mathcal{X} \to \{0, 1\}$ successfully tests the two hypotheses against each other if

$$\max_{i=0,1} \sup_{P \in H_i} \mathbb{P}_{S \sim P}(\psi(S) \neq i) \leq 1/3. \tag{6}$$

Remark 2. For our purposes, the constant $1/3$ above is unimportant and could be replaced by any number less than $1/2$. Throughout the paper we are interested in the asymptotic order of the sample complexity, and $\Omega(\log(1/\delta))$-way sample splitting followed by a majority vote decreases the overall error probability to $\mathcal{O}(\delta)$ of any successful tester, at the cost of inflating the sample complexity by a multiplicative $\mathcal{O}(\log(1/\delta))$ factor. Unfortunately, the resulting dependence on δ is sub-optimal except for binary hypothesis testing, see for example [36, Theorem 4.7]. Recent results for uniformity [37] and two-sample testing [38], and our follow-up work on LFHT [11] resolves the optimal dependence to be $\sqrt{\log(1/\delta)}$ or even $\sqrt[3]{\log(1/\delta)}$ in some regimes.

Throughout 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. Finally, $\mathbb{P}_0, \mathbb{P}_1 \in \mathcal{P}$ refer to distributions that are *known* to us. We now define five fundamental problems in statistics that we refer to throughout this paper.

Definition 1. *Binary hypothesis testing* is the problem of testing

$$H_0 : \mathbb{P}_{\mathsf{X}} = \mathbb{P}_0 \qquad \text{against} \qquad H_1 : \mathbb{P}_{\mathsf{X}} = \mathbb{P}_1 \tag{7}$$

based on the sample X. We use $n_{\mathsf{HT}}(\varepsilon, \mathcal{P})$ to denote the *minimax sample complexity of binary hypothesis testing*, which is the smallest number such that for all $n \geq n_{\mathsf{HT}}(\varepsilon, \mathcal{P})$ and all $\mathbb{P}_0, \mathbb{P}_1 \in \mathcal{P}$ with $\mathsf{TV}(\mathbb{P}_0, \mathbb{P}_1) \geq \varepsilon$ there exists a function $\psi : \mathcal{X}^n \to \{0, 1\}$, which given X as input successfully tests H_0 against H_1 in the sense of (6).

It is well known that the complexity of binary hypothesis testing is controlled by the Hellinger divergence.

Lemma 1. *For all ε and $\mathcal{P}$ with $|\mathcal{P}| \geq 2$, the relation*

$$n_{\mathsf{HT}}(\varepsilon, \mathcal{P}) = \Theta\Big(\sup_{\mathbb{P}_i \in \mathcal{P} : \mathsf{TV}(\mathbb{P}_0, \mathbb{P}_1) \geq \varepsilon} \mathsf{H}^{-2}(\mathbb{P}_0, \mathbb{P}_1) \Big) \tag{8}$$

holds, where the implied constant is universal.

Proof: We include the proof in Section D-A for completeness. ∎

For all $\mathcal{P}$ considered in this paper $n_{\mathsf{HT}} = \Theta(1/\varepsilon^2)$ holds. Therefore, going forward we usually refrain from the general notation n_{HT} and simply write $1/\varepsilon^2$.

Definition 2. *Goodness-of-fit testing* is the problem of testing

$$\begin{aligned} &H_0 : \mathbb{P}_{\mathsf{X}} = \mathbb{P}_0 \\ \text{against} \qquad &H_1 : \mathsf{TV}(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_0) \geq \varepsilon \text{ and } \mathbb{P}_{\mathsf{X}} \in \mathcal{P} \end{aligned} \tag{9}$$

based on the sample X. We write $n_{\mathsf{GoF}}(\varepsilon, \mathcal{P})$ for the *minimax sample complexity of goodness-of-fit testing*, which is the smallest value such that for all $n \geq n_{\mathsf{GoF}}(\varepsilon, \mathcal{P})$ and $\mathbb{P}_0 \in \mathcal{P}$ there exists a function $\psi : \mathcal{X}^n \to \{0, 1\}$, which given X as input successfully tests H_0 against H_1 in the sense of (6).

Definition 3. *Two-sample testing* is the problem of testing

$$\begin{aligned} &H_0 : \mathbb{P}_{\mathsf{X}} = \mathbb{P}_{\mathsf{Z}} \text{ and } \mathbb{P}_{\mathsf{X}} \in \mathcal{P} \\ \text{against } &H_1 : \mathsf{TV}(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Z}}) \geq \varepsilon \text{ and } \mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Z}} \in \mathcal{P} \end{aligned} \tag{10}$$

based on the samples X and Z. We write $\mathcal{R}_{\mathsf{TS}}(\varepsilon, \mathcal{P})$ for the maximal subset of $\mathbb{R}^2$ such that for any $(n, m) \in \mathbb{N}^2$ for which there exists $(x, y) \in \mathcal{R}_{\mathsf{TS}}(\varepsilon, \mathcal{P})$ with $(n, m) \geq (x, y)$ coordinate-wise, there also exists a function $\psi : \mathcal{X}^n \times \mathcal{X}^m \to \{0, 1\}$, which given X and Z as input successfully tests between H_0 and H_1 in the sense of (6). We will use the abbreviation $n_{\mathsf{TS}}(\varepsilon, \mathcal{P}) = \min\{\ell \in \mathbb{N} : (\ell, \ell) \in \mathcal{R}_{\mathsf{TS}}(\varepsilon, \mathcal{P})\}$ and refer to it as the *minimax sample complexity of two-sample testing*.

Definition 4. The *minimax sample complexity of* estimation is the smallest value $n_{\mathsf{Est}}(\varepsilon, \mathcal{P})$ such that for all $n \geq n_{\mathsf{Est}}(\varepsilon, \mathcal{P})$ there exists an estimator $\widehat{\mathbb{P}}_{\mathsf{X}}$, which given X as input satisfies

$$\mathbb{E}\mathsf{TV}(\widehat{\mathbb{P}}_{\mathsf{X}}, \mathbb{P}_{\mathsf{X}}) \leq \varepsilon. \tag{11}$$

In order to simplify the presentation of our final definition, let us temporarily write $\mathcal{P}_\varepsilon = \{(\mathbb{Q}_0, \mathbb{Q}_1) \in \mathcal{P}^2 : \mathsf{TV}(\mathbb{Q}_0, \mathbb{Q}_1) \geq \varepsilon\}$. That is, $\mathcal{P}_\varepsilon$ is the set of pairs of distributions in the class $\mathcal{P}$ which are ε separated in total variation.

Definition 5. *Likelihood-free hypothesis testing* is the problem of testing

$$\begin{aligned} H_0 &: \mathbb{P}_{\mathsf{Z}} = \mathbb{P}_{\mathsf{X}} \text{ and } (\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}) \in \mathcal{P}_\varepsilon \\ \text{against } H_1 &: \mathbb{P}_{\mathsf{Z}} = \mathbb{P}_{\mathsf{Y}} \text{ and } (\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}) \in \mathcal{P}_\varepsilon \end{aligned} \tag{12}$$

based on the samples X, Y and Z. Write $\mathcal{R}_{\mathsf{LF}}(\varepsilon, \mathcal{P})$ for the maximal subset of $\mathbb{R}^2$ such that for any $(n, m) \in \mathbb{N}^2$ for which there exists $(x, y) \in \mathcal{R}_{\mathsf{LF}}(\varepsilon, \mathcal{P})$ with $(n, m) \geq (x, y)$ coordinate-wise, there also exists a function $\psi : \mathcal{X}^n \times \mathcal{X}^n \times \mathcal{X}^m \to \{0, 1\}$, which given X, Y and Z as input successfully tests H_0 against H_1 in the sense of (6).

Requiring $\mathcal{R}_{\mathsf{TS}}(\varepsilon, \mathcal{P})$ to be maximal is well defined, because for any $(n_0, m_0) \in \mathcal{R}_{\mathsf{TS}}(\varepsilon, \mathcal{P})$ and $(n, m) \in \mathbb{N}^2$ with $(n_0, m_0) \leq (n, m)$ coordinate-wise, it must also hold that $(n, m) \in \mathcal{R}_{\mathsf{LF}}$, since ψ can simply disregard the extra samples. Clearly the same applies also to $\mathcal{R}_{\mathsf{LF}}(\varepsilon, \mathcal{P})$.

Remark 3. All five definitions above can be modified to measure separation with respect to an arbitrary function d instead of TV. We will write $n_{\mathsf{GoF}}(\varepsilon, \mathsf{d}, \mathcal{P})$ et cetera for the corresponding values.

B. Four Classes of Distributions

All of our definitions in the previous section assumed that we have some class of distributions $\mathcal{P}$ at hand. Below

we introduce the classes that we study throughout the rest of the paper.

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

$$\|f\|_{\mathcal{C}_\beta} \triangleq \max\Big\{ \max_{0 \le |\alpha| \le \underline{\beta}} \|f^{(\alpha)}\|_\infty, \sup_{x \neq y \in [0,1]^d, |\alpha| = \underline{\beta}} \frac{|f^{(\alpha)}(x) - f^{(\alpha)}(y)|}{\|x - y\|_2^{\beta - \underline{\beta}}} \Big\} \le C, \tag{13}$$

where we write $|\alpha| = \sum_{i=1}^d \alpha_i$ for the multiindex $\alpha \in \mathbb{N}^d$ as usual. We further define $\mathcal{P}_{\mathsf{H}}(\beta, d, C)$ to be the class of distributions with Lebesgue-densities in $\mathcal{C}(\beta, d, C)$.

(ii) *Gaussian sequence model on the Sobolev ellipsoid.* Given $C > 0$ and a smoothness parameter $s > 0$, we define the Sobolev ellipsoid

$$\mathcal{E}(s, C) \triangleq \Big\{ \theta \in \mathbb{R}^{\mathbb{N}} : \sum_{j=1}^{\infty} j^{2s} \theta_j^2 \le C \Big\}. \tag{14}$$

Our second distribution class is given by

$$\mathcal{P}_{\mathsf{G}}(s, C) \triangleq \{ \mu_\theta \, : \theta \in \mathcal{E}(s, C) \}, \tag{15}$$

where $\mu_\theta = \otimes_{i=1}^\infty \mathcal{N}(\theta_i, 1)$. It is well known that this class models an s-smooth signal under Gaussian white noise, see for example [39, Section 1.7.1] for an exposition of this connection.

(iii) *Distributions on a finite alphabet.* For $k \ge 2$, let

$$\mathcal{P}_{\mathsf{D}}(k) \triangleq \{\text{all distributions on } \{1, 2, \dots, k\}\} \tag{16}$$

denote the class of all discrete distributions.

(iv) *Bounded distributions on a finite alphabet.* Our final class is defined as

$$\mathcal{P}_{\mathsf{Db}}(k, C) \triangleq \{ p \in \mathcal{P}_{\mathsf{D}}(k) : \|p\|_\infty \le C/k \} \tag{17}$$

for $C > 1$. In other words, $\mathcal{P}_{\mathsf{Db}}$ are those distributions with support in $\{1, 2, \dots, k\}$ that are bounded by a constant multiple of the uniform distribution.

Note that depending on the choice of C some of the above distribution classes may be empty. To avoid such issues, throughout the rest of paper we implicitly operate under the following assumption.

Assumption 1. We always assume that $C > 1$ when referring to $\mathcal{P}_{\mathsf{H}}(\beta, d, C)$ and $\mathcal{P}_{\mathsf{Db}}(k, C)$.

As we shall see in Section III-B when discussing our results, the behaviour of $\mathcal{P}_{\mathsf{D}}$ is qualitatively different from the other three classes introduced above. Consequently, we will sometimes refer to $\mathcal{P}_{\mathsf{Db}}$ as the "regular discrete" class, and we will see that its minimax sample complexities are similar to $\mathcal{P}_{\mathsf{H}}$ and $\mathcal{P}_{\mathsf{G}}$ but different from $\mathcal{P}_{\mathsf{D}}$. More generally we will call the classes $\mathcal{P}_{\mathsf{H}}, \mathcal{P}_{\mathsf{G}}, \mathcal{P}_{\mathsf{Db}}$ "regular", characterized by the fact that $n_{\mathsf{GoF}} \asymp n_{\mathsf{TS}}$, or equivalently, by the fact that $\mathcal{R}_{\mathsf{TS}} \asymp \{(n, m) : \min\{n, m\} \ge n_{\mathsf{TS}}\}$.

C. Tests for LFHT

We start this section by reintroducing the difference of L^2-distances statistic that our results are based on, and which we've already seen in (2). Then, in Section II-C2 we mention some natural alternative approaches to the problem, which we however do not study further. Therefore, the reader that wishes to proceed to our results without delay may safely skip over Section II-C2.

1) Ingster's L^2-Distance Test: For simplicity we focus on the case of discrete distributions. This case is more general than may first appear: for example in the case of smooth densities on $[0,1]^d$ one can simply take a regular grid (whose resolution is determined by the smoothness of the densities) and count the number of datapoints falling in each cell. Let $\widehat{p}_{\mathsf{X}}, \widehat{p}_{\mathsf{Y}}, \widehat{p}_{\mathsf{Z}}$ denote the empirical probability mass functions of the finitely supported distributions $\widehat{\mathbb{P}}_{\mathsf{X}}, \widehat{\mathbb{P}}_{\mathsf{Y}}, \widehat{\mathbb{P}}_{\mathsf{Z}}$. The test proceeds via the comparison

$$\|\widehat{p}_{\mathsf{X}} - \widehat{p}_{\mathsf{Z}}\|_2 \lesseqgtr \|\widehat{p}_{\mathsf{Y}} - \widehat{p}_{\mathsf{Z}}\|_2. \tag{18}$$

Squaring both sides and rearranging, we arrive at the form

$$\frac{1}{m} \sum_{i=1}^m (\widehat{p}_{\mathsf{Y}}(Z_i) - \widehat{p}_{\mathsf{X}}(Z_i)) \lesseqgtr \gamma, \tag{19}$$

where $\gamma = (\|\widehat{p}_{\mathsf{Y}}\|^2 - \|\widehat{p}_{\mathsf{X}}\|^2)/2$. As mentioned in the introduction, variants of this L^2-distance based test have been invented and re-invented multiple times for goodness-of-fit [12], [13] and two-sample testing [40], [41]. The exact statistic (18) with application to $\mathcal{P}_{\mathsf{Db}}$ has appeared in [14], [24], and Huang and Meyn [22] proposed an ingenious improvement restricting attention exclusively to bins whose counts are one of $(2,0), (1,1), (0,2)$ for the samples (X, Z) or (Y, Z). We attribute (18) to Ingster because his work on goodness-of-fit testing for smooth densities is the first occurence of the idea of comparing empirical L^2 norms, but we note that [14] and [13] arrive at this influential idea apparently independently.

We emphasize the following subtlety. Let us rewrite (18) as

$$\|\widehat{p}_{\mathsf{X}} - \widehat{p}_{\mathsf{Z}}\|_2^2 - \|\widehat{p}_{\mathsf{Y}} - \widehat{p}_{\mathsf{Z}}\|_2^2 \lessgtr 0\,. \tag{20}$$

As we shall see from our proofs, this difference results in an optimal test for the full range of possible values of n and m for $\mathcal{P}_{\mathsf{Db}}$. However, this does not mean that each term by itself is a meaningful estimate of the corresponding distance: rejecting the null by thresholding just $\|\widehat{p}_{\mathsf{X}} - \widehat{p}_{\mathsf{Z}}\|_2^2$ would not work. Indeed, the variance of $\|p_{\mathsf{X}} - \widehat{p}_{\mathsf{Z}}\|_2^2$ is so large that one needs $m \gtrsim n_{\mathsf{GoF}} \gg 1/\varepsilon^2$ observations to obtain a reliable estimate of $\|p_{\mathsf{X}} - p_{\mathsf{Z}}\|_2^2$. The "magic" of the L^2-difference test is that the two terms in (20) separately have high variance, and thus are not good estimators of their means, but their difference cancels the high-variance terms.

Remark 4. While testing (12), practitioners are usually interested in obtaining a p-value, rather than purely a decision whether to reject the null hypothesis. For this we propose the following scheme. Let $\sigma_1, \ldots, \sigma_P$ be i.i.d. uniformly random permutations on $n+m$ elements. Let $\widehat{T} = \|\widehat{p}_{\mathsf{X}} - \widehat{p}_{\mathsf{Z}}\|_2^2 - \|\widehat{p}_{\mathsf{Y}} - \widehat{p}_{\mathsf{Z}}\|_2^2$ be our statistic, and write $\widehat{T}_i$ for the statistic $\widehat{T}$ evaluated on the permuted dataset where $\{X_1, \ldots, X_n, Z_1, \ldots, Z_m\}$ are shuffled according to σ_i. Under the null the random variables $\widehat{T}, \widehat{T}_1, \ldots, \widehat{T}_P$ are exchangeable, thus reporting the empirical upper quantile of $\widehat{T}$ in this sample yields an estimate of the p-value. Studying the variance of this estimate or the power of the test that rejects when the estimated p-value is less than some threshold, is beyond the scope of this work.

2) Alternative Tests for LFHT: In this section we discuss a variety of alternative tests that may be considered for (12) instead of (20). These are included only to provide additional context for our problem, and the reader may safely skip it and proceed to our results in Section III. The approaches we consider are

(i) Scheffé's test,
(ii) Likelihood-free Neyman-Pearson test and
(iii) Huber's and Birgé's robust tests.

The tests $(i\text{-}ii)$ are based on the idea of using the simulated samples to learn a set or a function that separates $\mathbb{P}_{\mathsf{X}}$ from $\mathbb{P}_{\mathsf{Y}}$. The test (iii) and (20) use the simulated samples to obtain density estimates of $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}$ directly. All of them, however, are of the form

$$\sum_{i=1}^{m} s(Z_i) \lessgtr 0 \tag{21}$$

with only the function s varying.

Variants of *Scheffé's test* using machine-learning enabled classifiers are the subject of current research in two-sample testing [29]–[33] and are used in practice for LFI specifically in high energy physics, see also our discussion of the Higgs boson discovery in Section I. Thus, understanding the performance of Scheffé's test in the context of (12) is of great practical importance. Suppose that using the simulated samples we train a probabilistic classifier $C : \mathcal{X} \to [0,1]$ on the labeled data $\cup_{i=1}^{n}\{(X_i, 0), (Y_i, 1)\}$. The specific form of the classifier here is arbitrary and can be anything from logistic regression to a deep neural network. Given thresholds $t, \gamma \in [0,1]$ chosen to satisfy our risk appetite for type-I vs type-II errors, Scheffé's test proceeds via the comparison

$$\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{C(Z_i) \geq t\} \lessgtr \gamma. \tag{22}$$

We see that (22) is of the form (21) with $s(z) = (\mathbb{1}\{C(z) \geq t\} - \gamma)/m$. The follow-up work [11] studies the performance of Scheffé's test in great detail, finding that it is (near-)minimax optimal in all cases considered in this paper. It is found that the optimal classifier C must be trained *not* purely to minimize misclassification error, but rather must also keep the variance of its output small.

If the distributions $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}$ are fully known, then the likelihood-ratio test corresponds to

$$\sum_{i=1}^{m} s_{\mathsf{NP}}(Z_i) \lessgtr \gamma \text{ with } s_{\mathsf{NP}}(z) = \log\left(\frac{\mathrm{d}\mathbb{P}_{\mathsf{X}}}{\mathrm{d}\mathbb{P}_{\mathsf{Y}}}(z)\right), \tag{23}$$

where γ is again chosen to satisfy our type-I vs type-II error trade-off preferences. It is well known that the above procedure is optimal due to the Neyman-Pearson lemma. Recall that in our setting $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}$ are known only up to i.i.d. samples, and therefore it seems natural to try to estimate s_{NP} from samples. It is not hard to see that s_{NP} minimizes the population *cross-entropy/logistic loss*, that is

$$s_{\mathsf{NP}} = \arg\min_{s} \mathbb{E}[\ell(s(X), 1)] + \mathbb{E}[\ell(s(Y), 0)], \tag{24}$$

where $\ell(s, y) = \log(1 + e^s) - ys$ and X, Y are random draws from $\mathbb{P}_{\mathsf{X}}$ and $\mathbb{P}_{\mathsf{Y}}$ respectively. In practice, the majority of today's classifiers are obtained by running some form of gradient descent on the problem

$$\widehat{s} = \arg\min_{s \in \mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \Big(\ell(s(X_i), 1) + \ell(s(Y_i), 0)\Big), \tag{25}$$

where $\mathcal{G}$ is, say, a parametric class of neural networks. Given such an estimate $\widehat{s}$, we can replace the unknown s_{NP} in (23) by $\widehat{s}$ to obtain the *likelihood-free Neyman-Pearson test*. For recent work on this approach in LFI see for example [42]. Studying properties of this test is outside the scope of this paper.

The final approach is based on the idea of *robust testing*, first proposed by Huber [43], [44]. Huber's seminal result implies that if one has approximately correct distributions $\widehat{\mathbb{P}}_{\mathsf{X}}, \widehat{\mathbb{P}}_{\mathsf{Y}}$ satisfying

$$\begin{aligned} \max\left\{\mathsf{TV}(\widehat{\mathbb{P}}_{\mathsf{X}}, \mathbb{P}_{\mathsf{X}}), \mathsf{TV}(\widehat{\mathbb{P}}_{\mathsf{Y}}, \mathbb{P}_{\mathsf{Y}})\right\} \le \varepsilon/3 \\ \text{and } \mathsf{TV}(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}) \ge \varepsilon, \end{aligned} \tag{26}$$

then for some $c_1 < c_2$ the test $\sum_{i=1}^m s_{\mathsf{H}}(Z_i) \lessgtr 0$ where

$$s_{\mathsf{H}}(z) = \min\left\{\max\left\{c_1, \log\left(\frac{\mathrm{d}\widehat{\mathbb{P}}_{\mathsf{X}}}{\mathrm{d}\widehat{\mathbb{P}}_{\mathsf{Y}}}(z)\right)\right\}, c_2\right\} \tag{27}$$

has type-I and type-II error bounded by $\exp(-\Omega(m\varepsilon^2))$, and is in fact minimax optimal for all sample sizes analogously to the likelihood-ratio test in the case of binary hypothesis testing. From the above formula we can see that Scheffé's test can be interpreted as an approximation of the maximally robust Huber's test. Let $\widehat{\mathcal{L}}(z) = (\mathrm{d}\widehat{\mathbb{P}}_{\mathsf{Y}}/\mathrm{d}\widehat{\mathbb{P}}_{\mathsf{X}})(z)$ denote the likelihood-ratio of the estimates. The values of c_1, c_2 are given as the solution to

$$\varepsilon/3 = \mathbb{E}_{z\sim\widehat{\mathbb{P}}_{\mathsf{X}}}\left[\mathbb{1}\left\{\widehat{\mathcal{L}}(z) \le c_1\right\}\frac{c_1 - \widehat{\mathcal{L}}(z)}{1 + c_1}\right] \tag{28}$$

$$= \mathbb{E}_{z\sim\widehat{\mathbb{P}}_{\mathsf{Y}}}\left[\mathbb{1}\left\{\widehat{\mathcal{L}}(z) \ge c_2\right\}\frac{\widehat{\mathcal{L}}(z) - c_2}{1 + c_2}\right], \tag{29}$$

which can be easily approximated to high accuracy given samples from $\widehat{\mathbb{P}}_{\mathsf{X}}, \widehat{\mathbb{P}}_{\mathsf{Y}}$. This suggests both a theoretical construction, since $\widehat{\mathbb{P}}_{\mathsf{X}}, \widehat{\mathbb{P}}_{\mathsf{Y}}$ can be obtained with high probability from simulation samples via the general estimator of Yatracos [45], and a practical rule: instead of the possibly brittle likelihood-free Neyman-Pearson test (ii), one should try clamping the estimated log-likelihood ratio from above and below.

Similar results hold due to Birgé [46], [47] in the case when distance is measured by Hellinger divergence:

$$\begin{aligned} \max\left\{\mathsf{H}(\widehat{\mathbb{P}}_{\mathsf{X}}, \mathbb{P}_{\mathsf{X}}), \mathsf{H}(\widehat{\mathbb{P}}_{\mathsf{Y}}, \mathbb{P}_{\mathsf{Y}})\right\} \le \varepsilon/3 \\ \text{and}\mathsf{H}(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}) \ge \varepsilon. \end{aligned} \tag{30}$$

For ease of notation, let $\widehat{p}_{\mathsf{X}}, \widehat{p}_{\mathsf{Y}}$ denote the densities of $\widehat{\mathbb{P}}_{\mathsf{X}}, \widehat{\mathbb{P}}_{\mathsf{Y}}$ with respect to some base measure μ. Regarding $\sqrt{\widehat{p}_{\mathsf{X}}}$ and $\sqrt{\widehat{p}_{\mathsf{Y}}}$ as unit vectors of the Hilbert space $L^2(\mu)$, let $\gamma : [0,1] \to L^2(\mu)$ be the constant speed geodesic on the unit sphere of $L^2(\mu)$ with $\gamma(0) = \sqrt{\widehat{p}_{\mathsf{X}}}$ and $\gamma(1) = \sqrt{\widehat{p}_{\mathsf{Y}}}$. It is easily checked that each γ_t is positive, and Birgé showed that the test

$$\sum_{i=1}^m \log\left(\frac{\gamma_{1/3}^2}{\gamma_{2/3}^2}(Z_i)\right) \lessgtr 0 \tag{31}$$

has both type-I and type-II errors bounded by $\exp(-\Omega(m\varepsilon^2))$. For an exposition of this result see also [48, Theorem 7.1.2]

III. Results

In this section we describe our results on the sample complexity of likelihood-free hypothesis testing.

A. General Reductions

First, we give reductions that hold in great generality and show the relationship of our problem with other classical testing and estimation problems that were introduced in Section II-A. The result below holds for a generic class $\mathcal{P}$ of distributions and a generic measure of separation d, see also Remark 3.

Proposition 1. *Let $\mathcal{P}$ be a generic family of distributions and $\mathsf{d} : \mathcal{P}^2 \to \mathbb{R}$ be any function used to measure separation. There exists a universal constant $c > 0$ such that for $n, m \in \mathbb{N}$ the following implications hold.*

$$(n,m) \in \mathcal{R}_{\mathsf{LF}} \implies m \ge n_{\mathsf{HT}}, \tag{32}$$

$$(n,m) \in \mathcal{R}_{\mathsf{TS}} \implies n \wedge m \ge n_{\mathsf{GoF}} \tag{33}$$

$$(n,m) \in \mathcal{R}_{\mathsf{LF}} \implies cn \ge n_{\mathsf{GoF}}, \tag{34}$$

$$(n,m) \in \mathcal{R}_{\mathsf{TS}} \implies (n,m) \in \mathcal{R}_{\mathsf{LF}}, \tag{35}$$

$$m \ge n \text{ and } (n,m) \in \mathcal{R}_{\mathsf{LF}} \implies (cn, cm) \in \mathcal{R}_{\mathsf{TS}}, \tag{36}$$

where we omit the argument $(\varepsilon, \mathsf{d}, \mathcal{P})$ throughout for simplicity. In particular,

$$\mathbb{N}^2_{n\le m} \cap \mathcal{R}_{\mathsf{LF}} \asymp \mathbb{N}^2_{n\le m} \cap \mathcal{R}_{\mathsf{TS}}, \tag{37}$$

where $\mathbb{N}^2_{n\le m} = \{(n,m) \in \mathbb{N}^2 : n \le m\}$.

Proof: In what follows, let $\Psi_{\mathsf{LF}}, \Psi_{\mathsf{TS}}$ be minimax optimal tests for (12) and (10) respectively. Throughout the proof we omit the arguments $(\varepsilon, \mathsf{d}, \mathcal{P})$ for notational simplicty.

We start by *reducing hypothesis testing to* (12). Suppose $(n,m) \in \mathcal{R}_{\mathsf{LF}}$. Let $\mathbb{P}_0, \mathbb{P}_1 \in \mathcal{P}$ be given with $\mathsf{d}(\mathbb{P}_0, \mathbb{P}_1) \ge \varepsilon$ and suppose Z is an i.i.d. sample with m observations. We wish to test the hypothesis $H_0 : Z_i \sim$

$\mathbb{P}_0$ against $H_1 : Z_i \sim \mathbb{P}_1$. To this end generate n i.i.d. observations X, Y from $\mathbb{P}_0, \mathbb{P}_1$ respectively, and simply output $\Psi_{\mathsf{LF}}(X, Y, Z)$. This shows that if $(n, m) \in \mathcal{R}_{\mathsf{LF}}$ then $m \geq n_{\mathsf{HT}}$ and concludes the proof of (32).

Next, we *reduce goodness-of-fit testing to two-sample testing*. Suppose $(n, m) \in \mathcal{R}_{\mathsf{TS}}$. Then obviously $(n \wedge m, \infty) \in \mathcal{R}_{\mathsf{TS}}$. However, two-sample testing with sample sizes $n \wedge m, \infty$ is equivalent to goodness-of-fit testing with a sample size of $n \wedge m$. Therefore, $n \wedge m \geq n_{\mathsf{GoF}}$ must hold, concluding the proof of (33).

Next we *reduce goodness-of-fit testing to* (12). Suppose $(n, m) \in \mathcal{R}_{\mathsf{LF}}$ with $m \leq n$. Let a distribution $\mathbb{P}_0 \in \mathcal{P}$ be given as well as an i.i.d. sample X of size cn with unknown distribution $\mathbb{P}_{\mathsf{X}}$, where $c \in \mathbb{N}$ is a large integer. We want to test $H_0 : \mathbb{P}_{\mathsf{X}} = \mathbb{P}_0$ against $H_1 : \mathbb{P}_{\mathsf{X}} \in \mathcal{P}, \mathsf{d}(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_0) \geq \varepsilon$. Generate $c \times 2$ i.i.d. samples $Y^{(i)}, Z^{(i)}$ for $i = 1, \ldots, c$ of size n, m respectively, all from $\mathbb{P}_0$. Split the sample X into c batches $X^{(i)}, i = 1, \ldots, c$ of size n each and form the variables

$$A_i = \Psi_{\mathsf{LF}}(X^{(i)}, Y^{(i)}, Z^{(i)}) - \Psi_{\mathsf{LF}}(X^{(i)}, Y^{(i)}, X^{(i+1)}_{1:m}) \tag{38}$$

for $i = 1, 3, \ldots, 2\lfloor c/2 \rfloor - 1$, where $X^{(i)}_{1:m}$ denotes the first m observations in the batch $X^{(i)}$. Note that the A_i are i.i.d. and bounded random variables. Under the null hypothesis we have $\mathbb{E}A_i = 0$, while under the alternative they have mean $\mathbb{E}A_i \geq 1/3$ (since Ψ_{LF} is a successful tester in the sense of (6)). Therefore, a constant number $c/2$ observations suffice to decide whether $\mathbb{P}_{\mathsf{X}} = \mathbb{P}_0$ or not. In particular, $cn \geq n_{\mathsf{GoF}}$ which concludes the proof of (34) for the case $m \leq n$. The case $n \leq m$ follows from (36) and (33).

Next we *reduce* (12) *to two-sample testing*. Suppose $(n, m) \in \mathcal{R}_{\mathsf{TS}}$. Let three samples X, Y, Z be given, of sizes a, a, b from the unknown distributions $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}, \mathbb{P}_{\mathsf{Z}}$ respectively, where $\{a, b\} = \{n, m\}$. We want to test the hypothesis $H_0 : \mathbb{P}_{\mathsf{X}} = \mathbb{P}_{\mathsf{Z}}$ against $H_1 : \mathbb{P}_{\mathsf{Y}} = \mathbb{P}_{\mathsf{Z}}$, where $\mathsf{d}(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}) \geq \varepsilon$ under both. Then, the test

$$\widetilde{\Psi}_{\mathsf{LF}}(X, Y, Z) \triangleq \Psi_{\mathsf{TS}}(X, Z) \tag{39}$$

shows that $(n, m), (m, n) \in \mathcal{R}_{\mathsf{LF}}$ and concludes the proof of (35).

Next we *reduce two-sample testing to* (12). Suppose $(n, m) \in \mathcal{R}_{\mathsf{LF}}$ where $m \geq n$. Let two samples X, Y be given, from the unknown distributions $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}} \in \mathcal{P}$ and of sample size cn, cm respectively, where $c \in \mathbb{N}$ is a large integer. We wish to test the hypothesis $H_0 : \mathbb{P}_{\mathsf{X}} = \mathbb{P}_{\mathsf{Y}}$ against $H_1 : \mathsf{d}(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}) \geq \varepsilon$. Split the samples X, Y into $2 \times c$ batches $X^{(i)}, Y^{(i)}, i = 1, \ldots, c$ of sizes n, m respectively, and form the variables

$$A_i = \Psi_{\mathsf{LF}}(X^{(i)}, Y^{(i)}_{1:n}, Y^{(i+1)}) - \Psi_{\mathsf{LF}}(Y^{(i)}_{1:n}, X^{(i)}, Y^{(i+1)}) \tag{40}$$

for $i = 1, 3, \ldots, 2\lfloor c/2 \rfloor - 1$, where $Y^{(i)}_{1:n}$ denotes the first n observations in the batch $Y^{(i)}$. The variables A_i are i.i.d. and bounded. Under the null hypothesis we have $\mathbb{E}A_i = 0$ while under the alternative $\mathbb{E}A_i \geq 1/3$ holds. Therefore a constant number $c/2$ observations suffice to decide whether $\mathbb{P}_{\mathsf{X}} = \mathbb{P}_{\mathsf{Y}}$ or not. In particular, $(cn, cm) \in \mathcal{R}_{\mathsf{TS}}$ which concludes the proof of (36).

Finally, we show the *equivalence between two-sample testing and* (12). Equation (37) follows immediately from (36) and (35). ∎

Equation (37) tells us that the problems of likelihood-free hypothesis testing and two-sample testing are equivalent, *but only for* $m \geq n$, that is, when we have more real data than simulated data. We will see in the next section, and on Fig. 1 visually, that this distinction is necessary.

B. Sample Complexity of Likelihood-Free Hypothesis Testing

In this section we present our results on the sample complexity of (12) for the specific classes $\mathcal{P}$ that were introduced in Section II-A, with separation measured by TV. In all results below the parameters β, s, d, C are regarded as constants, we only care about the dependence on the separation ε and the alphabet size k (in the case of $\mathcal{P}_{\mathsf{D}}, \mathcal{P}_{\mathsf{Db}}$). Where convenient we omit the arguments of $n_{\mathsf{GoF}}, n_{\mathsf{TS}}, \mathcal{R}_{\mathsf{TS}}, n_{\mathsf{Est}}, \mathcal{R}_{\mathsf{LF}}$ to ease notation, whose value should be clear from the context.

Theorem 1. *Under* TV*-separation, for each choice* $\mathcal{P} \in \{\mathcal{P}_{\mathsf{H}}, \mathcal{P}_{\mathsf{G}}, \mathcal{P}_{\mathsf{Db}}\}$*, we have*

$$\mathcal{R}_{\mathsf{LF}} \asymp \left\{ (n, m) : \begin{array}{c} m \geq 1/\varepsilon^2 \\ \& \quad n \geq n_{\mathsf{GoF}} \\ \& \quad mn \geq n^2_{\mathsf{GoF}} \end{array} \right\}, \tag{41}$$

where the implied constants do not depend on k *(in the case of* $\mathcal{P}_{\mathsf{Db}}$*) or* ε*.*

For each class $\mathcal{P}$ in Theorem 1, the entire region $\mathcal{R}_{\mathsf{LF}}$ (within universal constant) is attained by a suitable modification of Ingster's L^2-distance test from Section II-C1. The region $\mathcal{R}_{\mathsf{LF}}$ is visualized on Fig. 1 on a log-log scale, with each corner point $\{\mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D}\}$ having a special interpretation, as per the reductions presented

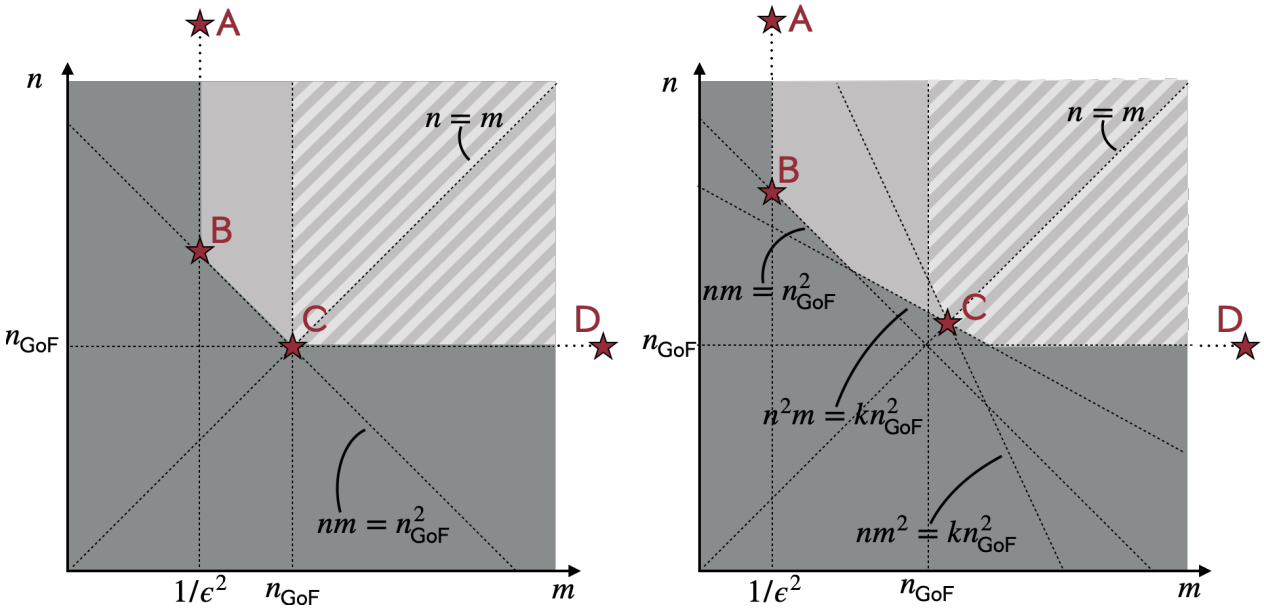

Fig. 1. Light and dark gray show $\mathcal{R}_{\mathsf{LF}}$ and its complement resp. on log scale; the striped region depicts $\mathcal{R}_{\mathsf{TS}} \subsetneq \mathcal{R}_{\mathsf{LF}}$. Left plot is valid for $\mathcal{P} \in \{\mathcal{P}_{\mathsf{H}}, \mathcal{P}_{\mathsf{G}}, \mathcal{P}_{\mathsf{Db}}\}$ for all settings of ε, k. For $\mathcal{P}_{\mathsf{D}}$ the left plot applies when $k \lesssim \varepsilon^{-4}$ and the right plot otherwise.

in Proposition 1. The point A corresponds to binary hypothesis testing and D can be reduced to goodness-of-fit testing. Similarly, B and C can be reduced to the well-known problems of estimation followed by robust hypothesis testing and two-sample testing respectively. In other words, (12) allows us to naturally interpolate between multiple statistical problems. Finally, we make an interesting observation: since the product of n and m remains constant on the line segment $[\mathsf{B}, \mathsf{C}]$ on the left plot of Fig. 1, it follows that

$$n_{\mathsf{Est}}(\varepsilon, \mathcal{P}) \asymp n^2_{\mathsf{GoF}}(\varepsilon, \mathcal{P})\, \varepsilon^2 \tag{42}$$

for each class $\mathcal{P}$ treated in Theorem 1. This relation between the sample complexity of estimation and goodness-of-fit testing has not been observed before to our knowledge, and understanding the scope of validity of this relationship is an exciting future direction.[5]

Turning to our results on $\mathcal{P}_{\mathsf{D}}$ the picture is less straightforward. As first identified in [50] and fully resolved in [51], the sample complexity of two-sample testing undergoes a phase transition when $k \gtrsim 1/\varepsilon^4$. This phase transition appears also in likelihood-free hypothesis testing.

Theorem 2. *Let* $\alpha = \max\left\{1, \min\left\{\frac{k}{n}, \frac{k}{m}\right\}\right\}$. *Then* $\mathcal{R} - \mathsf{LF}(\epsilon, \mathcal{P}_{\mathsf{D}}(k))$ *is proportional, up to a logarithmic factor in the alphabet size* k*, to the set*

$$\left\{ (n, m) \; : \; \begin{array}{c} m \geq 1/\varepsilon^2 \\ \& \; n \geq n_{\mathsf{GoF}}(\varepsilon, \mathcal{P}_{\mathsf{D}}(k)) \cdot \sqrt{\alpha} \\ \& \; mn \geq n_{\mathsf{GoF}}(\varepsilon, \mathcal{P}_{\mathsf{D}}(k))^2 \cdot \alpha \end{array} \right\}. \tag{43}$$

The $\log k$ factor in our analysis originates from a union bound, and it is possible that it may be removed. It follows from follow up work [11] and past results on two-sample testing [38] that the $\log(k)$ factor can be removed in all regimes, thus fully characterizing the sample complexity of (12), but using a different test from ours.

Table I summarizes previously known tight results for the values of $n_{\mathsf{GoF}}, n_{\mathsf{TS}}, \mathcal{R}_{\mathsf{TS}}$ and n_{Est}. The fact that $n_{\mathsf{HT}} = \Theta(1/\varepsilon^2)$ for reasonable classes is classical, see Lemma 1. The study of goodness-of-fit testing within a minimax framework was pioneered by Ingster [12], [34] for $\mathcal{P}_{\mathsf{H}}, \mathcal{P}_{\mathsf{G}}$, and independently studied by the computer science community [13], [52] for $\mathcal{P}_{\mathsf{D}}, \mathcal{P}_{\mathsf{Db}}$ under the name *identity testing*. Two-sample testing (a.k.a. *closeness testing*) was solved in [51] for $\mathcal{P}_{\mathsf{D}}$ (with the optimal result for $\mathcal{P}_{\mathsf{Db}}$ implicit) and [12], [41], [53] consider $\mathcal{P}_{\mathsf{H}}$.

[5] *Added in print:* for example in [49] it is demonstrated that for the Gaussian sequence model (see definition (ii) in Section II-B) with the Sobolev ellipsoid replaced by the set $\Theta = \{\theta \in \ell^2 : \sum_{i=1}^{\infty} i|\theta_i| \leq 1\}$, it holds that $n_{\mathsf{Est}} \ll n^2_{\mathsf{GoF}}/\varepsilon^2$.

TABLE I
PRIOR RESULTS ON TESTING AND ESTIMATION

	n_{HT}	n_{GoF}	$\mathcal{R}_{\mathsf{TS}}$	n_{Est}
$\mathcal{P}_{\mathsf{G}}$	$1/\varepsilon^2$	$1/\varepsilon^{(2s+1/2)/s}$	$n \wedge m \geq n_{\mathsf{GoF}}$	$\varepsilon^2\, n^2_{\mathsf{GoF}}$
$\mathcal{P}_{\mathsf{H}}$	$1/\varepsilon^2$	$1/\varepsilon^{(2\beta+d/2)/\beta}$	$n \wedge m \geq n_{\mathsf{GoF}}$	$\varepsilon^2\, n^2_{\mathsf{GoF}}$
$\mathcal{P}_{\mathsf{Db}}$	$1/\varepsilon^2$	$\sqrt{k}/\varepsilon^2$	$n \wedge m \geq n_{\mathsf{GoF}}$	$\varepsilon^2\, n^2_{\mathsf{GoF}}$
$\mathcal{P}_{\mathsf{D}}$	$1/\varepsilon^2$	$\sqrt{k}/\varepsilon^2$	$n \vee m \geq \frac{\sqrt{k}}{\varepsilon^2} \vee \frac{k^{2/3}}{\varepsilon^{4/3}} \asymp n_{\mathsf{TS}},\ n \wedge m \geq n_{\mathsf{GoF}}\sqrt{\alpha}$	$\varepsilon^2\, n^2_{\mathsf{GoF}}$

The study of the rate of estimation n_{Est} is older, see [39], [48], [54], [55] and references for $\mathcal{P}_{\mathsf{H}}, \mathcal{P}_{\mathsf{G}}$ and [56] for $\mathcal{P}_{\mathsf{D}}, \mathcal{P}_{\mathsf{Db}}$.

C. L^2-Robust Likelihood-Free Hypothesis Testing

Even before seeing Theorems 1 and 2 one might guess that estimation in TV followed by a robust hypothesis test should work whenever $m \gtrsim 1/\varepsilon^2$ and $n \geq n_{\mathsf{Est}}(c\varepsilon)$ for a small enough constant c. This strategy does indeed work, which can be deduced from the work of Huber and Birgé [43], [47] for total variation and Hellinger separation respectively, see also Section II-C for a brief discussion of these robust tests. In other words, we have the informal theorem that if separation is measured by TV or H, then

$$\{n \geq n_{\mathsf{Est}} \text{ and } m \geq n_{\mathsf{HT}}\} \implies (cn, cm) \in \mathcal{R}_{\mathsf{LF}}. \quad (44)$$

In the case of total variation separation, in fact an even simpler approach succeeds: if $\widehat{p}_{\mathsf{X}}$ and $\widehat{p}_{\mathsf{Y}}$ are minimax optimal density estimators with respect to TV, then Scheffé's test using the classifier $C(x) = \mathbb{1}\{\widehat{p}_{\mathsf{Y}}(x) \geq \widehat{p}_{\mathsf{X}}(x)\}$ can be shown to achieve the optimal sample complexity by Chebyshev's inequality.

The upshot of these observations is that they provide a solution to (12) that is robust to model misspecification, specifically at the corner point B on Fig. 1. This naturally leads us to the question of robust likelihood-free hypothesis testing: can we construct robust tests for the full m vs n trade-off?

As before, suppose we observe samples X, Y, Z of size n, n, m from distributions belonging to the class $\mathcal{P}$ with densities f, g, h with respect to some base measure μ. Given any $u \in \mathcal{P}$, let $\mathsf{B}_u(\varepsilon, \mathcal{P}) \subseteq \mathcal{P}$ denote a region around u against which we wish to be robust. Recall the notation $\mathcal{P}_\varepsilon = \{(\mathbb{Q}_0, \mathbb{Q}_1) \in \mathcal{P}^2 : \mathsf{TV}(\mathbb{Q}_0, \mathbb{Q}_1) \geq \varepsilon\}$ from Definition 5. We compare the hypotheses

$$\begin{aligned} H_0 &: h \in \mathsf{B}_f(\varepsilon, \mathcal{P}), (f, g) \in \mathcal{P}_\varepsilon \\ \text{versus } H_1 &: h \in \mathsf{B}_g(\varepsilon, \mathcal{P}), (f, g) \in \mathcal{P}_\varepsilon, \end{aligned} \quad (45)$$

and write $\mathcal{R}_{\mathsf{rLF}}(\varepsilon, \mathcal{P}, \mathsf{B}.)$ for the region of (n, m)-values for which (45) can be performed successfully, defined analogously to $\mathcal{R}_{\mathsf{LF}}(\varepsilon, \mathcal{P})$. Note that $\mathcal{R}_{\mathsf{rLF}} \subseteq \mathcal{R}_{\mathsf{LF}}$ provided $u \in \mathsf{B}_u$ for all $u \in \mathcal{P}$, that is, the range of sample sizes n, m for which robustly testing (12) is possible ought to be a subset of $\mathcal{R}_{\mathsf{LF}}$.

Theorem 3. *Theorems 1 and 2 remain true if we replace $\mathcal{R}_{\mathsf{LF}}(\varepsilon, \mathcal{P})$ by $\mathcal{R}_{\mathsf{rLF}}(\varepsilon, \mathcal{P}, \mathsf{B}.)$ for the following choices:*

(i) for $\mathcal{P}_{\mathsf{H}}(\beta, d, C)$ and $\mathsf{B}_u = \{v \in \mathcal{P}_{\mathsf{H}}(\beta, d, C) : \|u - v\|_2 \leq c\varepsilon\}$ for a constant $c > 0$ independent of ε,

(ii) for $\mathcal{P}_{\mathsf{G}}(s, C)$ and $\mathsf{B}_{\mu_\theta} = \{\mu_{\theta'} : \theta' \in \mathcal{E}(s, C), \|\theta - \theta'\|_2 \leq \varepsilon/4\}$,

(iii) for $\mathcal{P}_{\mathsf{Db}}(k, C)$ and $\mathsf{B}_u = \{v : \|u - v\|_2 \leq \varepsilon/(2\sqrt{k})\}$,

(iv) for $\mathcal{P}_{\mathsf{D}}(k)$ and $\mathsf{B}_u = \{v : \|u - v\|_2 \leq c\varepsilon/\sqrt{k}, \|v/u\|_\infty \leq c\}$ for a constant $c > 0$ independent of k and ε.

D. Beyond Total Variation

Recall from Remark 3 the notation $n_{\mathsf{GoF}}(\varepsilon, \mathsf{d}, \mathcal{P})$ etc. that is applicable when separation is measured with respect to a general measure of discrepancy d instead of TV. In recent work [57, Theorem 1] and [58, Lemma 3.6] it is shown that any test that first quantizes the data by a map $\Phi : \mathcal{X} \to \{1, 2, \ldots, M\}$ for some $M \geq 2$ must decrease the Hellinger distance between the two hypotheses by a $\log$ factor in the worst case. This implies that for every class $\mathcal{P}$ rich enough to contain such worst case examples, a quantizing test, such as Scheffé's, can hope to achieve $m \asymp \log(1/\varepsilon)/\varepsilon^2$ at best, as opposed to the optimal $m \asymp 1/\varepsilon^2$. Thus, if separation is assumed with respect to Hellinger distance, Scheffé's test should be avoided. This example shows that the choice of d can have surprising effects on the performance of specific tests that would be optimal under other circumstances. Understanding the sample complexity of (12) for d other than TV might lead to new algorithms and insights.

This motivates us to pose the question: does a trade-off analogous to that identified in Theorem 1 hold for

TABLE II
PRIOR RESULTS ON TESTING AND ESTIMATION FOR $\mathsf{d} = \mathsf{H}$.

	n_{HT}	n_{GoF}	n_{TS}	n_{Est}
$\mathcal{P}_{\mathsf{D}}$	$1/\varepsilon^2$	$\sqrt{k}/\varepsilon^2$	$k^{2/3}/\varepsilon^{8/3} \wedge k^{3/4}/\varepsilon^2$	$n_{\mathsf{GoF}}^2\,\varepsilon^2$
$\mathcal{P}_{\mathsf{H}}$	$1/\varepsilon^2$	?	?	$1/\varepsilon^{2(\beta+d)/\beta}$

other choices of d, and H in particular? In the case of $\mathcal{P}_{\mathsf{G}}$ we obtain a simple, almost vacuous answer. From Lemma 2 it follows immediately that the results of Table I and Theorem 1 continue to hold for $\mathcal{P}_{\mathsf{G}}$ for any of $\mathsf{d} \in \{\mathsf{H}, \sqrt{\mathsf{KL}}, \sqrt{\chi^2}\}$, to name a few.

Lemma 2. *Let $C > 0$ be a constant. For any $\theta \in \ell^2$ with $\|\theta\|_2 \le C$*

$$\begin{aligned}\mathsf{TV}(\mu_\theta, \mu_0) \asymp \mathsf{H}(\mu_\theta, \mu_0) &\asymp \sqrt{\mathsf{KL}(\mu_\theta \| \mu_0)} \\ &\asymp \sqrt{\chi^2(\mu_\theta \| \mu_0)} \asymp \|\theta\|_2,\end{aligned} \tag{46}$$

where $\mu_\theta \triangleq \otimes_{i=1}^{\infty} \mathcal{N}(\theta_i, 1)$ and the implied constant depends on C.

The case of $\mathcal{P}_{\mathsf{D}}$ is more intricate. Substantial recent progress [25], [56], [59], [60] has been made, where among others, the complexities $n_{\mathsf{GoF}}, n_{\mathsf{TS}}, n_{\mathsf{Est}}$ for Hellinger separation are identified, see Table II.

Since our algorithm for (12) is $\|\cdot\|_2$-based, we could immediately derive achievability bounds for $\mathcal{R}_{\mathsf{LF}}(\varepsilon, \mathsf{H}, \mathcal{P}_{\mathsf{D}})$ via the inequality $\|\cdot\|_2 \ge \mathsf{H}^2/\sqrt{k}$, however such a naive technique yields suboptimal results, and thus we omit it. Studying (12) under Hellinger separation for $\mathcal{P}_{\mathsf{D}}$ and $\mathcal{P}_{\mathsf{Db}}$ is beyond the scope of this work.

Finally, we turn to $\mathcal{P}_{\mathsf{H}}$. Due to the nature of our proofs, the results of Theorem 1 easily generalize to $\mathsf{d} = \|\cdot\|_p$ for any $p \in [1, 2]$. The simple reason for this is that (i) our algorithm is $\|\cdot\|_2$-based and $\|\cdot\|_2 \ge \|\cdot\|_p$ by Jensen's inequality and (ii) the lower bound construction involves perturbations near 1, where all said norms are equivalent. In the important case $\mathsf{d} = \mathsf{H}$ the estimation rate $n_{\mathsf{Est}}(\varepsilon, \mathsf{H}, \mathcal{P}_{\mathsf{H}}) \asymp 1/\varepsilon^{2(\beta+d)/\beta}$ was obtained by Birgé [61], our contribution here is the study of n_{GoF}.

Theorem 4. *For any $\beta > 0, C > 1$ and $d \ge 1$ there exists a constant $c > 0$ such that*

$$n_{\mathsf{GoF}}(\varepsilon, \mathsf{H}, \mathcal{P}(\beta, d, C)) \ge c/\varepsilon^{2(\beta+d/2)/\beta}. \tag{47}$$

If in addition we assume that $\beta \in (0, 1]$, c can be chosen such that

$$c n_{\mathsf{GoF}}(\varepsilon, \mathsf{H}, \mathcal{P}) \le 1/\varepsilon^{2(\beta+d/2)/\beta}. \tag{48}$$

In particular, $n_{\mathsf{Est}} \asymp n_{\mathsf{GoF}}^2\,\varepsilon^2$.

IV. SKETCH PROOF OF MAIN RESULTS

In this section we briefly sketch the proofs of the main results of the paper.

A. Upper Bounds for Theorems 1 to 4

1) Bounded Discrete Distributions: Consider first the case when $\mathbb{P}_{\mathsf{X}}$ and $\mathbb{P}_{\mathsf{Y}}$ belong to the class $\mathcal{P}_{\mathsf{Db}}$, that is, they are supported on the discrete set $\{1, 2, \ldots, k\}$ and bounded by the uniform distribution. Let $\widehat{p}_{\mathsf{X}}, \widehat{p}_{\mathsf{Y}}, \widehat{p}_{\mathsf{Z}}$ denote empirical probability mass functions based on the samples X, Y, Z of size n, n, m from $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}, \mathbb{P}_{\mathsf{Z}}$ respectively. Define the test statistic

$$T_{\mathsf{LF}} = \|\widehat{p}_{\mathsf{X}} - \widehat{p}_{\mathsf{Z}}\|_2^2 - \|\widehat{p}_{\mathsf{Y}} - \widehat{p}_{\mathsf{Z}}\|_2^2 \tag{49}$$

and the corresponding test $\psi(X, Y, Z) = \mathbb{1}\{T_{\mathsf{LF}} \ge 0\}$. The proof of Theorems 1 and 2 hinge on the precise calculation of the mean and variance of T_{LF}. Due to symmetry it is enough to compute these under the null. The proof of the upper bound is then completed via Chebyshev's inequality: if n, m are such that $(\mathbb{E}T_{\mathsf{LF}})^2 \gtrsim \mathrm{var}(T_{\mathsf{LF}})$ for large enough implied constant on the right then ψ tests (12) successfully in the sense of (6).

Proposition 2 (informal)**.** *Suppose $\|p_{\mathsf{X}} + p_{\mathsf{Y}} + p_{\mathsf{Z}}\|_\infty \le C_\infty/k$. Then ψ successfully tests* (12) *if*

$$\underbrace{\frac{\varepsilon^4}{k^2}}_{(\mathbb{E}T_{\mathsf{LF}})^2} \gtrsim \underbrace{\frac{C_\infty \varepsilon^2}{k^2}\left(\frac{1}{n} + \frac{1}{m}\right) + \frac{C_\infty}{k}\left(\frac{1}{n^2} + \frac{1}{nm}\right)}_{\mathrm{var}(T_{\mathsf{LF}})}. \tag{50}$$

From (50) one can immediately see where each constraint in the region $\mathcal{R}_{\mathsf{LF}}(\varepsilon, \mathcal{P}_{\mathsf{Db}}(k, C))$ in Theorem 1 emerges. The first two terms in the variance require that both m and n be larger than $\Omega(1/\varepsilon^2)$. The $1/n^2$ term in the variance requires that n be at least $\Omega(\sqrt{k}/\varepsilon^2) \asymp n_{\mathsf{GoF}}$, and the $1/(nm)$ term requires that the product nm be at least $\Omega(n_{\mathsf{GoF}}^2)$.

2) Smooth Densities: Next we describe how Proposition 2 can be applied to the class $\mathcal{P}_{\mathsf{H}}$ of smooth densities. Divide $[0, 1]^d$ into into κ^d regular grid cells for some $\kappa \in \mathbb{N}$. Discretize the three samples X, Y, Z over this grid and simply apply the optimal test for $\mathcal{P}_{\mathsf{Db}}$, observing the crucial fact that this discretization belongs to $\mathcal{P}_{\mathsf{Db}}$. The following lemma, originally due to Ingster [12] controls the approximation error of the discretization.

Lemma 3 ([41, Lemma 7.2])**.** *Let P_κ denote the L^2 projection onto the space of functions constant on each*

grid cell. For any $\beta > 0, C > 1$ and $d \geq 1$ there exist constants $c, c' > 0$ such that for any $f, g \in \mathcal{P}_{\mathsf{H}}(\beta, d, C)$ the following holds:

$$\|f - g\|_2 \geq \|P_\kappa(f - g)\|_2 \geq c\|f - g\|_2 - c'\kappa^{-\beta}. \quad (51)$$

Based on Lemma 3 we set $\kappa \asymp \varepsilon^{-1/\beta}$. This resolution is chosen to ensure that the discrete approximation to any β-smooth density is sufficiently accurate, that is, approximate ε-separation is maintained even after discretization. We see now that our problem is reduced entirely to testing over $\mathcal{P}_{\mathsf{Db}}$, so we may apply Proposition 2 with $k = \kappa^d \asymp \varepsilon^{-d/\beta}$, which yields the minimax optimal rates from Theorems 1 and 3.

Our proof of the achieavability direction in Theorem 4 follows similarly by reduction to goodness-of-fit testing for discrete distributions [60] under Hellinger separation, where it is known that $n_{\mathsf{GoF}}(\varepsilon, \mathsf{H}, \mathcal{P}_{\mathsf{D}}) \asymp \sqrt{k}/\varepsilon^2$. The key step is to prove a result similar to Lemma 3 but for H instead of $\|\cdot\|_2$.

Proposition 3. *For any $\beta \in (0, 1]$, $C > 1$ and $d \geq 1$ there exists a constant $c > 0$ such that*

$$c\mathsf{H}(f, g) \leq \mathsf{H}(P_\kappa f, P_\kappa g) \leq \mathsf{H}(f, g) \quad (52)$$

holds for any $f, g \in \mathcal{P}_{\mathsf{H}}(\beta, d, C)$, provided we set $\kappa = (c\varepsilon)^{-2/\beta}$.

3) Gaussian Sequence Model: Let us briefly discuss the Gaussian sequence class $\mathcal{P}_{\mathsf{G}}(s, C)$. Here our approach is not to discretize the distributions, but conceptually the test is very similar to the cases we've already covered. Let us write $\mathbb{P}_{\mathsf{X}} = \mu_{\theta_{\mathsf{X}}}$ and define $\theta_{\mathsf{Y}}, \theta_{\mathsf{Z}}$ analogously. For a given cutoff r, we simply reject the null if

$$T_{\mathsf{LF},\mathsf{G}} \triangleq \sum_{i=1}^{r} \left\{ \left(\widehat{\theta}_{\mathsf{X},i} - \widehat{\theta}_{\mathsf{Z},i}\right)^2 - \left(\widehat{\theta}_{\mathsf{Y},i} - \widehat{\theta}_{\mathsf{Z},i}\right)^2 \right\} \geq 0, \quad (53)$$

where $\widehat{\theta}_{\mathsf{X},i} = \frac{1}{n}\sum_{j=1}^{n} X_{ji}$ and $\widehat{\theta}_{\mathsf{Y}}, \widehat{\theta}_{\mathsf{Z}}$ are defined analogously. Once again, a precise calculation of the mean and the variance of the sum above, yields the following result.

Proposition 4 (informal)**.** *Set $r \asymp \varepsilon^{-1/s}$. The test* (53) *succeeds if*

$$\underbrace{\varepsilon^4}_{(\mathbb{E}T_{\mathsf{LF},\mathsf{G}})^2} \gtrsim \underbrace{\varepsilon^2\left(\frac{1}{n} + \frac{1}{m}\right) + \varepsilon^{-1/s}\left(\frac{1}{n^2} + \frac{1}{nm}\right)}_{\mathrm{var}(T_{\mathsf{LF},\mathsf{G}})}. \quad (54)$$

Similarly to (50), we can again read of the constraints that define the region $\mathcal{R}_{\mathsf{LF}}(\varepsilon, \mathcal{P}_{\mathsf{G}}(s, C))$ from (54). The first and second terms in the variance ensure that $n, m = \Omega(1/\varepsilon^2)$ and $n^2, mn = \Omega(n_{\mathsf{GoF}}^2) = \Omega(\varepsilon^{-(4s+1)/s})$ respectively.

4) General Discrete Distributions: Finally, we comment on $\mathcal{P}_{\mathsf{D}}$. Here we can no longer assume that $C_\infty = \mathcal{O}(1)$ in Proposition 2, in fact $C_\infty = \Omega(k)$ is possible. We get around this by utilizing the reduction based approach of [25], [35]. We take the first half of the data and compute

$$\begin{aligned} B_i = 1 &+ \#\left\{ j \leq \frac{\min\{k, n\}}{2} : X_j = i \right\} \\ &+ \#\left\{ j \leq \frac{\min\{k, n\}}{2} : Y_j = i \right\} \\ &+ \#\left\{ j \leq \frac{\min\{k, m\}}{2} : Z_j = i \right\} \end{aligned} \quad (55)$$

for each $i \in [k]$. Then, we divide the i'th support element into B_i bins, uniformly. This transformation preserves pairwise total variation, but reduces the ℓ^∞-norms of $p_{\mathsf{X}}, p_{\mathsf{Y}}, p_{\mathsf{Z}}$ with high probability, to order $1/(k \wedge (n \vee m))$, after an additional step that we omit here. We can then perform the usual test with these new "flattened" distributions, using the untouched half of the data.

It is insightful to interpret the "flattening" procedure followed by L^2-distance comparison as a one-step procedure that simply compares a different divergence of the empirical measures. Intuitively, in contrast to the regular classes, one needs to mitigate the effect of potentially massive differences in the empirical counts on bins $i \in [k]$ where both $p_{\mathsf{X}}(i)$ and $p_{\mathsf{Y}}(i)$ are large but their difference $|p_{\mathsf{X}}(i) - p_{\mathsf{Y}}(i)|$ is moderate. Let LC_λ be the "weighted Le-Cam divergence" which we define as $\mathsf{LC}_\lambda(p\|q) = \sum_i (p_i - q_i)^2/(p_i + \lambda q_i)$ for two probability mass functions p, q. One may interpret the two step procedure (flattening followed by comparing L^2 distances) as approximately comparing empirical weighted Le-Cam divergences. Performing the test in two steps is a proof device, and we expect the test that directly compares, say, the Le-Cam divergence of the empirical probability mass functions to have the same minimax optimal sample complexity. Such a one-shot approach is used for example in the paper [51] for two-sample testing. While Ingster [12] only considers goodness-of-fit testing to the uniform distribution, his notation also suggests the idea of normalizing by the bin mass under the null.

B. Lower Bounds for Theorems 1 to 4

The reductions given in Proposition 1 immediately yields a number of tight lower bounds on n and m. Namely, (32) gives $m \gtrsim 1/\varepsilon^2$ and (34) gives $n \gtrsim n_{\mathsf{GoF}}$. Obtaining the lower bound on the product term mn proves more challenging. First we introduce the well known information theoretic tools we use to prove our minimax lower bounds.

Suppose that we have two (potentially composite) hypotheses H_0, H_1 that we test against each other. Our strategy relies on the method of two fuzzy hypotheses [39], which is a generalization of Le-Cam's two point method. Write $\mathcal{M}(\mathcal{X})$ for the set of probability measures on the set $\mathcal{X}$.

Lemma 4. *Take two hypotheses $H_i \subseteq \mathcal{M}(\mathcal{X})$ and random $P_i \in \mathcal{M}(\mathcal{X})$. Then*

$$\begin{aligned} 2 \inf_{\psi} \max_{i=0,1} \sup_{P \in H_i} & P(\psi \neq i) \\ & \geq 1 - \mathsf{TV}(\mathbb{E}P_0, \mathbb{E}P_1) - \sum_i \mathbb{P}(P_i \notin H_i), \end{aligned} \tag{56}$$

where the infimum is over all tests $\psi : \mathcal{X} \to \{0, 1\}$.

Proof: We may assume without loss of generality that $\mathbb{P}(P_i \in H_i) > 0$ for both $i = 0$ and $i = 1$, as otherwise the claim is vacuous. Let $\tilde{P}_i$ be distributed as $P_i | \{P_i \in H_i\}$. Then for any set $A \subset \mathcal{X}$ we have

$$\begin{aligned} \left| \mathbb{E}\tilde{P}_i(A) - \mathbb{E}P_i(A) \right| &= \mathbb{P}(P_i \notin H_i) \\ &\quad \times \left| \mathbb{E}[P_i(A) | P_i \in H_i] - \mathbb{E}[P_i(A) | P_i \notin H_i] \right| && (57) \\ &\leq \mathbb{P}(P_i \notin H_i). && (58) \end{aligned}$$

In particular, $\mathsf{TV}(\mathbb{E}\tilde{P}_0, \mathbb{E}\tilde{P}_1) \leq \mathsf{TV}(\mathbb{E}P_0, \mathbb{E}P_1) + \sum_i \mathbb{P}(P_i \notin H_i)$. Therefore, for any ψ

$$\max_{i=0,1} \sup_{\mathbb{P}_i \in H_i} \mathbb{P}_i(\psi \neq i) \geq \frac{1}{2}(1 - \mathsf{TV}(\mathbb{E}\tilde{P}_0, \mathbb{E}\tilde{P}_1)) \tag{59}$$

$$\geq \frac{1}{2}\left(1 - \mathsf{TV}(\mathbb{E}P_0, \mathbb{E}P_1) - \sum_i \mathbb{P}(P_i \notin H_i)\right). \tag{60}$$

■

For clarity, we formally state (12) as testing between the null hypothesis

$$\left\{ \mathbb{P}_{\mathsf{X}}^{\otimes n} \otimes \mathbb{P}_{\mathsf{Y}}^{\otimes n} \otimes \mathbb{P}_{\mathsf{X}}^{\otimes m} \; : \; \begin{array}{c} \mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}} \in \mathcal{P} \\ \& \; \mathsf{TV}(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}) \geq \varepsilon \end{array} \right\} \tag{61}$$

versus the alternative hypothesis

$$\left\{ \mathbb{P}_{\mathsf{X}}^{\otimes n} \otimes \mathbb{P}_{\mathsf{Y}}^{\otimes n} \otimes \mathbb{P}_{\mathsf{Y}}^{\otimes m} \; : \; \begin{array}{c} \mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}} \in \mathcal{P} \\ \& \; \mathsf{TV}(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}) \geq \varepsilon \end{array} \right\} \tag{62}$$

The lower bounds of Theorem 3 follow from those for Theorems 1 and 2 so we may focus on the latter.

1) Smooth Densities: For concreteness let us focus on the case of $\mathcal{P} = \mathcal{P}_{\mathsf{H}}$. We take $\mathbb{P}_0$ to be uniform on $[0, 1]^d$ and $\mathbb{P}_\eta$ to have density

$$p_\eta = 1 + \sum_{j \in [\kappa]^d} \eta_j h_j \tag{63}$$

with respect to $\mathbb{P}_0$. Here $\kappa \in \mathbb{N}$, each $\eta \in \{\pm 1\}^{\kappa^d}$ is uniform and h_j is a bump function supported on the j'th cell of the regular grid of size κ^d on $[0, 1]^d$. The parameters κ, h_j of the construction are set in a way to ensure $\mathbb{P}_\eta \in \mathcal{P}_{\mathsf{H}}$ and $\mathsf{TV}(\mathbb{P}_0, \mathbb{P}_\eta) \geq \varepsilon$ with probability 1 over η. We have

$$\begin{aligned} 1 + \chi^2(\mathbb{E}_\eta \mathbb{P}_\eta^{\otimes m} \| \mathbb{P}_0^{\otimes m}) & \\ &= \int_{[0,1]^{dm}} \left(\mathbb{E}_\eta \prod_{i=1}^n p_\eta(x_i) \right)^2 \mathrm{d}x && (64) \\ &= \mathbb{E}_{\eta\eta'} \langle p_\eta, p_{\eta'} \rangle_{L^2}^m && (65) \\ &= \mathbb{E}(1 + \|h_{\mathbb{1}}\|_2^2 \langle \eta, \eta' \rangle)^m && (66) \\ &\leq \exp(m^2 \|h_{\mathbb{1}}\|_2^4 \kappa^d), && (67) \end{aligned}$$

where η, η' are i.i.d. uniform and we assume $\|h_{\mathbb{1}}\|_2 = \|h_j\|_2$ for all $j \in [\kappa]^d$. The above approach is what Ingster used in his seminal paper [12] on goodness-of-fit testing, which we adapt to likelihood-free hypothesis testing (61), (62). Take $P_0 = \mathbb{P}_\eta^{\otimes n} \otimes \mathbb{P}_0^{\otimes n} \otimes \mathbb{P}_\eta^{\otimes m}$ and $P_1 = \mathbb{P}_\eta^{\otimes n} \otimes \mathbb{P}_0^{\otimes n} \otimes \mathbb{P}_0^{\otimes m}$ in Lemma 4. Bounding $\mathsf{TV}(\mathbb{E}P_0, \mathbb{E}P_1)$ proceeds in multiple steps: first, we drop the Y-sample using the data-processing inequality. Then, we use Pinsker's inequality and the chain rule to bound TV by the KL divergence of Z conditioned on X. We bound KL by χ^2, arriving at the same equation (66). However, the mixing parameters η, η' are no longer independent, instead, given X they're independent from the posterior. In the remaining steps we use the fact that the posterior factorizes over the bins and the calculation is reduced to just a single bin where it can be done explicitly.

Let us now turn to the lower bound in Theorem 4. The difference in the rate is a consequence of the fact that H and TV behave differently for densities near zero. Inspired by this, we slightly modify the construction (63) by putting the perturbations at density level ε^2 as opposed to 1. Bounding TV then proceeds analogously to the steps outlined above.

2) Bounded Discrete Distributions: The construction is entirely analogous to the case of $\mathcal{P}_{\mathsf{H}}$ and we refer to the appendix for details. In the computer science community the construction of p_η is attributed to Paninski [62].

3) Gaussian Sequence Model: The null distribution $\mathbb{P}_0$ is the no signal case $\otimes_{i=1}^{\infty}\mathcal{N}(0,1)$ while the alternative is $\mathbb{P}_\theta = \otimes_{i=1}^{\infty}\mathcal{N}(\theta_i, 1)$ where θ has prior distribution $\otimes_{i=1}^{\infty}\mathcal{N}(0,\gamma_i)$ for an appropriate sequence $\gamma \in \mathbb{R}^{\mathbb{N}}$. We refer to the appendix for more details.

4) General Discrete Distributions: Once again, the irregular case $\mathcal{P}_{\mathsf{D}}$ requires special consideration. Clearly the lower bound for $\mathcal{P}_{\mathsf{Db}}$ carries over. However, in the regime $k \gtrsim 1/\varepsilon^4$ said lower bound becomes suboptimal, and we need a new construction, for which we utilize the moment-matching based approach of Valiant [63] as a black-box. The construction is derived from that used for two-sample testing by Valiant, namely the pair $(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}})$ is chosen uniformly at random from $\{(p\circ\pi, q\circ\pi)\}_{\pi\in S_k}$. Here we write S_k for the symmetric group on $[k]$ and

$$p(i) = \begin{cases} \frac{1-\varepsilon}{n} & \text{for } i \in [n] \\ \frac{4\varepsilon}{k} & \text{for } i \in [\frac{k}{2}, \frac{3k}{4}] \\ 0 & \text{otherwise,} \end{cases} \tag{68}$$

where we assume that $m \le n \le k/2$ and define $q(i) = p(i)$ for $i \in [k/2-1]$ and $q(i) = p(3k/2-i)$ for $i \in [k/2, k]$. This construction gives a lower bound matching our upper bound in the regime $m \lesssim n \lesssim k$. The final piece of the puzzle follows by the reduction from two-sample testing with unequal sample size (37), as this shows that likelihood-free hypothesis testing is at least as hard as two-sample testing in the $n \le m$ regime, and known lower bounds on the sample complexity of two-sample testing [26] (see also Table I) let us conclude.

V. Open Problems

A natural follow-up direction to the present paper would be to study multiple hypothesis testing where $\mathbb{P}_{\mathsf{X}}$ and $\mathbb{P}_{\mathsf{Y}}$ are replaced by $\mathbb{P}_{\mathsf{X}_1}, \dots, \mathbb{P}_{\mathsf{X}_M}$ with corresponding hypotheses $H_1, \dots, H_M$. The geometry of the family $\{\mathbb{P}_{\mathsf{X}_j}\}_{j\in[M]}$ might have interesting effects on the sample complexities.

Open problem 1. *Study the dependence on $M > 2$ of likelihood-free testing with M hypotheses.*

Another possible avenue of research is the study of local minimax/instance optimal rates, which is the focus of recent work [52], [64]–[67] in the case of goodness-of-fit and two-sample testing.

Open problem 2. *Define and study the local minimax rates of likelihood-free hypothesis testing.*

Our discussion of the Hellinger case in Section III-D is quite limited, natural open problems in this direction include the following.

Open problem 3. *Let $\mathcal{P}$ be one of $\mathcal{P}_{\mathsf{H}}(\beta, d, C)$, $\mathcal{P}_{\mathsf{Db}}(k, C_{\mathsf{Db}})$ or $\mathcal{P}_{\mathsf{D}}(k)\}$.*

(i) Study n_{GoF} and n_{TS} under Hellinger separation.
(ii) Study $\mathcal{R}_{\mathsf{LF}}$ under Hellinger separation.

More ambitiously, one might ask for a characterization of 'regular' models $(\mathcal{P}, \mathsf{d})$ for which goodness-of-fit testing and two-sample testing are equally hard and the region $\mathcal{R}_{\mathsf{LF}}$ is given by the trade-off in Theorem 1.

Open problem 4. *Find a general family of "regular" models $(\mathcal{P}, \mathsf{d})$ for which*

$$n_{\mathsf{GoF}}(\varepsilon, \mathsf{d}, \mathcal{P}) \asymp n_{\mathsf{TS}}(\varepsilon, \mathsf{d}, \mathcal{P}) \tag{69}$$

and

$$\mathcal{R}_{\mathsf{LF}}(\varepsilon, \mathsf{d}, \mathcal{P}) \asymp \left\{ (m,n) : \begin{array}{c} m \ge 1/\varepsilon^2 \\ \& \quad n \ge n_{\mathsf{GoF}} \\ \& \quad mn \ge n_{\mathsf{GoF}}^2 \end{array} \right\}. \tag{70}$$

Recent follow-up work [11] showed that Scheffé's test is also minimax optimal and achieves the entire trade-off in Fig. 1. It appears that the optimality of Scheffé's test is a consequence of the minimax point of view. Basically, in the worst-case the log-likelihood ratio between the hypotheses is close to being binary, hence quantizing it to $\{0,1\}$ does not lose optimality. Consequently, an important future direction is to better understand the competitive properties of various tests and studying some notion of regret, see [68] for prior related work.

Open problem 5. *Study the competitive optimality of likelihood-free hypothesis testing algorithms, and Scheffé's test in particular.*

Appendix A
Proof of Achievability in Theorem 1 and 2

Let μ be a measure on the measurable space $(\mathcal{X}, \mathcal{F})$. Let $\{\phi_i\}_{i\in[r]}$ be a sequence of orthonormal functions in $L^2(\mu)$, where we use the notation $[r] \triangleq \{1, 2, \dots, r\}$. For $f \in L^2(\mu)$, define its projection onto the span of $\{\phi_1, \dots, \phi_r\}$ as

$$P_r(f) \triangleq \sum_{i\in[r]} \langle f\phi_i\rangle \phi_i, \tag{71}$$

where we write $\langle\cdot\rangle$ for integration with respect to μ and $\|\cdot\|_p$ for $\|\cdot\|_{L^p(\mu)}$. Given an i.i.d. sample $X = (X_1, \ldots, X_n)$ from some density f, define its empirical projection as

$$\widehat{P}_r[X] \triangleq \sum_{i\in[r]} \left(\frac{1}{n}\sum_{j=1}^{n}\phi_i(X_j)\right)\phi_i. \tag{72}$$

We define the difference in L^2-distances statistics to be

$$T_{\mathsf{LF}} = \|\widehat{P}_r[X] - \widehat{P}_r[Z]\|_2^2 - \|\widehat{P}_r[Y] - \widehat{P}_r[Z]\|_2^2, \tag{73}$$

for an appropriate choice of μ and $\{\phi_j\}_{j\geq 1}$ depending on the class $\mathcal{P}$. Before calculating the mean and variance, we separate out the diagonal terms in T_{LF} thereby decomposing the statistic into two terms:

$$T_{\mathsf{LF}} \triangleq T_{\mathsf{LF}}^{-\mathsf{d}} + \underbrace{\frac{1}{n^2}\sum_{i\in[r]}\sum_{j\in[n]}\big(\phi_i^2(X_j) - \phi_i^2(Y_j)\big)}_{\triangleq D}, \tag{74}$$

which will simplify our proofs somewhat.

To ease notation in the results below, we define the quantities

$$\begin{aligned} A_{fgh} &= \langle f\big[P_r(g-h)\big]^2\rangle \\ B_{fg} &= \sum_{i=1}^{r}\langle f\phi_i P_r(g\phi_i)\rangle \end{aligned} \tag{75}$$

for $f, g, h \in L^2(\mu)$, assuming the quantities involved are well-defined. We are ready to state our meta-result from which we derive all our likelihood-free hypothesis testing upper bounds.

Proposition 5. *Let f, g, h denote probability densities on $\mathcal{X}$ with respect to μ, and suppose we observe independent samples X, Y, Z of size n, n, m from f, g, h respectively. Then*

$$\begin{aligned} \mathbb{E}T_{\mathsf{LF}}^{-\mathsf{d}} &= \|P_r(f-h)\|_2^2 - \|P_r(g-h)\|_2^2 \\ &\quad + \frac{1}{n}(\|P_r(g)\|_2^2 - \|P_r(f)\|_2^2) \end{aligned} \tag{76}$$

$$\begin{aligned} \operatorname{var}(T_{\mathsf{LF}}^{-\mathsf{d}}) \lesssim{}& \frac{A_{ffh} + A_{ggh}}{n} + \frac{A_{hfg}}{m} \\ &+ \frac{\|f+g+h\|_2^4 + |B_{fh}| + |B_{gh}|}{nm} \\ &+ \frac{|B_{ff}| + |B_{gg}| + \|f+g+h\|_2^4}{n^2} \\ &+ \frac{\sqrt{A_{ff0}A_{ffh} + A_{gg0}A_{ggh}}}{n^2} \\ &+ \frac{|B_{ff}| + |B_{gg}| + \|f+g+h\|_2^4}{n^3} \\ &+ \frac{A_{ff0} + A_{gg0}}{n^3}, \end{aligned} \tag{77}$$

where the implied constant is universal.

Proposition 5 is used to test (12) by rejecting the null whenever $T_{\mathsf{LF}}^{-\mathsf{d}} \geq 0$. To prove that this procedure performs well we show that $T_{\mathsf{LF}}^{-\mathsf{d}}$ concentrates around its mean by Chebyshev's inequality. For this we find sufficient conditions on the sample sizes n, m so that $(\mathbb{E}T_{\mathsf{LF}}^{-\mathsf{d}})^2 \gtrsim \operatorname{var}(T_{\mathsf{LF}}^{-\mathsf{d}})$ for a small enough implied constant on the left.

While Proposition 5 is enough to conclude the proof of our main theorems, notice that it uses the statistic $T_{\mathsf{LF}}^{-\mathsf{d}}$ which has the diagonal terms removed. For completeness we show that rejecting when $T_{\mathsf{LF}} \geq 0$ is also minimax optimal, that is, the diagonal term D in (74) can be included without degrading performance.

A. The Class $\mathcal{P}_{\mathsf{Db}}$

Proposition 6. *For any $C > 1$ there exists a constant $c > 0$ such that*

$$\begin{aligned} &\mathcal{R}_{\mathsf{rLF}}(\varepsilon, \mathcal{P}_{\mathsf{Db}}(k, C), \mathsf{B}_\cdot) \\ &\qquad \supset c\left\{(m,n) : \begin{array}{c} m \geq 1/\varepsilon^2 \\ \& \quad n \geq \sqrt{k}/\varepsilon^2 \\ \& \; mn \geq k/\varepsilon^4 \end{array}\right\}, \end{aligned} \tag{78}$$

where $\mathsf{B}_u = \{u \in \mathcal{P}_{\mathsf{Db}}(k, C) : \|u - v\|_2 \leq \varepsilon/(2\sqrt{k})\}$.

Proof: Choice of μ and ϕ. Take $\mathcal{X} = [k]$ and let $\mu = \sum_{i=1}^{k}\delta_i$ be the counting measure. Let $\phi_i(j) = \mathbb{1}_{\{i=j\}}$ and choose $r = k$ so that $P_r = P_k$ is the identity. By the Cauchy-Schwarz inequality $\|u\|_1 \leq \sqrt{k}\|u\|_2$ for all $u \in \mathbb{R}^k$.

Applying Proposition 5. Recall the notation of Proposition 5, so that f, g, h are the pmfs of $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}, \mathbb{P}_{\mathsf{Z}}$ respectively. We analyse the performance of the test $\mathbb{1}\{T_{\mathsf{LF}}^{-\mathsf{d}} \geq 0\}$ under the null hypothesis, the proof under the alternative is analogous. The inequality

$$\|f-h\|_2 \leq \frac{\varepsilon}{2\sqrt{k}} \leq \frac{\|f-g\|_1}{4\sqrt{k}} \leq \frac{\|f-g\|_2}{4} \tag{79}$$

along with the reverse triangle inequality yields

$$\begin{aligned} &\|g-h\|_2^2 - \|f-h\|_2^2 \\ &\quad \geq (\|f-g\|_2 - \|f-h\|_2)^2 - \|f-h\|_2^2 && (80)\\ &\quad = \|f-g\|_2^2 - 2\|f-g\|_2\|f-h\|_2 && (81)\\ &\quad \geq \|f-g\|_2^2/2. && (82) \end{aligned}$$

Combining the above inequality with Proposition 5, we get that $-\mathbb{E}T_{\mathsf{LF}}^{-\mathsf{d}} \geq \|f-g\|_2^2/2 + R$, where the residual term R can be bounded as

$$|R| = \left|\frac{\|f\|_2^2 - \|g\|_2^2}{n}\right| \tag{83}$$

$$\leq 2C\frac{\|f-g\|_2}{n\sqrt{k}}. \tag{84}$$

Therefore, $-\mathbb{E}T_{\mathsf{LF}}^{-\mathsf{d}} \geq \frac{\|f-g\|_2^2}{4}$ holds provided $2C\|f-g\|_2/(n\sqrt{k}) \leq \|f-g\|_2^2/4$, which in turn is implied by $n \gtrsim 1/\varepsilon$ and is thus always satisfied.

Turning towards the variance, we apply Proposition 5 to see that

$$\frac{\mathrm{var}(T_{\mathsf{LF}}^{-\mathsf{d}})}{n+m} \lesssim \frac{\|f-g\|_2^2}{knm} + \frac{1}{kn^2m}, \tag{85}$$

where we use the trivial bounds

$$\|f+g+h\|_2 \lesssim \sqrt{\frac{C}{k}} \lesssim \sqrt{\frac{1}{k}} \tag{86}$$

$$|B_{ff}| + |B_{gg}| + |B_{fh}| + |B_{gh}| \lesssim \frac{C}{k} \lesssim \frac{1}{k} \tag{87}$$

$$A_{ffh} + A_{ggh} + A_{hfg} \lesssim \frac{C}{k}\|f-g\|_2^2 \lesssim \frac{1}{k}\|f-g\|_2^2 \tag{88}$$

$$A_{ff0} + A_{gg0} \lesssim \left(\frac{C}{k}\right)^2 \lesssim \frac{1}{k^2}. \tag{89}$$

Applying Chebyshev's inequality and looking at each term separately in (85) and using that $\|f-g\|_2 \geq \varepsilon/(2\sqrt{k})$ yields the desired bounds on n, m.

The diagonal. While the above test using $T_{\mathsf{LF}}^{-\mathsf{d}}$ already achieves the minimax optimal sample complexity, here we show for completeness that the diagonal D, defined in (74), can be included without degrading the test's performance. Indeed, we always have

$$D = \frac{1}{n^2}\sum_{i\in[r]}\sum_{j\in[n]}\left(\mathbb{1}\{X_j = i\}^2 - \mathbb{1}\{Y_j = i\}^2\right) \tag{90}$$

$$= 0. \tag{91}$$

Therefore, trivially, the test $\mathbb{1}\{T_{\mathsf{LF}} \geq 0\}$ has the same performance as the one analyzed above. ∎

B. The Class $\mathcal{P}_{\mathsf{H}}$

Proposition 7. *For every $C > 1, \beta > 0$ and $d \geq 1$ there exist two constants $c, c_{\mathsf{r}} > 0$ such that*

$$\mathcal{R}_{\mathsf{rLF}}(\varepsilon, \mathcal{P}_{\mathsf{H}}(\beta, d, C), \mathsf{B}_\cdot)$$

$$\supset c\left\{(m,n): \begin{array}{c} m \geq 1/\varepsilon^2 \\ \& \ n \geq 1/\varepsilon^{(2\beta+d/2)/\beta} \\ \& \ mn \geq 1/\varepsilon^{2(2\beta+d/2)/\beta} \end{array}\right\}, \tag{92}$$

where $\mathsf{B}_u = \{v \in \mathcal{P}_{\mathsf{H}}(\beta, d, C) : \|v-u\|_2 \leq c_{\mathsf{r}}\varepsilon\}$.

Proof: Choice of μ and ϕ. Take $\mathcal{X} = [0,1]^d$, let μ the Lebesgue measure on $\mathcal{X}$. Let $\{\phi_i\}_{1\leq i\leq \kappa^d}$ be the indicators of the cells of the regular grid with κ^d bins, normalized to have $L^2(\mu)$-norm equal to 1, that is, the indicator is multiplied by κ^d, which is one over the volume of one grid cell. By [41, Lemma 7.2] for any resolution $r = \kappa^d$ and $u \in \mathcal{C}(\beta, d, 2C)$ we have

$$\|P_r(u)\|_2 \geq c_1\|u\|_2 - c_2\kappa^{-\beta} \tag{93}$$

for constants $c_1, c_2 > 0$ that don't depend on r. In particular, the inequalities

$$\|P_r(u)\|_2 \geq \frac{c_1}{2}\|u\|_2 \tag{94}$$

holds for any $\|u\|_2 \geq \varepsilon$ provided we choose $\kappa = \left(\frac{2c_2}{c_1\varepsilon}\right)^{1/\beta}$.

Applying Proposition 5. Recall the notation of Proposition 5 so that f, g, h are the μ-densities of $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}, \mathbb{P}_{\mathsf{Z}}$. We analyse the performance of the test $\mathbb{1}\{T_{\mathsf{LF}}^{-\mathsf{d}} \geq 0\}$ under the null hypothesis, the proof under the alternative is analogous. Let the radius of robustness be $c_{\mathsf{r}} = c_1/4$, and set $\kappa = \left(\frac{2c_2}{c_1\varepsilon}\right)^{1/\beta}$. Then we have

$$\|P_r(f-h)\|_2 \leq c_{\mathsf{r}}\varepsilon \tag{95}$$

$$= \frac{c_{\mathsf{r}}}{2}\|f-g\|_2 \tag{96}$$

$$\leq \frac{c_{\mathsf{r}}}{c_1}\|P_r(f-g)\|_2 \tag{97}$$

by taking $u = f-g$ in (94). Using the reverse triangle inequality we obtain

$$\|P_r(g-h)\|_2^2 - \|P_r(f-h)\|_2^2$$

$$\geq \left(\|P_r(f-g)\|_2 - \|P_r(f-h)\|_2\right)^2 - \|P_r(f-h)\|_2^2 \tag{98}$$

$$= \|P_r(f-g)\|_2^2 - 2\|P_r(f-g)\|_2\|P_r(f-h)\|_2 \tag{99}$$

$$\geq \|P_r(f-g)\|_2^2(1 - 2\frac{c_{\mathsf{r}}}{c_1}) \tag{100}$$

$$= \|P_r(f-g)\|_2^2/2 \tag{101}$$

Combining the above inequality with Proposition 5, we see that $-\mathbb{E}T_{\mathsf{LF}}^{-\mathsf{d}} \geq \|P_r(f-g)\|_2^2/2 + R$ where the residual term R can be bounded as

$$|R| = \left|\frac{\|f\|_2^2 - \|g\|_2^2}{n}\right| \tag{102}$$

$$\leq 2C\frac{\|f-g\|_2}{n}. \tag{103}$$

Therefore, the inequality $-\mathbb{E}T_{\mathsf{LF}}^{-\mathsf{d}} \geq \|P_r(f-g)\|_2^2/4$ holds provided $2C\|f-g\|_2/n \leq \|P_r(f-g)\|_2^2/4$, which in turn is implied by $n \gtrsim 1/\varepsilon$ and is thus always satisfied.

Turning to the variance, using Proposition 5 we obtain

$$\frac{\mathrm{var}(T_{\mathsf{LF}}^{-\mathsf{d}})}{n+m} \lesssim \frac{\|P_r(f-g)\|_2^2}{nm} + \frac{\varepsilon^{-d/\beta}}{n^2 m}, \tag{104}$$

where we apply the trivial inequalities

$$\|f+g+h\|_2 \lesssim \sqrt{C} \lesssim 1 \tag{105}$$

$$|B_{ff}| + |B_{gg}| + |B_{fh}| + |B_{gh}| \lesssim Cr \tag{106}$$

$$= C\kappa^d \asymp \varepsilon^{-d/\beta} \tag{107}$$

$$A_{ffh} + A_{ggh} + A_{hfg} \lesssim C\|P_r(f-g)\|_2^2 \tag{108}$$

$$\lesssim \|P_r(f-g)\|_2^2 \tag{109}$$

$$A_{ff0} + A_{gg0} \lesssim C^2 \lesssim 1. \tag{110}$$

Applying Chebyshev's inequality and looking at each term separately in (104) and using that $\|P_r(f-g)\|_2 \gtrsim \|f-g\|_2 \geq \|f-g\|_1 \geq 2\varepsilon$ yields the desired bounds on n, m.

The diagonal. While the above test using $T_{\mathsf{LF}}^{-\mathsf{d}}$ already achieves the minimax optimal sample complexity, for completeness we also note that including the diagonal terms D defined in (74) doesn't degrade performance. This follows from the simple fact that $D = 0$, which is true for reasons analogous to the case of $\mathcal{P}_{\mathsf{Db}}$ that we already covered. ∎

C. The Class $\mathcal{P}_{\mathsf{G}}$

Proposition 8. *For all $s, C > 0$ there exists a constant $c > 0$ such that*

$$\mathcal{R}_{\mathsf{rLF}}(\varepsilon, \mathcal{P}_{\mathsf{G}}(s,C), \mathsf{B}_\cdot) \supset c\left\{(m,n) : \begin{array}{c} m \geq 1/\varepsilon^2 \\ \& \; n \geq 1/\varepsilon^{(2s+1/2)/s} \\ \& \; mn \geq 1/\varepsilon^{2(2s+1/2)/s} \end{array}\right\}, \tag{111}$$

where $\mathsf{B}_{\mu_\theta} = \{\mu_{\theta'} : \theta' \in \mathcal{E}(s,C), \|\theta-\theta'\|_2 \leq \varepsilon/4\}$.

Proof: Choosing μ and ϕ. Let $\mathcal{X} = \mathbb{R}^{\mathbb{N}}$ be the set of infinite sequences and take as the base measure $\mu = \otimes_{d=1}^{\infty}\mathcal{N}(0,1)$, the infinite dimensional standard Gaussian. For $\theta \in \ell^2$ write $\mu_\theta = \otimes_{d=1}^{\infty}\mathcal{N}(\theta_i, 1)$ so that $\mu_0 = \mu$. Take the orthonormal functions $\phi_i(x) = x_i$ in $L^2(\mu)$ for $i \geq 1$, so that

$$P_r\left(\frac{\mathrm{d}\mu_\theta}{\mathrm{d}\mu}\right) = \sum_{i=1}^{r} x_i\theta_i. \tag{112}$$

Let $\theta, \theta' \in \mathcal{E}(s,C)$ with $\mathsf{TV}(\mu_\theta, \mu_{\theta'}) \geq \varepsilon$. By direct computation we obtain

$$\left\|P_r\left(\frac{\mathrm{d}\mu_\theta}{\mathrm{d}\mu} - \frac{\mathrm{d}\mu_{\theta'}}{\mathrm{d}\mu}\right)\right\|_2^2 = \sum_{i=1}^{r}(\theta_i - \theta_i')^2 \tag{113}$$

$$\geq \|\theta-\theta'\|_2^2 - r^{-2s}\sum_{i>r}(\theta_i-\theta_i')^2 i^{2s} \tag{114}$$

$$\geq \|\theta-\theta'\|_2^2 - 4C^2 r^{-2s}. \tag{115}$$

In particular, the inequality

$$\left\|P_r\left(\frac{\mathrm{d}\mu_\theta}{\mathrm{d}\mu} - \frac{\mathrm{d}\mu_{\theta'}}{\mathrm{d}\mu}\right)\|_2^2 \geq \frac{1}{2}\|\theta-\theta'\right\|_2^2 \tag{116}$$

holds for all $\theta, \theta' \in \mathcal{E}(s,C)$ with $\|\theta-\theta'\|_2 \geq \varepsilon$, provided we take $r = (4C/\varepsilon)^{1/s}$.

Applying Proposition 5. Recall the notation of Proposition 5, and let f, g, h be the μ-densities of $\mathbb{P}_{\mathsf{X}} = \mu_{\theta_{\mathsf{X}}}, \mathbb{P}_{\mathsf{Y}} = \mu_{\theta_{\mathsf{Y}}}, \mathbb{P}_{\mathsf{Z}} = \mu_{\theta_{\mathsf{Z}}}$ respectively. We analyse the test $\mathbb{1}\{T_{\mathsf{LF}}^{-\mathsf{d}} \geq 0\}$ only under the null hypothesis, as the analysis under the alternative is analogous. Note also that by Lemma 5 the inequality

$$\mathsf{TV}(\mu_\theta, \mu_{\theta'}) \leq \mathsf{H}(\mu_\theta, \mu_{\theta'}) \tag{117}$$

$$= \sqrt{2(1-\exp(-\|\theta-\theta'\|_2^2/8))} \tag{118}$$

$$\leq \frac{\|\theta-\theta'\|_2}{2} \tag{119}$$

holds for any $\theta, \theta' \in \ell^2$. Therefore, we have

$$\|P_r(f-h)\|_2 \leq \frac{\varepsilon}{4} \leq \frac{\mathsf{TV}(\mu_\theta, \mu_{\theta'})}{4} \tag{120}$$

$$\leq \frac{\|\theta-\theta'\|_2}{8} \leq \frac{\|P_r(f-g)\|_2}{4} \tag{121}$$

by (116).

By the reverse triangle inequality we have

$$\|P_r(g-h)\|_2^2 - \|P_r(f-h)\|_2^2 \tag{122}$$

$$\geq \left(\|P_r(f-g)\|_2 - \|P_r(f-h)\|_2\right)^2 - \|P_r(f-h)\|_2^2 \tag{123}$$

$$= \|P_r(f-g)\|_2^2 - 2\|P_r(f-g)\|_2\|P_r(f-h)\|_2 \tag{124}$$

$$\geq \|P_r(f-g)\|_2^2/2 \tag{125}$$

Combining the inequality above with Proposition 5, we see that $-\mathbb{E}T_{\mathsf{LF}}^{-\mathsf{d}} \geq \|P_r(f-g)\|_2^2/2 + R$, where the residual term R can be bounded as

$$|R| = \left|\frac{\|P_r(f)\|_2^2 - \|P_r(g)\|_2^2}{n}\right| \tag{126}$$

$$\leq 2C\frac{\|P_r(f-g)\|_2}{n}. \tag{127}$$

Therefore, $-\mathbb{E}T_{\mathsf{LF}}^{-\mathsf{d}} \geq \|P_r(f-g)\|_2^2/4$ provided $2C\|P_r(f-g)\|_2/n \leq \|P_r(f-g)\|_2^2/4$, which in turn is implied by $n \gtrsim 1/\varepsilon$ and is therefore always satisfied.

Let us turn to the variance of the statistic. Let u, v, t be the μ-densities of the distributions $\mu_\theta, \mu_{\theta'}, \mu_{\theta''}$ for some vectors $\theta, \theta', \theta'' \in \mathcal{E}(s, C)$ in the Sobolev ellipsoid. Straightforward calculations involving Gaussian random variables produce

$$A_{uvt} = \sum_{ij}^{r}(\mathbb{1}(i=j) + \theta_i\theta_j)(\theta'_i - \theta''_i)(\theta'_j - \theta''_j) \tag{128}$$

$$\leq (1+C^2)\|P_r(v-t)\|_2^2 \tag{129}$$

$$\lesssim \|P_r(v-t)\|_2^2 \lesssim C^2 \lesssim 1 \tag{130}$$

$$\|u\|_2 = \exp\left(\frac{1}{2}\|\theta\|_2^2\right) \leq \exp(C^2/2) \lesssim 1 \tag{131}$$

$$B_{uv} = \sum_{i=1}^{r}\left(1 + \theta_i^2 + {\theta'}_i^2 + \theta_i\theta'_i\sum_{j=1}^{r}\theta_j{\theta'}_j\right) \tag{132}$$

$$\leq r + 2C^2 + C^4 \lesssim r. \tag{133}$$

Applying Proposition 5 tells us that

$$\frac{\mathrm{var}(T_{\mathsf{LF}}^{-\mathsf{d}})}{n+m} \lesssim \frac{\|P_r(f-g)\|_2^2}{nm} + \frac{\varepsilon^{-1/s}}{n^2 m} \tag{134}$$

Applying Chebyshev's inequality and looking at each term separately in (134) and using that $\mathsf{TV}(\mu_\theta, \mu_{\theta'}) \lesssim \|P_r(f-g)\|$ yields the desired bounds on n, m.

The diagonal. While the above test using $T_{\mathsf{LF}}^{-\mathsf{d}}$ already achieves the minimax optimal sample complexity, for completeness we show that including the diagonal terms D defined in (74) doesn't degrade performance. To this end we compute

$$\mathbb{E}D = \mathbb{E}\frac{1}{n^2}\sum_{i\in[r]}\sum_{j\in[n]}\left(\phi_i^2(X_j) - \phi_i^2(Y_j)\right) \tag{135}$$

$$= \frac{1}{n}\sum_{i\in[r]}(\theta_{\mathsf{X},i}^2 - \theta_{\mathsf{Y},i}^2) \tag{136}$$

$$\leq \frac{1}{n}\|\theta_{\mathsf{X}} + \theta_{\mathsf{Y}}\|_2\sqrt{\sum_{i\in[r]}(\theta_{\mathsf{X},i} - \theta_{\mathsf{Y},i}{}^2)} \tag{137}$$

$$\leq 2C\frac{\|P_r(f-g)\|_2}{n}. \tag{138}$$

We see that $|\mathbb{E}T_{\mathsf{LF}}^{-\mathsf{d}}| \gtrsim |\mathbb{E}D|$ as soon as $n \gtrsim 1/\varepsilon$. Turning to the variance, we have

$$\mathrm{var}(D) = \frac{1}{n^3}\sum_{i\in[r]}\left(\mathrm{var}(\phi_i^2(X_1)) + \mathrm{var}(\phi_i^2(Y_1))\right) \tag{139}$$

$$\lesssim \frac{rC^2}{n^3}, \tag{140}$$

and so the diagonal terms do not inflate the variance by more than a constant factor. Therefore, the sample complexity of the test is unchanged. ∎

D. The Class $\mathcal{P}_{\mathsf{D}}$

Proposition 9. *Let $\alpha = 1 \vee \left(\frac{k}{n} \wedge \frac{k}{m}\right)$. There exist constants $c, c', c_{\mathsf{r}} > 0$ such that*

$$\mathcal{R}_{\mathsf{rLF}}(\varepsilon, \mathcal{P}_{\mathsf{D}}(k), \mathsf{B}_\cdot) \supset \frac{c}{\log(k)}\left\{(m,n) : \begin{array}{c} m \geq 1/\varepsilon^2 \\ \& \ n \geq \sqrt{k\alpha}/\varepsilon^2 \\ \& \ mn \geq k\alpha/\varepsilon^4 \end{array}\right\}, \tag{141}$$

where $\mathsf{B}_u = \{v : \|u-v\|_2 \leq c_{\mathsf{r}}\varepsilon/\sqrt{k}, \|v/u\|_\infty \leq c'\}$.

Proof: Choosing μ and ϕ. As for $\mathcal{P}_{\mathsf{Db}}$, we take $\mathcal{X} = [k]$, $\mu = \sum_{i=1}^{k}\delta_i$, $\phi_i(j) = \mathbb{1}_{\{i=j\}}$ and $r = k$. By the Cauchy-Schwarz inequality $\|h\|_1 \leq \sqrt{k}\|h\|_2$ for all $h \in \mathbb{R}^k$.

Reducing to the small-norm case. Before applying Proposition 5 we need to 'pre-process' our distributions. For an in-depth explanation of this technique see [25], [35]. Recall that we write f, g, h for the probability mass functions of $\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}, \mathbb{P}_{\mathsf{Z}}$ respectively, from which we observe the samples X, Y, Z of size n, n, m respectively. Recall also that the null hypothesis is that $\|f-h\|_2 \leq c_{\mathsf{r}}\varepsilon/\sqrt{k}$ while the alternative says that $\|g-h\|_2 \leq c_{\mathsf{r}}\varepsilon/\sqrt{k}$, with $\|f-g\|_2 \geq 2\varepsilon/\sqrt{k}$ guaranteed under both. In the following section we use the standard inequality $\mathbb{P}(\lambda - x \geq \mathrm{Poi}(\lambda)) \leq \exp(-\frac{x^2}{2(\lambda+x)})$ valid for all $x \geq 0$ repeatedly. We also utilize the identity

$$\mathbb{E}\left[\frac{1}{\mathrm{Poi}(\lambda)+1}\right] = \begin{cases} 1 & \text{if } \lambda = 0 \\ \frac{1-e^{-\lambda}}{\lambda} & \text{if } \lambda > 0, \end{cases} \tag{142}$$

which is easily verified by direct calculation. Finally, the following Lemma will come handy.

Proposition 10. *[35, Corollary 11.6] Given t samples from an unknown discrete distribution p, there exists*

an algorithm that produces an estimate $\widehat{\|p\|_2^2}$ with the property

$$\mathbb{P}\left(\widehat{\|p\|_2^2} \notin \left(\frac{1}{2}\|p\|_2^2, \frac{3}{2}\|p\|_2^2\right)\right) \lesssim \frac{1}{\|p\|_2 t}, \tag{143}$$

where the implied constant is universal.

First we describe a random "filter" $F : \mathcal{P}_{\mathsf{D}}(k) \to \mathcal{P}_{\mathsf{D}}(K)$ that maps distributions on $[k]$ to distributions on the inflated alphabet $[K]$. Let $(n_{\mathsf{X}}, n_{\mathsf{Y}}, n_{\mathsf{Z}}) = \frac{1}{2}(n\wedge k, n\wedge k, m\wedge k)$ and let $N^{\mathsf{X}} \sim \mathrm{Poi}(n_{\mathsf{X}}/2)$ independently of all other randomness, and define $N^{\mathsf{Y}}, N^{\mathsf{Z}}$ similarly. We take the first $N^{\mathsf{X}}, N^{\mathsf{Y}}, N^{\mathsf{Z}}$ samples from the data sets X, Y, Z respectively. In the event $N^{\mathsf{X}} \vee N^{\mathsf{Y}} > n$ or $N^{\mathsf{Z}} > m$ let our output to the likelihood-free hypothesis test be arbitrary, this happens with exponentially small probability. Let N_i^{X} be the number of the samples $X_1, \ldots, X_{N^{\mathsf{X}}}$ falling in bin i, so that $N_i^{\mathsf{X}} \sim \mathrm{Poi}(n_{\mathsf{X}} f_i/2)$ independently for each $i \in [k]$, and define $N_i^{\mathsf{Y}}, N_i^{\mathsf{Z}}$ analogously. The filter F is defined as follows: divide each support element $i \in \{1, 2, \ldots, k\}$ uniformly into $1+N_i^{\mathsf{X}}+N_i^{\mathsf{Y}}+N_i^{\mathsf{Z}}$ bins. The filter has the following properties trivially:

1) The construction succeeds with probability $\geq 1 - 3\exp(-n \wedge m \wedge k/16)$, focus on this event from here on.
2) The construction uses at most $n_{\mathsf{X}}, n_{\mathsf{Y}}, n_{\mathsf{Z}}$ samples from X, Y, Z respectively and satisfies $K \leq 5k/2$.
3) For any $u, v \in \mathcal{P}_{\mathsf{D}}(k)$ we have $\mathsf{TV}(F(u), F(v)) = \mathsf{TV}(u, v)$ and $\|F(u) - F(v)\|_2 \leq \|u - v\|_2$.
4) Given a sample from an unknown $u \in \mathcal{P}_{\mathsf{D}}(k)$ we can generate a sample from $F(u)$ and vice-versa.

Let $\tilde{f} \triangleq F(f)$ be the probability mass function after processing and define $\tilde{g}, \tilde{h}$ analogously. By properties $1-2$ of the filter, we may assume with probability 99% that the new alphabet's size is at most $5k/2$ and that we used at most half of our samples X, Y, Z. We immediately get $2\varepsilon \leq \|f - g\|_1 = \|\tilde{f} - \tilde{g}\|_1 \leq \sqrt{5k/2}\|\tilde{f} - \tilde{g}\|_2$ and $\|\tilde{f} - \tilde{h}\|_2 \leq \|f - h\|_2, \|\tilde{g} - \tilde{h}\|_2 \leq \|g - h\|_2$. Notice that

$$\sum_{i\in[K]} \tilde{f}_i \tilde{g}_i = \sum_{i\in[k]} \frac{f_i g_i}{1 + N_i^{\mathsf{X}} + N_i^{\mathsf{Y}} + N_i^{\mathsf{Z}}} \tag{144}$$

holds, and similar statements can be derived for the inner product between $\tilde{f}, \tilde{h}$ etc. Recall that we set

$$\alpha = \max\left\{1, \min\left\{\frac{k}{n}, \frac{k}{m}\right\}\right\}. \tag{145}$$

Adopting the convention $0/0 = 1$ and using (142) we can bound inner products between the mass functions as

$$\mathbb{E}\left[B_{\tilde{f}\tilde{h}} + B_{\tilde{g}\tilde{h}}\right] = \mathbb{E}\left[\langle \tilde{f}\tilde{h}\rangle + \langle \tilde{g}\tilde{h}\rangle\right] \tag{146}$$

$$\leq 4\sum_{i\in[k]} \frac{f_i h_i + g_i h_i}{(n\wedge k)(f_i + g_i) + (m \wedge k)h_i} \tag{147}$$

$$\leq \frac{8}{(n\vee m)\wedge k} = \frac{8\alpha}{k} \tag{148}$$

$$\mathbb{E}\left[B_{\tilde{f}\tilde{f}} + B_{\tilde{g}\tilde{g}}\right] = \mathbb{E}\left[\|\tilde{f}\|_2^2 + \|\tilde{g}\|_2^2\right] \tag{149}$$

$$\leq 4\sum_{i\in[k]} \frac{f_i^2 + g_i^2}{(n\wedge k)(f_i + g_i) + (m \wedge k)h_i} \tag{150}$$

$$\leq \frac{8}{n \wedge k} \tag{151}$$

$$\mathbb{E}\|\tilde{h}\|_2^2 \leq 4\sum_{i\in[k]} \frac{h_i^2}{(n\wedge k)(f_i + g_i) + (m \wedge k)h_i} \tag{152}$$

$$\leq \frac{4}{m \wedge k}. \tag{153}$$

By Markov's inequality we may assume that the inequalities in the preceding display hold not only in expectation but with 99% probability overall with universal constants. Notice that under the null hypothesis $\|\tilde{f} - \tilde{h}\|_2 \leq c_{\mathsf{r}}\varepsilon/\sqrt{k}$ and thus $\|\tilde{f}\|_2 \leq \|\tilde{h}\|_2 + c_{\mathsf{r}}\varepsilon/\sqrt{k} \leq \|\tilde{f}\|_2 + 2c_{\mathsf{r}}\varepsilon/\sqrt{k}$, and similarly with $\tilde{f}$ replaced by $\tilde{g}$ under the alternative. We restrict our attention to $c_{\mathsf{r}} \in (0, 1)$ so that c_{r} is treated as a constant where appropriate. Notice that $\varepsilon/\sqrt{k} \lesssim 1/\sqrt{(n\vee m)\wedge k}$ holds trivially. Thus, we obtain $\|\tilde{f}\|_2 \vee \|\tilde{h}\|_2 \leq c/\sqrt{(m\vee n)\wedge k}$ under the null and $\|\tilde{g}\|_2 \vee \|\tilde{h}\|_2 \leq c/\sqrt{(n\vee m)\wedge k}$ under the alternative for a universal constant c. We would like to ensure that

$$\|\tilde{f}\|_2 \vee \|\tilde{g}\|_2 \vee \|\tilde{h}\|_2 \lesssim \frac{1}{\sqrt{(m\vee n)\wedge k}} = \sqrt{\frac{\alpha}{k}}. \tag{154}$$

To this end we apply Proposition 10 using $(n/4, n/4)$ of the remaining, transformed but otherwise untouched X, Y samples. Let $\widehat{\|\tilde{f}\|_2^2}, \widehat{\|\tilde{g}\|_2^2}$ denote the estimates, which lie in $(\frac{1}{2}\|\tilde{f}\|_2^2, \frac{3}{2}\|\tilde{f}\|_2^2)$ and $(\frac{1}{2}\|\tilde{g}\|_2^2, \frac{3}{2}\|\tilde{g}\|_2^2)$ respectively, with probability at least $1 - \mathcal{O}((\|\tilde{f}\|_2^{-1} + \|\tilde{g}\|_2^{-1})/n) \geq 1 - \mathcal{O}(\sqrt{k}/n)$, since $\|\tilde{f}\|_2 \wedge \|\tilde{g}\|_2 \geq \sqrt{2/(5k)}$ by the Cauchy-Schwarz inequality. Assuming that $n \gtrsim \sqrt{k}$ this probability can be taken to be arbitrarily high, say 99%. Now we perform the following procedure: if $\widehat{\|\tilde{f}\|_2^2} > \frac{3}{2}c^2/((n\vee m)\wedge k)$ reject the null hypothesis, otherwise if $\widehat{\|\tilde{g}\|_2^2} > \frac{3}{2}c^2/((n\vee m)\wedge k)$ accept the null hypothesis, otherwise proceed with the assumption that (154) holds. By design this process, on our $97\% \leq$ probability event of interest, correctly identifies the hypothesis or correctly concludes that (154) holds. The last step of the reduction is ensuring that the

quantities $A_{\tilde{f}\tilde{f}\tilde{h}}, A_{\tilde{g}\tilde{g}\tilde{h}}, A_{\tilde{h}\tilde{f}\tilde{g}}, A_{\tilde{f}\tilde{f}0}, A_{\tilde{g}\tilde{g}0}$ are small. The first two and last two may be bounded easily as

$$\begin{aligned} A_{\tilde{f}\tilde{f}\tilde{h}} + A_{\tilde{g}\tilde{g}\tilde{h}} &= \langle \tilde{f}(\tilde{f}-\tilde{h})^2\rangle + \langle \tilde{g}(\tilde{g}-\tilde{h})^2\rangle && (155)\\ &\le \|\tilde{f}\|_2\|\tilde{f}-\tilde{h}\|_4^2 + \|\tilde{g}\|_2\|\tilde{g}-\tilde{h}\|_4^2 && (156)\\ &\lesssim \frac{\|\tilde{f}-\tilde{h}\|_2^2 + \|\tilde{g}-\tilde{h}\|_2^2}{\sqrt{(n\vee m)\wedge k}} && (157)\\ &\lesssim \frac{\|\tilde{f}-\tilde{g}\|_2^2 + c_{\mathsf{r}}^2\varepsilon^2/k}{\sqrt{(n\vee m)\wedge k}} && (158)\\ &\lesssim \frac{\|\tilde{f}-\tilde{g}\|_2^2}{\sqrt{(n\vee m)\wedge k}} && (159)\\ &= \sqrt{\frac{\alpha}{k}}\|\tilde{f}-\tilde{g}\|_2^2 && (160)\\ A_{\tilde{f}\tilde{f}0} + A_{\tilde{g}\tilde{g}0} &= \|\tilde{f}\|_3^3 + \|\tilde{g}\|_3^3 \le \|\tilde{f}\|_2^3 + \|\tilde{g}\|_2^3 && (161)\\ &\lesssim \frac{1}{((n\vee m)\wedge k)^{3/2}} = \Big(\frac{\alpha}{k}\Big)^{3/2}. && (162)\end{aligned}$$

To bound $A_{\tilde{h}\tilde{f}\tilde{g}}$ we need a more sophisticated method. Recall that by definition

$$A_{\tilde{h}\tilde{f}\tilde{g}} = \sum_{i\in[k]} \frac{h_i(f_i-g_i)^2}{(1+N_i^{\mathsf{X}}+N_i^{\mathsf{Y}}+N_i^{\mathsf{Z}})^2}. \tag{163}$$

Fix an $i\in[k]$ and let $P \triangleq N_i^{\mathsf{X}}+N_i^{\mathsf{Y}}+N_i^{\mathsf{Z}} \sim \mathrm{Poi}((n\wedge k)(f_i+g_i)/4 + (m\wedge k)h_i/4)$ and take a constant $c>0$ to be specified. Assuming that i is such that $\mathbb{E}P \ge c\log(k)$ and taking k large enough so that $c\log(k)\ge 2$, we clearly have

$$\begin{aligned} &\mathbb{P}\left(\frac{1}{1+P} > \frac{c\log(k)}{\mathbb{E}P}\right) && (164)\\ &\quad\le \exp\left(-\frac{1}{2}\frac{(\mathbb{E}P(1-\frac{1}{c\log(k)})+1)^2}{\mathbb{E}P(2-\frac{1}{c\log(k)})+1}\right) && (165)\\ &\quad\le \exp\left(-\frac{1}{16}\mathbb{E}P\right) && (166)\\ &\quad\le \frac{1}{k^{c/16}}. && (167)\end{aligned}$$

Choosing $c=32$ and taking a union bound, the inequality

$$\begin{aligned} A_{\tilde{h}\tilde{f}\tilde{g}} &\lesssim \frac{\log(k)}{m\wedge k}\sum_{i\in[k]}\frac{(f_i-g_i)^2}{1+N_i^{\mathsf{X}}+N_i^{\mathsf{Y}}+N_i^{\mathsf{Z}}} && (168)\\ &\asymp \frac{\log(k)}{m\wedge k}\|\tilde{f}-\tilde{g}\|_2^2 && (169)\end{aligned}$$

holds with probability at least $1-1/k$. Using that $\|h/f\|_\infty \wedge \|h/g\|_\infty \lesssim 1$ by assumption, we obtain $A_{\tilde{h}\tilde{f}\tilde{g}} \lesssim \frac{\log(k)}{n\wedge k}\|\tilde{f}-\tilde{g}\|_2^2$ similarly. Combining the two bounds yields

$$\begin{aligned} A_{\tilde{h}\tilde{f}\tilde{g}} &\lesssim \frac{\log(k)}{(m\vee n)\wedge k}\|\tilde{f}-\tilde{g}\|_2^2 && (170)\\ &= \frac{\log(k)\alpha}{k}\|\tilde{f}-\tilde{g}\|_2^2. && (171)\end{aligned}$$

To summarize, under the assumptions that $n\gtrsim\sqrt{k}$, and at the cost of inflating the alphabet size to at most $\frac{5}{2}k$ and a probability of error at most $3\%+\frac{1}{k}$, we may assume that the inequalities (154), (160), (162) and (170) hold with universal constants.

Applying Proposition 5. We only analyse the type-I error, as the type-II error follows analogously. As explained earlier, we apply the test $\mathbb{1}\{T_{\mathsf{LF}}^{-\mathsf{d}}\ge 0\}$ to the transformed samples with probability mass functions $\tilde{f},\tilde{g},\tilde{h}$. Note that taking c_{r} small eonugh shows that

$$\|\tilde{g}-\tilde{h}\|_2^2 - \|\tilde{f}-\tilde{h}\|_2^2 \gtrsim \|\tilde{f}-\tilde{g}\|_2^2 \tag{172}$$

for a universal implied constant. Therefore, by Proposition 5 we see that $-\mathbb{E}T_{\mathsf{LF}}^{-\mathsf{d}} \ge c\|\tilde{f}-\tilde{g}\|_2^2 + R$ for some universal constant $c>0$, where the residual term R can be bounded as

$$\begin{aligned} |R| &= \left|\frac{\|\tilde{f}\|_2^2-\|\tilde{g}\|_2^2}{n}\right| && (173)\\ &\lesssim \frac{\|\tilde{f}-\tilde{g}\|_2}{n\sqrt{k\wedge(m\vee n)}}, && (174)\end{aligned}$$

where we used (154). We have $-\mathbb{E}T_{\mathsf{LF}}^{-\mathsf{d}} \gtrsim \|\tilde{f}-\tilde{g}\|_2^2$ provided $n \gtrsim 1/(\|\tilde{f}-\tilde{g}\|_2\sqrt{k\wedge(m\vee n)}) \asymp \sqrt{\alpha}/\varepsilon$, which we assume from here on. Plugging in the bounds derived above, the test $\mathbb{1}\{T_{\mathsf{LF}}\ge 0\}$ on the transformed observations has type-I probability of error bounded by $1/3$ provided

$$\begin{aligned} \left\|\tilde{f}-\tilde{g}\right\|_2^4 \gtrsim{}& \frac{1}{n}\sqrt{\frac{\alpha}{k}}\|\tilde{f}-\tilde{g}\|_2^2 \\ &+ \frac{1}{m}\frac{\log(k)\alpha}{k}\|\tilde{f}-\tilde{g}\|_2^2 \\ &+ \frac{\alpha}{k}\left(\frac{1}{nm}+\frac{1}{n^2}\right) && (175)\end{aligned}$$

for a small enough implied constant on the left. Looking at each term separately yields the sufficient conditions

$$\underbrace{m \gtrsim \frac{\log(k)\alpha}{\varepsilon^2}}_{(I)} \;\&\; n \gtrsim \frac{\sqrt{k\alpha}}{\varepsilon^2} \;\&\; mn \gtrsim \frac{k\alpha}{\varepsilon^4}. \tag{176}$$

The final step is to check that the sufficient conditions in (176) are implied by what is indicated in the statement of Theorem 2. Recall from the statement of the Theorem, that it states that

$$m \gtrsim \frac{\log(k)}{\varepsilon^2} \;\&\; n \gtrsim \frac{\sqrt{k\alpha}}{\varepsilon^2} \;\&\; mn \gtrsim \frac{k\log(k)\alpha}{\varepsilon^4} \quad (177)$$

is sufficient to successfully perform the test, where we have replaced the generic $\gtrsim_{\log(k)}$ notation with the precise dependence on $\log(k)$ that we require. Note that the only difference between (176) and (177) is the condition on m, that is, the first term in the equations (176) and (177). Suppose now that (177) holds, and let us split this discussion into cases.

1) Suppose $\max\{m,n\} \geq k$. In this case $\alpha = 1$, and (I) is implied by $m \gtrsim \log(k)/\varepsilon^2$. For this the first condition of (177) is clearly sufficient.
2) Suppose $n \leq m \leq k$. In this case $\alpha = k/m$, and (I) is implied by $m \gtrsim \sqrt{k\log(k)}/\varepsilon$. By the third condition of (177) we know that $m^2 n \gtrsim k^2/\varepsilon^4$. Using that $n \leq m$, this implies that $m \gtrsim k^{2/3}/\varepsilon^{4/3}$, which is clearly sufficient.
3) Suppose $m \leq n \leq k$. In this case $\alpha = k/n$, and (I) is implied by $mn \gtrsim k\log(k)/\varepsilon^2$. By the third condition of (177) we know that $mn^2 \gtrsim k^2\log(k)/\varepsilon^4$. After noting that $n \leq k$ we get $mn \gtrsim k\log(k)/\varepsilon^4$, which is sufficient.

The diagonal. See the discussion at the end of the proof for $\mathcal{P}_{\mathsf{Db}}$. ∎

APPENDIX B
LOWER BOUNDS OF THEOREM 1 AND 2

Let $\mathcal{M}(\mathcal{X})$ be the set of all probability measures on some space $\mathcal{X}$, and $\mathcal{P} \subseteq \mathcal{M}(\mathcal{X})$ be some family of distributions. In this section we prove lower bounds for likelihood-free hypothesis testing problems. For clarity, let us formally state the problem as testing between the null hypothesis

$$\left\{ \mathbb{P}_{\mathsf{X}}^{\otimes n} \otimes \mathbb{P}_{\mathsf{Y}}^{\otimes n} \otimes \mathbb{P}_{\mathsf{X}}^{\otimes m} \;:\; \begin{array}{c} \mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}} \in \mathcal{P} \\ \& \; \mathsf{TV}(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}) \geq \varepsilon \end{array} \right\} \quad (178)$$

versus the alternative hypothesis

$$\left\{ \mathbb{P}_{\mathsf{X}}^{\otimes n} \otimes \mathbb{P}_{\mathsf{Y}}^{\otimes n} \otimes \mathbb{P}_{\mathsf{Y}}^{\otimes m} \;:\; \begin{array}{c} \mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}} \in \mathcal{P} \\ \& \; \mathsf{TV}(\mathbb{P}_{\mathsf{X}}, \mathbb{P}_{\mathsf{Y}}) \geq \varepsilon \end{array} \right\} \quad (179)$$

Our strategy for proving lower bounds relies on the following well known result that

$$\begin{aligned} &\inf_{\psi} \max_{i=0,1} \sup_{P \in H_i} P(\psi \neq i) \\ &\geq \frac{1}{2}\left(1 - \mathsf{TV}(\mathbb{E}P_0, \mathbb{E}P_1)\right) - \sum_i \mathbb{P}(P_i \notin H_i), \end{aligned} \quad (180)$$

which we recorded in Lemma 4. The following will also be used multiple times throughout:

Lemma 5 ([39, Lemmas 2.3 and 2.4])**.** *For any probability measures $\mathbb{P}_0, \mathbb{P}_1$,*

$$\begin{aligned} \frac{1}{4}\mathsf{H}^4(\mathbb{P}_0,\mathbb{P}_1) \leq \mathsf{TV}^2(\mathbb{P}_0,\mathbb{P}_1) \leq \mathsf{H}^2(\mathbb{P}_0,\mathbb{P}_1) \\ \leq \mathsf{KL}(\mathbb{P}_0\|\mathbb{P}_1) \leq \chi^2(\mathbb{P}_0\|\mathbb{P}_1). \end{aligned} \quad (181)$$

Note that some of the inequalities in Lemma 5 can be improved, but since such improvements have no effect on our results, we present their simplest available version. The inequalities between TV and H are attributed to Le Cam, while the bound $\mathsf{TV} \leq \sqrt{\mathsf{KL}/2}$ is due to Pinsker. The use of the χ^2-divergence for bounding the total variation distance between mixtures of products was pioneered by Ingster [69], and is sometimes referred to as the *Ingster-trick*.

In our bounds we will also rely on the following simple technical result.

Lemma 6. *Suppose that $a, b, c > 0$ and $N = (N_1, \ldots, N_k) \sim \mathrm{Multinomial}(n, (\frac{1}{k}, \ldots, \frac{1}{k}))$. Then*

$$\mathbb{E}_N \prod_{j \in [k]} (a + b(1+c)^{N_j}) \leq (a + be^{cn/k})^k. \quad (182)$$

Recall that the necessity of $m \gtrsim n_{\mathsf{HT}}(\varepsilon, \mathcal{P})$ and $n \gtrsim n_{\mathsf{GoF}}(\varepsilon, \mathcal{P})$ were shown in Proposition 1. Thus, most of our work lies in obtaining the lower bound on the product mn.

A. The Class $\mathcal{P}_{\mathsf{H}}$

Proposition 11. *For any $\beta > 0, C > 1$ and $d \geq 1$ there exists a finite c independent of ε such that*

$$\begin{aligned} c\left\{ (m,n) \;:\; \begin{array}{c} m \geq 1/\varepsilon^2 \\ \& \; n \geq \varepsilon^{-(2\beta+d/2)/\beta} \\ \& \; mn \geq \varepsilon^{-2(2\beta+d/2)/\beta} \end{array} \right\} \\ \supseteq \mathcal{R}_{\mathsf{LF}}(\varepsilon, \mathcal{P}_{\mathsf{H}}(\beta, d, C)) \end{aligned} \quad (183)$$

for all $\varepsilon \in (0,1)$.

Proof: Adversarial construction. Take a smooth function $h : \mathbb{R}^d \to \mathbb{R}$ supported on of $[0,1]^d$ with $\int_{[0,1]^d} h(x)\mathrm{d}x = 0$ and $\int_{[0,1]^d} h(x)^2\mathrm{d}x = 1$. Let $\kappa \geq 1$ be an integer, and for $j \in [\kappa]^d$ define the scaled and translated functions h_j as

$$h_j(x) = \kappa^{d/2} h(\kappa x - j + 1). \quad (184)$$

Then h_j is supported on the cube $[(j-1)/\kappa, j/\kappa]$ and $\int_{[0,1]^d} h_j(x)^2 \mathrm{d}x = 1$, where we write $j/\kappa = (j_1/\kappa, \ldots, j_d/\kappa)$. Let $\rho > 0$ be small and for each $\eta \in \{-1,0,1\}^{\kappa^d}$ define the function

$$f_\eta(x) = 1 + \rho \sum_{j \in [\kappa]^d} \eta_j h_j(x). \tag{185}$$

In particular, $f_0 = 1$ is the uniform density. Clearly $\int_{[0,1]^d} f_\eta(x)\mathrm{d}x = 1$, and to make it positive we choose ρ, κ such that $\rho\kappa^{d/2}\|h\|_\infty \le 1/2$. By [41], choosing

$$\rho\kappa^{d/2+\beta} \le C/(4\|h\|_{\mathcal{C}^{\lfloor\beta\rfloor}} \vee 2\|h\|_{\mathcal{C}^{\lfloor\beta\rfloor+1}}) \tag{186}$$

ensures that $f_\eta \in \mathcal{P}(\beta, d, C)$. Note also that $\|f_\eta - 1\|_1 = \rho\kappa^{d/2}$. For $\varepsilon \in (0,1)$ we set $\kappa \asymp \varepsilon^{-1/\beta}$ and $\rho \asymp \varepsilon^{(2\beta+d)/(2\beta)}$. These ensure that (186) and $\mathsf{TV}(f_\eta, f_0) \gtrsim \varepsilon$ hold, where as usual the constants may depend on (β, d, C). Noting that $\|\sqrt{f_\eta} - 1\|_2 \asymp \|f_\eta - 1\|_1 \gtrsim \varepsilon$, we immediately obtain that $m \gtrsim 1/\varepsilon^2$ is necessary for testing, by reduction from binary hypothesis testing (32). Observe also that for any η, η',

$$\int_{[0,1]^d} f_\eta(x) f_{\eta'}(x)\mathrm{d}x = 1 + \rho^2\langle \eta, \eta'\rangle \tag{187}$$

which will be used later.

Goodness-of-fit testing. Let η be drawn uniformly at random. We show that $\mathsf{TV}(f_0^{\otimes n}, \mathbb{E} f_\eta^{\otimes n})$ can be made arbitrarily small provided $n \lesssim \varepsilon^{-(2\beta+d/2)/\beta}$, which yields a lower bound on n via reduction from goodness-of-fit testing (34). By Lemma 5 we can focus on bounding the χ^2 divergence. Via Ingster's trick we have

$$\begin{aligned}
&\chi^2(\mathbb{E}_\eta[f_\eta^{\otimes n}], f_0^{\otimes n}) + 1 \\
&\quad = \int_{[0,1]^d \times \cdots \times [0,1]^d} \left(\mathbb{E}_\eta \prod_{i=1}^n f_\eta(x_i)\right)^2 \mathrm{d}x && (188)\\
&\quad = \mathbb{E}_{\eta\eta'} \prod_{i=1}^n \left(\int_{[0,1]^d} f_\eta(x) f_{\eta'}(x)\mathrm{d}x\right), && (189)
\end{aligned}$$

where η, η' are i.i.d.. By (187) and the inequalities $1 + x \le e^x, \cosh(x) \le \exp(x^2)$ for all $x \in \mathbb{R}$, we have

$$\begin{aligned}
&= \mathbb{E}_{\eta\eta'}\left(1 + \rho^2\langle\eta,\eta'\rangle\right)^n && (190)\\
&\le \mathbb{E}_{\eta\eta'} \exp(n\rho^2\langle\eta,\eta'\rangle) && (191)\\
&= \cosh(n\rho^2)^{\kappa^d} && (192)\\
&\le \exp(n^2\rho^4\kappa^d). && (193)
\end{aligned}$$

Thus, goodness-of-fit testing is impossible unless $n \gtrsim \rho^{-2}\kappa^{-d/2} \asymp 1/\varepsilon^{(2\beta+d/2)/\beta}$.

Likelihood-free hypothesis testing. We are now ready to show the lower bound on the product mn. Once again $\eta \in \{\pm 1\}^{\kappa^d}$ is drawn uniformly at random and we apply Lemma 4 with the choices $P_0 = f_\eta^{\otimes n} \otimes f_0^{\otimes n} \otimes f_\eta^{\otimes m}$ against $P_1 = f_\eta^{\otimes n} \otimes f_0^{\otimes n+m}$. Let $\mathbb{P}_{0,XYZ}, \mathbb{P}_{1,XYZ}$ denote the joint distribution of the samples X, Y, Z under the measures $\mathbb{E}P_0, \mathbb{E}P_1$ respectively. By Pinsker's inequality and the chain rule we have

$$\begin{aligned}
&\mathsf{TV}(\mathbb{P}_{0,XYZ}, \mathbb{P}_{1,XYZ})^2 \\
&\quad = \mathsf{TV}(\mathbb{P}_{0,XZ}, \mathbb{P}_{1,XZ})^2 && (194)\\
&\quad \le \mathsf{KL}(\mathbb{P}_{0,XZ} \| \mathbb{P}_{1,XZ}) && (195)\\
&\quad = \mathsf{KL}(\mathbb{P}_{0,Z|X} \| \mathbb{P}_{1,Z|X} | \mathbb{P}_{0,X}) \\
&\qquad + \underbrace{\mathsf{KL}(\mathbb{P}_{0,X} \| \mathbb{P}_{1,X})}_{=0}, && (196)
\end{aligned}$$

where the last line uses that the marginal of X is equal under both measures. Clearly $\mathbb{P}_{1,Z|X}$ is simply $\mathrm{Unif}([0,1]^d)^{\otimes m}$ and $\mathbb{P}_{0,X}, \mathbb{P}_{0,Z|X}$ have densities $\mathbb{E}_\eta f_\eta^{\otimes n}$ and $\mathbb{E}_{\eta|X} f_\eta^{\otimes m}$ respectively. Given X, let η' be an independent copy of η from the posterior given X. By Ingster's trick we have

$$\begin{aligned}
&\mathsf{KL}(\mathbb{P}_{0,Z|X} \| \mathbb{P}_{1,Z|X} | \mathbb{P}_{0,X}) \\
&\quad \le \chi^2(\mathbb{P}_{0,Z|X} \| \mathbb{P}_{1,Z|X} | \mathbb{P}_{0,X}) && (197)\\
&\quad = \mathbb{E}_X \int_{[0,1]^{md}} \mathbb{E}_{\eta|X}\mathbb{E}_{\eta'|X} \prod_{i=1}^m f_\eta(z_i) f_{\eta'}(z_i)\mathrm{d}z \\
&\qquad - 1 && (198)\\
&\quad = -1 + \mathbb{E}_{\eta\eta'}(1 + \rho^2\langle\eta,\eta'\rangle)^m, && (199)
\end{aligned}$$

where the last line uses (187). Let $N = (N_1, \ldots, N_{\kappa^d})$ be the vector of counts indicating the number of X_i that fall into each bin $\{[(j-1)/\kappa, j/\kappa]\}_{j \in [\kappa]^d}$. Clearly $N \overset{d}{\sim} \mathrm{Multinomial}(n, (\frac{1}{\kappa^d}, \ldots, \frac{1}{\kappa^d}))$. Using that $\eta_j\eta'_j$ depends on only those X_i that fall in bin j and the inequality $1 + x \le \exp(x)$ valid for all $x \in \mathbb{R}$, we can write

$$\begin{aligned}
&\chi^2(\mathbb{P}_{0,Z|X} \| \mathbb{P}_{1,Z|X} | \mathbb{P}_{0,X}) + 1 \\
&\quad \le \mathbb{E}_N \mathbb{E}_{\eta\eta'|N} \prod_{j \in [\kappa]^d} \exp(\rho^2 m \eta_j \eta'_j) && (200)\\
&\quad = \mathbb{E}_N \prod_{j \in [\kappa]^d} \mathbb{E}_{\eta_j\eta'_j|N_j} \exp(\rho^2 m \eta_j \eta'_j). && (201)
\end{aligned}$$

We now focus on a particular bin j. Define the bin-conditional densities

$$p_\pm = \kappa^d(1 \pm \rho h_j)\mathbb{1}_{[(j-1)/\kappa, j/\kappa]}, \tag{202}$$

where we drop the dependence on j in the notation. Let $X^{(j)} \triangleq (X_{i_1}, \ldots, X_{i_{N_j}})$ be those X_i that fall in bin j.

Note that $\{i_1, \dots, i_{N_j}\}$ is a uniformly distributed size N_j subset of $[n]$ and given N_j, the density of $X_{i_1}, \dots, X_{i_{N_j}}$ is $\frac{1}{2}(p_+^{\otimes N_j} + p_-^{\otimes N_j})$. We can calculate

$$\begin{aligned}
\mathbb{P}(\eta_j \eta_j' = 1 | N_j) &= \mathbb{E}_{X^{(j)}|N_j} \mathbb{P}(\eta_j \eta_j' = 1 | X^{(j)}) && (203)\\
&= \mathbb{E}_{X^{(j)}|N_j} \Big[\mathbb{P}(\eta_j = 1 | X^{(j)})^2 \\
&\qquad + \mathbb{P}(\eta_j = -1 | X^{(j)})^2 \Big] && (204)\\
&= \mathbb{E}_{X^{(j)}|N_j} \left[\frac{\frac{1}{4}(p_+^{\otimes N_j})^2 + \frac{1}{4}(p_-^{\otimes N_j})^2}{\frac{1}{4}(p_+^{\otimes N_j} + p_-^{\otimes N_j})^2} \right] && (205)\\
&= \frac{1}{2} + \frac{1}{4} \Big(\chi^2(p_+^{\otimes N_j} \| \frac{1}{2}(p_+^{\otimes N_j} + p_-^{\otimes N_j})) \\
&\qquad + \chi^2(p_-^{\otimes N_j} \| \frac{1}{2}(p_+^{\otimes N_j} + p_-^{\otimes N_j})) \Big). && (206)
\end{aligned}$$

By convexity of the χ^2 divergence in its arguments and tensorization, we have

$$\begin{aligned}
\mathbb{P}(\eta_j \eta_j' = 1 | N_j) &\le \frac{1}{2} + \frac{1}{8} \Big(\chi^2(p_+^{\otimes N_j} \| p_-^{\otimes N_j}) \\
&\qquad + \chi^2(p_-^{\otimes N_j} \| p_+^{\otimes N_j}) \Big) && (207)\\
&= \frac{1}{4} + \frac{1}{8} \sum_{\omega \in \{\pm 1\}} \left(\kappa^d \int_{[\frac{j-1}{\kappa}, \frac{j}{\kappa}]} \frac{(1 + \omega \rho h_j(x))^2}{1 - \omega \rho h_j(x)} \mathrm{d}x \right)^{N_j}. && (208)
\end{aligned}$$

Using that $\rho \|h_j\|_\infty \le 1/2$ by construction, we have

$$\begin{aligned}
&\int_{[(j-1)/\kappa, j/\kappa]} \frac{(1 + \rho h_j(x))^2}{1 - \rho h_j(x)} \mathrm{d}x \\
&\quad = \frac{1}{\kappa^d} + \int_{[\frac{j-1}{\kappa}, \frac{j}{\kappa}]} \frac{4\rho^2 h_j^2(x)}{1 - \rho h_j(x)} \mathrm{d}x && (209)\\
&\quad \le \frac{1}{\kappa^d} + 8\rho^2. && (210)
\end{aligned}$$

The same bound is obtained for the other integral term. We get

$$\begin{aligned}
&\chi^2(\mathbb{P}_{0,Z|X} \| \mathbb{P}_{1,Z|X} | \mathbb{P}_{0,X}) + 1 \\
&\le \mathbb{E}_N \prod_{j \in [\kappa]^d} \Big(\frac{\sinh(\rho^2 m)}{2} \big(1 + (1 + 8\rho^2 \kappa^d)^{N_j}\big) \\
&\qquad + e^{-\rho^2 m} \Big) = (\dagger). && (211)
\end{aligned}$$

The final step is to apply Lemma 6 to pass the expectation through the product. Assuming that $m \vee n \lesssim \rho^{-2} \asymp \varepsilon^{-(2\beta+d)/\beta}$ for a small enough implied constant, using the inequalities $e^x \le 1 + x + x^2, 1 - x \le e^{-x} \le 1 - x + x^2/2$ valid for all $x \in [0,1]$, and Lemma 6, we obtain

$$\begin{aligned}
(\dagger) &\le (e^{-\rho^2 m} + \frac{\sinh(\rho^2 m)}{2} (1 + e^{8\rho^2 n}))^{\kappa^d} && (212)\\
&\le (1 + c\rho^4 mn)^{\kappa^d} && (213)\\
&\le \exp(c\rho^4 \kappa^d mn) && (214)
\end{aligned}$$

for a universal constant $c > 0$. Therefore, if $m \vee n \lesssim \varepsilon^{-(2\beta+d)/\beta}$ likelihood-free hypothesis testing is impossible unless $mn \gtrsim \rho^{-4}\kappa^{-d} \asymp 1/\varepsilon^{2(2\beta+d/2)/\beta}$.

Suppose now that $m \vee n \gtrsim \varepsilon^{-(2\beta+d)/\beta}$ instead. We have two cases:

1) If $n \gtrsim \varepsilon^{-(2\beta+d)/\beta}$ then from Proposition 7 we know that $m \asymp 1/\varepsilon^2$ is enough for achievability. However, by the first part of the proof we know that $m \gtrsim 1/\varepsilon^2$ must always hold, which provides the matching lower bound in this case.
2) If $m \gtrsim \varepsilon^{-(2\beta+d)/\beta}$ then we can assume $m \gtrsim n$ also holds, otherwise the first case above would apply. From the goodness-of-fit testing lower bound we know that $n \gtrsim \varepsilon^{-(2\beta+d/2)/\beta}$ must always hold, and from Proposition 7 we know that $(m, n) \asymp (\varepsilon^{-(2\beta+d/2)/\beta}, \varepsilon^{-(2\beta+d/2)/\beta})$ is achievable, so we get matching bounds in this case too.

Summarizing, we've shown that for succesful testing $m \gtrsim 1/\varepsilon^2, n \gtrsim 1/\varepsilon^{(2\beta+d/2)/\beta}$ and $mn \gtrsim \varepsilon^{-2(2\beta+d/2)/\beta}$ must hold, which concludes our proof. ∎

B. The Class $\mathcal{P}_{\mathsf{G}}$

Proposition 12. *For any $s, C > 0$ there exists a finite constant c independent of ε such that*

$$c \left\{ (m, n) : \begin{array}{c} m \ge 1/\varepsilon^2 \\ \& \ n \ge \varepsilon^{-(2s+1/2)/s} \\ \& \ mn \ge \varepsilon^{-2(2s+1/2)/s} \end{array} \right\} \supseteq \mathcal{R}_{\mathsf{LF}}(\varepsilon, \mathcal{P}_{\mathsf{G}}(s, C)) \quad (215)$$

for all $\varepsilon \in (0,1)$.

Proof: Adversarial construction. Let $\gamma \in \ell^1$ be a non-negative sequence, and let $\theta \sim \otimes_{k=1}^\infty \mathcal{N}(0, \gamma_k)$. Define the random measure $\mu_\theta = \otimes_{j=1}^\infty \mathcal{N}(\theta_j, 1)$. Let $\varepsilon \in (0,1)$ be given. For our proofs we use

$$\gamma_k = \begin{cases} c_1 \varepsilon^{(2s+1)/s} & \text{for } 1 \le k \le c_2 \varepsilon^{-1/s} \\ 0 & \text{otherwise} \end{cases} \quad (216)$$

for appropriate constants c_1, c_2. Recall our definition of the Sobolev ellipsoid $\mathcal{E}(s, C)$ with associated sobolev norm $\|\cdot\|_s$. We have

$$(\mathbb{E}\|\theta\|_s)^2 \le \mathbb{E}\sum_{j=1}^{\infty} j^{2s}\theta_i^2 = \|\sqrt{\gamma}\|_s^2 \tag{217}$$

$$= c_1\varepsilon^{(2s+1)/s}\sum_{j=1}^{c_2\varepsilon^{-1/s}} j^{2s} \le c_1 c_2^{2s+1} \tag{218}$$

$$\mathsf{TV}(\mathbb{P}_\gamma, \mathbb{P}_0) \ge \frac{1 \wedge \|\theta\|_2}{200}, \tag{219}$$

where last line holds by [70, Theorem 1.2].

First, we need to verify that our construction is valid, that is, that $\mathbb{P}_\gamma \in \mathcal{P}_{\mathsf{G}}(s, C)$ and $\mathsf{TV}(\mathbb{P}_\gamma, \mathbb{P}_0) \ge \varepsilon$ with high probability. For standard Gaussian $Z \sim \mathcal{N}(0, 1)$ it holds that

$$\mathbb{E}\exp\left(\lambda(Z^2 - 1)\right) \le \exp(2\lambda^2) \tag{220}$$

for all $|\lambda| \le 1/4$. Therefore, for a sequence of independent standard Gaussians $Z_1, Z_2, \ldots$ we get

$$\mathbb{E}\exp\left(\lambda\sum_{j=1}^{\infty}\gamma_j(Z_j^2 - 1)\right) \le \exp(2\lambda^2\|\gamma\|_2^2) \tag{221}$$

for all $|\lambda| \le \min_j(4\gamma_j)^{-1} = c_1^{-1}\varepsilon^{-(2s+1)/s}/4$. Assuming that $c_1\varepsilon^{(2s+1)/s} \le \|\gamma\|_2$, standard sub-Exponential concentration bounds imply that there exists a universal constant $c_3 > 0$ such that

$$\mathbb{P}(\|\theta\|_2^2 - \mathbb{E}\|\theta\|_2^2 \le -t) \le \exp(-\frac{c_3 t}{\|\gamma\|_2}) \tag{222}$$

for all $t \ge 0$. Since $\mathbb{E}\|\theta\|_2^2 = \|\gamma\|_1 = c_1c_2\varepsilon^2$, and $\|\gamma\|_2^2 = c_2c_1^2\varepsilon^{\frac{4s+1}{s}}$, we can set $t = \frac{1}{2}\|\theta\|_2^2$ to get

$$\mathbb{P}(\|\theta\|_2^2 \le \frac{1}{2}c_1c_2\varepsilon^2) \le \exp(-\frac{1}{2}c_3\sqrt{c_2}\varepsilon^{-1/(2s)}). \tag{223}$$

Now choose c_1 and c_2 to satisfy

$$100c_1c_2^{2s+1} = C \qquad \text{and} \qquad c_1c_2 = 2. \tag{224}$$

and ε small enough to satisfy

$$c_1\varepsilon^{(2s+1)/s} \le \|\gamma\|_2 = \sqrt{c}_1c_1\varepsilon^{(2s+1/2)/s}, \tag{225}$$

$$\text{and} \quad \frac{1}{2}c_3\sqrt{c}_2\varepsilon^{-1/(2s)} \ge \log(100). \tag{226}$$

Long story short, these conditions ensure that $\mathbb{P}(\mu_\gamma \in \mathcal{P}_{\mathsf{G}}(s, C), \mathsf{TV}(\mu_\gamma, \mu_0) \ge \varepsilon) \ge 0.98$ for all ε small enough in terms of C and s, and therefore we can proceed to computation using Lemma 4.

Note that we immediately get the binary hypothesis testing lower bound $m \gtrsim 1/\varepsilon^2$ via our reduction (34), as $\mathsf{H}(\mu_0, \mu_{\sqrt{\gamma}}) \asymp \mathsf{TV}(\mu_0, \mu_{\sqrt{\gamma}}) = \sqrt{2}\varepsilon$ by Lemma 2 and the choice (224).

Goodness-of-fit testing. We show that $\mathsf{TV}(\mu_0^{\otimes n}, \mathbb{E}\mu_\gamma^{\otimes n})$ can be made arbitrarily small as long as $n \lesssim 1/\varepsilon^{(2s+1/2)/s}$, which yields a lower bound on n via reduction from goodness-of-fit testing (34). Let us compute the distribution $\mathbb{E}\mu_\gamma^{\otimes n}$. By independence clearly $\mathbb{E}\mu_\gamma^{\otimes n} = \otimes_{k=1}^{\infty}\mathbb{E}_{\theta\sim\mathcal{N}(0,\gamma_k)}\mathcal{N}(\theta, 1)^{\otimes n}$. Focusing on the inner term and and dropping the subscript k, for the density we have

$$\mathbb{E}_{\theta\sim\mathcal{N}(0,\gamma)}\left[\frac{1}{(2\pi)^{n/2}}\exp\left(-\frac{1}{2}\sum_{j=1}^{n}(x_j - \theta)^2\right)\right] \propto \exp\left(-\frac{\|x\|_2^2}{2}\right)\mathbb{E}\exp\left(-\frac{n}{2}(\theta^2 - 2\theta\bar{x})\right), \tag{227}$$

where we write $\bar{x} \triangleq \frac{1}{n}\sum_j x_j$. Looking at just the term involving θ, we have

$$\mathbb{E}\exp\left(-\frac{n}{2}(\theta^2 - 2\theta\bar{x})\right) \propto \int \exp\left(-\frac{1}{2}(\theta^2(n + \frac{1}{\gamma}) - 2\theta n\bar{x})\right)\mathrm{d}\theta \tag{228}$$

$$\propto \exp\left(\frac{1}{2}\frac{n^2\bar{x}^2}{n + \frac{1}{\gamma}}\right). \tag{229}$$

Putting everything together, we see that $\mathbb{E}\mu_\gamma^{\otimes n} = \otimes_{k=1}^{\infty}\mathcal{N}(0, \Theta_k^{-1})$, where

$$\Theta_k \triangleq I_n - \frac{\gamma_k}{1 + n\gamma_k}\mathbb{1}_n\mathbb{1}_n^{\mathsf{T}}. \tag{230}$$

Thus, using Lemma 5 we obtain

$$\mathsf{TV}^2(\mu_0^{\otimes n}, \mathbb{E}\mu_\gamma^{\otimes n}) \le \sum_{k=1}^{\infty}\mathsf{KL}\left(\mathcal{N}(0, I_n)\|\,\mathcal{N}\left(0, \Theta_k^{-1}\right)\right) \tag{231}$$

$$= \frac{1}{2}\sum_{k=1}^{\infty}\left(-\frac{n\gamma_k}{n\gamma_k + 1} + \log(1 + n\gamma_k)\right) \tag{232}$$

$$\le \frac{1}{2}\sum_{k=1}^{\infty}\frac{n^2\gamma_k^2}{1 + n\gamma_k} \lesssim \sum_{k=1}^{\infty} n^2\gamma_k^2. \tag{233}$$

Taking γ as in (216) gives

$$\mathsf{TV}^2(\mu_0^{\otimes n}, \mathbb{E}\mu_\gamma^{\otimes n}) \lesssim n^2\varepsilon^{2(2s+1/2)/s}. \tag{234}$$

Therefore, goodness-of-fit testing is impossible unless $n \gtrsim 1/\varepsilon^{(2s+1/2)/s}$ as desired.

Likelihood-free hypothesis testing. We apply Lemma 4 with measures $P_0 = \mu_\gamma^{\otimes n} \otimes \mu_0^{\otimes n} \otimes \mu_\gamma^{\otimes m}$ and $P_1 = \mu_\gamma^{\otimes n} \otimes \mu_0^{\otimes n} \otimes \mu_0^{\otimes m}$. Define the matrices

$$\Theta_0 = \begin{pmatrix} \mathbb{1}_n\mathbb{1}_n^\mathsf{T} & 0 & \mathbb{1}_n\mathbb{1}_m^\mathsf{T} \\ 0 & 0 & 0 \\ \mathbb{1}_m\mathbb{1}_n^\mathsf{T} & 0 & \mathbb{1}_m\mathbb{1}_m^\mathsf{T} \end{pmatrix}$$
$$\Theta_1 = \begin{pmatrix} \mathbb{1}_n\mathbb{1}_n^\mathsf{T} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

By an analogous calculation to that in the previous paragraph, we obtain

$$\mathbb{E}P_0 = \bigotimes_{k=1}^\infty \mathcal{N}\left(0, \left(I_{2n+m} - \frac{\Theta_0}{n+m+\frac{1}{\gamma_k}}\right)^{-1}\right) \quad (235)$$
$$\triangleq \otimes_{k=1}^\infty \mathcal{N}(0, \Sigma_{0k}) \quad (236)$$
$$\mathbb{E}P_1 = \bigotimes_{k=1}^\infty \mathcal{N}\left(0, \left(I_{2n+m} - \frac{\Theta_1}{n+\frac{1}{\gamma_k}}\right)^{-1}\right) \quad (237)$$
$$\triangleq \otimes_{k=1}^\infty \mathcal{N}(0, \Sigma_{1k}). \quad (238)$$

By the Sherman-Morrison formula, we have

$$\Sigma_{0k} = I_{2n+m} + \gamma_k \begin{pmatrix} \mathbb{1}_n\mathbb{1}_n^\mathsf{T} & 0 & \mathbb{1}_n\mathbb{1}_m^\mathsf{T} \\ 0 & 0 & 0 \\ \mathbb{1}_m\mathbb{1}_n^\mathsf{T} & 0 & \mathbb{1}_m\mathbb{1}_m^\mathsf{T} \end{pmatrix} \quad (239)$$

Therefore, by Pinsker's inequality and the closed form expression for the KL-divergence between centered Gaussians, we obtain

$$\mathsf{TV}^2(\mathbb{E}P_0, \mathbb{E}P_1) \le \mathsf{KL}(\mathbb{E}P_0 \| \mathbb{E}P_1) \quad (240)$$
$$= \frac{1}{2}\sum_{k=1}^\infty \left(\gamma_k m - \log\left(\frac{\gamma_k(n+2m)+1}{\gamma_k(n+m)+1}\right)\right). \quad (241)$$

Once again we choose γ as in (216). Using the inequality $\log(1+x) \ge x - x^2$ valid for all $x \ge 0$ we obtain

$$\mathsf{TV}^2(\mathbb{E}P_0, \mathbb{E}P_1) \lesssim \varepsilon^{-2(2s+1/2)/s}(m^2 + mn). \quad (242)$$

Therefore, likelihood-free hypothesis testing is impossible unless $m \gtrsim \varepsilon^{-(2s+1/2)/s}$ or $nm \gtrsim \varepsilon^{-2(2s+1/2)/s}$. Note that we already have the lower bound $n \gtrsim \varepsilon^{-(2s+1/2)/s}$ by reduction from goodness-of-fit testing (34), so that $m \gtrsim \varepsilon^{-(2s+1/2)/s}$ automatically implies $nm \gtrsim \varepsilon^{-2(2s+1/2)/s}$. Combining everything we get the desired bounds. ∎

C. The Classes $\mathcal{P}_{\mathsf{Db}}$ and $\mathcal{P}_{\mathsf{D}}$

Our first result in this section derives tight minimax lower bounds for the class $\mathcal{P}_{\mathsf{Db}}$. Since $\mathcal{P}_{\mathsf{D}} \supset \mathcal{P}_{\mathsf{Db}}$ these lower bounds immediately carry over to the larger class. However, to get tight lower bounds for all regimes for $\mathcal{P}_{\mathsf{D}}$, we have to prove additional results in Propositions 14 and 15 below.

Proposition 13. *For any $C > 1$ there exists a finite constant c independent of ε and k, such that*

$$c\left\{(m,n): \begin{array}{c} m \ge 1/\varepsilon^2 \\ \& \ n \ge \sqrt{k}/\varepsilon^2 \\ \& \ mn \ge k/\varepsilon^4 \end{array}\right\} \supseteq \mathcal{R}_{\mathsf{LF}}(\varepsilon, \mathcal{P}_{\mathsf{Db}}(k, C)) \supseteq \mathcal{R}_{\mathsf{LF}}(\varepsilon, \mathcal{P}_{\mathsf{D}}(k)) \quad (243)$$

for all $\varepsilon \in (0,1)$ and $k \ge 2$.

Proof: The second inclusion is trivial. For the first inclusion we proceed analogously to the case of $\mathcal{P}_{\mathsf{H}}$.
Adversarial construction. Let k be an integer and $\varepsilon \in (0,1)$. For $\eta \in \{-1,1\}^k$ define the distribution p_η on $[2k]$ by

$$p_\eta(2j-1) = \frac{1}{2k}(1+\eta_j\varepsilon) \quad (244)$$
$$p_\eta(2j) = \frac{1}{2k}(1-\eta_j\varepsilon), \quad (245)$$

for $j \in [k]$. Clearly $\mathsf{H}(p_\eta, p_0) \asymp \mathsf{TV}(p_\eta, p_0) = \varepsilon$, where $p_0 = \mathrm{Unif}[2k]$, so that by reduction from binary hypothesis testing (34) we get the lower bound $m \gtrsim 1/\varepsilon^2$. Observe also that for any $\eta, \eta' \in \{\pm 1\}^k$,

$$\sum_{j\in[2k]} p_\eta(j)p_{\eta'}(j) = \frac{1}{2k}\left(1 + \frac{\varepsilon^2\langle\eta,\eta'\rangle}{k}\right). \quad (246)$$

Goodness-of-fit testing. Let η be uniformly random. We show that $\mathsf{TV}(p_0^{\otimes n}, \mathbb{E}p_\eta^{\otimes n})$ can be made arbitrarily small as long as $n \lesssim \sqrt{k}/\varepsilon^2$, which yields the corresponding lower bound on n by reduction from goodness-of-fit testing (34). Once again, by Lemma 5 we focus on the χ^2 divergence. We have

$$\chi^2(\mathbb{E}p_\eta^{\otimes n} \| p_0^{\otimes n}) + 1$$
$$= (2k)^n \sum_{j\in[2k]^n} \mathbb{E}_{\eta\eta'} \prod_{i=1}^n p_\eta(j_i)p_{\eta'}(j_i) \quad (247)$$
$$= \mathbb{E}_{\eta\eta'}\left(1 + \frac{\varepsilon^2\langle\eta,\eta'\rangle}{k}\right)^n \quad (248)$$
$$\le \exp(n^2\varepsilon^4/k) \quad (249)$$

where the penultimate line follows from (246) and the last line via the same argument as in B-A. Thus, goodness-of-fit testing is impossible unless $n \gtrsim \sqrt{k}/\varepsilon^2$.

Likelihood-free hypothesis testing. We apply Lemma 4 with the two random measures $P_0 = p_\eta^{\otimes n} \otimes p_0^{\otimes n} \otimes p_\eta^{\otimes m}$ and $P_1 = p_\eta^{\otimes n} \otimes p_0^{\otimes (n+m)}$. Analogously to the case of $\mathcal{P}_{\mathsf{H}}$, let $\mathbb{P}_{0,XYZ}, \mathbb{P}_{1,XYZ}$ respectively denote the distribution of the observations X, Y, Z under $\mathbb{E}P_0, \mathbb{E}P_1$ respectively. As for $\mathcal{P}_{\mathsf{H}}$, we have

$$\begin{aligned}
&\mathsf{TV}^2(\mathbb{P}_{0,XYZ}, \mathbb{P}_{1,XYZ}) \\
&\qquad \le \mathsf{KL}(\mathbb{P}_{0,XYZ} \| \mathbb{P}_{1,XYZ}) && (250)\\
&\qquad \le \mathsf{KL}(\mathbb{P}_{0,Z|X} \| \mathbb{P}_{1,Z|X} | \mathbb{P}_{0,X}). && (251)
\end{aligned}$$

For any X the distribution $\mathbb{P}_{1,Z|X}$ is uniform, and $\mathbb{P}_{0,Z|X}, \mathbb{P}_{0,X}$ have pmf $\mathbb{E}_{\eta|X} p_\eta^{\otimes m}$ and $\mathbb{E}_\eta p_\eta^{\otimes n}$ respectively. Once again, by Lemma 5 we may turn our attention to the χ^2-divergence. Given X, let η' have the same distribution as η and be independent of it. Then

$$\begin{aligned}
&\chi^2(\mathbb{P}_{0,Z|X} \| \mathbb{P}_{1,Z|X} | \mathbb{P}_{0,X}) + 1 \\
&\quad = (2k)^m \mathbb{E}_X \sum_{j \in [2k]^m} \mathbb{E}_{\eta|X} \mathbb{E}_{\eta'|X} \prod_{i=1}^n p_\eta(j_i) p_{\eta'}(j_i) && (252)\\
&\quad = \mathbb{E}_{\eta\eta'} \left(1 + \frac{\varepsilon^2 \langle \eta, \eta' \rangle}{k}\right)^m && (253)\\
&\quad \le \mathbb{E}_{\eta\eta'} \prod_{j \in [k]} \exp\left(\frac{\varepsilon^2 m \eta_j \eta_j'}{k}\right), && (254)
\end{aligned}$$

where we used Lemma 246. Let $N = (N_1, \dots, N_k)$ be the vector of counts indicating the number of the $X_1, \dots, X_n$ that fall into the bins $\{2j-1, 2j\}$ for $j \in [k]$. Clearly $N \sim \mathrm{Mult}(n, (\frac{1}{k}, \dots, \frac{1}{k}))$. Let us focus on a specific bin $\{2j-1, 2j\}$ and define the bin-conditional pmf

$$p_\pm(x) = \begin{cases} \frac{1}{2}(1 \pm \varepsilon) & \text{if } x = 2j-1, \\ \frac{1}{2}(1 \mp \varepsilon) & \text{if } x = 2j \\ 0 & \text{otherwise,} \end{cases} \qquad (255)$$

where we drop the dependence on j in the notation. Let $X_{i_1}, \dots, X_{i_{N_j}}$ be the N_j observations falling in $\{2j-1, 2j\}$. Given N_j, the pmf of $X_{i_1}, \dots, X_{i_{N_j}}$ is $\frac{1}{2}(p_+^{\otimes N_j} + p_-^{\otimes N_j})$. We have $\eta_j \eta_j' \in \{\pm 1\}$ almost surely, and analogously to Section B-A we may compute

$$\begin{aligned}
&\mathbb{P}(\eta_j \eta_j' = 1 | N_j) = \mathbb{E}_{X|N_j} \mathbb{P}(\eta_j \eta_j' = 1 | X) && (256)\\
&\quad = \mathbb{E}_{X|N_j} \left[\mathbb{P}(\eta_j = 1|X)^2 + \mathbb{P}(\eta_j = -1|X)^2\right] && (257)\\
&\quad = \frac{1}{2} + \frac{1}{4}\Big(\chi^2(p_+^{\otimes N_j} \| \frac{1}{2}(p_+^{\otimes N_j} + p_-^{\otimes N_j})) \\
&\qquad\qquad + \chi^2(p_-^{\otimes N_j} \| \frac{1}{2}(p_+^{\otimes N_j} + p_-^{\otimes N_j}))\Big) && (258)\\
&\quad \le \frac{1}{2} + \frac{1}{8}\Big(\chi^2(p_-^{\otimes N_j} \| p_+^{\otimes N_j}) \\
&\qquad\qquad + \chi^2(p_+^{\otimes N_j} \| p_-^{\otimes N_j})\Big). && (259)
\end{aligned}$$

We can bound the two χ^2-divergences by

$$\begin{aligned}
\chi^2(p_\pm^{\otimes N_j} \| p_\mp^{\otimes N_j}) + 1 &= \left(\frac{1 + \frac{3}{2}\varepsilon^2}{1 - \varepsilon^2}\right)^{N_j} && (260)\\
&\le (1 + 3\varepsilon^2)^{N_j}, && (261)
\end{aligned}$$

provided $\varepsilon \le c$ for some universal constant $c > 0$. Using Lemma 6, we obtain the bound

$$\begin{aligned}
&\mathbb{E}_N \prod_{j \in [k]} \mathbb{E}_{\eta\eta'|N_j} \exp\left(\frac{\varepsilon^2 m \eta_j \eta_j'}{k}\right) \\
&\quad \le \mathbb{E}_N \prod_{j \in [k]} \left(\frac{\sinh\left(\frac{\epsilon^2 m}{k}\right)}{2}(1 + (1 + 2\varepsilon^2)^{N_j}) \right. \\
&\qquad\qquad \left. + \exp\left(-\frac{\varepsilon^2 m}{k}\right)\right) && (262)\\
&\quad \le \left(\frac{\sinh\left(\frac{\epsilon^2 m}{k}\right)}{2}(1 + e^{\frac{2\varepsilon^2 n}{k}}) + e^{-\frac{\varepsilon^2 m}{k}}\right)^k. && (263)
\end{aligned}$$

Now, under the assumption that $m \vee n \lesssim k/\varepsilon^2$ for some small enough implied constant, the above can be further bounded by

$$\begin{aligned}
&\le (1 + c\frac{\varepsilon^4 mn}{k^2})^k && (264)\\
&\le \exp(\frac{c\varepsilon^4 mn}{k}), && (265)
\end{aligned}$$

for a universal constant $c > 0$. In other words, for $n \vee m \lesssim k/\varepsilon^2$ likelihood-free hypothesis testing is impossible unless $mn \gtrsim k/\varepsilon^4$. The treatment of the case $m \vee n \gtrsim k/\varepsilon^2$ is straightforward, and entirely analogous to our discussion at the end of the proof of Proposition 11, so we won't repeat it here. This completes the proof. ∎

This takes care of the class $\mathcal{P}_{\mathsf{Db}}$. To prove tight bounds for $\mathcal{P}_{\mathsf{D}}$ in the large k regime, we have to work harder. Our second lower bound, Proposition 14 below, proves tight bounds in the regime $n \le m$ and follows by reduction to two-sample testing Proposition 1.

Proposition 14. *There exists a finite constant c independent of ε and k,*

$$c\left\{(m,n):\begin{array}{c} m\geq 1/\varepsilon^2 \\ \& \ n^2m\geq k^2/\varepsilon^4 \\ \& \ n\leq m\end{array}\right\} \supseteq \mathcal{R}_{\mathsf{LF}}(\varepsilon,\mathsf{TV},\mathcal{P}_{\mathsf{D}})\cap\mathbb{N}^2_{n\leq m} \quad (266)$$

for all $k\geq 2, \varepsilon\in(0,2)$, where $\mathbb{N}^2_{n\leq m}=\{(n,m)\in\mathbb{N}^2 : n\leq m\}$.

Proof: Follows from (37) and the lower bound construction in [26]. ∎

1) Valiant's Wishful Thinking Theorem: For our third and final lower bound, which is tight in the regime $m\leq n$, we apply a method developed by Valiant, which we describe below.

Definition 6. For distributions $p_1,\dots,p_\ell$ on $[k]$ and $(n_1,\dots,n_\ell)\in\mathbb{N}^\ell$, we define the $(n_1,\dots,n_\ell)$-based moments of $(p_1,\dots,p_\ell)$ as

$$m(a_1,\dots,a_\ell)=\sum_{i=1}^{k}\prod_{j=1}^{\ell}(n_jp_j(i))^{a_j} \quad (267)$$

for $(a_1,\dots,a_\ell)\in\mathbb{N}^\ell$.

Let $p^+=(p_1^+,\dots,p_\ell^+)$ and $p^-=(p_1^-,\dots,p_\ell^-)$ be ℓ-tuples of distributions on $[k]$ and suppose we observe samples $\{X^{(i)}\}_{i\in[\ell]}$, where the number of observations in $X^{(i)}$ is $\mathrm{Poi}(n_i)$. Let $H^\pm$ denote the hypothesis that the samples came from $p^\pm$, up to an arbitrary relabeling of the alphabet $[k]$. It can be shown that to test H^+ against H^-, we may assume without loss of generality that our test is invariant under relabeling of the support, or in other words, is a function of the *fingerprints*. The fingerprint f of a sample $\{X^{(i)}\}_{i\in[\ell]}$ is the function $f:\mathbb{N}^\ell\to\mathbb{N}$ which given $(a_1,\dots,a_\ell)\in\mathbb{N}^\ell$ counts the number of bins in $[k]$ which have exactly a_i occurences in the sample $X^{(i)}$.

Theorem 5 ([63, Wishful thinking theorem])**.** *Suppose that $|p_i^\pm|_\infty\leq\eta/n_i$ for all $i\in[\ell]$ for some $\eta>0$, and let m^+ and m^- denote the $(n_1,\dots,n_\ell)$-based moments of p^+,p^- respectively. Let $f^\pm$ denote the distribution of the fingerprint under $H^\pm$ respectively. Then*

$$\mathsf{TV}(f^+,f^-)\leq 2(e^{\eta\ell}-1) + e^{\ell(\eta/2+\log 3)}\sum_{a\in\mathbb{N}^\ell}\frac{|m^+(a)-m^-(a)|}{\sqrt{1+m^+(a)\vee m^-(a)}}. \quad (268)$$

Proof: The proof is a straightforward adaptation of [63] and thus we omit it. ∎

Although Theorem 5 assumes a random (Poisson distributed) number of samples, the results carry over to the deterministic case with no modification, due to the subexponential concentration of the Poisson distribution. We are ready to prove our likelihood-free hypothesis testing lower bound using Theorem 5.

Proposition 15. *There exists a finite constant c independent of ε and k, such that*

$$c\left\{(m,n):\begin{array}{c} m\geq 1/\varepsilon^2 \\ \& \ n^2m\geq k^2/\varepsilon^4 \\ \& \ m\leq n\end{array}\right\} \supseteq \mathcal{R}_{\mathsf{LF}}(\varepsilon,\mathsf{TV},\mathcal{P}_{\mathsf{D}})\cap\mathbb{N}^2_{m\leq n} \quad (269)$$

for all $\varepsilon\in(0,1)$ and $k\geq 2$, where $\mathbb{N}^2_{m\leq n}=\{(n,m)\in\mathbb{N}^2 : m\leq n\}$.

Proof: We focus on the regime $n\leq k$, as otherwise the result is subsumed by Proposition 13. Suppose that $\varepsilon\in(0,1/2)$, $\eta=0.01$ (say) and $n/\eta\leq k/2$. Define $\gamma=n/\eta$ and let p,q be pmfs on $[k]$ with weight $(1-\varepsilon)/\gamma$ on $[\gamma]$ and $k/4$ light elements with weight $4\varepsilon/k$ on $[k/2,3k/4]$ and $[3k/4,k]$ respectively. To apply Valiant's wishful thinking theorem, we take $p^+=(p,q,p)$ and $p^-=(p,q,q)$ with corresponding hypotheses $H^\pm$. The (n,n,m)-based moments of $p^\pm$ are given by

$$\frac{1}{n^{a+b}m^c}m^+(a,b,c) \quad (270)$$
$$=\begin{cases} k, \text{ if } a+c=0 \text{ and } b=0, \\ \left(\frac{1-\varepsilon}{\alpha}\right)^{a+b+c}\alpha, \text{ if } a+c\geq 1 \text{ and } b\geq 1, \\ \left(\frac{1-\varepsilon}{\alpha}\right)^{a+b+c}\alpha+\left(\frac{4\varepsilon}{k}\right)^{a+b+c}\frac{k}{4}, \text{ otherwise, and}\end{cases}$$

$$\frac{1}{n^{a+b}m^c}m^-(a,b,c) \quad (271)$$
$$=\begin{cases} k, \text{ if } a=0 \text{ and } b+c=0, \\ \left(\frac{1-\varepsilon}{\alpha}\right)^{a+b+c}\alpha, \text{ if } a\geq 1 \text{ and } b+c\geq 1, \\ \left(\frac{1-\varepsilon}{\alpha}\right)^{a+b+c}\alpha+\left(\frac{4\varepsilon}{k}\right)^{a+b+c}\frac{k}{4}, \text{ otherwise.}\end{cases}$$

By the wishful thinking theorem we know that

$$\mathsf{TV}(f^+,f^-)\leq 0.061 \quad (272)$$
$$+27.41\sum_{a,b,c\in\mathbb{N}}\frac{|m^+(a,b,c)-m^-(a,b,c)|}{\sqrt{1+\max(m^+,m^-)}}. \quad (273)$$

Let us consider the possible values of $|m^+(a,b,c)-m^-(a,b,c)|$. It is certainly zero if $a\wedge b\geq 1$ or

$a=b=c=0$. Suppose that $a=0$ so that necessarily $b+c\geq 1$. Then

$$\frac{1}{n^b m^c}|m^+(0,b,c)-m^-(0,b,c)| = \left(\frac{4\varepsilon}{k}\right)^{b+c}\frac{k}{4}\mathbb{1}(\min\{b,c\}\geq 1). \tag{274}$$

Using the symmetry between a and b and that $1+m^+\vee m^-\geq n^b m^c((1-\varepsilon)/\gamma)^{b+c}\gamma$ (for $m^+\neq m^-$), we can bound the infinite sum above as

$$\lesssim \sum_{b,c\geq 1}\frac{n^b m^c k^{1-(b+c)}\varepsilon^{b+c}}{\sqrt{n^b m^c \gamma^{1-(b+c)}(1-\varepsilon)^{b+c}}} \tag{275}$$

$$\lesssim \sum_{b,c\geq 1} n^{b/2}m^{c/2}\left(\frac{\sqrt{\gamma}}{k}\right)^{b+c-1}\varepsilon^{b+c} \tag{276}$$

Plugging in $\gamma=n/\eta\asymp n$, and using $m\leq n\leq k$, we obtain

$$\begin{aligned}
&\mathsf{TV}(f^+,f^-)-0.061\\
&\lesssim \sum_{b,c\geq 1} n^{b+\frac{c}{2}-\frac{1}{2}}m^{c/2}\frac{1}{k^{b+c-1}}\varepsilon^{b+c} &(277)\\
&= \frac{n\sqrt{m}\varepsilon^2}{k}\sum_{b,c\geq 0}\left(\frac{n}{k}\right)^{b+\frac{c}{2}}\left(\frac{m}{k}\right)^{\frac{c}{2}}\varepsilon^{b+c} &(278)\\
&\leq \frac{n\sqrt{m}\varepsilon^2}{k}\sum_{b,c\geq 0}\varepsilon^{b+c} &(279)\\
&\lesssim \frac{n\sqrt{m}\varepsilon^2}{k}, &(280)
\end{aligned}$$

where we use that $\varepsilon<1/2$. Thus, likelihood-free hypothesis testing is impossible for $m\leq n$ unless $n^2 m\gtrsim k^2/\varepsilon^4$. ∎

Appendix C
Proof of Theorem 4

A. Upper Bound

We deduce the upper bound by applying the corresponding result for $\mathcal{P}_{\mathsf{D}}$ as a black-box procedure.

Theorem 6 ([60])**.** *For a constant independent of ε and k,*

$$n_{\mathsf{GoF}}(\varepsilon,\mathsf{H},\mathcal{P}_{\mathsf{D}})\asymp\sqrt{k}/\varepsilon^2. \tag{281}$$

Write $\mathcal{G}_\ell$ for the regular grid of size ℓ^d on $[0,1]^d$ and let P_ℓ denote the L^2-projector onto the space of functions piecewise constant on the cells of $\mathcal{G}_\ell$. For convenience let us re-state Proposition 3.

Proposition 16. *For any $\beta\in(0,1]$, $C>1$ and $d\geq 1$ there exists a constant $c>0$ such that*

$$c\mathsf{H}(f,g)\leq \mathsf{H}(P_\kappa f,P_\kappa g)\leq \mathsf{H}(f,g) \tag{282}$$

holds for any $f,g\in\mathcal{P}_{\mathsf{H}}(\beta,d,C)$, provided we set $\kappa=(c\varepsilon)^{-2/\beta}$.

With the above approximation result, the proof of Theorem 4 is straightforward.

Proof of Theorem 4: Suppose we are testing goodness-of-fit to $f_0\in\mathcal{P}_{\mathsf{H}}$ based on an i.i.d. sample $X_1,\ldots,X_n$ from $f\in\mathcal{P}_{\mathsf{H}}$. Take $\kappa\asymp\varepsilon^{-2/\beta}$ and bin the observations on $\mathcal{G}_\kappa$, denoting the pmf of the resulting distribution as p_f. Then, under the alternative hypothesis that $\mathsf{H}(f,f_0)\geq\varepsilon$, by Proposition 3

$$\varepsilon\lesssim \mathsf{H}(P_\kappa f_0,P_\kappa f)=\mathsf{H}(p_{f_0},p_f). \tag{283}$$

In particular, applying the algorithm achieving the upper bound in Theorem 6 to the binned observations, we see that $n\gtrsim\sqrt{\kappa^d}/\varepsilon^2=\varepsilon^{-(2\beta+d)/\beta}$ samples suffice. ∎

B. Lower Bound

The proof is extremely similar to the TV case, except we put the perturbations at density level ε^2 instead of 1.

Proof: Let $\phi:[0,1]\to[0,1]$ be a smooth function such that $\phi(x)=0$ for $x\leq 1/3$ and $\phi(x)=1$ for $x\geq 2/3$. Let $h:\mathbb{R}^d\to\mathbb{R}$ be smooth, supported in $[0,1]^d$, and satisfy $\int_{[0,1]^d}h(x)\mathrm{d}x=0$ and $\int_{[0,1]^d}h(x)^2\mathrm{d}x=1$. Given $\varepsilon\in(0,1)$ let

$$f_0(x)=\varepsilon^2+\frac{\phi(x_1)}{\|\phi\|_1}(1-\varepsilon^2), \tag{284}$$

which is a density on $[0,1]^d$. For a large integer κ and $j\in[\kappa/3]\times[\kappa]^{d-1}$ let

$$h_j(x)=\kappa^{d/2}h(\kappa x-j+1) \tag{285}$$

for $x\in[0,1]^d$. Then h_j is supported on $[(j-1)/\kappa,j/\kappa]\subseteq[0,1/3]\times[0,1]^{d-1}$ and $\int h_j^2=1$. For $\eta\in\{\pm 1\}^{[\kappa/3]\times[\kappa]^{d-1}}$ and $\rho>0$ let

$$f_\eta(x)=f_0+\rho\sum_{j\in[\kappa/3]\times[\kappa]^{d-1}}\eta_j h_j(x). \tag{286}$$

Then f_η is positive provided that $\varepsilon^2\geq\rho\kappa^{d/2}|h|_\infty\asymp\rho\kappa^{d/2}$. Further, $\|f_\eta\|_{\mathcal{C}^\beta}$ is of constant order provided $\rho\kappa^{d/2+\beta}\lesssim 1$. Under these assumptions $f_\eta\in\mathcal{P}_{\mathsf{H}}$. Note that the Hellinger distance between f_η and f_0 is

$$H^2(f_0,f_\eta)$$

$$= \sum_{j:j_1\le\kappa/3} \int_{[\frac{j-1}{\kappa},\frac{j}{\kappa}]} \left(\sqrt{f_0(x)} - \sqrt{f_\eta(x)}\right)^2 \mathrm{d}x \quad (287)$$

$$= \sum_{j:j_1\le\kappa/3} \int_{[\frac{j-1}{\kappa},\frac{j}{\kappa}]} \frac{\rho^2 h_j^2(x)}{(\sqrt{f_0(x)} + \sqrt{f_\eta(x)})^2}\mathrm{d}x \quad (288)$$

$$\ge \sum_{j:j_1\le\kappa/3} \int_{[\frac{j-1}{\kappa},\frac{j}{\kappa}]} \frac{\rho^2 h_j^2(x)}{4\varepsilon^2}\mathrm{d}x \quad (289)$$

$$\gtrsim \frac{\rho^2\kappa^d}{\varepsilon^2}. \quad (290)$$

Suppose we draw η uniformly at random. Via Ingster's trick we compute

$$\chi^2(\mathbb{E}_\eta f_\eta^{\otimes n}\|f_0^{\otimes n}) + 1$$

$$= \int \mathbb{E}_{\eta\eta'} \prod_{i=1}^n \frac{f_\eta(x_i)f_{\eta'}(x_i)}{f_0(x_i)}\mathrm{d}x_1\ldots\mathrm{d}x_n \quad (291)$$

$$= \mathbb{E}_{\eta\eta'}\left(\int \frac{f_\eta(x)f_{\eta'}(x)}{f_0(x)}\mathrm{d}x\right)^n. \quad (292)$$

Looking at the integral term on the inside we get

$$\int \frac{f_\eta(x)f_{\eta'}(x)}{f_0(x)}\mathrm{d}x$$

$$= 1 + \rho\sum_j(\eta_j+\eta'_j)\int h_j(x)\mathrm{d}x \quad (293)$$

$$+ \rho^2\sum_j \eta_j\eta'_j\int\frac{h_j(x)^2}{f_0(x)}\mathrm{d}x$$

$$= 1 + \frac{\rho^2}{\varepsilon^2}\sum_j\eta_j\eta'_j\int h_j(x)^2\mathrm{d}x \quad (294)$$

$$= 1 + \frac{\rho^2}{\varepsilon^2}\langle\eta,\eta'\rangle, \quad (295)$$

where we've used that h_j and $h_{j'}$ have disjoint support unless $j = j'$, $\int h_j = 0$, $\int h_j^2 = 1$, and that $f_0(x) = \varepsilon^2$ for all x with $x_1 \le 1/3$. Plugging in, using the inequalities $1 + x \le \exp(x)$ and $\cosh(x) \le \exp(x^2)$ we obtain

$$\chi^2(\mathbb{E}_\eta f_\eta^{\otimes n}\|f_0^{\otimes n}) + 1 \le \mathbb{E}_{\eta\eta'}\left(1 + \frac{\rho^2}{\varepsilon^2}\langle\eta,\eta'\rangle\right)^n \quad (296)$$

$$\le \mathbb{E}_{\eta\eta'}\exp\left(\frac{\rho^2 n}{\varepsilon^2}\langle\eta,\eta'\rangle\right) \quad (297)$$

$$= \cosh(\frac{\rho^2 n}{\varepsilon^2})^{\kappa^d/3} \quad (298)$$

$$\le \exp\left(\frac{\rho^4 n^2\kappa^d}{3\varepsilon^4}\right). \quad (299)$$

Choosing $\kappa = \varepsilon^{-2/\beta}$ and $\rho = \varepsilon^{(2\beta+d)/\beta}$ we see that goodness-of-fit testing of f_0 is impossible unless

$$n \gtrsim \frac{\varepsilon^2}{\rho^2\kappa^{d/2}} = \varepsilon^{-\frac{2\beta+d}{\beta}}. \quad (300)$$

■

APPENDIX D
AUXILIARY TECHNICAL RESULTS

A. Proof of Lemma 1

Proof: We prove the upper bound first. Let $\mathbb{P}_0, \mathbb{P}_1 \in \mathcal{P}$ be arbitrary. Then by Lemma 5,

$$\inf_\psi \max_{i=0,1} \mathbb{P}_i^{\otimes m}(\psi \ne i)$$

$$\le \inf_\psi \left(\mathbb{P}_0^{\otimes m}(\psi = 1) + \mathbb{P}_1^{\otimes m}(\psi = 0)\right) \quad (301)$$

$$= 1 - \mathsf{TV}(\mathbb{P}_0^{\otimes m}, \mathbb{P}_1^{\otimes m}) \quad (302)$$

$$\le 1 - \frac{1}{2}\mathsf{H}^2(\mathbb{P}_0^{\otimes m}, \mathbb{P}_1^{\otimes m}) \triangleq (\dagger). \quad (303)$$

By tensorization of the Hellinger affinity, we have

$$\mathsf{H}^2(\mathbb{P}_0^{\otimes m}, \mathbb{P}_1^{\otimes m}) = 2 - 2\left(1 - \frac{1}{2}\mathsf{H}^2(\mathbb{P}_0, \mathbb{P}_1)\right)^m. \quad (304)$$

Plugging in, along with $1 + x \le e^x$ gives

$$(\dagger) \le \exp(-\frac{m}{2}\mathsf{H}^2\left(\mathbb{P}_0^{\otimes m}, \mathbb{P}_1^{\otimes m}\right)). \quad (305)$$

Taking $m > 2\log(3)/\mathsf{H}^2(\mathbb{P}_0, \mathbb{P}_1)$ shows the existence of a successful test. Let us turn to the lower bound. Using Lemma 5 we have

$$\inf_\psi \max_{i=0,1} \mathbb{P}_i^{\otimes m}(\psi \ne i)$$

$$\ge \frac{1}{2}\left(1 - \mathsf{TV}(\mathbb{P}_0^{\otimes m}, \mathbb{P}_1^{\otimes m})\right) \quad (306)$$

$$\ge \frac{1}{2}\left(1 - \mathsf{H}(\mathbb{P}_0^{\otimes m}, \mathbb{P}_1^{\otimes m})\right). \quad (307)$$

Note that it is enough to restrict the maximization in Lemma 1 to $\mathbb{P}_0, \mathbb{P}_1 \in \mathcal{P}$ with $\mathsf{H}^2(\mathbb{P}_0, \mathbb{P}_1) < 1$. Now, by (304) and the inequalities $e^{-2x} \le 1 - x$ valid for all $x \in [0, 1/2]$ and $1 - x \le e^{-x}$ valid for all $x \in \mathbb{R}$, we obtain

$$\mathsf{H}^2(\mathbb{P}_0^{\otimes m}, \mathbb{P}_1^{\otimes m}) = 2 - 2\left(1 - \frac{1}{2}\mathsf{H}^2(\mathbb{P}_0, \mathbb{P}_1)\right)^m \quad (308)$$

$$\le 2 - 2\exp(-m\mathsf{H}^2(\mathbb{P}_0, \mathbb{P}_1)) \quad (309)$$

$$\le 2m\mathsf{H}^2(\mathbb{P}_0, \mathbb{P}_1). \quad (310)$$

Taking $m = 1/(18\mathsf{H}^2(\mathbb{P}_0, \mathbb{P}_1))$ concludes the proof via Lemma 4. ■

B. Proof of Lemma 2

Proof: By standard inequalities between divergences (see Lemma 5), omitting the argument (μ_θ, μ_0) for simplicity we have

$$\mathsf{TV} \le \mathsf{H} \le \sqrt{\mathsf{KL}} \le \sqrt{\chi^2} \tag{311}$$

$$= \sqrt{\exp(\|\theta\|_2^2) - 1} \lesssim \|\theta\|_2. \tag{312}$$

For the lower bound we obtain $\mathsf{TV}(\mu_\theta, \mu_0) \ge \min\{1, \|\theta\|_2/200\} \gtrsim \|\theta\|_2$ by [70, Theorem 1.2]. ∎

C. Proof of Proposition 3

Let us write $a_+ \triangleq a \vee 0$ for both functions and real numbers. We start with some known results of approximation theory.

Definition 7. For $f : [0,1]^d \to \mathbb{R}$ define the modulus of continuity as

$$\omega(\delta; f) = \sup_{\|x-y\|_2 \le \delta} |f(x) - f(y)|. \tag{313}$$

Lemma 7. *For any real-valued function f and $\delta \ge 0$,*

$$\omega(\delta; \sqrt{f_+}) \le \omega(\delta; f)^{1/2}. \tag{314}$$

Proof: Follows from the inequality $|\sqrt{a_+} - \sqrt{b_+}|^2 \le |a-b|$ valid for all $a, b \in \mathbb{R}$. ∎

Lemma 8. *Let $f : [0,1]^d \to \mathbb{R}$ be β-smooth for $\beta \in (0,1]$. Then*

$$\omega(\delta; f) \le c\,\delta^\beta \tag{315}$$

for a constant c depending only on $\|f\|_{\mathcal{C}^\beta}$.

Proof: Follows by the definition of Hölder continuity. ∎

Lemma 9 ([71, Theorem 4])**.** *For any continuous function $f : [0,1]^d \to \mathbb{R}$ the best polynomial approximation p_n of degree n satisfies*

$$\|p_n - f\|_\infty \le c\,\omega\left(\frac{d^{3/2}}{n}; f\right) \tag{316}$$

for a universal constant $c > 0$.

Definition 8. Given a function $f : [0,1]^d \to \mathbb{R}$, $\ell \ge 1$ and $j \in [\ell]^d$, let $\pi_{j,\ell} f : [0,1]^d \to \mathbb{R}$ denote the function

$$\pi_{j,\ell} f(x) \triangleq f\left(\frac{x + j - 1}{\ell}\right). \tag{317}$$

In other words, $\pi_{j,\ell} f$ is equal to f zoomed in on the j'th bin of the regular grid $\mathcal{G}_\ell$.

Recall that here P_ℓ denotes the L^2 projector onto the space of functions piecewise constant on the bins of $\mathcal{G}_\ell$. We are ready for the proof of Proposition 3.

Proof: Let $\kappa \ge r \ge 1$ whose values we specify later. We treat the parameters $\beta, d, \|f\|_{\mathcal{C}^\beta}, \|g\|_{\mathcal{C}^\beta}$ as constants in our analysis. Let $u_f : [0,1]^d \to \mathbb{R}$ denote the (piecewise polynomial) function that is equal to the best polynomial approximation of $\sqrt{f}$ on each bin of $\mathcal{G}_{\kappa/r}$ with maximum degree α. By lemmas 7 and 8 for any $\ell \ge 1$ and $j \in [\ell]^d$

$$\omega(\delta; \pi_{j,\ell}\sqrt{f}) \le \omega(\delta/\ell; \sqrt{f}) \lesssim (\delta/\ell)^{\beta/2}, \tag{318}$$

so that by Lemma 9

$$|u_f - \sqrt{f}|_\infty = \sup_{j \in [\kappa/r]^d} |\pi_{j,\kappa/r}(u_f - \sqrt{f})|_\infty \tag{319}$$

$$\lesssim \sup_{j \in [\kappa/r]^d} \omega(d^{3/2}/\alpha; \pi_{j,\kappa/r}\sqrt{f}) \tag{320}$$

$$\lesssim (\alpha\kappa/r)^{-\beta/2}. \tag{321}$$

Regarding r as a constant independent of κ, α can be chosen large enough independently of κ such that $|u_f - \sqrt{f}|_\infty \le c_1 \kappa^{-\beta/2}$ for c_1 arbitrarily small. Define u_g analogously to u_f. We have the inequalities

$$\mathsf{H}(f,g) = \|\sqrt{f} - \sqrt{g}\|_2 \tag{322}$$

$$\le \|\sqrt{f} - u_f\|_2 + \|u_f - u_g\|_2 \tag{323}$$

$$+ \|u_g - \sqrt{g}\|_2$$

$$\le 2c_1\kappa^{-\beta/2} + \|u_f - u_g\|_2. \tag{324}$$

We can write

$$\|u_f - u_g\|_2^2 = \frac{1}{(\kappa/r)^d} \sum_{j \in [\kappa/r]^d} \|\pi_{j,\kappa/r}(u_f - u_g)\|_2^2 \tag{325}$$

Now, by [41, Lemma 7.4] we can take r large enough (depending only on $\beta, d, \|f\|_{\mathcal{C}^\beta}, \|g\|_{\mathcal{C}^\beta}$) such that

$$\|\pi_{j,\kappa/r}(u_f - u_g)\|_2 \le c_2 \|P_r \pi_{j,\kappa/r}(u_f - u_g)\|_2 \tag{326}$$

where the implied constant depends on the same parameters as r. Thus, we get

$$\mathsf{H}^2(f,g) \le 8c_1^2\kappa^{-\beta}$$

$$+ \frac{2c_2^2}{(\kappa/r)^d} \sum_{j \in [\kappa/r]^d} \|P_r \pi_{j,\kappa/r}(u_f - u_g)\|_2^2 \tag{327}$$

$$\le 8c_1^2\kappa^{-\beta} \tag{328}$$

$$+ \frac{6c_2^2}{(\kappa/r)^d} \sum_{j \in [\kappa/r]^d} \Big(\|P_r \pi_{j,\kappa/r} u_f - \sqrt{P_r \pi_{j,\kappa/r} f}\|_2^2$$

$$+ \|P_r\pi_{j,\kappa/r}u_g - \sqrt{P_r\pi_{j,\kappa/r}g}\|_2^2\Big) + 6c_2^2\mathsf{H}^2(P_\kappa f, P_\kappa f), \tag{329}$$

where c_1, c_2 depend only on the unimportant parameters, and c_1 can be taken arbitrarily small compared to c_2. We also used the fact that $P_r\pi_{j,\kappa/r} = \pi_{j,\kappa/r}P_\kappa$. Looking at the terms separately, we have

$$\begin{aligned}
&\|P_r\pi_{j,\kappa/r}u_f - \sqrt{P_r\pi_{j,\kappa/r}f}\|_2 \\
&\quad\le \|P_r\pi_{j,\kappa/r}u_f - P_r\sqrt{\pi_{j,\kappa/r}f}\|_2 \\
&\qquad + \|P_r\sqrt{\pi_{j,\kappa/r}f} - \sqrt{P_r\pi_{j,\kappa/r}f}\|_2 &(330)\\
&\quad\le c\kappa^{-\beta/2} \\
&\qquad + \|P_r\sqrt{\pi_{j,\kappa/r}f} - \sqrt{P_r\pi_{j,\kappa/r}f}\|_2, &(331)
\end{aligned}$$

since P_r is a contraction by Lemma 10. We can decompose the second term as

$$\begin{aligned}
&\|P_r\sqrt{\pi_{j,\kappa/r}f} - \sqrt{P_r\pi_{j,\kappa/r}f}\|_2^2 &(332)\\
&= \sum_{\ell\in[r]^d}\int_{\left[\frac{\ell-1}{r},\frac{\ell}{r}\right]}\Big(r^d\int_{\left[\frac{\ell-1}{r},\frac{\ell}{r}\right]}\sqrt{\pi_{j,\kappa/r}f(x)}\mathrm{d}x \\
&\qquad - \sqrt{r^d\int_{\left[\frac{\ell-1}{r},\frac{\ell}{r}\right]}\pi_{j,\kappa/r}f(x)\mathrm{d}x}\Big)^2 = (\dagger).
\end{aligned}$$

For $x \in [(\ell-1)/r, \ell/r]$ we always have

$$\begin{aligned}
&|\pi_{j,\kappa/r}f(x) - \pi_{j,\kappa/r}f(\ell/r)| &(333)\\
&\quad\le \omega\left(\frac{\sqrt{d}}{r}; \pi_{j,\kappa/r}f\right) \quad \lesssim \left(\frac{\sqrt{d}/r}{\kappa/r}\right)^\beta \lesssim \kappa^{-\beta}.
\end{aligned}$$

Using the inequality $\sqrt{a+b} - \sqrt{(a-b)_+} \le 2\sqrt{b}$ valid for all $a, b \ge 0$, we can bound $(\dagger)$ by $\kappa^{-\beta}$ up to constant and the result follows. ∎

D. Proof of Proposition 5

For $f \in L^2(\mu)$ write $f_i = \langle f\phi_i\rangle$ and $f_{ii'} = \langle f\phi_i\phi_{i'}\rangle$, assuming that the quantities involved are well-defined. We record some useful identities related to P_r that will be instrumental in our proof of Proposition 5.

Lemma 10. *P_r is self-adjoint and has operator norm*

$$\|P_r\| \triangleq \sup_{f\in L^2(\mu):\|f\|_2\le 1}\|P_r(f)\|_2 \le 1. \tag{334}$$

Suppose that $f, g, h, t \in L^2(\mu)$ and that each quantity below is finite. Then

$$\sum_{ii'} f_i g_{i'} h_{ii'} = \langle hP_r(f)P_r(g)\rangle, \tag{335}$$

$$\sum_{ii'} f_i g_i h_{i't_{i'}} = \langle fP_r(g)\rangle\langle hP_r(t)\rangle \tag{336}$$

$$\sum_{ii'} f_{ii'}g_{ii'} = \sum_i \langle f\phi_i P_r(g\phi_i)\rangle, \tag{337}$$

where the summation is over $i, i' \in [r]$.

Proof: Let $P_r^\perp$ denote the orthogonal projection onto the orthogonal complement of $\mathrm{span}(\{\phi_1, \dots, \phi_r\})$. Then for any $f, g \in L^2(\mu)$ we have

$$\begin{aligned}
\langle fP_r(g)\rangle &= \langle (P_r(f) + P_r^\perp(f))P_r(g)\rangle &(338)\\
&= \langle P_r(f)P_r(g)\rangle &(339)\\
&= \langle P_r(f)g\rangle, &(340)
\end{aligned}$$

where the last equality is by symmetry. We also have

$$\begin{aligned}
\|P_r(f)\|_2^2 &\le \|P_r(f)\|_2^2 + \|P_r^\perp(f)\|_2^2 &(341)\\
&= \|P_r(f) + P_r^\perp(f)\|^2 &(342)\\
&= \|f\|_2^2. &(343)
\end{aligned}$$

Let $f, g, h, t \in L^2(\mu)$. Then

$$\begin{aligned}
\sum_{ii'} f_i g_{i'} h_{ii'} &= \sum_i f_i \sum_{i'} g_{i'} h_{ii'} &(344)\\
&= \sum_i f_i \sum_{i'} \langle gP_r(h\phi_i)\rangle &(345)\\
&= \sum_i f_i \langle P_r(g)h\phi_i\rangle &(346)\\
&= \langle P_r(f)hP_r(g)\rangle &(347)
\end{aligned}$$

$$\begin{aligned}
\sum_{ii'} f_i g_i h_{i't_{i'}} &= (\sum_i f_i g_i)(\sum_{i'} h_{i't_{i'}}) &(348)\\
&= \langle fP_r(g)\rangle\langle hP_r(t)\rangle &(349)\\
\sum_{ii'} f_{ii'}g_{ii'} &= \sum_i \langle f\phi_i \sum_{i'}\langle g\phi_i\phi_{i'}\rangle\phi_{i'}\rangle &(350)\\
&= \sum_i \langle f\phi_i P_r(g\phi_i)\rangle. &(351)
\end{aligned}$$

∎

Proof of Proposition 5: Let us label the different terms of the statistic $T_{\mathsf{LF}}^{-\mathsf{d}}$:

$$\begin{aligned}
T_{\mathsf{LF}}^{-\mathsf{d}} = \sum_{i=1}^r \Bigg\{ &\frac{2}{n^2}\sum_{j<j'}^n \phi_i(X_j)\phi_i(X_{j'}) \\
&- \frac{2}{n^2}\sum_{j<j'}^n \phi_i(Y_j)\phi_i(Y_{j'}) \\
&- \frac{2}{nm}\sum_{j=1}^n\sum_{u=1}^m \phi_i(X_j)\phi_i(Z_u)
\end{aligned}$$

$$+\frac{2}{nm}\sum_{j=1}^{n}\sum_{u=1}^{m}\phi_i(Y_j)\phi_i(Z_u)\Bigg\} \tag{352}$$

$$=\frac{2}{n^2}\mathsf{I}-\frac{2}{n^2}\mathsf{II}-\frac{2}{nm}\mathsf{III}+\frac{2}{nm}\mathsf{IV}. \tag{353}$$

Recall that $X, Y, Z \sim f^{\otimes n}, g^{\otimes n}, h^{\otimes m}$ respectively. A straightforward computation yields

$$\begin{aligned}\mathbb{E}T_{\mathsf{LF}} &= \|P_r(f-h)\|_2^2 - \|P_r(g-h)\|_2^2 \\ &\quad - \frac{1}{n}\big(\|P_r(f)\|_2^2 - \|P_r(g)\|_2^2\big).\end{aligned} \tag{354}$$

We decompose the variance as

$$\begin{aligned}\operatorname{var}(T_{\mathsf{LF}}) &= \frac{4}{n^4}\operatorname{var}(\mathsf{I}) + \frac{4}{n^4}\operatorname{var}(\mathsf{II}) \\ &\quad + \frac{4}{n^2m^2}\operatorname{var}(\mathsf{III}) + \frac{4}{n^2m^2}\operatorname{var}(\mathsf{IV}) \\ &\quad - \frac{8}{n^3m}\operatorname{Cov}(\mathsf{I},\mathsf{III}) - \frac{8}{n^3m}\operatorname{Cov}(\mathsf{II},\mathsf{IV}) \\ &\quad - \frac{8}{n^2m^2}\operatorname{Cov}(\mathsf{III},\mathsf{IV}),\end{aligned} \tag{355}$$

where we used independence of the pairs $(\mathsf{I},\mathsf{II}), (\mathsf{I},\mathsf{IV}), (\mathsf{II},\mathsf{III})$. Expanding the variances we obtain

$$\begin{aligned}\operatorname{var}(\mathsf{I}) = \sum_{ii'}\Bigg\{&\binom{n}{2}(f_{ii'}^2 - f_i^2 f_{i'}^2) \\ &+ \Big(\binom{n}{2}^2 - \binom{n}{2} - \binom{4}{2}\binom{n}{4}\Big) \\ &\times \big(f_i f_{i'} f_{ii'} - f_i^2 f_{i'}^2\big)\Bigg\}\end{aligned} \tag{356}$$

$$\begin{aligned}\operatorname{var}(\mathsf{II}) = \sum_{ii'}\Bigg\{&\binom{n}{2}(g_{ii'}^2 - g_i^2 g_{i'}^2) \\ &+ \left(\binom{n}{2}^2 - \binom{n}{2} - \binom{4}{2}\binom{n}{4}\right) \\ &\times \big(g_i g_{i'} g_{ii'} - g_i^2 g_{i'}^2\big)\Bigg\}\end{aligned} \tag{357}$$

$$\begin{aligned}\operatorname{var}(\mathsf{III}) = \sum_{ii'}\Bigg\{&nm(f_{ii'}h_{ii'} - f_i f_{i'} h_i h_{i'}) \\ &+ nm(m-1)(f_{ii'}h_i h_{i'} - f_i f_{i'} h_i h_{i'}) \\ &+ mn(n-1)(f_i f_{i'} h_{ii'} - f_i f_{i'} h_i h_{i'})\Bigg\}\end{aligned} \tag{358}$$

$$\begin{aligned}\operatorname{var}(\mathsf{IV}) = \sum_{ii'}\Bigg\{&nm(h_{ii'}g_{ii'} - h_i h_{i'} g_i g_{i'}) \\ &+ mn(n-1)(h_{ii'}g_i g_{i'} - h_i h_{i'} g_i g_{i'}) \\ &+ nm(m-1)(g_{ii'}h_i h_{i'} - h_i h_{i'} g_i g_{i'})\Bigg\}.\end{aligned} \tag{359}$$

For the covariance terms we obtain

$$\operatorname{Cov}(\mathsf{I},\mathsf{III}) = \sum_{ii'} 2m\binom{n}{2}(f_{ii'}f_i h_{i'} - f_i^2 f_{i'} h_{i'}) \tag{360}$$

$$\operatorname{Cov}(\mathsf{II},\mathsf{IV}) = \sum_{ii'} 2m\binom{n}{2}(g_{ii'}g_i h_{i'} - g_i^2 g_{i'} h_{i'}) \tag{361}$$

$$\operatorname{Cov}(\mathsf{III},\mathsf{IV}) = \sum_{ii'} mn^2(h_{ii'}f_i g_{i'} - f_i g_{i'} h_i h_{i'}). \tag{362}$$

We can now start collecting the terms, applying the calculation rules from Lemma 10 repeatedly. Note that $\binom{n}{2}^2 - \binom{n}{2} - \binom{4}{2}\binom{n}{4} = n^3 - 3n^2 + 2n$, and by inspection we can conclude that $1/n, 1/m, 1/nm, 1/n^2$ and $1/n^3$ are the only terms with nonzero coefficients. We look at each of them one-by-one:

$$\begin{aligned}\operatorname{Coef}\left(\frac{1}{n}\right) = \sum_{ii'}^{r}\Bigg\{&\underbrace{4(f_i f_{i'} f_{ii'} - f_i^2 f_{i'}^2)}_{\operatorname{var}(I)} \\ &+ \underbrace{4(g_i g_{i'} g_{ii'} - g_i^2 g_{i'}^2)}_{\operatorname{var}(II)} \\ &+ \underbrace{4(h_i h_{i'} f_{ii'} - f_i f_{i'} h_i h_{i'})}_{\operatorname{var}(III)} \\ &+ \underbrace{4(g_{ii'} h_i h_{i'} - h_i h_{i'} g_i g_{i'})}_{\operatorname{var}(IV)} \\ &- \underbrace{8(f_{ii'} f_i h_{i'} - f_i^2 f_{i'} h_{i'})}_{\operatorname{Cov}(I,III)} \\ &- \underbrace{8(g_{ii'} g_i h_{i'} - g_i^2 g_{i'} h_{i'})}_{\operatorname{Cov}(II,IV)}\Bigg)\end{aligned} \tag{363}$$

$$\begin{aligned}&= 4\langle f P_r(f)^2\rangle - 4\langle f P_r(f)\rangle^2 \\ &\quad + 4\langle g P_r(g)^2\rangle - 4\langle g P_r(g)\rangle^2 \\ &\quad + 4\langle f P_r(h)^2\rangle - 4\langle f P_r(h)\rangle^2 \\ &\quad + 4\langle g P_r(h)^2\rangle - 4\langle h P_r(g)\rangle^2 \\ &\quad - 8\langle f P_r(f) P_r(h)\rangle + 8\langle f P_r(f)\rangle\langle f P_r(h)\rangle \\ &\quad - 8\langle g P_r(g) P_r(h)\rangle + 8\langle g P_r(g)\rangle\langle g P_r(h)\rangle\end{aligned} \tag{364}$$

$$\begin{aligned}&= 4\langle f(P_r(f-h))^2\rangle \\ &\quad + 4\langle g(P_r(g-h))^2\rangle \\ &\quad - 4\langle P_r(f-h)\rangle^2\end{aligned} \tag{365}$$

$$
\begin{aligned}
&\quad - 4\langle P_r(g-h)\rangle^2 \\
&\le 4A_{ffh} + 4A_{ggh}, \qquad (366)
\end{aligned}
$$

recalling the definition $A_{uvt} = \langle u\big[P_r(v-t)\big]^2\rangle$ for $u, v, t \in L^2(\mu)$. Similarly, we get

$$
\begin{aligned}
\mathrm{Coef}\left(\frac{1}{m}\right) &= \sum_{ii'}^{r}\Bigg\{ \underbrace{4(h_{ii'}f_if_{i'} - f_if_{i'}h_ih_{i'})}_{\mathrm{var}(III)} \qquad (367)\\
&\quad + \underbrace{4(h_{ii'}g_ig_{i'} - h_ih_{i'}g_ig_{i'})}_{\mathrm{var}(IV)} \\
&\quad - \underbrace{8(h_{ii'}f_ig_{i'} - f_ih_ih_{i'}g_{i'})}_{\mathrm{Cov}(III,IV)}\Bigg\} \\
&= 4\langle h(P_r(f-g))^2\rangle \\
&\quad - 4\langle hP_r(f-g)\rangle^2 \qquad (368)\\
&\le 4A_{hfg}. \qquad (369)
\end{aligned}
$$

For the lower order terms we obtain

$$
\begin{aligned}
\mathrm{Coef}\left(\frac{1}{nm}\right) &= \sum_{ii'}^{r}\Bigg\{ \underbrace{4(f_{ii'}h_{ii'} - f_if_{i'}h_ih_{i'})}_{\mathrm{var}(III)} \qquad (370)\\
&\quad - \underbrace{4(f_{ii'}h_ih_{i'} - f_if_{i'}h_ih_{i'})}_{\mathrm{var}(III)} \\
&\quad - \underbrace{4(f_if_{i'}h_{ii'} - f_if_{i'}h_ih_{i'})}_{\mathrm{var}(III)} \\
&\quad + \underbrace{4(h_{ii'}g_{ii'} - h_ih_{i'}g_ig_{i'})}_{\mathrm{var}(IV)} \\
&\quad - \underbrace{4(h_{ii'}g_ig_{i'} - h_ih_{i'}g_ig_{i'})}_{\mathrm{var}(IV)} \\
&\quad - \underbrace{4(g_{ii'}h_ih_{i'} - h_ih_{i'}g_ig_{i'})}_{\mathrm{var}(IV)}\Bigg\} \\
&= 4B_{fh} - 4\langle fP_r(h)\rangle^2 \qquad (371)\\
&\quad - 4\langle fP_r(h)^2\rangle + 4\langle fP_r(h)\rangle^2 \\
&\quad - 4\langle hP_r(f)^2\rangle + 4\langle fP_r(h)\rangle^2 \\
&\quad + 4B_{gh} - 4\langle gP_r(h)\rangle^2 \\
&\quad - 4\langle hP_r(g)^2\rangle + 4\langle gP_r(h)\rangle^2 \\
&\quad - 4\langle gP_r(h)^2\rangle + 4\langle gP_r(h)\rangle^2 \\
&\le 4\langle fP_r(h)\rangle^2 + 4\langle gP_r(h)\rangle^2 \qquad (372)\\
&\quad + 4B_{fh} + 4B_{gh} \\
&\lesssim |B_{fh}| + |B_{gh}| + \|f+g+h\|_2^4 \qquad (373)
\end{aligned}
$$

where we recall the definition $B_{uv} = \sum_i \langle u\phi_i P_r(v\phi_i)\rangle$ for $u, v \in L^2(\mu)$ and apply the Cauchy-Schwarz inequality. Next, we look at the coefficient of $1/n^2$ and find

$$
\begin{aligned}
\mathrm{Coef}\left(\frac{1}{n^2}\right) &= \sum_{ii'}\Bigg\{ \underbrace{2(f_{ii'}^2 - f_i^2f_{i'}^2)}_{\mathrm{var(I)}} \qquad (374)\\
&\quad - \underbrace{12(f_{ii'}f_if_{i'} - f_i^2f_{i'}^2)}_{\mathrm{var(I)}} \\
&\quad + \underbrace{2(g_{ii'}^2 - g_i^2g_{i'}^2)}_{\mathrm{var(II)}} \\
&\quad - \underbrace{12(g_{ii'}g_ig_{i'} - g_i^2g_{i'}^2)}_{\mathrm{var(II)}} \\
&\quad + \underbrace{8(f_{ii'}f_ih_{i'} - f_i^2f_{i'}h_{i'})}_{\mathrm{Cov(I,III)}} \\
&\quad + \underbrace{8(g_{ii'}g_ih_{i'} - g_i^2g_{i'}h_{i'})}_{\mathrm{Cov(II,IV)}}\Bigg\} \\
&= 2B_{ff} - 2\langle fP_r(f)\rangle^2 \qquad (375)\\
&\quad - 12\langle fP_r(f)^2\rangle + 12\langle fP_r(f)\rangle^2 \\
&\quad + 2B_{gg} - 2\langle gP_r(g)\rangle^2 \\
&\quad - 12\langle gP_r(g)^2\rangle + 12\langle gP_r(g)\rangle^2 \\
&\quad + 8\langle fP_r(f)P_r(h)\rangle \\
&\quad - 8\langle fP_r(f)\rangle\langle fP_r(h)\rangle \\
&\quad + 8\langle gP_r(g)P_r(h)\rangle \\
&\quad - 8\langle gP_r(g)\rangle\langle gP_r(h)\rangle \\
&\le 2B_{ff} + 2B_{gg} \qquad (376)\\
&\quad + 8\langle fP_r(f)P_r(h-f)\rangle \\
&\quad + 8\langle gP_r(g)P_r(h-g)\rangle \\
&\quad + 40\|f+g+h\|_2^4 \\
&\lesssim |B_{ff}| + |B_{gg}| + \|f+g+h\|_2^4 \qquad (377)\\
&\quad + \sqrt{A_{ff0}A_{ffh} + A_{gg0}A_{ggh}}.
\end{aligned}
$$

Finally, we look at the coefficient of $1/n^3$:

$$
\begin{aligned}
\mathrm{Coef}\left(\frac{1}{n^3}\right) &= \sum_{ii'}\Bigg\{ \underbrace{-2(f_{ii'}^2 - f_i^2f_{i'}^2)}_{\mathrm{Cov(I,III)}} \qquad (378)\\
&\quad + \underbrace{8(f_{ii'}f_if_{i'} - f_i^2f_{i'}^2)}_{\mathrm{Cov(I,III)}} \\
&\quad - \underbrace{2(g_{ii'}^2 - g_i^2g_{i'}^2)}_{\mathrm{Cov(I,III)}}
\end{aligned}
$$

$$+\underbrace{8(g_{ii'}g_ig_{i'}-g_i^2g_{i'}^2)}_{\mathrm{Cov(I,III)}}\Bigg\}$$

$$= -2B_{ff} + 2\langle fP_r(f)\rangle^2 \tag{379}$$
$$+ 8\langle fP_r(f)^2\rangle - 8\langle fP_r(f)\rangle^2$$
$$- 2B_{gg} + 2\langle gP_r(g)\rangle^2$$
$$+ 8\langle gP_r(g)^2\rangle - 8\langle gP_r(g)\rangle^2$$
$$\lesssim |B_{ff}| + |B_{gg}| \tag{380}$$
$$+ \|f+g+h\|_2^4 + A_{ff0} + A_{gg0}.$$

■

E. Proof of Lemma 6

Proof: Expanding via the binomial formula and using the fact that sums of N_j's are binomial random variables, we get

$$\mathbb{E}_N \prod_{j=1}^{k}(a+b(1+c)^{N_j})$$
$$= \mathbb{E}\sum_{\ell=0}^{k}\binom{k}{\ell}b^\ell(1+c)^{\mathrm{Bin}(n,\ell/k)}a^{k-\ell} \tag{381}$$
$$= \sum_{\ell=0}^{k}\binom{k}{\ell}b^\ell\left(1+\frac{c\ell}{k}\right)^n a^{k-\ell} \tag{382}$$
$$\leq (a+be^{cn/k})^k, \tag{383}$$

where we used $1+x \leq e^x$ for all $x \in \mathbb{R}$. ■

ACKNOWLEDGMENTS

We thank Julien Chhor for pointing out the connection between flattening and the Jensen-Shannon divergence.

REFERENCES

[1] S. Chatrchyan, V. Khachatryan, A. M. Sirunyan, A. Tumasyan, W. Adam, E. Aguilo, T. 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*, vol. 716, no. 1, pp. 30–61, 2012.

[2] C. Adam-Bourdarios, G. Cowan, C. Germain, I. Guyon, B. Kégl, and D. Rousseau, "The higgs boson machine learning challenge," in *NIPS 2014 workshop on high-energy physics and machine learning*. PMLR, 2015, pp. 19–55.

[3] S. Agostinelli, J. Allison, K. a. Amako, J. Apostolakis, H. Araujo, P. Arce, M. Asai, D. Axen, S. Banerjee, G. Barrand *et al.*, "Geant4—a simulation toolkit," *Nuclear instruments and methods in physics research section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 506, no. 3, pp. 250–303, 2003.

[4] S. Frixione, P. Nason, and C. Oleari, "Matching nlo qcd computations with parton shower simulations: the powheg method," *Journal of High Energy Physics*, vol. 2007, no. 11, p. 070, 2007.

[5] G. Corcella, I. G. Knowles, G. Marchesini, S. Moretti, K. Odagiri, P. Richardson, M. H. Seymour, and B. R. Webber, "Herwig 6: an event generator for hadron emission reactions with interfering gluons (including supersymmetric processes)," *Journal of High Energy Physics*, vol. 2001, no. 01, p. 010, 2001.

[6] T. Sjöstrand, S. Mrenna, and P. Skands, "Pythia 6.4 physics and manual," *Journal of High Energy Physics*, vol. 2006, no. 05, p. 026, 2006.

[7] J. Alwall, P. Demin, S. De Visscher, R. Frederix, M. Herquet, F. Maltoni, T. Plehn, D. L. Rainwater, and T. Stelzer, "Madgraph/madevent v4: the new web generation," *Journal of High Energy Physics*, vol. 2007, no. 09, p. 028, 2007.

[8] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, "Asymptotic formulae for likelihood-based tests of new physics," *The European Physical Journal C*, vol. 71, no. 2, pp. 1–19, 2011.

[9] L. Lista, *Statistical methods for data analysis in particle physics*. Springer, 2017, vol. 941.

[10] L. Devroye and G. Lugosi, *Combinatorial methods in density estimation*. Springer Science & Business Media, 2001.

[11] P. R. Gerber, Y. Han, and Y. Polyanskiy, "Minimax optimal testing by classification," in *The Thirty Sixth Annual Conference on Learning Theory*. PMLR, 2023, pp. 5395–5432.

[12] Y. I. Ingster, "Minimax testing of nonparametric hypotheses on a distribution density in the l_p metrics," *Theory of Probability & Its Applications*, vol. 31, no. 2, pp. 333–337, 1987.

[13] O. Goldreich and D. Ron, "On testing expansion in bounded-degree graphs," *Electronic Colloquium on Computational Complexity*, 2000.

[14] B. G. Kelly, T. Tularak, A. B. Wagner, and P. Viswanath, "Universal hypothesis testing in the learning-limited regime," in *2010 IEEE International Symposium on Information Theory*. IEEE, 2010, pp. 1478–1482.

[15] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Transactions on Information Theory*, vol. 35, no. 2, pp. 401–408, 1989.

[16] J. Ziv, "On classification with empirically observed statistics and universal data compression," *IEEE Trans. Inf. Theory*, vol. 34, pp. 278–286, 1988.

[17] L. Zhou, V. Y. F. Tan, and M. Motani, "Second-order asymptotically optimal statistical classification," in *2019 IEEE International Symposium on Information Theory (ISIT)*, 2019, pp. 231–235.

[18] H.-W. Hsu and I.-H. Wang, "On binary statistical classification from mismatched empirically observed statistics," in *2020 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2020, pp. 2533–2538.

[19] H. He, L. Zhou, and V. Y. Tan, "Distributed detection with empirically observed statistics," *IEEE Transactions on Information Theory*, vol. 66, no. 7, pp. 4349–4367, 2020.

[20] M. Haghifam, V. Y. Tan, and A. Khisti, "Sequential classification with empirically observed statistics," *IEEE Transactions on Information Theory*, vol. 67, no. 5, pp. 3095–3113, 2021.

[21] P. Boroumand and A. Guillén i Fàbregas, "Universal neyman-pearson classification with a known hypothesis," *arXiv preprint arXiv:2206.11700*, 2022.

[22] D. Huang and S. Meyn, "Classification with high-dimensional sparse samples," in *2012 IEEE International Symposium on Information Theory Proceedings*. IEEE, 2012, pp. 2586–2590.

[23] ——, "Generalized error exponents for small sample universal hypothesis testing," *IEEE transactions on information theory*, vol. 59, no. 12, pp. 8157–8181, 2013.

[24] B. G. Kelly, A. B. Wagner, T. Tularak, and P. Viswanath, "Classification of homogeneous data with large alphabets," *IEEE transactions on information theory*, vol. 59, no. 2, pp. 782–795, 2012.

[25] I. Diakonikolas and D. M. Kane, "A new approach for testing properties of discrete distributions," in *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2016, pp. 685–694.

[26] B. Bhattacharya and G. Valiant, "Testing closeness with unequal sized samples," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[27] L. Devroye, L. Gyorfi, and G. Lugosi, "A note on robust hypothesis testing," *IEEE Transactions on Information Theory*, vol. 48, no. 7, pp. 2111–2114, 2002.

[28] J. Friedman, "On multivariate goodness-of-fit and two-sample testing," Citeseer, Tech. Rep., 2004.

[29] D. Lopez-Paz and M. Oquab, "Revisiting classifier two-sample tests," *International Conference on Learning Representations*, 2017.

[30] M. U. Gutmann, R. Dutta, S. Kaski, and J. Corander, "Likelihood-free inference via classification," *Statistics and Computing*, vol. 28, no. 2, pp. 411–425, 2018.

[31] F. Liu, W. Xu, J. Lu, G. Zhang, A. Gretton, and D. J. Sutherland, "Learning deep kernels for non-parametric two-sample tests," in *International conference on machine learning*. PMLR, 2020, pp. 6316–6326.

[32] I. Kim, A. Ramdas, A. Singh, and L. Wasserman, "Classification accuracy as a proxy for two-sample testing," *The Annals of Statistics*, vol. 49, no. 1, pp. 411–434, 2021.

[33] S. Hediger, L. Michel, and J. Näf, "On the use of random forest for two-sample testing," *Computational Statistics & Data Analysis*, vol. 170, p. 107435, 2022.

[34] Y. I. Ingster, "On the minimax nonparametric detection of signals in white gaussian noise," *Problemy Peredachi Informatsii*, vol. 18, no. 2, pp. 61–73, 1982.

[35] O. Goldreich, *Introduction to property testing*. Cambridge University Press, 2017.

[36] Z. Bar-Yossef, *The complexity of massive data set computations*. University of California, Berkeley, 2002.

[37] I. Diakonikolas, T. Gouleakis, J. Peebles, and E. 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.

[38] I. Diakonikolas, T. Gouleakis, D. M. Kane, J. Peebles, and E. Price, "Optimal testing of discrete distributions with high probability," in *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, 2021, pp. 542–555.

[39] A. B. Tsybakov, *Introduction to Nonparametric Estimation*, 1st ed. Springer Publishing Company, Incorporated, 2008.

[40] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White, "Testing closeness of discrete distributions," *Journal of the ACM (JACM)*, vol. 60, no. 1, pp. 1–25, 2013.

[41] E. Arias-Castro, B. Pelletier, and V. Saligrama, "Remember the curse of dimensionality: The case of goodness-of-fit testing in arbitrary dimension," *Journal of Nonparametric Statistics*, vol. 30, no. 2, pp. 448–471, 2018.

[42] N. Dalmasso, R. Izbicki, and A. Lee, "Confidence sets and hypothesis testing in a likelihood-free inference setting," in *International Conference on Machine Learning*. PMLR, 2020, pp. 2323–2334.

[43] P. J. Huber, "A robust version of the probability ratio test," *The Annals of Mathematical Statistics*, pp. 1753–1758, 1965.

[44] P. J. Huber and V. Strassen, "Minimax tests and the neyman-pearson lemma for capacities," *The Annals of Statistics*, pp. 251–263, 1973.

[45] Y. G. Yatracos, "Rates of convergence of minimum distance estimators and kolmogorov's entropy," *The Annals of Statistics*, vol. 13, no. 2, pp. 768–774, 1985.

[46] L. Birgé, *Sur un théoreme de minimax et son application aux tests*. Univ. de Paris-Sud, Dép. de Mathématique, 1979.

[47] L. Birgé *et al.*, "Robust tests for model selection," *From probability to statistics and back: high-dimensional models and processes–A Festschrift in honor of Jon A. Wellner*, pp. 47–64, 2013.

[48] E. Giné and R. Nickl, *Mathematical foundations of infinite-dimensional statistical models*. Cambridge university press, 2021.

[49] Y. Polyanskiy and Z. Jia, Personal communication, Feb. 2024.

[50] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White, "Testing that distributions are close," in *Proceedings 41st Annual Symposium on Foundations of Computer Science*. IEEE, 2000, pp. 259–269.

[51] S.-O. Chan, I. Diakonikolas, P. Valiant, and G. Valiant, "Optimal algorithms for testing closeness of discrete distributions," in *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*. SIAM, 2014, pp. 1193–1203.

[52] G. Valiant and P. Valiant, "An automatic inequality prover and instance optimal identity testing," *SIAM Journal on Computing*, vol. 46, no. 1, pp. 429–455, 2017.

[53] T. Li and M. Yuan, "On the optimality of gaussian kernel based nonparametric tests against smooth alternatives," *arXiv preprint arXiv:1909.03302*, 2019.

[54] I. A. Ibragimov and R. Z. Khas'minskii, "On the estimation of an infinite-dimensional parameter in gaussian white noise," in *Doklady Akademii Nauk*, vol. 236, no. 5. Russian Academy of Sciences, 1977, pp. 1053–1055.

[55] I. M. Johnstone, "Gaussian estimation: Sequence and wavelet models," 2019.

[56] C. L. Canonne, "A short note on learning discrete distributions," *arXiv preprint arXiv:2002.11457*, 2020.

[57] B. Nazer, O. Ordentlich, and Y. Polyanskiy, "Information-distilling quantizers," in *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2017, pp. 96–100.

[58] A. Pensia, V. Jog, and P.-L. Loh, "Communication-constrained hypothesis testing: Optimality, robustness, and reverse data processing inequalities," *IEEE Transactions on Information Theory*, 2023.

[59] G. C. Kamath, "Modern challenges in distribution testing," Ph.D. dissertation, Massachusetts Institute of Technology, 2018.

[60] C. Daskalakis, G. C. Kamath, and J. Wright, "Which distribution distances are sublinearly testable?" in *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2018, pp. 2747–2764.

[61] L. Birgé, "On estimating a density using hellinger distance and some other strange facts," *Probability theory and related fields*, vol. 71, no. 2, pp. 271–291, 1986.

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

[63] P. Valiant, "Testing symmetric properties of distributions," *SIAM Journal on Computing*, vol. 40, no. 6, pp. 1927–1968, 2011.

[64] S. Balakrishnan and L. Wasserman, "Hypothesis testing for densities and high-dimensional multinomials: Sharp local minimax rates," *The Annals of Statistics*, vol. 47, no. 4, pp. 1893–1927, 2019.

[65] J. Chhor and A. Carpentier, "Sharp local minimax rates for goodness-of-fit testing in multivariate binomial and poisson families and in multinomials," *Mathematical Statistics and Learning*, vol. 5, no. 1/2, pp. 1–54, 2022.

[66] ——, "Goodness-of-fit testing for Hölder-continuous densities: Sharp local minimax rates," *arXiv preprint arXiv:2109.04346*, 2021.

[67] J. Lam-Weil, A. Carpentier, and B. K. Sriperumbudur, "Local minimax rates for closeness testing of discrete distributions," *Bernoulli*, vol. 28, no. 2, pp. 1179–1197, 2022.
[68] J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, S. Pan, and A. Suresh, "Competitive classification and closeness testing," in *Conference on Learning Theory*. JMLR Workshop and Conference Proceedings, 2012, pp. 22–1.
[69] Y. I. Ingster and I. A. Suslina, *Nonparametric goodness-of-fit testing under Gaussian models*. Springer Science & Business Media, 2003, vol. 169.
[70] L. Devroye, A. Mehrabian, and T. Reddad, "The total variation distance between high-dimensional gaussians," *arXiv preprint arXiv:1810.08693*, 2018.
[71] D. Newman and H. Shapiro, "Jackson's theorem in higher dimensions," in *On Approximation Theory/Über Approximationstheorie*. Springer, 1964, pp. 208–219.

Patrik Róbert Gerber was born in Kecskemét, Bács-Kiskun megye, Hungary in 1996. He received the MMath in Mathematics and Statistics in 2019 from University of Oxford, Oxford, UK in 2019, and the Ph.D. in Mathematics and Statistics from the Massachusetts Institute of Technology, Cambridge, USA in 2024. He is currently working as a quantitative researcher at Citadel Securities in New York City.

Yury Polyanskiy (S'08-M'10-SM'14-F'24) is a Professor of Electrical Engineering and Computer Science, a member of IDSS and LIDS at MIT. Yury received M.S. degree in applied mathematics and physics from the Moscow Institute of Physics and Technology, Moscow, Russia in 2005 and Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ in 2010. His research interests span information theory, machine learning and statistics. Dr. Polyanskiy won the 2020 IEEE Information Theory Society James Massey Award, 2013 NSF CAREER award and 2011 IEEE Information Theory Society Paper Award.