The Annals of Statistics
2019, Vol. 47, No. 4, 1893–1927
https://doi.org/10.1214/18-AOS1729

HYPOTHESIS TESTING FOR DENSITIES AND HIGH-DIMENSIONAL MULTINOMIALS: SHARP LOCAL MINIMAX RATES[1]

BY SIVARAMAN BALAKRISHNAN AND LARRY WASSERMAN

Carnegie Mellon University

We consider the goodness-of-fit testing problem of distinguishing whether the data are drawn from a specified distribution, versus a composite alternative separated from the null in the total variation metric. In the discrete case, we consider goodness-of-fit testing when the null distribution has a possibly growing or unbounded number of categories. In the continuous case, we consider testing a Hölder density with exponent $0 < s \leq 1$, with possibly unbounded support, in the low-smoothness regime where the Hölder parameter is not assumed to be constant. In contrast to existing results, we show that the minimax rate and critical testing radius in these settings depend strongly, and in a precise way, on the null distribution being tested and this motivates the study of the (local) minimax rate as a function of the null distribution. For multinomials, the local minimax rate has been established in recent work. We revisit and extend these results and develop two modifications to the χ^2-test whose performance we characterize. For testing Hölder densities, we show that the usual binning tests are inadequate in the low-smoothness regime and we design a spatially adaptive partitioning scheme that forms the basis for our locally minimax optimal tests. Furthermore, we provide the first local minimax lower bounds for this problem which yield a sharp characterization of the dependence of the critical radius on the null hypothesis being tested. In the low-smoothness regime, we also provide adaptive tests that adapt to the unknown smoothness parameter. We illustrate our results with a variety of simulations that demonstrate the practical utility of our proposed tests.

1. Introduction. Hypothesis testing is one of the pillars of modern mathematical statistics with a vast array of scientific applications. There is a well-developed theory of hypothesis testing starting with the work of Neyman and Pearson [29], and their framework plays a central role in the theory and practice of statistics. In this paper we revisit the classical goodness-of-fit testing problem of distinguishing the hypotheses:

$$H_0 : Z_1, \dots, Z_n \sim P_0 \quad \text{versus} \quad H_1 : Z_1, \dots, Z_n \sim P \in \mathcal{A} \tag{1.1}$$

for some set of distributions $\mathcal{A}$. This fundamental problem has been widely studied (see, for instance, [26] and references therein).

Received June 2017; revised May 2018.
[1]Supported in part by the NSF Grant DMS-17-13003.
MSC2010 subject classifications. 60K35.
Key words and phrases. Local-minimax, nonparametric goodness-of-fit testing.

A natural choice of the composite alternative, one that has a clear probabilistic interpretation, excludes a total variation neighborhood around the null, that is, we take $\mathcal{A} = \{P : \mathrm{TV}(P, P_0) \geq \varepsilon/2\}$. This is equivalent to $\mathcal{A} = \{P : \|P - P_0\|_1 \geq \varepsilon\}$, and we use this latter representation in the rest of this paper. However, there exist no consistent tests that can distinguish an arbitrary distribution P_0 from alternatives separated in ℓ_1; see [6, 24]. Hence, we impose structural restrictions on P_0 and $\mathcal{A}$. We focus on two cases:

1. *Multinomial testing*: When the null and alternate distributions are multinomials.
2. *Hölder testing*: When the null and alternate distributions have Hölder densities with Hölder exponent $0 < s \leq 1$.

The problem of goodness-of-fit testing for multinomials has a rich history in statistics and popular approaches are based on the χ^2-test [32] or the likelihood ratio test [11, 29, 39]; see, for instance, [15, 17, 28, 31, 33] and references therein. Motivated by connections to property testing [34], there is also a recent literature developing in computer science; see [7, 16, 19, 37]. Testing Hölder densities is one of the basic non-parametric hypothesis testing problems and tests are often based on the Kolmogorov–Smirnov or Cramér–von Mises statistics [13, 35, 38]. This problem was originally studied from the minimax perspective in the work of Ingster and coauthors [20, 22]. See [3, 18, 22] for further references.

In the goodness-of-fit testing problem in (1.1), previous results use the (global) critical radius as a benchmark. Roughly, this global critical radius is a measure of the minimal separation between the null and alternate hypotheses that ensures distinguishability, as the null hypothesis is varied over a large class of distributions (for instance over the class of distributions with Hölder densities or over the class of all multinomials on d categories). Remarkably, as shown in the work of Valiant and Valiant [37] for the case of multinomials and as we show in this paper for the case of Hölder densities, there is considerable heterogeneity in the critical radius as a function of the null distribution P_0. In other words, even within the class of Hölder densities, testing certain null hypotheses can be much easier than testing others. Consequently, the *local minimax rate* which describes the critical radius for each individual null distribution provides a much more nuanced picture. In this paper we provide (near) matching upper and lower bounds on the critical radii for Hölder testing as a function of the null distribution, that is, we precisely upper and lower bound the critical radius for each individual Hölder null hypothesis. Our upper bounds are based on χ^2-type tests, performed on a carefully chosen spatially adaptive binning, and highlight the fact that the standard prescriptions of choosing bins with a fixed width or with a fixed probability content [36] can yield suboptimal tests.

The distinction between local and global perspectives is reminiscent of similar effects that arise in some estimation problems, for instance in shape-constrained inference [9], in constrained least-squares problems [12] and in classical Fisher information-Cramér–Rao bounds [25].

The remainder of this paper is organized as follows. In Section 2, we provide some background on the minimax perspective on hypothesis testing, and formally describe the local and global minimax rates. We provide a detailed discussion of the problem of study and finally provide an overview of our main results. In Section 3, we review the results of Valiant and Valiant [37] and present a new globally-minimax test for testing multinomials, as well as a (nearly) locally-minimax test. In Section 4, we consider the problem of testing a Hölder density against a total variation neighborhood. We present the body of our main technical result in Section 4.3 and defer technical aspects of this proof to the Supplementary Material [4]. In each of Sections 3 and 4, we present simulation results that demonstrate the superiority of the tests we propose and their potential practical applicability. In the Supplementary Material [4], we also present several other results including a brief study of limiting distributions of the test statistics under the null, as well as tests that are adaptive to various parameters.

2. Background and problem setup. We begin with some basic background on hypothesis testing, the testing risk and minimax rates, before providing a detailed treatment of some related work.

2.1. *Hypothesis testing and minimax rates.* Our focus in this paper is on the one sample goodness-of-fit testing problem. We observe samples $Z_1, \dots, Z_n \in \mathcal{X}$, where $\mathcal{X} \subset \mathbb{R}^d$, which are independent and identically distributed with distribution P. For a fixed distribution P_0, we want to test the hypotheses:

$$H_0 : P = P_0 \quad \text{versus} \quad H_1 : \|P - P_0\|_1 \geq \varepsilon_n. \tag{2.1}$$

Throughout this paper, we use P_0 to denote the null distribution and P to denote an arbitrary alternate distribution. We use the total variation distance (or equivalently the ℓ_1 distance) between two distributions P and Q, defined by

$$\mathrm{TV}(P, Q) = \sup_A \big|P(A) - Q(A)\big|, \tag{2.2}$$

where the supremum is over all measurable sets. If P and Q have densities p and q with respect to a common dominating measure ν, then

$$\mathrm{TV}(P, Q) = \frac{1}{2}\int |p - q|\, d\nu = \frac{1}{2}\|p - q\|_1 \equiv \frac{1}{2}\|P - Q\|_1. \tag{2.3}$$

We consider the total variation distance because it has a clear probabilistic meaning and because it is invariant under one-to-one transformations [14]. The ℓ_2 metric is often easier to work with but in the context of distribution testing its interpretation is less intuitive. Of course, other metrics (for instance Hellinger, χ^2 or Kullback–Leibler) can be used as well but we focus on TV (or ℓ_1) throughout this paper. It is well understood [6, 24] that without further restrictions there are no uniformly consistent tests for distinguishing these hypotheses. Consequently, we focus on two restricted variants of this problem:

1. Multinomial testing: In the multinomial testing problem, the domain of the distributions is $\mathcal{X} = \{1, \ldots, d\}$ and the distributions P_0 and P are equivalently characterized by vectors $p_0, p \in \mathbb{R}^d$. Formally, we define

$$\mathcal{M} = \left\{ p : p \in \mathbb{R}^d, \sum_{i=1}^{d} p_i = 1, p_i \geq 0 \; \forall \; i \in \{1, \ldots, d\} \right\},$$

and consider the multinomial testing problem of distinguishing:

$$H_0 : P = P_0, P_0 \in \mathcal{M} \quad \text{versus} \quad H_1 : \|P - P_0\|_1 \geq \varepsilon_n, \qquad P \in \mathcal{M}. \tag{2.4}$$

In contrast to classical "fixed-cells" asymptotic theory [33], we focus on high-dimensional multinomials where d can grow with, and potentially exceed the sample size n.

2. Hölder testing: In the Hölder density testing problem the set $\mathcal{X} \subset \mathbb{R}^d$, and we restrict our attention to distributions with Hölder densities, that is, for a fixed Hölder exponent $0 < s \leq 1$, letting p_0 and p denote the densities of P_0 and P with respect to the Lebesgue measure, we consider the set of densities:

$$\mathcal{L}_s(L_n) = \left\{ p : \int_{\mathcal{X}} p(x)\, dx = 1, p(x) \geq 0 \; \forall x, \right.$$
$$\left. \left|p(x) - p(y)\right| \leq L_n \|x - y\|_2^s \; \forall x, y \in \mathbb{R}^d \right\},$$

and consider the Hölder testing problem of distinguishing:

$$\begin{aligned} &H_0 : P = P_0, P_0 \in \mathcal{L}_s(L_n) \quad \text{versus} \\ &H_1 : \|P - P_0\|_1 \geq \varepsilon_n, \qquad P \in \mathcal{L}_s(L_n). \end{aligned} \tag{2.5}$$

Throughout the paper, we refer to the fixed quantity s as the Hölder exponent, the parameter L_n as the Hölder parameter and to the testing problem described above as the Hölder testing problem (deferring a discussion of the case when $s > 1$ to Section 5). We emphasize that unlike prior work [3, 20] we do not require p_0 to be uniform. We also do not restrict the domain of the densities and we consider the low-smoothness regime where the Hölder parameter L_n is allowed to grow with the sample size.

Hypothesis testing and risk. Returning to the setting described in (2.1), we define a test ϕ as a Borel measurable map, $\phi : \mathcal{X}^n \mapsto \{0, 1\}$. For a fixed null distribution P_0, we define the set of level α tests:

$$\Phi_{n,\alpha} = \{\phi : P_0^n(\phi = 1) \leq \alpha\}. \tag{2.6}$$

The worst-case risk (type II error) of a test ϕ over a restricted class $\mathcal{C}$ which contains P_0 is

$$R_n(\phi; P_0, \varepsilon_n, \mathcal{C}) = \sup\{\mathbb{E}_P[1-\phi] : \|P - P_0\|_1 \geq \varepsilon_n, P \in \mathcal{C}\}.$$

The local minimax risk is:[2]

$$R_n(P_0, \varepsilon_n, \mathcal{C}) = \inf_{\phi \in \Phi_{n,\alpha}} R_n(\phi; P_0, \varepsilon_n, \mathcal{C}). \tag{2.7}$$

It is common to study the minimax risk via a coarse lens by studying instead the critical radius or the minimax separation. The critical radius is the smallest value ε_n for which a hypothesis test has nontrivial power to distinguish P_0 from the set of alternatives. Formally, we define the local critical radius as

$$\varepsilon_n(P_0, \mathcal{C}) = \inf\{\varepsilon_n : R_n(P_0, \varepsilon_n, \mathcal{C}) \leq 1/2\}. \tag{2.8}$$

The constant 1/2 is arbitrary; we could use any number in $(0, 1-\alpha)$.

The local minimax risk and critical radius depend on the null distribution P_0. A more common quantity of interest is the *global* minimax risk

$$R_n(\varepsilon_n, \mathcal{C}) = \sup_{P_0 \in \mathcal{C}} R_n(P_0, \varepsilon_n, \mathcal{C}). \tag{2.9}$$

The corresponding global critical radius is

$$\varepsilon_n(\mathcal{C}) = \inf\{\varepsilon_n : R_n(\varepsilon_n, \mathcal{C}) \leq 1/2\}. \tag{2.10}$$

In typical nonparametric problems, the local minimax risk and the global minimax risk match up to constants and this has led researchers in past work to focus on the global minimax risk. We show that for the distribution testing problems we consider, the local critical radius in (2.8) can vary considerably as a function of the null distribution P_0. As a result, the global critical radius, provides only a partial understanding of the intrinsic difficulty of this family of hypothesis testing problems. In this paper we focus on producing tight bounds on the local minimax separation. These bounds yield as a simple corollary, sharp bounds on the global minimax separation, but are in general considerably more refined.

Notation: For two sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, we write $a_n \asymp b_n$ if $0 < \liminf_{n\to\infty} |a_n|/|b_n| \leq \limsup_{n\to\infty} |a_n|/|b_n| < \infty$.

Poissonization: In constructing upper bounds on the minimax risk—we work under a simplifying assumption that the sample size is random: $n_0 \sim \text{Poisson}(n)$. This assumption is standard in the literature, and simplifies several calculations. When the sample size is chosen to be distributed as $\text{Poisson}(n)$, it is straightforward to verify that for any fixed set $A, B \subset \mathcal{X}$ with $A \cap B = \varnothing$, under P the number of samples falling in A and B are distributed independently as $\text{Poisson}(nP(A))$ and $\text{Poisson}(nP(B))$, respectively.

[2]Although our proofs are explicit in their dependence on α, we suppress this dependence in our notation and in our main results, treating α as a fixed strictly positive universal constant.

In the Poissonized setting, we consider the averaged minimax risk, where we additionally average the risk in (2.7) over the random sample size. The Poisson distribution is tightly concentrated around its mean and this additional averaging only affects constant factors in the minimax risk and we ignore this averaging in the rest of the paper.

2.2. *Overview of our results*. With the basic framework in place, we now provide a high-level overview of the main results of this paper. In the context of testing multinomials, the results of Valiant and Valiant [37] characterize the local and global minimax rates. We provide the following additional results:

- In Theorem 3.2, we characterize a simple and practical globally minimax test. In Theorem 3.4 building on the results of Diakonikolas and Kane [16] we provide a simple (near) locally minimax test.

In the context of testing Hölder densities, we make advances over classical results [18, 22] by eliminating several unnecessary assumptions (uniform null, bounded support, fixed Hölder parameter). We provide the first characterization of the local minimax rate for this problem. In studying the Hölder testing problem in its full generality, we find that the critical testing radius can exhibit a wide range of possible behaviours, based roughly on the tail behaviour of the null hypothesis.

- In Theorem 4.1, we provide a characterization of the local minimax rate for Hölder density testing. In Section 4.1, we consider a variety of concrete examples that demonstrate the rich scaling behaviour exhibited by the critical radius in this problem.
- Our upper and lower bounds are based on a novel spatially adaptive partitioning scheme. We describe this scheme and derive some of its useful properties in Section 4.2.

In the Supplementary Material, we provide the technical details of the proofs. We briefly consider the limiting behaviour of our test statistics under the null in the Supplementary Material [4] (Appendix A). Our results show that the critical radius is determined by a certain functional of the null hypothesis. In the Supplementary Material [4] (Appendix D), we study certain important properties of this functional pertaining to its stability. Finally, in the Supplementary Material we also study tests which are adaptive to various parameters (Appendix F).

3. Testing high-dimensional multinomials. Given a sample $Z_1, \dots, Z_n \sim P$ define the counts $X = (X_1, \dots, X_d)$ where $X_j = \sum_{i=1}^n I(Z_i = j)$. The local minimax critical radii for the multinomial problem have been found in Valiant and Valiant [37]. We begin by summarizing these results.

Without loss of generality, we assume that the entries of the null multinomial p_0 are sorted so that $p_0(1) \geq p_0(2) \geq \cdots \geq p_0(d)$. For any $0 \leq \sigma \leq 1$, we denote

σ-tail of the multinomial by

$$\mathcal{Q}_\sigma(p_0) = \left\{ i : \sum_{j=i}^{d} p_0(j) \le \sigma \right\}. \tag{3.1}$$

The σ-bulk is defined to be

$$\mathcal{B}_\sigma(p_0) = \{ i > 1 : i \notin \mathcal{Q}_\sigma(p_0) \}. \tag{3.2}$$

Note that $i = 1$ is excluded from the σ-bulk. The minimax rate depends on the functional

$$V_\sigma(p_0) = \left(\sum_{i \in \mathcal{B}_\sigma(p_0)} p_0(i)^{2/3} \right)^{3/2}. \tag{3.3}$$

For a given multinomial p_0, our goal is to upper and lower bound the local critical radius $\varepsilon_n(p_0, \mathcal{M})$ in (2.8). We define, ℓ_n and u_n to be the solutions to the equations[3]

$$\begin{aligned} \ell_n(p_0) &= \max\left\{ \frac{1}{n}, \sqrt{\frac{V_{\ell_n(p_0)}(p_0)}{n}} \right\}, \\ u_n(p_0) &= \max\left\{ \frac{1}{n}, \sqrt{\frac{V_{u_n(p_0)/16}(p_0)}{n}} \right\}. \end{aligned} \tag{3.4}$$

With these definitions in place, we are now ready to state the result of Valiant and Valiant [37]. We use $c_1, c_2, C_1, C_2 > 0$ to denote positive universal constants.

THEOREM 3.1 (Valiant and Valiant [37]). *The local critical radius $\varepsilon_n(p_0, \mathcal{M})$ for multinomial testing is upper and lower bounded as*

$$c_1 \ell_n(p_0) \le \varepsilon_n(p_0, \mathcal{M}) \le C_1 u_n(p_0). \tag{3.5}$$

Furthermore, the global critical radius $\varepsilon_n(\mathcal{M})$ is bounded as

$$\frac{c_2 d^{1/4}}{\sqrt{n}} \le \varepsilon_n(\mathcal{M}) \le \frac{C_2 d^{1/4}}{\sqrt{n}}.$$

REMARKS.

- The local critical radius is roughly determined by the (truncated) 2/3rd norm of the multinomial p_0. This norm is maximized when p_0 is uniform and is small when p_0 is sparse, and at a high-level captures the "effective sparsity" of p_0.

[3]These equations always have a unique solution since the right-hand side monotonically decreases to 0 as the left-hand side monotonically increases from 0 to 1.

- The global critical radius can shrink to zero even when $d \gg n$. When $d \asymp n^2$ almost all categories of the multinomial are unobserved but it is still possible to reliably distinguish any p_0 from an ℓ_1-neighborhood. This phenomenon is noted for instance in the work of [31]. We also note the work of [6] that shows that when $d = \omega(n)$, no test can have power that approaches 1 at an exponential rate.
- The local critical radius can be much smaller than the global minimax radius. If the multinomial p_0 is nearly (or exactly) s-sparse, then the critical radius is upper and lower bounded up to constants by $s^{1/4}/\sqrt{n}$. Furthermore, these results also show that it is possible to design consistent tests for sufficiently structured null hypotheses: in cases when $\sqrt{d} \gg n$, and even in cases when d is infinite.
- Except for certain pathological multinomials, the upper and lower critical radii match up to constants. We revisit this issue in the Supplementary Material [4] (Appendix D), in the context of Hölder densities, where we present examples where the solutions to critical equations similar to (3.4) are stable and examples where they are unstable.

In the remainder of this section, we consider a variety of tests, including the test presented in [37] and several alternatives. The test of Valiant and Valiant [37] is a composite test that requires knowledge of ε_n and the analysis of their test is quite intricate. We present an alternative, simple test that is globally minimax, and then present an alternative composite test that is locally minimax but simpler to analyze. Finally, we present a few illustrative simulations.

3.1. *The truncated χ^2 test.* We begin with a simple globally minimax test. From a practical standpoint, the most popular test for multinomials is Pearson's χ^2 test. However, in the high-dimensional regime where the dimension of the multinomial d is not treated as fixed the χ^2 test can have bad power due to the fact that the variance of the χ^2 statistic is dominated by small entries of the multinomial (see [27, 37]).

A natural thought then is to truncate the normalization factors of the χ^2 statistic in order to limit the contribution to the variance from each cell of the multinomial. Recalling that $(X_1, \dots, X_d)$ denote the observed counts, we propose the test statistic:

$$T_{\text{trunc}} = \sum_{i=1}^{d} \frac{(X_i - np_0(i))^2 - X_i}{\max\{1/d, p_0(i)\}} := \sum_{i=1}^{d} \frac{(X_i - np_0(i))^2 - X_i}{\theta_i} \tag{3.6}$$

and the corresponding test,

$$\phi_{\text{trunc}} = \mathbb{I}\left(T_{\text{trunc}} > n\sqrt{\frac{2}{\alpha}\sum_{i=1}^{d} \frac{p_0(i)^2}{\theta_i^2}}\right). \tag{3.7}$$

This test statistic truncates the usual normalization factor for the χ^2 test for any entry which falls below $1/d$, and thus ensures that very small entries in p_0 do not have a large effect on the variance of the statistic. We emphasize the simplicity and practicality of this test. We have the following result which bounds the power and size of the truncated χ^2 test. We use $C > 0$ to denote a positive universal constant.

THEOREM 3.2. *Consider the testing problem in* (2.4). *The truncated* χ^2 *test has size at most* α, *i.e.* $P_0(\phi_{\text{trunc}} = 1) \leq \alpha$. *Furthermore*, *there is a universal constant* $C > 0$ *such that if for any* $0 < \zeta \leq 1$ *we have that*

$$\varepsilon_n^2 \geq \frac{C\sqrt{d}}{n}\left[\frac{1}{\sqrt{\alpha}} + \frac{1}{\zeta}\right], \tag{3.8}$$

then the Type II error of the test is bounded by ζ, *that is*, $P(\phi_{\text{trunc}} = 0) \leq \zeta$.

REMARKS.

- A straightforward consequence of this result together with the result in Theorem 3.1 is that the truncated χ^2 test is globally minimax optimal.
- The classical χ^2 and likelihood ratio tests are not generally consistent (and thus not globally minimax optimal) in the high-dimensional regime (see also, Figure 2).
- At a high-level the proof follows by verifying that when the alternate hypothesis is true, under the condition on the critical radius in (3.8), the test statistic is larger than the threshold in (3.7). To verify this, we lower bound the mean and upper bound the variance of the test statistic under the alternate and then use standard concentration results. We defer the details to the Supplementary Material [4] (Appendix B).

3.2. *The* $2/3rd$ + *tail test.* The truncated χ^2 test described in the previous section, although globally minimax, is not locally adaptive. The test from [37], achieves the local minimax upper bound in Theorem 3.1. We refer to this as the $2/3$rd + tail test. We use a slightly modified version of their test when testing Hölder goodness-of-fit in Section 4, and provide a description here.

The test is a composite two-stage test, and has a tuning parameter σ. Recalling the definitions of $\mathcal{B}_\sigma(p_0)$ and $\mathcal{Q}_\sigma(p_0)$ [see (3.1)], we define two test statistics T_1, T_2 and corresponding test thresholds t_1, t_2:

$$T_1(\sigma) = \sum_{j \in \mathcal{Q}_\sigma(p_0)} \big(X_j - np_0(j)\big), \qquad t_1(\alpha, \sigma) = \sqrt{\frac{nP_0(\mathcal{Q}_\sigma(p_0))}{\alpha}},$$

$$T_2(\sigma) = \sum_{j \in \mathcal{B}_\sigma(p_0)} \frac{(X_j - np_0(j))^2 - X_j}{p_0(j)^{2/3}}, \qquad t_2(\alpha, \sigma) = \sqrt{\frac{\sum_{j \in \mathcal{B}_\sigma} 2n^2 p_0(j)^{2/3}}{\alpha}}.$$

We define two tests:

1. The tail test: $\phi_{\text{tail}}(\sigma, \alpha) = \mathbb{I}(T_1(\sigma) > t_1(\alpha, \sigma))$.
2. The 2/3-test: $\phi_{2/3}(\sigma, \alpha) = \mathbb{I}(T_2(\sigma) > t_2(\alpha, \sigma))$.

The composite test $\phi_V(\sigma, \alpha)$ is then given as

$$\phi_V(\sigma, \alpha) = \max\{\phi_{\text{tail}}(\sigma, \alpha/2), \phi_{2/3}(\sigma, \alpha/2)\}. \tag{3.9}$$

With these definitions in place, the following result is essentially from the work of Valiant and Valiant [37]. We use $C > 0$ to denote a positive universal constant.

THEOREM 3.3. *Consider the testing problem in* (2.4). *The composite test* $\phi_V(\sigma, \alpha)$ *has size at most* α, *that is*, $P_0(\phi_V = 1) \leq \alpha$. *Furthermore, if we choose* $\sigma = \varepsilon_n(p_0, \mathcal{M})/8$, *and* $u_n(p_0)$ *as in* (3.4), *then for any* $0 < \zeta \leq 1$, *if*

$$\varepsilon_n(p_0, \mathcal{M}) \geq C u_n(p_0) \max\{1/\alpha, 1/\zeta\}, \tag{3.10}$$

then the Type II error of the test is bounded by ζ, *that is*, $P(\phi_V = 0) \leq \zeta$.

REMARKS.

- The test ϕ_V is also motivated by deficiencies of the χ^2 test. In particular, the test includes two main modifications to the χ^2 test which limit the contribution of the small entries of p_0: some of the small entries of p_0 are dealt with via a separate tail test and further the normalization of the χ^2 test is changed from $p_0(i)$ to $p_0(i)^{2/3}$.
- This result provides the upper bound of Theorem 3.1. It requires that the tuning parameter σ is chosen as $\varepsilon_n(p_0, \mathcal{M})/8$. In the Supplementary Material [4] (Appendix F), we discuss adaptive choices for σ.
- The proof essentially follows from the paper of Valiant and Valiant [37], but we maintain an explicit bound on the power and size of the test, which we use in later sections. We provide the details in the Supplementary Material [4] (Appendix B).

While the 2/3rd norm test is locally minimax optimal its analysis is quite challenging. In the next section, we build on results from a recent paper of Diakonikolas and Kane [16] to provide an alternative (nearly) locally minimax test with a simpler analysis.

3.3. *The max test.* An important insight, one that is seen for instance in Figure 1, is that many simple tests are optimal when p_0 is uniform and that careful modifications to the χ^2 test are required only when p_0 is far from uniform. This suggests the following strategy: partition the multinomial into nearly uniform groups, apply a simple test within each group and combine the tests with an appropriate Bonferroni correction. We refer to this as the max test. Such a strategy was used by Diakonikolas and Kane [16], but their test is quite complicated and involves many constants. Furthermore, it is not clear how to ensure that their test controls the Type I error at level α. Motivated by their approach, we present a simple test that controls the type I error as required and is (nearly) locally minimax.

As with the test in the previous section, the test has to be combined with the tail test. The test is defined to be

$$\phi_{\max}(\sigma,\alpha) = \max\{\phi_{\text{tail}}(\sigma,\alpha/2), \phi_M(\sigma,\alpha/2)\},$$

where ϕ_M is defined as follows. We partition $\mathcal{B}_\sigma(p_0)$ into sets S_j for $j \geq 1$, where

$$S_j = \left\{t : \frac{p_0(2)}{2^j} < p_0(t) \leq \frac{p_0(2)}{2^{j-1}}\right\}.$$

We can bound the total number of sets S_j by noting that for any $i \in \mathcal{B}_\sigma(p_0)$, we have that $p_0(i) \geq \sigma/d$, so that the number of sets k is bounded by $\lceil \log_2(d/\sigma)\rceil$. Within each set, we use a modified ℓ_2 statistic. Let

$$T_j = \sum_{t\in S_j} \left[\left(X_t - np_0(t)\right)^2 - X_t\right] \tag{3.11}$$

for $j \geq 1$. Unlike the traditional ℓ_2 statistic, each term in this statistic is centered around X_t. As observed in [37], this results in the statistic having smaller variance in the $n \ll d$ regime. Let

$$\phi_M(\sigma,\alpha) = \bigvee_j \mathbb{I}(T_j > t_j), \tag{3.12}$$

where

$$t_j = \sqrt{\frac{2kn^2[\sum_{t\in S_j} p_0(t)^2]}{\alpha}}. \tag{3.13}$$

THEOREM 3.4. *Consider the testing problem in* (2.4). *Suppose we choose* $\sigma = \varepsilon_n(p_0,\mathcal{M})/8$, *then the composite test* $\phi_{\max}(\sigma,\alpha)$ *has size at most* α, *that is,* $P_0(\phi_{\max} = 1) \leq \alpha$. *Furthermore, there is a universal constant* $C > 0$, *such that for* $u_n(p_0)$ *as in* (3.4), *if for any* $0 < \zeta \leq 1$ *we have that*

$$\varepsilon_n(p_0,\mathcal{M}) \geq Ck^2 u_n(p_0) \max\left\{\frac{\sqrt{k}}{\alpha}, \frac{1}{\zeta}\right\}, \tag{3.14}$$

then the Type II error of the test is bounded by ζ, *that is,* $P(\phi_{\max} = 0) \leq \zeta$.

REMARKS.

- Comparing the critical radii in equations (3.14) and (3.5), and noting that $k \leq \lceil \log_2(8d/\varepsilon_n)\rceil$, we conclude that the max test is locally minimax optimal, up to a logarithmic factor.

- In contrast to the analysis of the 2/3rd + tail test in [37], the analysis of the max test involves mostly elementary calculations. We provide the details in the Supplementary Material [4] (Appendix B). As emphasized in the work of Diakonikolas and Kane [16], the reduction of testing problems to simpler testing problems (in this case, testing uniformity) is a more broadly useful idea. Our upper bound for the Hölder testing problem (in Section 4) proceeds by reducing it to a multinomial testing problem through a spatially adaptive binning scheme.

3.4. *Simulations*. In this section, we report some simulation results that demonstrate the practicality of the proposed tests. We focus on two simulation scenarios and compare the globally-minimax truncated χ^2 test, and the 2/3rd + tail test [37] with the classical χ^2-test, the likelihood ratio test, the ℓ_1 test and the ℓ_2 test. The test statistics are

$$T_{\chi^2} = \sum_{i=1}^{d} \frac{(X_i - np_0(i))^2 - np_0(i)}{np_0(i)}, \qquad T_{\mathrm{LRT}} = 2\sum_{i=1}^{d} X_i \log\left(\frac{X_i}{np_0(i)}\right),$$

$$T_{\ell_1} = \sum_{i=1}^{d} |X_i - np_0(i)|, \qquad T_{\ell_2} = \sum_{i=1}^{d} (X_i - np_0(i))^2.$$

In the Supplementary Material [4] (Appendix G), we consider a few additional simulations.

In each setting described below, we set the α level threshold via simulation (by sampling from the null 1000 times) and we calculate the power under particular alternatives by averaging over a 1000 trials.

1. Figure 1 considers a high-dimensional setting where $n = 300$, $d = 2000$, the null distribution is uniform, and the alternate is either dense (perturbing each coordinate by a scaled Rademacher) or sparse (perturbing only two coordinates).

In each case, we observe that all the tests perform comparably indicating that a variety of tests are optimal around the uniform distribution, a fact that we exploit in designing the max test. The test from [37] performs slightly worse than others due to the Bonferroni correction from applying a two-stage test.

2. Figure 2 considers a power-law null where $p_0(i) \propto 1/i$. Again we take $n = 300$, $d = 2000$ and compare against a dense and sparse alternative. In this setting, we choose the sparse alternative to only perturb the first two coordinates of the distribution.

We observe two notable effects. First, we see that when the alternate is dense, the truncated χ^2 test, although consistent in the high-dimensional regime, is outperformed by the other tests highlighting the need to study the local-minimax properties of tests. Perhaps more surprising is that in the setting where the alternate is sparse, the classical χ^2 and likelihood ratio tests can fail dramatically.

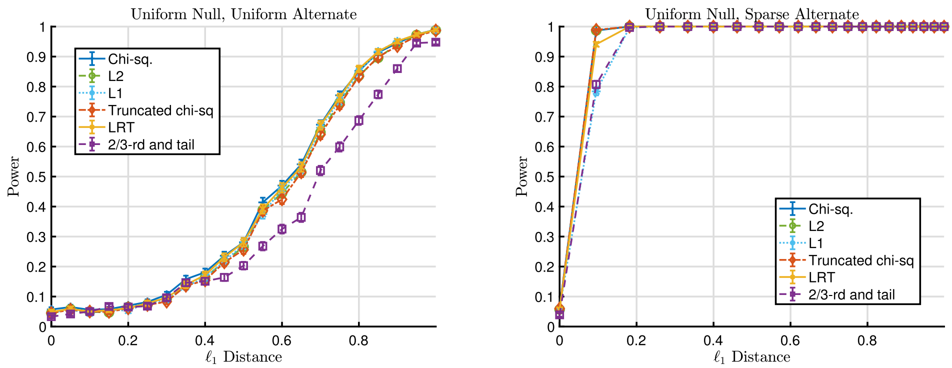

FIG. 1. *A comparison between the truncated χ^2 test, the 2/3rd + tail test* [37], *the χ^2-test and the likelihood ratio test. The null is chosen to be uniform, and the alternate is either a dense or sparse perturbation of the null. The power of the tests are plotted against the ℓ_1 distance between the null and alternate. Each point in the graph is an average over* 1000 *trials. Despite the high-dimensionality* (*i.e.*, $n = 300$, $d = 2000$) *the tests have high-power, and perform comparably.*

The locally minimax test is remarkably robust across simulation settings. However, it requires that we specify ε_n, a drawback shared by the max test. In the Supplementary Material [4] (Appendix F), we provide adaptive alternatives that adapt to unknown ε_n.

4. Testing Hölder densities. In this section, we focus our attention on the Hölder testing problem (2.5). As is standard in nonparametric problems, through-

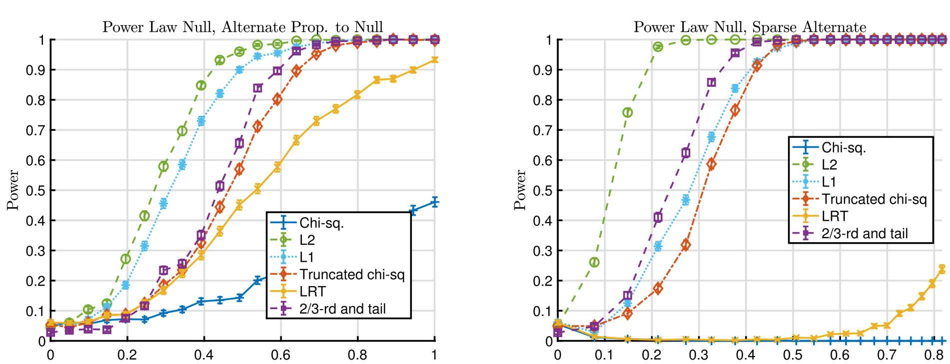

FIG. 2. *A comparison between the truncated χ^2 test, the 2/3rd + tail test* [37], *the χ^2-test and the likelihood ratio test. The null is chosen to be a power law, and the alternate is either a dense or sparse perturbation of the null. The power of the tests are plotted against the ℓ_1 distance between the null and alternate. Each point in the graph is an average over* 1000 *trials. The truncated χ^2 test despite being globally minimax optimal, can perform poorly for any particular fixed null. The χ^2 and likelihood ratio tests can fail to be consistent even when ε_n is quite large, and $n \gg \sqrt{d}$.*

out this section, we treat the dimension d as a fixed (universal) constant. Our emphasis is on understanding the local critical radius while making minimal assumptions. In contrast to past work, we do not assume that the null is uniform or even that its support is compact. We would like to be able to detect more subtle deviations from the null as the sample size gets large, and hence we do not assume that the Hölder parameter L_n is fixed as n grows.

The classical method, due to [20, 21] to constructing lower and upper bounds on the critical radius, is based on binning the domain of the density. In particular, upper bounds were obtained by considering χ^2 tests applied to the multinomial that results from binning the null distribution. Ingster focused on the case when the null distribution P_0 was taken to be uniform on [0, 1], noting that the testing problem for a general null distribution could be "reduced" to testing uniformity by modifying the observations via the quantile transformation corresponding to the null distribution P_0 (see also [18]). We emphasize that such a reduction *alters the smoothness class* tailoring it to the null distribution P_0. The quantile transformation forces the deviations from the null distribution to be more smooth in regions where P_0 is small and less smooth where P_0 is large, that is, we need to reinterpret smoothness of the alternative density p as an assumption about the function $p(F_0^{-1}(t))$, where F_0^{-1} is the quantile function of the null distribution P_0. We find this assumption to be unnatural and instead aim to directly test the hypotheses in (2.5). We note that some upper bounds for directly testing nonuniform densities against ℓ_2-alternatives without appealing to a quantile transform appear, for instance, in [18].

We begin with some high-level intuition for our upper and lower bounds.

- *Upper bounding the critical radius*: The strategy of binning domain of p_0, and then testing the resulting multinomial against an appropriate ℓ_1 neighborhood using a locally minimax test is natural even when p_0 is not uniform. However, there is considerable flexibility in how precisely to bin the domain of p_0. Essentially, the only constraint in the choice of bin-widths is that the approximation error (of approximating the density by its piecewise constant, histogram approximation) is sufficiently well controlled. When the null is not uniform the choice of fixed bin-widths is arbitrary and as we will see, suboptimal. A bulk of the technical effort in constructing our optimal tests is then in determining the optimal inhomogeneous, spatially adaptive partition of the domain in order to apply a multinomial test.
- *Lower bounding the critical radius*: At a high-level the construction of Ingster is similar to standard lower bounds in nonparametric problems. Roughly, we create a collection of possible alternate densities, by evenly partitioning the domain of p_0, and then perturbing each cell of the partition by adding or subtracting a small (sufficiently smooth) bump. We then analyze the optimal likelihood ratio test for the (simple versus simple) testing problem of distinguishing p_0 from a uniform mixture of the set of possible alternate densities. We observe that when p_0 is

not uniform once again creating a fixed bin-width partition is not optimal. The optimal choice of bin-widths is to choose larger bin-widths when p_0 is large and smaller bin-widths when p_0 is small. Intuitively, this choice allows us to perturb the null distribution p_0 more when the density is large, without violating the constraint that the alternate distributions remain sufficiently smooth. Once again, we create an inhomogeneous, spatially adaptive partition of the domain, and then use this partition to construct the optimal perturbation of the null.

Define,

$$\gamma := \frac{2s}{3s+d}, \tag{4.1}$$

and for any $0 \le \sigma \le 1$ denote the collection of sets of probability mass at least $1-\sigma$ as $\mathcal{B}_\sigma$, that is, $\mathcal{B}_\sigma := \{B : P_0(B) \ge 1-\sigma\}$. Define the functional,

$$T_\sigma(p_0) := \inf_{B \in \mathcal{B}_\sigma} \left(\int_B p_0^\gamma(x)\,dx \right)^{1/\gamma}. \tag{4.2}$$

We refer to this as the truncated T-functional.[4] The functional $T_\sigma(p_0)$ is the analog of the functional $V_\sigma(p_0)$ in (3.3), and roughly characterizes the local critical radius. We return to study this functional in light of several examples, in Section 4.1 and the Supplementary Material [4], Appendix D.

In constructing lower bounds, we will assume that the null density lies in the interior of the Hölder ball, that is, we assume that for some constant $0 \le c_{\text{int}} < 1$, we have that, $p_0 \in \mathcal{L}_s(c_{\text{int}} L_n)$. This assumption avoids certain technical issues that arise in creating perturbations of the null density when it lies on the boundary of the Hölder ball.

Finally, we define for two universal constants $C \ge c > 0$ (that are explicit in our proofs) the upper and lower critical radii:

$$\begin{aligned} v_n(p_0) &= \left(\frac{L_n^{d/(2s)} T_{C v_n(p_0)}(p_0)}{n} \right)^{\frac{2s}{4s+d}}, \\ w_n(p_0) &= \left(\frac{L_n^{d/(2s)} T_{c w_n(p_0)}(p_0)}{n} \right)^{\frac{2s}{4s+d}}. \end{aligned} \tag{4.3}$$

With these preliminaries in place, we now state our main result on testing Hölder densities. We let $c, C > 0$ denote two positive universal constants (different from the ones above).

THEOREM 4.1. *The local critical radius $\varepsilon_n(p_0, \mathcal{L}_s(L_n))$ for testing Hölder densities is upper bounded as*

$$\varepsilon_n\big(p_0, \mathcal{L}_s(L_n)\big) \le C w_n(p_0). \tag{4.4}$$

[4] Although the set B that achieves the minimum in the definition of $T_\sigma(p_0)$ need not be unique, the functional itself is well defined.

Furthermore, if for some constant $0 \le c_{int} < 1$ *we have that,* $p_0 \in \mathcal{L}_s(c_{int} L_n)$*, then the critical radius is lower bounded as*

$$c v_n(p_0) \le \varepsilon_n\big(p_0, \mathcal{L}_s(L_n)\big). \tag{4.5}$$

REMARKS.

- A natural question of interest is to understand the worst-case rate for the critical radius, or equivalently to understand the largest that the T-functional can be. Since the T-functional can be infinite if the support is unrestricted, we restrict our attention to Hölder densities with a bounded support S. In this case, letting $\mu(S)$ denote the Lebesgue measure of S and using Hölder's inequality (see the Supplementary Material [4], Appendix D) we have that for any $\sigma > 0$,

$$T_\sigma(p_0) \le (1-\sigma)\mu(S)^{\frac{1-\gamma}{\gamma}}. \tag{4.6}$$

  Up to constants involving γ, σ this is attained when p_0 is uniform on the set S. In other words, the critical radius is maximal for testing the uniform density against a Hölder, ℓ_1 neighborhood. In this case, we simply recover a generalization of the result of [20] for testing when p_0 is uniform on [0, 1].
- The main discrepancy between the upper and lower bounds is in the truncation level, that is, the upper and lower bounds depend on the functional $T_\sigma(p_0)$ for different values of the parameter σ. This is identical to the situation in Theorem 3.1 for testing multinomials. In most nonpathological examples this functional is stable with respect to constant factor discrepancies in the truncation level and consequently our upper and lower bounds are typically tight (see the examples in Section 4.1). In the Supplementary Material (see Appendix D), we formally study the stability of the T-functional. We provide general bounds and relate the stability of the T-functional to the stability of the level-sets of p_0.

The remainder of this section is organized as follows: we first consider various examples, calculate the T-functional and develop the consequences of Theorem 4.1 for these examples. We then turn our attention to our adaptive binning, describing both a recursive partitioning algorithm for constructing it as well as developing some of its useful properties. Finally, we provide the body of our proof of Theorem 4.1 and defer more technical aspects to the Supplementary Material. We conclude with a few illustrative simulations.

4.1. *Examples.* The result in Theorem 4.1 provides a general characterization of the critical radius for testing any density p_0, against a Hölder, ℓ_1 neighborhood. In this section, we consider several concrete examples. Although our theorem is more generally applicable, for ease of exposition we focus on the setting where $d = 1$ and $s = 1$ (i.e., the Lipschitz case) highlighting the variability of the T-functional and consequently of the critical radius as the null density is changed. Our examples have straightforward d-dimensional extensions.

When $d = 1$, $s = 1$, we have that $\gamma = 1/2$ so the T-functional is simply

$$T_\sigma(p_0) = \inf_{B \in \mathcal{B}_\sigma} \left(\int_B \sqrt{p_0(x)}\, dx \right)^2,$$

where $\mathcal{B}_\sigma$ is as before. Our interest in general is in the setting where $\sigma \to 0$ (which happens as $n \to \infty$), so in some examples we will simply calculate $T_0(p_0)$. In other examples, however, the truncation at level σ will play a crucial role and in those cases we will compute $T_\sigma(p_0)$.

EXAMPLE 4.1 (Uniform null). Suppose that the null distribution p_0 is Uniform$[a, b]$ then

$$T_0(p_0) = |b - a|.$$

EXAMPLE 4.2 (Gaussian null). Suppose that the null distribution p_0 is a Gaussian, that is, for some $\nu > 0$, $\mu \in \mathbb{R}$,

$$p_0(x) = \frac{1}{\sqrt{2\pi}\,\nu} \exp\bigl(-(x - \mu)^2/(2\nu^2)\bigr).$$

In this case, a simple calculation (see the Supplementary Material [4], Appendix C) shows that

$$T_0(p_0) = (8\pi)^{1/2}\nu.$$

EXAMPLE 4.3 (Beta null). Suppose that the null density is a Beta distribution:

$$p_0(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} x^{\beta-1} = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} x^{\beta-1},$$

where Γ and B denote the gamma and beta functions, respectively. It is easy to verify that

$$\begin{aligned} T_0(p_0) &= \left(\int_0^1 \sqrt{p_0(x)}\, dx \right)^2 \\ &= \frac{B^2((\alpha + 1)/2, (\beta + 1)/2)}{B(\alpha, \beta)}. \end{aligned}$$

To get some sense of the behaviour of this functional, we consider the case when $\alpha = \beta = t \to \infty$. In this case, we show (see the Supplementary Material [4], Appendix C) that for $t \ge 1$,

$$\frac{\pi^2}{4e^4} t^{-1/2} \le T_0(p_0) \le \frac{e^4}{4} t^{-1/2}.$$

In particular, we have that $T_0(p_0) \asymp t^{-1/2}$.

REMARK.

- These examples illustrate that in the simplest settings when the density p_0 is close to uniform, the T-functional is roughly the effective support of p_0. In each of these cases, it is straightforward to verify that the truncation of the T-functional simply affects constants so that the critical radius scales as
$$\varepsilon_n \asymp \left(\frac{\sqrt{L_n} T_0(p_0)}{n}\right)^{2/5},$$
where $T_0(p_0)$ in each case scales as roughly the size of the $(1-\varepsilon_n)$-support of the density p_0, that is, as the Lebesgue measure of the smallest set that contains $(1-\varepsilon_n)$ probability mass. This motivates understanding the Lipschitz density with smallest effective support, and we consider this next.

EXAMPLE 4.4 (Spiky null). Suppose that the null hypothesis is
$$p_0(x) = \begin{cases} L_n x, & 0 \le x \le \frac{1}{\sqrt{L_n}}, \\ 2\sqrt{L_n} - L_n x, & \frac{1}{\sqrt{L_n}} \le x \le \frac{2}{\sqrt{L_n}}, \\ 0, & \text{otherwise,} \end{cases}$$
then we have that $T_0(p_0) \asymp \frac{1}{\sqrt{L_n}}$.

REMARK.

- For the spiky null distribution we obtain an extremely fast rate, that is, we have that the critical radius $\varepsilon_n \asymp n^{-2/5}$, and is independent of the Lipschitz parameter L_n (although, we note that the null p_0 is more spiky as L_n increases). This is the fastest rate we obtain for Lipschitz testing. In settings where the tail decay is slow, the truncation of the T-functional can be crucial and the rates can be much slower. We consider these examples next.

EXAMPLE 4.5 (Cauchy distribution). The mean zero, Cauchy distribution with parameter α has pdf:
$$p_0(x) = \frac{1}{\pi\alpha}\frac{\alpha^2}{x^2+\alpha^2}.$$
As we show (see the Supplementary Material [4] [Appendix C]), the T-functional without truncation is infinite, that is, $T_0(p_0) = \infty$. However, the truncated T-functional is finite. In the Supplementary Material, we show that for any $0 \le \sigma \le 0.5$ (recall that our interest is in cases where $\sigma \to 0$),
$$\frac{4\alpha}{\pi}\left[\ln^2\left(\frac{1}{\sigma}\right)\right] \le T_\sigma(p_0) \le \frac{4\alpha}{\pi}\left[\ln^2\left(\frac{2e}{\pi\sigma}\right)\right],$$
that is, we have that $T_\sigma(p_0) \asymp \ln^2(1/\sigma)$.

REMARK.

- When the null distribution is Cauchy as above, we note that the rate for the critical radius is no longer the typical $\varepsilon_n \asymp n^{-2/5}$, even when the other problem specific parameters (L_n and the Cauchy parameter α) are held fixed. We instead obtain a slower $\varepsilon_n \asymp (n/\log^2 n)^{-2/5}$ rate. Our final example shows that we can obtain an entire spectrum of slower rates.

EXAMPLE 4.6 (Pareto null). For a fixed $x_0 > 0$ and for $0 < \alpha < 1$, suppose that the null distribution is

$$p_0(x) = \begin{cases} \dfrac{\alpha x_0^\alpha}{x^{\alpha+1}} & \text{for } x \geq x_0, \\ 0 & \text{for } x < x_0. \end{cases}$$

This distribution for $0 < \alpha < 1$ has thicker tails than the Cauchy distribution. The T-functional without truncation is infinite, that is, $T_0(p_0) = \infty$, and we can further show that (see the Supplementary Material [4], Appendix C):

$$\frac{4\alpha x_0}{(1-\alpha)^2}\big(\sigma^{-\frac{1-\alpha}{2\alpha}} - 1\big)^2 = T_\sigma(p_0) \leq \frac{4\alpha x_0}{(1-\alpha)^2}\sigma^{-\frac{1-\alpha}{\alpha}}.$$

In the regime of interest when $\sigma \to 0$, we have that $T_\sigma(p_0) \asymp \sigma^{-\frac{1-\alpha}{\alpha}}$.

REMARK.

- We observe that once again, the critical radius no longer follows the typical rate: $\varepsilon_n \asymp n^{-2/5}$. Instead we obtain the rate, $\varepsilon_n \asymp n^{-2\alpha/(2+3\alpha)}$, and indeed have much slower rates as $\alpha \to 0$, indicating the difficulty of testing heavy-tailed distributions against a Lipschitz, ℓ_1 neighborhood.

We conclude this section by emphasizing the value of the local minimax perspective and of studying the goodness-of-fit problem beyond the uniform null. We are able to provide a sharp characterization of the critical radius for a broad class of interesting examples, and we obtain faster (than at uniform) rates when the null is spiky and nonstandard rates in cases when the null is heavy-tailed.

4.2. *A recursive partitioning scheme.* For the remainder of this section, we encourage the reader to focus on the case when $s = 1$ (i.e., the Lipschitz setting) in their first reading. At the heart of our upper and lower bounds are spatially adaptive partitions of the domain of p_0. The partitions used in our upper and lower bounds are similar but not identical. In this section, we describe an algorithm for producing the desired partitions and then briefly describe some of the main properties of the partition that we leverage in our upper and lower bounds.

We begin by describing the desiderata for the partition from the perspective of the upper bound. Our goal is to construct a test for the hypotheses in (2.5), and we do so by constructing a partition (consisting of $N+1$ cells) $\{A_1, \ldots, A_N, A_\infty\}$ of $\mathbb{R}^d$. Each cell A_i for $i \in \{1, \ldots, N\}$ will be a cube, while the cell A_∞ will be arbitrary but will have small total probability content. We let

$$K := \bigcup_{i=1}^{N} A_i. \tag{4.7}$$

We form the multinomial corresponding to the partition $\{P_0(A_1), \ldots, P_0(A_N), P_0(A_\infty)\}$, where $P_0(A_i) = \int_{A_i} p_0(x)\,dx$. We then test this multinomial using the counts of the number of samples falling in each cell of the partition.

REQUIREMENT 1. A basic requirement of the partition is that it must ensure that a density p that is at least ε_n far away in ℓ_1 distance from p_0 should remain roughly ε_n away from p_0 when converted to a multinomial. Formally, for any p such that $\|p - p_0\|_1 \geq \varepsilon_n$, $p \in \mathcal{L}_s(L_n)$ we require that for some small constant $c > 0$,

$$\sum_{i=1}^{N} \bigl|P_0(A_i) - P(A_i)\bigr| + \bigl|P_0(A_\infty) - P(A_\infty)\bigr| \geq c\varepsilon_n. \tag{4.8}$$

Of course, there are several ways to ensure this condition is met. In particular, supposing that we restrict attention to densities supported on [0, 1] then it suffices for instance to choose roughly $(L_n/\varepsilon_n)^{1/s}$ even-width bins. This is precisely the partition considered in prior work [3, 20, 21]. When we do not restrict attention to compactly supported, uniform densities an even-width partition is no longer optimal and a careful optimization of the upper and lower bounds with respect to the partition yields the optimal choice. The optimal partition has bin-widths that are roughly taken proportional to $p_0^{\gamma/s}(x)$, where the constant of proportionality is chosen to ensure that the condition in (4.8) is satisfied. Precisely determining the constant of proportionality turns out to be quite subtle so we defer a discussion of this to the end of this section.

REQUIREMENT 2. A second requirement that arises in both our upper and lower bounds is that the cells of our partition (except A_∞) are not chosen too wide. In particular, we must choose the cells small enough to ensure that the density is roughly constant on each cell, that is, on each cell we need that for any $i \in \{1, \ldots, N\}$,

$$\frac{\sup_{x \in A_i} p_0(x)}{\inf_{x \in A_i} p_0(x)} \leq 2. \tag{4.9}$$

Using the Hölder property of p_0, this condition is satisfied if any point x is in a cell of diameter at most $(p_0(x_j)/(3L_n))^{1/s}$, where x_j denotes the centroid of the cell containing x.

Taken together the first two requirements suggest that we need to create a partition such that: for every point $x \in K$, the diameter of the cell A containing the point x, should be roughly

$$[\mathrm{diam}(A)]^s \approx \min\{\theta_1 p_0(x), \theta_2 p_0^\gamma(x)\},$$

where θ_1 is to be chosen to be smaller than $1/(3L_n)$, and θ_2 is chosen to ensure that Requirement 1 is satisfied.

Algorithm 1 constructively establishes the existence of a partition satisfying these requirements. The upper and lower bounds use this algorithm with slightly different parameters. The key idea is to recursively partition cells that are too large by halving each side. This is illustrated in Figure 3. The proof of correctness of the algorithm uses the smoothness of p_0 in an essential fashion. Indeed, were the density p_0 not sufficiently smooth then such a partition would likely not exist.

In order to ensure that the algorithm has a finite termination, we choose two parameters $a, b \ll \varepsilon_n$ (these are chosen sufficiently small to not affect subsequent results):

- We restrict our attention to the a-effective support of p_0, that is, we define S_a to be the smallest cube centered at the mean of p_0 such that, $P_0(S_a) \geq 1 - a$. We begin with $A_\infty = S_a^c$.
- If the density in any cell is sufficiently small, we do not split the cell further, that is, for a parameter b, if $\sup_{x \in A} p_0(x) \leq b/\mathrm{vol}(S_a)$ then we do not split it, rather we add it to A_∞. By construction, such cells have total probability content at most b.

For each cube A_i for $i \in \{1, \dots, \widetilde{N}\}$, we let x_i denote its centroid, and we let $\widetilde{N}$ denote the number of cubes created by Algorithm 1.

REQUIREMENT 3. The final major requirement is two-fold: (1) we require that the γ-norm of the density over the support of the partition should be upper bounded by the truncated T-functional, and (2) that the density over the cells of the partition be sufficiently large. This necessitates a further pruning of the partition, where we order cells by their probability content and successively eliminate (adding them to A_∞) cells of low probability until we have eliminated mass that is close to the desired truncation level. This is accomplished by Algorithm 2.

It remains to specify a precise choice for the parameter θ_2. We do so indirectly by defining a function $\mu : \mathbb{R} \mapsto \mathbb{R}$ that is closely related to the truncated T-functional. For $x \in \mathbb{R}$, we define $\mu(x)$ as the smallest positive number that satisfies the equation

$$\varepsilon_n = \int_{\mathbb{R}^d} \min\left\{\frac{p_0(y)}{x}, \frac{\varepsilon_n p_0(y)^\gamma}{\mu(x)}\right\} dy. \tag{4.12}$$

If $x < 1/\varepsilon_n$, then we obtain a finite value for $\mu(x)$, otherwise we take $\mu(x) = \infty$. The following result, relates μ to the truncated T-functional.

Algorithm 1 Adaptive Partition

1. *Input*: Parameters θ_1, θ_2, a, b.
2. Set $A_\infty = \varnothing$ and $A_1 = S_a$.
3. For each cube A_i do:

   - If

$$\sup_{x \in A_i} p_0(x) \leq \frac{b}{\mathrm{vol}(S_a)}, \tag{4.10}$$

     then remove A_i from the partition and let $A_\infty = A_\infty \cup A_i$.
   - If

$$\left[\mathrm{diam}(A_i)\right]^s \leq \min\{\theta_1 p_0(x_i), \theta_2 p_0^\gamma(x_i)\}, \tag{4.11}$$

     then do nothing to A_i.
   - If A_i fails to satisfy (4.10) or (4.11), then replace A_i by a set of $\lceil 2^{1/s} \rceil^d$ cubes that are obtained dividing the original A_i into $\lceil 2^{1/s} \rceil$ equal length pieces along each of its axes.

4. If no cubes are split or removed, STOP. Else go to step 3.
5. *Output*: Partition $\mathcal{P} = \{A_1, \dots, A_{\widetilde{N}}, A_\infty\}$.

LEMMA 4.1. *For any* $0 \leq x < 1/\varepsilon_n$,

$$T^{\gamma}_{x\varepsilon_n}(p_0) \leq \mu(x) \leq 2T^{\gamma}_{x\varepsilon_n/2}(p_0). \tag{4.13}$$

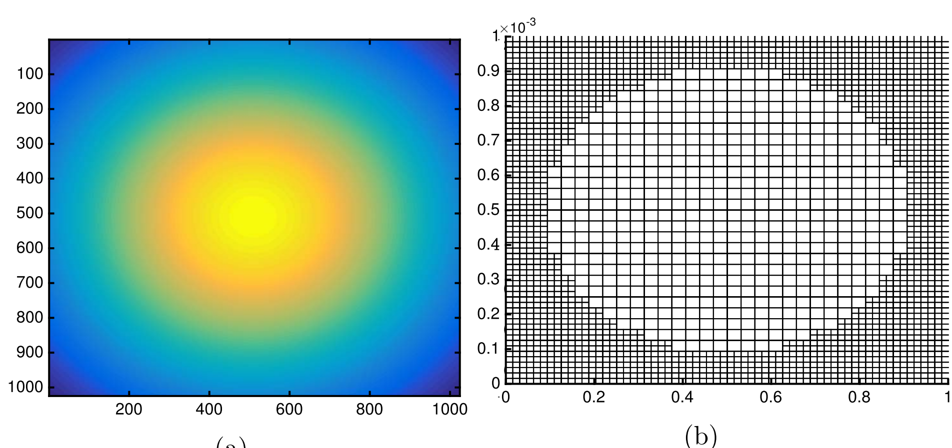

FIG. 3. (a) *A density* p_0 *on* $[0, 1]^2$ *evaluated on a* 1000×1000 *grid.* (b) *The corresponding spatially adaptive partition* $\mathcal{P}$ *produced by Algorithm* 1. *Cells of the partition are larger in regions where the density* p_0 *is higher.*

Algorithm 2 Prune Partition

1. *Input*: Unpruned partition $\mathcal{P} = \{A_1, \dots, A_{\widetilde{N}}, A_\infty\}$ and a target pruning level c. Without loss of generality, we assume $P_0(A_1) \geq P_0(A_2) \geq \cdots \geq P_0(A_{\widetilde{N}})$.
2. For any $j \in \{1, \dots, \widetilde{N}\}$, let $\mathcal{Q}(j) = \sum_{i=j}^{\widetilde{N}} P_0(A_i)$. Let j^* denote the smallest positive integer such that, $\mathcal{Q}(j^*) \leq c$.
3. If $\mathcal{Q}(j^*) \geq c/5$:
   - Set $N = j^* - 1$, and $A_\infty = A_\infty \cup A_{j^*} \cup \cdots \cup A_{\widetilde{N}}$.
4. If $\mathcal{Q}(j^*) \leq c/5$:
   - Set $N = j^*$, $\alpha = \min\{c/(5P_0(A_N)), 1/5\}$, and $A_\infty = A_\infty \cup A_{j^*} \cup \cdots \cup A_{\widetilde{N}}$.
   - A_N is a cube, that is, for some $\delta > 0$, $A_N = [a_1, a_1 + \delta] \times \cdots \times [a_d, a_d + \delta]$. Let $D_1 = [a_1, (1-\alpha)(a_1+\delta)] \times \cdots \times [a_d, (1-\alpha)(a_d+\delta)]$ and $D_2 = A_N - D_1$. Set: $A_N = D_1$ and $A_\infty = A_\infty \cup D_2$.
5. *Output*: $\mathcal{P}^\dagger = \{A_1, \dots, A_N, A_\infty\}$.

With the definition of μ in place, we now state our main result regarding the partitions produced by Algorithms 1 and 2. We let $\mathcal{P}$ denote the unpruned partition obtained from Algorithm 1 and $\mathcal{P}^\dagger$ denote the pruned partition obtained from Algorithm 2. For each cell A_i, we denote its centroid by x_i. We have the following result summarizing some of the important properties of $\mathcal{P}$ and $\mathcal{P}^\dagger$.

LEMMA 4.2. *Suppose we choose*, $\theta_1 = 1/(3L_n)$, $\theta_2 = \varepsilon_n/(8L_n\mu(3/8))$, $a = b = \varepsilon_n/1024$, $c = \varepsilon_n/512$, *then the partition* $\mathcal{P}^\dagger$ *satisfies the following properties*:

1. [*Diameter control.*] *The partition has the property that*

$$\frac{1}{5}\min\{\theta_1 p_0(x_i), \theta_2 p_0^\gamma(x_i)\} \leq [\mathrm{diam}(A_i)]^s \leq \min\{\theta_1 p_0(x_i), \theta_2 p_0^\gamma(x_i)\}. \tag{4.14}$$

2. [*Multiplicative control.*] *The density is multiplicatively controlled on each cell*, *that is*, *for* $i \in \{1, \dots, N\}$ *we have that*

$$\frac{\sup_{x \in A_i} p_0(x)}{\inf_{x \in A_i} p_0(x)} \leq 2. \tag{4.15}$$

3. [*Properties of* A_∞.] *The cell* A_∞ *has probability content roughly* ε_n, *that is*,

$$\frac{\varepsilon_n}{2560} \leq P_0(A_\infty) \leq \frac{\varepsilon_n}{256}. \tag{4.16}$$

4. [ℓ_1 *distance*.] *The partition preserves the* ℓ_1 *distance*, *that is*, *for any* p *such that* $\|p - p_0\|_1 \geq \varepsilon_n$, $p \in \mathcal{L}_s(L_n)$,

$$\sum_{i=1}^{N} \big|P_0(A_i) - P(A_i)\big| + \big|P_0(A_\infty) - P(A_\infty)\big| \geq \frac{\varepsilon_n}{8}. \tag{4.17}$$

5. [*Truncated* T*-functional*.] *Recalling the definition of* K *in* (4.7), *we have that*

$$\int_K p_0^\gamma(x)\,dx \leq T^\gamma_{\varepsilon_n/5120}(p_0). \tag{4.18}$$

6. [*Density Lower Bound*.] *The density over* K *is lower bounded as*

$$\inf_{x\in K} p_0(x) \geq \left(\frac{\varepsilon_n}{5120\mu(1/5120)}\right)^{1/(1-\gamma)}. \tag{4.19}$$

Furthermore, *for any choice of the parameter* θ_2 *the unpruned partition* $\mathcal{P}$ *of Algorithm* 1 *satisfies* (4.14) *with the constant* 5 *sharpened to* 4, (4.15) *and the upper bound in* (4.16).

The proof of this result is technical and we defer it to the Supplementary Material [4] (Appendix E).

While we focused our discussions on the properties of the partition from the perspective of establishing the upper bound in Theorem 4.1 it turns out that several of these properties are crucial in proving the lower bound as well. The optimal adaptive partition creates larger cells in regions where the density p_0 is higher, and smaller cells where p_0 is lower. This might seem counter-intuitive from the perspective of the upper bound since we create many low-probability cells which are likely to be empty in a small finite-sample, and indeed this construction is in some sense opposite to the quantile transformation suggested by previous work [18, 20]. However, from the perspective of the lower bound this is completely natural. It is intuitive that our perturbation be large in regions where the density is large since the likelihood ratio is relatively stable in these regions, and hence these changes are more difficult to detect. The requirement of smoothness constrains the amount by which we can perturb the density on any given cell, that is, for a large perturbation the corresponding cell should have a large diameter leading to the conclusion that we must use larger cells in regions where p_0 is higher.

In this section, we have focused on providing intuition for our adaptive partitioning scheme. In the next section, we provide the body of the proof of Theorem 4.1, and defer the remaining technical aspects to the Supplementary Material [4].

4.3. *Proof of Theorem* 4.1. We consider the lower and upper bounds in turn.

4.3.1. *Proof of lower bound.* We note that the lower bound in (4.5) is trivial when $\varepsilon_n \geq 1/C$ so throughout the proof we focus on the case when ε_n is smaller than a universal constant, that is, when $\varepsilon_n \leq \frac{1}{C}$.

Preliminaries: We begin by briefly introducing the lower bound technique due to Ingster (see for instance [22]). Let $\mathcal{P}$ be a set of distributions and let Φ_n be the set of level α tests based on n observations where $0 < \alpha < 1$ is fixed. We want to bound the minimax type II error

$$\zeta_n(\mathcal{P}) = \inf_{\phi \in \Phi_n} \sup_{P \in \mathcal{P}} P^n(\phi = 0).$$

Define Q as $Q(A) = \int P^n(A)\, d\pi(P)$, where π is a prior distribution whose support is contained in $\mathcal{P}$. In particular, if π is uniform on a finite set $P_1, \dots, P_N$ then

$$Q(A) = \frac{1}{N} \sum_j P_j^n(A).$$

Given n observations, we define the likelihood ratio

$$W_n(Z_1, \dots, Z_n) = \frac{dQ}{dP_0^n} = \int \frac{p(Z_1, \dots, Z_n)}{p_0(Z_1, \dots, Z_n)}\, d\pi(p) = \int \prod_j \frac{p(Z_j)}{p_0(Z_j)}\, d\pi(p).$$

LEMMA 4.3. *Let* $0 < \zeta < 1 - \alpha$. *If*

$$\mathbb{E}_0\big[W_n^2(Z_1, \dots, Z_n)\big] \leq 1 + 4(1 - \alpha - \zeta)^2 \tag{4.20}$$

then $\zeta_n(\mathcal{P}) \geq \zeta$.

Roughly, this result asserts that in order to produce a minimax lower bound on the Type II error, it suffices to appropriately upper bound the second moment under the null of the likelihood ratio. The proof is standard but presented in the Supplementary Material [4] (Appendix E) for completeness. A natural way to construct the prior π on the set of alternatives, is to partition the domain of p_0 and then to locally perturb p_0 by adding or subtracting sufficiently smooth "bumps". In the setting where the partition has fixed-width cells, this construction is standard [3, 20] and we provide a generalization to allow for variable width partitions and to allow for non-uniform p_0. Formally, let ψ be a smooth bounded function on the hypercube $\mathcal{I} = [-1/2, 1/2]^d$ such that

$$\int_{\mathcal{I}} \psi(x)\, dx = 0 \quad \text{and} \quad \int_{\mathcal{I}} \psi^2(x)\, dx = 1.$$

Let $\mathcal{P} = \{A_1, \dots, A_N, A_\infty\}$ be the partition obtained from Algorithm 1 that satisfies the condition in (4.15), and further let $\{x_1, \dots, x_N\}$ denote the centroids of the cells $\{A_1, \dots, A_N\}$. Each cell A_j for $j \in \{1, \dots, N\}$ is a cube with side-length $c_j h_j$ for some constants $1/4 \leq c_j \leq 1$, and

$$(\sqrt{d} h_j)^s = \min\{\theta_1 p_0(x_j), \theta_2 p_0^\gamma(x_j)\},$$

where $\theta_1 = 1/(3L_n)$ and we let $\theta_2 > 0$ be arbitrary. Let $\eta = (\eta_1, \eta_2, \dots, \eta_N)$ be a Rademacher sequence and define

$$p_\eta = p_0 + \sum_{j=1}^{N} \rho_j \eta_j \psi_j, \tag{4.21}$$

where each $\rho_j \geq 0$ and

$$\psi_j(t) = \frac{1}{c_j^{d/2} h_j^{d/2}} \psi\left(\frac{t - x_j}{c_j h_j}\right)$$

for $t \in A_j$. Hence, $\int_{A_j} \psi_j(t) = 0$ and $\int_{A_j} \psi_j^2(t) = 1$. Finally, let us denote:

$$\omega_1 := \max\left\{\frac{4\|\psi\|_\infty}{(1 - c_{\text{int}})}, \frac{8\|\psi'\|_\infty}{(1 - c_{\text{int}})}\right\} \quad \text{and} \quad \omega_2 := \|\psi\|_1.$$

With these definitions in place, we state a result that gives a lower bound for a sequence of perturbations $\{\rho_j\}_{j=1}^N$ that satisfy certain conditions.

LEMMA 4.4. *Let α, ζ and ε_n be nonnegative numbers with $1 - \alpha - \zeta > 0$. Let $C_0 = 1 + 4(1 - \alpha - \zeta)^2$. Assume that for each $j \in \{1, \dots, N\}$, ρ_j and h_j satisfy*:

$$\rho_j \leq \frac{c_j^{d/2}}{\omega_1} L_n h_j^{s + \frac{d}{2}}, \tag{4.22}$$

$$\sum_{j=1}^{N} \rho_j c_j^{d/2} h_j^{d/2} \geq \frac{\varepsilon_n}{\omega_2}, \tag{4.23}$$

$$\sum_{j=1}^{N} \frac{\rho_j^4}{p_0^2(x_j)} \leq \frac{\log C_0}{4n^2}, \tag{4.24}$$

then the Type II error of any test is at least ζ.

Effectively, this lemma generalizes the result of [20] to allow for nonuniform p_0 and further allows for variable width bins. The proof proceeds by verifying that under the conditions of the lemma, p_η is sufficiently smooth, and separated from p_0 by at least ε_n in the ℓ_1 metric. We let the prior be uniform on the the set of possible distributions p_η and directly analyze the second moment of the likelihood ratio, and obtain the result by applying Lemma 4.3. See the Supplementary Material [4] (Appendix E) for the proof of this lemma. It is worth noting the condition in (4.22), which ensures smoothness of p_η, allows for larger perturbations ρ_j for bins where h_j is large, which is one of the key benefits of using variable bin-widths in the lower bound.

With this result in place, to produce the desired minimax lower bound it only remains to specify the partition, select a sequence of perturbations $\{\rho_j\}_{j=1}^N$ and verify that the conditions of Lemma 4.4 are satisfied.

Final Steps: We begin by specifying the partition. We define

$$\nu = \min\left\{\frac{\omega_2}{\omega_1 4^{d+1} d^{1/(2s)}}, 1\right\}.$$

For the lower bound, we do not need to prune the partition, rather we simply apply Algorithm 1 with $\theta_1 = 1/(3L_n)$, and $\theta_2 = \varepsilon_n/(L_n \nu \mu(2/\nu))$. We choose $a = b = \varepsilon_n/1024$, and denote the resulting partition $\mathcal{P} = \{A_1, \dots, A_N, A_\infty\}$. Using Lemma 4.2, we have that the partition satisfies (4.14) with the constant 5 replaced by 4, (4.15) and the upper bound in (4.16). We now choose a sequence $\{\rho_1, \dots, \rho_N\}$, and proceed to verify that the conditions of Lemma 4.4 are satisfied. We choose

$$\rho_j = \frac{c_j^{d/2}}{\omega_1} L_n h_j^{s+\frac{d}{2}},$$

thus ensuring the condition in (4.22) is satisfied.

Verifying the condition in (4.23): Recall the definition of μ in (4.12),

$$\frac{\varepsilon_n}{\nu} = \int_{\mathbb{R}^d} \min\left\{\frac{p_0(y)}{2}, \frac{\varepsilon_n p_0(y)^\gamma}{\nu \mu(2/\nu)}\right\} dy,$$

provided that $\varepsilon_n < \nu/2$ which is true by our assumption on the critical radius. Recalling the definition of K in (4.7), we have that

$$\int_K \min\left\{\frac{p_0(y)}{2}, \frac{\varepsilon_n p_0(y)^\gamma}{\nu \mu(2/\nu)}\right\} dy \geq \frac{\varepsilon_n}{\nu} - \frac{P_0(A_\infty)}{2}.$$

We define the function

$$h^s(y) := \frac{1}{d^{1/(2s)}} \min\left\{\frac{p_0(y)}{3L_n}, \frac{\varepsilon_n p_0(y)^\gamma}{L_n \nu \mu(2/\nu)}\right\},$$

and as a consequence of the property (4.15) we obtain that for any $y \in A_j$ for $j \in \{1, \dots, N\}$,

$$h_j^s \geq \frac{h^s(y)}{2}.$$

This in turn yields that

$$\begin{aligned} L_n \sum_{j=1}^N h_j^{d+s} &\geq \frac{1}{2(\sqrt{d})^s} \int_K \min\left\{\frac{p_0(y)}{2}, \frac{\varepsilon_n p_0(y)^\gamma}{\nu \mu(2/\nu)}\right\} dy \\ &\geq \frac{1}{2(\sqrt{d})^s}\left(\frac{\varepsilon_n}{\nu} - \frac{P_0(A_\infty)}{2}\right) \\ &\geq \frac{\varepsilon_n}{4(\sqrt{d})^s \nu}, \end{aligned}$$

where the final step uses the upper bound in (4.16). We then have that

$$\sum_{j=1}^{N} \rho_j c_j^{d/2} h_j^{d/2} = \sum_{j=1}^{N} \frac{L_n c_j^d h_j^{d+s}}{\omega_1} \geq \sum_{j=1}^{N} \frac{L_n h_j^{d+s}}{4^d \omega_1} \geq \frac{\varepsilon_n}{\omega_2},$$

which establishes the condition in (4.23).

Verifying the condition in (4.24): We note the inequality (which can be verified by simple case analysis) that for $a, b, u, v \geq 0$,

$$\min\{a, b\} \leq \min\{a^{\frac{u}{u+v}} b^{\frac{v}{u+v}}, b\},$$

in particular for $u = s$, $v = 3s + d$ we obtain

$$\min\{a, b\} \leq \min\{(a^s b^{3s+d})^{\frac{1}{4s+d}}, b\}. \tag{4.25}$$

Returning to the condition in (4.24), we have that

$$\begin{aligned}
\sum_{j=1}^{N} \frac{\rho_j^4}{p_0(x_j)^2} &\leq \frac{L_n^4}{\omega_1^4} \sum_{j=1}^{N} \frac{c_j^{2d} h_j^{4s+2d}}{p_0(x_j)^2} \\
&\leq \frac{L_n^4}{\omega_1^4} \sum_{j=1}^{N} \frac{h_j^d h_j^{4s+d}}{p_0(x_j)^2},
\end{aligned}$$

using the fact that $c_j \leq 1$. Using the chosen values for h_j we obtain that

$$\begin{aligned}
&\sum_{j=1}^{N} \frac{\rho_j^4}{p_0(x_j)^2} \\
&\quad\leq \frac{L_n^4}{\omega_1^4 d^{(4s+d)/(2s)}} \sum_{j=1}^{N} \frac{h_j^d}{p_0(x_j)^2} \\
&\qquad \times \min\left\{\left[\frac{p_0(x_j)}{2L_n}\right]^{\frac{4s+d}{s}}, \left[\frac{\varepsilon_n p_0^{\gamma}(x_j)}{L_n \nu \mu(2/\nu)}\right]^{\frac{4s+d}{s}}\right\} \\
&\quad\overset{\text{(i)}}{\leq} \frac{L_n^4}{\omega_1^4 d^{(4s+d)/(2s)}} \sum_{j=1}^{N} \frac{h_j^d}{p_0(x_j)^2} \\
&\qquad \times \min\left\{\frac{p_0(x_j)^3 \varepsilon_n^{\frac{3s+d}{s}}}{3L_n (L_n \nu \mu(2/\nu))^{\frac{3s+d}{s}}}, \left[\frac{\varepsilon_n p_0^{\gamma}(x_j)}{L_n \nu \mu(2/\nu)}\right]^{\frac{4s+d}{s}}\right\} \\
&\quad= \frac{\varepsilon_n^{\frac{3s+d}{s}}}{L_n^{d/s} \mu(2/\nu)^{3+d/s} \nu^{4+d/s} \omega_1^4 d^{(4s+d)/(2s)}} \sum_{j=1}^{N} h_j^d
\end{aligned}$$

$$
\begin{aligned}
&\times \min\left\{\frac{p_0(x_j)}{2/\nu}, \frac{\varepsilon_n p_0^\gamma(x_j)}{\mu(2/\nu)}\right\} \\
&\leq \frac{2\varepsilon_n^{\frac{3s+d}{s}}}{L_n^{d/s}\mu(2/\nu)^{3+d/s}\nu^{4+d/s}\omega_1^4 d^{(4s+d)/(2s)}} \\
&\quad\times \int_K \min\left\{\frac{p_0(x)}{2/\nu}, \frac{\varepsilon_n p_0(x)^\gamma}{\mu(2/\nu)}\right\} dx \\
&\leq \frac{2\varepsilon_n^{\frac{3s+d}{s}}}{L_n^{d/s}\mu(2/\nu)^{3+d/s}\nu^{4+d/s}\omega_1^4 d^{(4s+d)/(2s)}} \\
&\quad\times \int_{\mathbb{R}^d} \min\left\{\frac{p_0(x)}{2/\nu}, \frac{\varepsilon_n p_0(x)^\gamma}{\mu(2/\nu)}\right\} dx \\
&\overset{\text{(ii)}}{\leq} \frac{2\varepsilon_n^{\frac{4s+d}{s}}}{L_n^{d/s}\mu(2/\nu)^{3+d/s}\nu^{4+d/s}\omega_1^4 d^{(4s+d)/(2s)}},
\end{aligned}
$$

where (i) follows from the inequality in (4.25), and (ii) uses (4.12). Using Lemma 4.1, we obtain

$$\mu(2/\nu) \geq T^\gamma_{2\varepsilon_n/\nu}(p_0),$$

provided that $\varepsilon_n < \nu/2$. This yields that

$$\sum_{j=1}^N \frac{\rho_j^4}{p_0(x_j)^2} \leq \frac{2\varepsilon_n^{\frac{4s+d}{s}}}{L_n^{d/s} T^2_{2\varepsilon_n/\nu}(p_0)\nu^{4+d/s}\omega_1^4 d^{(4s+d)/(2s)}},$$

and we require that this quantity is upper bounded by $\frac{\log C_0}{4n^2}$. For constants c_1, c_2 that depend only on the dimension d, it suffices to choose ε_n as the solution to the equation

$$\varepsilon_n = \left(\frac{L_n^{d/(2s)} T_{c_2\varepsilon_n}(p_0)\sqrt{\log C_0}}{c_1 n}\right)^{2s/(4s+d)}$$

and an application of Lemma 4.4 yields the lower bound of Theorem 4.1.

4.3.2. *Proof of upper bound.* In order to establish the upper bound, we construct an adaptive partition using Algorithms 1 and 2, and utilize the test analyzed in Theorem 3.3 from [37] to test the resulting multinomial. For the upper bound, we use the partition $\mathcal{P}^\dagger$ studied in Lemma 4.2, that is, we take $\theta_1 = 1/(3L_n), \theta_2 = \varepsilon_n/(8L_n\mu(3/8)), a = b = \varepsilon_n/1024$ and $c = \varepsilon_n/512$. Using the property in (4.17), it suffices to upper bound the V-functional in (3.3), for $\sigma = \varepsilon_n/128$.

The following technical lemma shows that the truncated V-functional is upper bounded by the V-functional over the partition excluding A_∞. For the partition

$\mathcal{P}^\dagger$, we have the associated multinomial $q := \{P_0(A_1), \ldots, P_0(A_\infty)\}$. With these definitions in place, we have the following result.

LEMMA 4.5. *For the multinomial q defined above, the truncated V-functional is upper bounded as*

$$V_{\varepsilon_n/128}^{2/3}(q) \leq \sum_{i=1}^{N} P_0(A_i)^{2/3} := \kappa.$$

We prove this result in the Supplementary Material [4] (Appendix E). Roughly, this lemma asserts that our pruning is less aggressive than the tail truncation of the multinomial test from the perspective of controlling the 2/3rd norm. With this claim in place, it only remains to upper bound κ. Using the property in (4.15), we have that

$$\begin{aligned}\kappa &\leq \sum_{i=1}^{N} \bigl(2p_0(x_i)\operatorname{vol}(A_i)\bigr)^{2/3} \\ &\leq 2^{2/3} \sum_{i=1}^{N} \frac{p_0(x_i)^{2/3}}{h_i^{d/3}} h_i^d.\end{aligned}$$

Using the condition in (4.19), verify that for all $x \in K$ we have that

$$\theta_1 p_0(x) \geq \frac{\varepsilon_n p_0^\gamma(x)}{10{,}240 L_n \mu(1/5120)},$$

and this yields that for a constant $c > 0$ for each $i \in \{1, \ldots, N\}$,

$$h_i^s \geq \frac{c\varepsilon_n p_0^\gamma(x_i)}{L_n \mu(1/5120)}.$$

Using the property in (4.15), we then obtain that for a constant $C > 0$,

$$\kappa \leq C \left(\frac{L_n \mu(1/5120)}{\varepsilon_n} \right)^{d/(3s)} \int_K p_0^\gamma(x)\, dx,$$

and using the property (4.18) and Lemma 4.1, we obtain that for constants $c, C > 0$ that

$$\kappa \leq C \left(\frac{L_n}{\varepsilon_n} \right)^{d/(3s)} T_{c\varepsilon_n}^{2/3}(p_0).$$

With Lemma 4.5, we obtain that for the multinomial q,

$$V_{\varepsilon_n/128}(q) \leq C^{3/2} \left(\frac{L_n}{\varepsilon_n} \right)^{d/(2s)} T_{c\varepsilon_n}(p_0),$$

which together with the upper bound of Theorem 3.1 yields the desired upper bound for Theorem 4.1. We note that a direct application of Theorem 3.1 yields a bound on the critical radius that is the maximum of two terms, one scaling as $1/n$ and the other being the desired term in Theorem 4.1. In Hölder testing, the $1/n$ term is always dominated by the term involving the truncated functional. This follows from lower bounds on the truncated functional [see, for instance, (F.4) for such a lower bound].

4.4. *Simulations*. In this section, we report some simulation results on Lipschitz testing. We focus on the case when $d = 1$ and $s = 1$. In Figure 4, we compare the following tests:

1. 2/3rd + Tail Test: This is the locally minimax test studied in Theorem 4.1, where we use our binning algorithm followed by the locally minimax multinomial test from [37].
2. Chi-sq. Test: Here, we use our binning algorithm followed by the standard χ^2 test.
3. Kolmogorov–Smirnov (KS) Test: Since we focus on the case when $d = 1$, we also compare to the standard KS test based on comparing the CDF of p_0 to the empirical CDF.
4. Naive Binning: Finally, we compare to the approach of using fixed-width bins, together with the χ^2 test. Following the prescription of Ingster [20] (for the case when p_0 is uniform), we choose the number of bins so that the ℓ_1-distance between the null and alternate is approximately preserved, that is, denoting the effective support to be S we choose the bin-width as $\varepsilon_n/(L_n\mu(S))$.

We focus on two simulation scenarios: when the null distribution is a standard Gaussian, and when the null distribution has a heavier tail, that is, a Pareto distribution with parameter $\alpha = 0.5$. We create the alternate density by smoothly perturbing the null after binning, and choose the perturbation weights as in our lower bound construction in order to construct a near worst-case alternative.

We set the α-level threshold via simulation (by sampling from the null 1000 times) and we calculate the power under particular alternatives by averaging over a 1000 trials. We observe several notable effects. First, we see that the locally minimax test can significantly out perform the KS test as well the test based on fixed bin-widths. The failure of the fixed bin-width test is more apparent in the setting where the null is Pareto as the distribution has a large effective support and the naive binning is far less parsimonious than the adaptive binning. On the other hand, we also observe that at least in these simulations the χ^2 test and the locally minimax test from [37] perform comparably when based on our adaptive binning indicating the crucial role played by the binning procedure.

5. Discussion. In this paper we studied the goodness-of-fit testing problem in the context of testing multinomials and Hölder densities. For testing multinomials, we built on prior works [16, 37] to provide new globally and locally minimax tests.

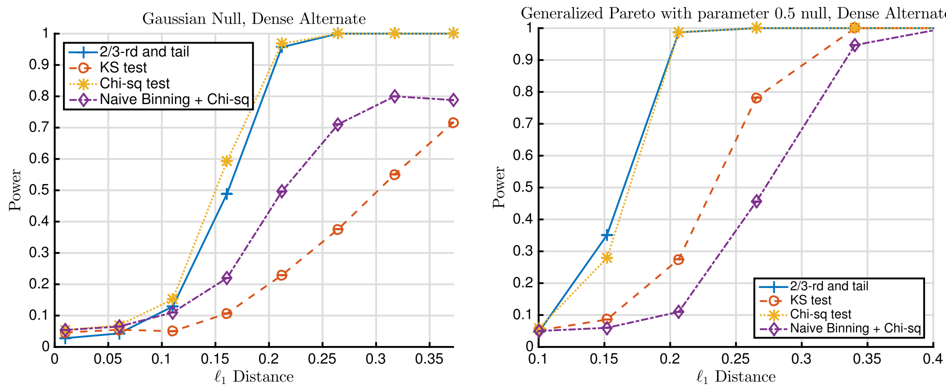

FIG. 4. *A comparison between the KS test, multinomial tests on an adaptive binning and multinomial tests on a fixed bin-width binning. In the figure on the left, we choose the null to be standard Gaussian and on the right we choose the null to be Pareto. The alternate is chosen to be a dense near worst-case, smooth perturbation of the null. The power of the tests are plotted against the ℓ_1 distance between the null and alternate. Each point in the graph is an average over* 1000 *trials.*

For testing Hölder densities, we provide the first results that give a characterization of the critical radius under mild conditions.

Our work highlights the heterogeneity of the critical radius in the goodness-of-fit testing problem and the importance of understanding the local critical radius. In the multinomial testing problem, it is particularly noteworthy that classical tests can perform quite poorly in the high-dimensional setting, and that simple modifications of these tests can lead to more robust inference. In the density testing problem, carefully constructed spatially adaptive partitions play a crucial role.

Our work motivates several open questions, and we conclude by highlighting a few of them. First, in the context of density testing we focused on the case when the density is Hölder with $0 < s \leq 1$. An important extension would be to consider higher-order smoothness. Surprisingly, [21] shows that bin-based tests continue to be optimal for higher-order smoothness classes when the null is uniform on $[0, 1]$. We conjecture that bin-based tests are no longer optimal when the null is not uniform, and further that the local critical radius continues to be determined by (4.3) even when $s > 1$. Second, it is possible to invert our locally minimax tests in order to construct confidence intervals. We believe that these intervals might also have some local adaptive properties that are worthy of further study. In the Supplementary Material [4], we provide some basic results on the limiting distributions of the multinomial test statistics under the null when the null is uniform, and it would be interesting to consider the extension to settings where the null is arbitrary. Finally, it would also be interesting to further explore the extent to which the local-minimax perspective can lead to a better understanding of composite-null inference problems [1, 2, 5, 8, 10, 23, 30].

Acknowledgments. The authors would like to thank the participants of the Oberwolfach workshop on "Statistical Recovery of Discrete, Geometric and Invariant Structures" for their generous feedback. Suggestions by various participants including David Donoho, Richard Nickl, Markus Reiss, Vladimir Spokoiny, Alexandre Tsybakov, Martin Wainwright, Yuting Wei and Harry Zhou have been incorporated in various parts of this manuscript. We are also grateful to the referees and Associate Editor for their valuable comments and suggestions.

SUPPLEMENTARY MATERIAL

Supplement to "Hypothesis testing for densities and high-dimensional multinomials: Sharp local minimax rates." (DOI: 10.1214/18-AOS1729SUPP; .pdf). The Supplementary Material contains detailed technical proofs. It also includes a brief study of limiting distributions of the test statistics we study. Finally, the Supplementary Material includes the design and analysis of tests that are adaptive to various parameters.

REFERENCES

[1] ADDARIO-BERRY, L., BROUTIN, N., DEVROYE, L. and LUGOSI, G. (2010). On combinatorial testing problems. *Ann. Statist.* **38** 3063–3092. MR2722464

[2] ARIAS-CASTRO, E., CANDÈS, E. J. and DURAND, A. (2011). Detection of an anomalous cluster in a network. *Ann. Statist.* **39** 278–304. MR2797847

[3] ARIAS-CASTRO, E., PELLETIER, B. and SALIGRAMA, V. (2018). Remember the curse of dimensionality: The case of goodness-of-fit testing in arbitrary dimension. *J. Nonparametr. Stat.* **30** 448–471. MR3794401

[4] BALAKRISHNAN, S. and WASSERMAN, L. (2019). Supplement to "Hypothesis testing for densities and high-dimensional multinomials: Sharp local minimax rates." DOI:10.1214/18-AOS1729SUPP.

[5] BALAKRISHNAN, S. and WASSERMAN, L. (2018). Hypothesis testing for high-dimensional multinomials: A selective review. *Ann. Appl. Stat.* To appear.

[6] BARRON, A. R. (1989). Uniformly powerful goodness of fit tests. *Ann. Statist.* **17** 107–124. MR0981439

[7] BATU, T., FISCHER, E., FORTNOW, L., KUMAR, R., RUBINFELD, R. and WHITE, P. (2001). Testing random variables for independence and identity. In 42*nd IEEE Symposium on Foundations of Computer Science* (*Las Vegas*, *NV*, 2001) 442–451. IEEE Computer Soc., Los Alamitos, CA. MR1948733

[8] BERTHET, Q. and RIGOLLET, P. (2013). Optimal detection of sparse principal components in high dimension. *Ann. Statist.* **41** 1780–1815. MR3127849

[9] CAI, T. T. and LOW, M. G. (2015). A framework for estimation of convex functions. *Statist. Sinica* **25** 423–456. MR3379081

[10] CARPENTIER, A. (2015). Testing the regularity of a smooth signal. *Bernoulli* **21** 465–488. MR3322327

[11] CASELLA, G. and BERGER, R. L. (2002). *Statistical Inference*. Duxbury, Pacific Grove, CA.

[12] CHATTERJEE, S. (2014). A new perspective on least squares under convex constraint. *Ann. Statist.* **42** 2340–2381. MR3269982

[13] CRAMÉR, H. (1928). On the composition of elementary errors. *Scand. Actuar. J.* **1928** 13–74.

[14] DEVROYE, L. and GYÖRFI, L. (1985). *Nonparametric Density Estimation*: *The L_1 View*. Wiley, New York. MR0780746

[15] DIACONIS, P. and MOSTELLER, F. (2006). Methods for studying coincidences. In *Selected Papers of Frederick Mosteller* (S. E. Fienberg and D. C. Hoaglin, eds.) 605–622. Springer, New York.

[16] DIAKONIKOLAS, I. and KANE, D. M. (2016). A new approach for testing properties of discrete distributions. In 57*th Annual IEEE Symposium on Foundations of Computer Science—FOCS* 2016 685–694. IEEE Computer Soc., Los Alamitos, CA. MR3631031

[17] FIENBERG, S. E. (1979). The use of chi-squared statistics for categorical data problems. *J. Roy. Statist. Soc. Ser. B* **41** 54–64. MR0535545

[18] GINÉ, E. and NICKL, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models. Cambridge Series in Statistical and Probabilistic Mathematics* **40**. Cambridge Univ. Press, New York. MR3588285

[19] GOLDREICH, O. and RON, D. (2011). On testing expansion in bounded-degree graphs. In *Studies in Complexity and Cryptography. Lecture Notes in Computer Science* **6650** 68–75. Springer, Heidelberg. MR2844253

[20] INGSTER, Y. I. (1990). Minimax detection of a signal in ℓ_p-metrics. *J. Math. Sci.* **68** 503–515.

[21] INGSTER, Y. I. (1997). Adaptive chi-square tests. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov.* (*POMI*) **244** 150–166, 333. MR1700386

[22] INGSTER, Y. I. and SUSLINA, I. A. (2003). *Nonparametric Goodness-of-Fit Testing Under Gaussian Models. Lecture Notes in Statistics* **169**. Springer, New York. MR1991446

[23] INGSTER, Y. I., TSYBAKOV, A. B. and VERZELEN, N. (2010). Detection boundary in sparse regression. *Electron. J. Stat.* **4** 1476–1526. MR2747131

[24] LECAM, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1** 38–53. MR0334381

[25] LEHMANN, E. L. and CASELLA, G. (1998). *Theory of Point Estimation*, 2nd ed. Springer, New York. MR1639875

[26] LEHMANN, E. L. and ROMANO, J. P. (2005). *Testing Statistical Hypotheses*, 3rd ed. Springer, New York. MR2135927

[27] MARRIOTT, P., SABOLOVA, R., VAN BEVER, G. and CRITCHLEY, F. (2015). Geometry of goodness-of-fit testing in high dimensional low sample size modelling. In *Geometric Science of Information. Lecture Notes in Computer Science* **9389** 569–576. Springer, Cham. MR3442239

[28] MORRIS, C. (1975). Central limit theorems for multinomial sums. *Ann. Statist.* **3** 165–188. MR0370871

[29] NEYMAN, J. and PEARSON, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. Roy. Soc. Lond. Ser. A* **231** 289–337.

[30] NICKL, R. and VAN DE GEER, S. (2013). Confidence sets in sparse regression. *Ann. Statist.* **41** 2852–2876. MR3161450

[31] PANINSKI, L. (2008). A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Trans. Inform. Theory* **54** 4750–4755. MR2591136

[32] PEARSON, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos. Mag. Ser.* 5 **50** 157–175.

[33] READ, T. R. C. and CRESSIE, N. A. C. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer, New York. MR0955054

[34] RON, D. (2008). Property testing: A learning theory perspective. *Found. Trends Mach. Learn.* **1** 307–402.

[35] SMIRNOFF, N. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Moscow Univ. Math. Bull.* **2** 3–14. MR0002062

[36] SNEDECOR, G. W. and COCHRAN, W. G. (1980). *Statistical Methods*, 7th ed. Iowa State Univ. Press, Ames, IA. MR0614143

[37] VALIANT, G. and VALIANT, P. (2014). An automatic inequality prover and instance optimal identity testing. In *55th Annual IEEE Symposium on Foundations of Computer Science—FOCS* 2014 51–60. IEEE Computer Soc., Los Alamitos, CA. MR3344854

[38] VON MISES, R. (1951). *Wahrscheinlichkeit*, *Statistik und Wahrheit*. Springer, Vienna. MR0041364

[39] WILKS, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann*. *Math*. *Stat*. **9** 60–62.

DEPARTMENT OF STATISTICS
CARNEGIE MELLON UNIVERSITY
PITTSBURGH, PENNSYLVANIA 15213
USA
E-MAIL: siva@stat.cmu.edu
larry@stat.cmu.edu

Submitted to the Annals of Statistics
arXiv: arXiv:0000.0000

SUPPLEMENT TO: HYPOTHESIS TESTING FOR DENSITIES AND HIGH-DIMENSIONAL MULTINOMIALS: SHARP LOCAL MINIMAX RATES

BY SIVARAMAN BALAKRISHNAN[†] AND LARRY WASSERMAN[†]

Carnegie Mellon University[†]

APPENDIX A: LIMITING BEHAVIOUR OF TEST STATISTICS UNDER THE NULL

In this section, we consider the problem of finding the asymptotic distribution of the multinomial test statistics under the null. Broadly, there is a dichotomy between classical asymptotics where the null distribution is kept fixed and a high-dimensional asymptotic where the number of cells is growing and the null distribution can vary with the number of cells. We present a few simple results on the limiting behaviour of our test statistics when the null is uniform and highlight some open problems. Although our techniques generalize in a straightforward way to non-uniform null distributions, they do not necessarily yield tight results.

We focus on the family of test statistics that we use in our paper, that are weighted χ^2-type statistics:

$$T(w) = \sum_{i=1}^{d} \frac{(X_i - np_0(i))^2 - X_i}{w_i}, \tag{A.1}$$

where each w_i is a positive weight that is a fixed function of $p_0(i)$. This family includes the 2/3-rd statistic from [4], the truncated χ^2 statistic that we propose, and the usual χ^2 and ℓ_2 statistics. When the null is uniform, this family of test statistics reduces to simple re-scalings of the ℓ_2 statistic in (3.11):

$$T_{\ell_2} = \sum_{i=1}^{d} \left[(X_i - np_0(i))^2 - X_i\right].$$

Our results are summarized in the following lemma.

LEMMA A.1. *1. Classical Asymptotics: For any fixed p_0, the statistic $T(w)$ under the null converges in distribution to a weighted sum of χ^2*

distributions, i.e. for $Z_1, \ldots, Z_d \sim \chi^2_1$,

$$T(w) \xrightarrow{d} \sum_{i=1}^{d} \frac{p_i}{w_i} (Z_i - 1)\,. \tag{A.2}$$

2. *High-dimensional Asymptotics: Suppose* p_0 *is uniform and* $d \to \infty$*, then we have that,*
   - *If* $n/\sqrt{d} \to \infty$*, then*
   $$\frac{T_{\ell_2}}{\sqrt{Var_0(T_{\ell_2})}} \xrightarrow{d} N(0,1).$$
   - *If* $n/\sqrt{d} \to 0$*, then*
   $$\frac{T_{\ell_2}}{\sqrt{Var_0(T_{\ell_2})}} \xrightarrow{d} \delta_0.$$

Remarks:

- The behaviour of the χ^2-type statistics under classical asymptotics is well understood and we do not prove the claim in (A.2).
- Focusing on the high-dimensional setting, the asymptotic distribution of the test statistic is Gaussian in the regime where the risk of the optimal test tends to 0 as $n \to \infty$, and is degenerate in the regime where there are no consistent tests. In the most interesting regime when, $n/\sqrt{d} \to c$, the optimal test can have non-trivial risk, and the limiting distribution is neither Gaussian nor degenerate.
- More broadly, an important open question is to characterize the limiting distribution of the test statistic, under both the null and the alternate in the high-dimensional asymptotic.

PROOF. The first part follows, by checking the Lyapunov conditions. We denote

$$\zeta_i = (X_i - np_0(i))^2 - X_i.$$

and can calculate the sum of the variances as:

$$s_d^2 = \sum_{i=1}^{d} \mathrm{var}(\zeta_i) = \frac{2n^2}{d}.$$

The Lyapunov condition then requires that,

$$\lim_{d\to\infty} \frac{1}{s_d^4} \sum_{i=1}^{d} \mathbb{E}\zeta_i^4 = 0.$$

A straightforward computation gives that,

$$\mathbb{E}\zeta_i^4 = 8\frac{n^2}{d^2} + 144\frac{n^3}{d^3} + 60\frac{n^4}{d^4},$$

so that the Lyapunov condition is satisfied provided that,

$$\lim_{d\to\infty} \frac{d^3}{n^6} \to 0,$$

which is indeed the case.

In order to verify the degenerate limit it suffices to show that when $n/\sqrt{d} \to 0$, then the number of categories that have strictly larger than one occurrence converges to 0. When each observed category is observed only once we have that the test statistic is deterministic, i.e.,

$$T_{\ell_2} = \sum_{i=1}^{d} \zeta_i = (d-n)\frac{n^2}{d^2} + n\left(\frac{n^2}{d^2} - \frac{2n}{d}\right).$$

When rescaled by the standard deviation we obtain that,

$$\frac{T_{\ell_2}}{\sqrt{\mathrm{var}_0(T_{\ell_2})}} = \sqrt{\frac{d}{2n^2}}\left[(d-n)\frac{n^2}{d^2} + n\left(\frac{n^2}{d^2} - \frac{2n}{d}\right)\right] \to 0.$$

Finally, we can bound the probability that any category is observed more than once as:

$$\begin{aligned}
P(\exists\, i, X_i \geq 2) &\leq \sum_{i=1}^{d} P(X_i \geq 2) \\
&\leq \sum_{i=1}^{d} \exp(-\lambda) \sum_{k=2}^{\infty} \left(\frac{n}{d}\right)^k \\
&\leq \frac{Cn^2}{d} \to 0.
\end{aligned}$$

Taken together these facts give the desired degenerate limit. $\square$

APPENDIX B: ANALYSIS OF MULTINOMIAL TESTS

B.1. Proof of Theorem 3.2. In this section we analyze the truncated χ^2 test. For convenience, throughout this proof we work with a scaled version of the statistic in (3.6), i.e. we let $T := T_{\text{trunc}}/d$ and abusing notation slightly we redefine θ_i appropriately, i.e. we take $\theta_i = \max\{1, dp_0(i)\}$.

We begin by controlling the size of the truncated χ^2 test. Fix any multinomial p on $\{1, \dots, d\}$, and suppose we denote $\Delta_i = p_0(i) - p(i)$, then a straightforward computation shows that,

$$\mathbb{E}_p[T] = n^2 \sum_{i=1}^{d} \frac{\Delta_i^2}{\theta_i}, \tag{B.1}$$

(B.2)

$$\mathrm{Var}_p[T] = \sum_{i=1}^{d} \frac{1}{\theta_i^2} \left[2n^2 p_0(i)^2 + 2n^2\Delta_i^2 - 4n^2\Delta_i p_0(i) + 4n^3\Delta_i^2 p_0(i) - 4n^3\Delta_i^3\right].$$

This yields that the null variance of T is given by:

$$\mathrm{Var}_0[T] = \sum_{i=1}^{d} \frac{2n^2 p_0(i)^2}{\theta_i^2}, \tag{B.3}$$

which together with Chebyshev's inequality yields the desired bound on the size. Turning our attention to the power of the test we fix a multinomial $p \in \mathcal{P}_1$. Denote the α level threshold of the test by

$$t_\alpha = n\sqrt{\frac{2}{\alpha}\sum_{i=1}^{d} \frac{p_0(i)^2}{\theta_i^2}}.$$

We observe that, if we can verify the following two conditions:

$$t_\alpha \leq \frac{\mathbb{E}_p[T]}{2} \tag{B.4}$$

$$\mathbb{E}_p[T] \geq 2\sqrt{\frac{\mathrm{Var}_p[T]}{\zeta}}, \tag{B.5}$$

then we obtain that $P(\phi_{\text{trunc}} = 0) \leq \zeta$. To see this, observe that

$$
\begin{aligned}
P(\phi_{\text{trunc}} = 0) &\leq P(T < t_\alpha) \\
&\leq P\Big(T < \frac{1}{2}\mathbb{E}_p[T]\Big) \\
&\leq P\Big(T < \mathbb{E}_p[T] - \sqrt{\frac{\text{Var}_p[T]}{\zeta}}\Big) \\
&\leq \zeta, \qquad \text{(B.6)}
\end{aligned}
$$

by Chebyshev's inequality.

Condition in Equation (B.4): This condition reduces to verifying the following,

$$2t_\alpha \leq n^2 \sum_{i=1}^{d} \frac{\Delta_i^2}{\theta_i},$$

and as a result we focus on lower bounding the mean under the alternate. By Cauchy-Schwarz we obtain that,

$$\sum_{i=1}^{d} \frac{\Delta_i^2}{\theta_i} \geq \frac{\|\Delta\|_1^2}{\sum_{i=1}^{d}\theta_i} \geq \frac{\epsilon_n^2}{\sum_{i=1}^{d}\{1 + dp_0(i)\}} \geq \frac{\epsilon_n^2}{2d}. \qquad \text{(B.7)}$$

We can further upper bound t_α as

$$t_\alpha = n\sqrt{\frac{2}{\alpha}\sum_{i=1}^{d}\frac{p_0(i)^2}{\theta_i^2}} \leq n\sqrt{\frac{2}{d\alpha}},$$

using the fact that $p_0(i)/\theta_i \leq \frac{1}{d}$. This yields that Equation (B.4) is satisfied if:

$$\frac{\epsilon_n^2}{2d} \geq \frac{2\sqrt{2}}{\sqrt{d\alpha}n},$$

which is indeed the case.

Condition in Equation (B.5): We can upper bound the variance under the alternate as:

$$
\begin{aligned}
\text{Var}_p[T] &\leq \sum_{t=1}^{d} \frac{1}{\theta_t^2}\left[4n^2p_0(t)^2 + 4n^2\Delta_t^2 + 4n^3\Delta_t^2 p_0(t) - 4n^3\Delta_t^3\right] \\
&= \underbrace{\sum_{t=1}^{d}\frac{4n^2p_0(t)^2}{\theta_t^2}}_{U_1} + \underbrace{\sum_{t=1}^{d}\frac{4n^2\Delta_t^2}{\theta_t^2}}_{U_2} + \underbrace{\sum_{t=1}^{d}\frac{4n^3\Delta_t^2p_0(t)}{\theta_t^2}}_{U_3} + \underbrace{\sum_{t=1}^{d}\frac{-4n^3\Delta_t^3}{\theta_t^2}}_{U_4}.
\end{aligned}
$$

Consequently, it suffices to verify that,

$$\sum_{i=1}^{4} \frac{2\sqrt{U_i/\zeta}}{\mathbb{E}_p[T]} \leq 1,$$

for $i = \{1, 2, 3, 4\}$ and we do this by bounding each of these terms in turn. For the first term we follow a similar argument to the one dealing with the first condition,

$$\frac{2\sqrt{U_1/\zeta}}{\mathbb{E}_p[T]} \leq \frac{8d\sqrt{\sum_{t=1}^{d} \frac{p_0(t)^2}{\theta_t^2}}}{\sqrt{\zeta} n \epsilon_n^2} \leq \frac{8\sqrt{d}}{\sqrt{\zeta} n \epsilon_n^2} \leq \frac{1}{4}.$$

For the second term,

$$\frac{2\sqrt{U_2/\zeta}}{\mathbb{E}_p[T]} \leq \frac{4\sqrt{\frac{1}{\zeta}\sum_{t=1}^{d} \frac{n^2\Delta_t^2}{\theta_t^2}}}{\mathbb{E}_p[T]} \leq \frac{4\sqrt{\frac{1}{\zeta}\sum_{t=1}^{d} \frac{n^2\Delta_t^2}{\theta_t}}}{\mathbb{E}_p[T]} = \frac{4}{\sqrt{\zeta \mathbb{E}_p[T]}}.$$

Using Equation (B.7) we obtain that,

$$\mathbb{E}_p[T] \geq \frac{n^2\epsilon_n^2}{2d},$$

which in turn yields that,

$$\frac{2\sqrt{U_2/\zeta}}{\mathbb{E}_p[T]} \leq \frac{8\sqrt{d}}{n\epsilon_n\sqrt{\zeta}} \leq \frac{1}{4}.$$

Turning our attention to the third term we obtain that,

$$\frac{2\sqrt{U_3/\zeta}}{\mathbb{E}_p[T]} = \frac{4\sqrt{\frac{1}{\zeta}\sum_{t=1}^{d} \frac{n^3\Delta_t^2 p_0(t)}{\theta_t^2}}}{\mathbb{E}_p[T]} \leq \frac{4\sqrt{\frac{n}{d\zeta}\sum_{t=1}^{d} \frac{n^2\Delta_t^2}{\theta_t}}}{\mathbb{E}_p[T]} = \frac{4\sqrt{\frac{n}{d\zeta}}}{\sqrt{\mathbb{E}_p[T]}}.$$

Using the lower bound on the mean we obtain that,

$$\frac{2\sqrt{U_3/\zeta}}{\mathbb{E}_p[T]} \leq \frac{8}{n\epsilon_n\sqrt{\zeta}} \leq \frac{1}{4}.$$

For the final term,

$$\frac{2\sqrt{U_4/\zeta}}{\mathbb{E}_p[T]} \leq \frac{4\sqrt{\frac{1}{\zeta}\sum_{t=1}^{d} \frac{n^3|\Delta_t^3|}{\theta_t^2}}}{\mathbb{E}_p[T]} \leq \frac{4\sqrt{\frac{n^3}{\zeta}\sum_{t=1}^{d} \frac{|\Delta_t^3|}{\theta_t^2}}}{\mathbb{E}_p[T]} \leq \frac{4\sqrt{\frac{1}{\zeta}\left(\sum_{i=1}^{d} \frac{n^2\Delta_i^2}{\theta_i^{4/3}}\right)^{3/2}}}{\mathbb{E}_p[T]},$$

where the last step uses the monotonicity of the ℓ_p norms. Observing that $\theta_i \geq 1$, we have

$$\frac{2\sqrt{U_4/\zeta}}{\mathbb{E}_p[T]} \leq \frac{4\sqrt{\frac{1}{\zeta}\left(\sum_{i=1}^{d} \frac{n^2\Delta_i^2}{\theta_i}\right)^{3/2}}}{\mathbb{E}_p[T]} = \frac{4\sqrt{\frac{1}{\zeta}}}{\mathbb{E}_p[T]^{1/4}} \leq \frac{8d^{1/4}}{\sqrt{\zeta}\epsilon_n^{1/2}\sqrt{n}} \leq \frac{1}{4}.$$

This completes the proof.

B.2. Proof of Theorem 3.3. Recall the definition of $\mathcal{B}_\sigma$ in (3.2). We define:

$$\Delta_{\mathcal{B}_\sigma} = \sum_{i \in \mathcal{B}_\sigma} |p_0(i) - p(i)|, \tag{B.8}$$

and

$$p_{\min,\sigma} = \min_{i \in \mathcal{B}_\sigma} p_0(i). \tag{B.9}$$

Our main results concern the combined test ϕ_V in (3.9). It is easy to verify that the size of this test is at most α so it only remains to control its power. We first provide a general result that allows for a range of possible values for the parameter σ.

Lemma B.1. *For any $\sigma \leq \frac{\epsilon_n}{8}$, if*

$$n \geq 2\max\left\{\frac{2}{\alpha}, \frac{1}{\zeta}\right\} \max\left\{\frac{1}{\sigma}, \frac{4096 V_{\sigma/2}(p_0)}{\epsilon_n^2}\right\},$$

then the Type II error $P(\phi_V = 0) \leq \zeta$.

Taking this lemma as given, it is straightforward to verify the result of Theorem 3.3. In particular, if we take $\sigma = \epsilon_n/8$, then we recover the result of the theorem.

Proof of Lemma B.1: As a preliminary, we state two technical results from [4]. The following result is Lemma 6 in [4].

Lemma B.2. *For any $c \geq 1$, suppose that $n \geq c\max\left\{\frac{V_\sigma(p_0)^{1/3}}{p_{\min,\sigma}^{1/3}\Delta_{\mathcal{B}_\sigma}}, \frac{V_\sigma(p_0)}{\Delta_{\mathcal{B}_\sigma}^2}\right\}$, then we have that*

$$Var_p(T_2(\sigma)) \leq \frac{16}{c}\left[\mathbb{E}_p(T_2(\sigma))\right]^2.$$

The following result appears in the proof of Proposition 1 of [4].

LEMMA B.3. *For any $c \geq 1$, suppose that,*

$$n \geq 2c \frac{V_{\sigma/2}(p_0)}{\Delta^2_{\mathcal{B}_\sigma}},$$

then we have that,

$$n \geq c \max \left\{ \frac{V_\sigma(p_0)^{1/3}}{p_{\min,\sigma}^{1/3} \Delta_{\mathcal{B}_\sigma}}, \frac{V_\sigma(p_0)}{\Delta^2_{\mathcal{B}_\sigma}} \right\}.$$

With these two results in place, we can now complete the proof. We divide the space of alternatives into two sets:

$$\mathcal{S}_1 = \left\{ p : \|p - p_0\|_1 \geq \epsilon_n, \sum_{i \in \mathcal{Q}_\sigma(p_0)} |p_0(i) - p(i)| \geq 3\sigma \right\}$$

$$\mathcal{S}_2 = \left\{ p : \|p - p_0\|_1 \geq \epsilon_n, \sum_{i \in \mathcal{Q}_\sigma(p_0)} |p_0(i) - p(i)| < 3\sigma \right\}.$$

In order to show desired result it then suffices to show that when $p \in \mathcal{S}_1$, $P(\phi_{\text{tail}} = 0) \leq \zeta$, and that when $p \in \mathcal{S}_2$, $P(\phi_{2/3} = 0) \leq \zeta$. We verify each of these claims in turn.

When $p \in \mathcal{S}_1$: In this case, we have that $P(\mathcal{Q}_\sigma(p_0)) \geq 2\sigma$. Under the alternate we have that $T_1(\sigma) \sim \text{Poi}(nP(\mathcal{Q}_\sigma(p_0))) - nP_0(\mathcal{Q}_\sigma(p_0))$. This yields,

$$P(\phi_{\text{tail}} = 0) \leq P\left(\text{Poi}(nP(\mathcal{Q}_\sigma(p_0))) < \rho n P(\mathcal{Q}_\sigma(p_0))\right), \tag{B.10}$$

where

$$\rho = \frac{P_0(\mathcal{Q}_\sigma(p_0))}{P(\mathcal{Q}_\sigma(p_0))} + \frac{1}{P(\mathcal{Q}_\sigma(p_0))} \sqrt{\frac{P_0(\mathcal{Q}_\sigma(p_0))}{n\alpha}}.$$

Provided $\rho \leq 1$ we obtain via Chebyshev's inequality that,

$$P(\phi_{\text{tail}} = 0) \leq \frac{1}{n(1-\rho)^2 P(\mathcal{Q}_\sigma(p_0))}.$$

We further have that,

$$\rho \leq \frac{1}{2}\left[1 + \frac{1}{\sqrt{n\alpha\sigma}}\right].$$

Under the conditions that,

$$n \geq \frac{4}{\alpha\sigma},$$

we obtain that $\rho \leq 1/2$, which yields that,

$$P(\phi_{\text{tail}} = 0) \leq \frac{2}{n\sigma} \leq \zeta,$$

where the final inequality uses the condition on n.

When $p \in \mathcal{S}_2$: In this case, we first observe that the bulk deviation must be sufficiently large. Concretely, at most $\epsilon_n/2$ deviation can occur in the largest element and at most 3σ occurs in the tail, i.e.

$$\Delta_{\mathcal{B}_\sigma} \geq \frac{\epsilon_n}{2} - 3\sigma \geq \frac{\epsilon_n}{8}.$$

Our next goal will be to upper bound the test threshold, $t_2(\alpha/2, \sigma)$. In particular, we claim that,

$$t_2(\alpha/2, \sigma) \leq \sqrt{\frac{2\text{Var}(T_2(\sigma))}{\alpha}} \tag{B.11}$$

Taking this claim as given for now and supposing that our sample size can be written as $n = c \max\left\{\frac{T_\sigma(p_0)^{1/3}}{p_{\min,\sigma}^{1/3}\Delta_{\mathcal{B}_\sigma}}, \frac{T_\sigma(p_0)}{\Delta_{\mathcal{B}_\sigma}^2}\right\}$, for some $c \geq 1$, we can use Lemma B.2 and Chebyshev's inequality to obtain that,

$$P(\phi_{2/3} = 0) \leq \frac{1}{(\sqrt{\frac{c}{16}} - \sqrt{\frac{2}{\alpha}})^2},$$

provided that $\sqrt{\frac{c}{16}} \geq \sqrt{\frac{2}{\alpha}}$. Thus, it suffices to ensure that,

$$n \geq 64 \max\left\{\frac{2}{\alpha}, \frac{1}{\zeta}\right\} \max\left\{\frac{V_\sigma(p_0)^{1/3}}{p_{\min,\sigma}^{1/3}\Delta_{\mathcal{B}_\sigma}}, \frac{V_\sigma(p_0)}{\Delta_{\mathcal{B}_\sigma}^2}\right\},$$

to obtain that $P(\phi_{2/3} = 0) \leq \zeta$ as desired. Using Lemma B.3, we have that this holds under the condition on n. It remains to verify the claim in (B.11). In order to do so we just note that the variance of the statistic is minimized at the null, i.e.

$$\text{Var}(T_2(\sigma)) \stackrel{\text{(i)}}{\geq} \sum_{i \in \mathcal{B}_\sigma} 2n^2 p_0(i)^{2/3} = \alpha t_2^2(\alpha/2, \sigma),$$

as desired. The inequality (i) follows similar reasoning to the Equation (B.3).

B.3. Proof of Theorem 3.4. Fix any multinomial p on $\{1, \ldots, d\}$, and suppose we denote $\Delta_i = p_0(i) - p(i)$, then a straightforward computation shows that,

$$\mathbb{E}_p[T_j] = n^2 \sum_{t \in S_j} \Delta_t^2, \tag{B.12}$$

$$\mathrm{Var}_p[T_j] = \sum_{t \in S_j} \left[2n^2 p_0(t)^2 + 2n^2 \Delta_t^2 - 4n^2 \Delta_t p_0(t) + 4n^3 \Delta_t^2 p_0(t) - 4n^3 \Delta_t^3\right]. \tag{B.13}$$

This in turn yields that the null variance of T_j is simply $\mathrm{Var}_0[T_j] = 2n^2 \sum_{t \in S_j} p_0(t)^2$. By Chebyshev's inequality we then obtain that:

$$P_0(T_j > t_j) \leq \alpha/k,$$

which together with the union bound yields,

$$P_0(\phi_{\max} = 1) \leq \alpha.$$

As in the proof of Theorem 3.3 we consider two cases: when $p \in \mathcal{S}_1$ and when $p \in \mathcal{S}_2$. Since the composite test includes the tail test, the analysis of the case when $p \in \mathcal{S}_1$ is identical to before. Now, we consider the case when $p \in \mathcal{S}_2$.

We have further partitioned the bulk of the distribution into at most k sets, so that at least one of the sets S_j must witness a discrepancy of at least $\epsilon_n/(8k)$, i.e. when $p \in \mathcal{S}_2$ we have that,

$$\sup_j \sum_{i \in S_j} |p_0(i) - p(i)| \geq \frac{\epsilon_n}{8k}.$$

Let j^* denote the set that witnesses this discrepancy. We focus the rest of the proof on this fixed set S_{j^*} and show that under the alternate $T_{j^*} > t_{j^*}$ with sufficiently high probability. Suppose that for j^* we can verify the following two conditions:

$$t_{j^*} \leq \frac{\mathbb{E}_p[T_{j^*}]}{2} \tag{B.14}$$

$$\mathbb{E}_p[T_{j^*}] \geq 2\sqrt{\frac{\mathrm{Var}_p[T_{j^*}]}{\zeta}}, \tag{B.15}$$

then we obtain that $P(\phi_{\max} = 0) \le \zeta$ (see (B.6)). Consequently, we focus the rest of the proof on showing the above two conditions. We let d_{j^*} denote the size of S_{j^*}.

Condition in Equation (B.14)**:** Observe that,

$$\sum_{i \in S_{j^*}} \Delta_i^2 \ge \frac{\left(\sum_{i\in S_{j^*}} |\Delta_i|\right)^2}{d_{j^*}} \ge \frac{\epsilon_n^2}{64k^2 d_{j^*}}. \tag{B.16}$$

Using Equations (3.13) and (B.12), it suffices to check that,

$$n\sqrt{\frac{2k\sum_{i\in S_{j^*}} p_0(i)^2}{\alpha}} \le n^2 \sum_{i\in S_{j^*}} \Delta_i^2,$$

and applying the lower bound in Equation (B.16) it suffices if,

$$\frac{\epsilon_n^2}{64k^2 d_{j^*}} \ge \frac{1}{n}\sqrt{\frac{2k\sum_{i\in S_{j^*}} p_0(i)^2}{\alpha}}.$$

Denote the maximum and minimum entry of the multinomial on S_{j^*} as b_{j^*} and a_{j^*}, respectively. Noting that on each bin the multinomial is roughly uniform one can further observe that,

$$d_{j^*}\sqrt{\sum_{i\in S_{j^*}} p_0(i)^2} \le d_{j^*}^{3/2} b_{j^*} \le 2d_{j^*}^{3/2} a_{j^*} \le 2V_{\epsilon_n/8}(p_0).$$

This yields that the first condition is satisfied if,

$$\epsilon_n^2 \ge \frac{256k^{5/2}}{n}\frac{V_{\epsilon_n/8}(p_0)}{\sqrt{\alpha}},$$

which is indeed the case.

Condition in Equation (B.15)**:** We proceed by upper bounding the variance under the alternate. Using Equation (B.13) we have,

$$\begin{aligned}
\mathrm{Var}_p[T_{j^*}] &= \sum_{t\in S_{j^*}} \left[2n^2 p_0(t)^2 + 2n^2\Delta_t^2 - 4n^2\Delta_t p_0(t) + 4n^3\Delta_t^2 p_0(t) - 4n^3\Delta_t^3\right] \\
&\le \sum_{t\in S_{j^*}} \left[4n^2 p_0(t)^2 + 4n^2\Delta_t^2 + 4n^3\Delta_t^2 p_0(t) - 4n^3\Delta_t^3\right] \\
&\le \underbrace{4n^2 b_{j^*}^2 d_{j^*}}_{U_1} + \underbrace{4n^2 \sum_{t\in S_{j^*}} \Delta_t^2}_{U_2} + \underbrace{4n^3 b_{j^*} \sum_{t\in S_{j^*}} \Delta_t^2}_{U_3} \underbrace{-4n^3 \sum_{t\in S_{j^*}} \Delta_t^3}_{U_4}.
\end{aligned}$$

In order to check the desired condition, it suffices to verify that

$$\mathbb{E}_p[T_{j^*}] \geq 8\sqrt{\frac{U_i}{\zeta}},$$

for each $i \in \{1, 2, 3, 4\}$. We consider these tasks in sequence. For the first term we obtain that it suffices if,

$$\sum_{i \in S_{j^*}} \Delta_i^2 \geq \left(\frac{16 b_{j^*} d_{j^*}^{1/2}}{n\sqrt{\zeta}} \right),$$

and applying the lower bound in Equation (B.16), and from some straightforward algebra it is sufficient to ensure that,

$$\epsilon_n^2 \geq \frac{2048 k^2 V_{\epsilon_n/8}(p_0)}{n\sqrt{\zeta}},$$

which is indeed the case. For the second term, some simple algebra yields that it suffices to have that,

$$\sum_{i \in S_{j^*}} \Delta_i^2 \geq \left(\frac{144}{n^2 \zeta} \right). \tag{B.17}$$

In order to establish this, we need to appropriately lower bound n. Let $p_{\min}$ denote the smallest entry in $\mathcal{B}_{\epsilon_n/8}(p_0)$. For a sufficiently large universal constant $C > 0$, let us denote:

$$\theta_{k,\alpha} := Ck^2 \left[\sqrt{\frac{k}{\alpha}} + \frac{1}{\zeta} \right].$$

Then using the lower bound on ϵ_n we obtain,

$$n \geq \frac{\theta_{k,\alpha} V_{\epsilon_n/16}(p_0)}{\epsilon_n^2} = \frac{\theta_{k,\alpha} V_{\epsilon_n/16}(p_0)^{1/3} \left[\sum_{i \in \mathcal{B}_{\epsilon_n/16}(p_0)} p_0(i)^{2/3} \right]}{\epsilon_n^2}.$$

Now denote $B = \mathcal{B}_{\epsilon_n/16}(p_0) \backslash \mathcal{B}_{\epsilon_n/8}(p_0)$, then we have that,

$$p_{\min} + \sum_{i \in B} p_0(i) \geq \epsilon_n/16,$$

so that,

$$\sum_{i \in \mathcal{B}_{\epsilon_n/16}(p_0)} p_0(i)^{2/3} \geq \sum_{i \in B} p_0(i)^{2/3} + p_{\min}^{2/3} = \frac{1}{p_{\min}^{1/3}} \left[\sum_{i \in B} p_0(i)^{2/3} p_{\min}^{1/3} + p_{\min} \right] \geq \frac{\epsilon_n}{16 p_{\min}^{1/3}},$$

where the final inequality uses the fact that for $i \in B$, $p_0(i) \leq p_{\min}$. This gives the lower bound,

$$n \geq \frac{\theta_{k,\alpha} V_{\epsilon_n/16}(p_0)^{1/3}}{16\epsilon_n p_{\min}^{1/3}} \geq \frac{\theta_{k,\alpha} \left(\sum_{t \in S_{j^*}} p_0(t)^{2/3}\right)^{1/2}}{16\epsilon_n p_{\min}^{1/3}} \geq \frac{\theta_{k,\alpha}\sqrt{d_{j^*}}}{16\epsilon_n}.$$

Returning to the bound in Equation (B.17), and using the lower bound in Equation (B.16) we obtain that it suffices to ensure that

$$\frac{\epsilon_n}{8k\sqrt{d_{j^*}}} \geq \left(\frac{192\epsilon_n}{\theta_{k,\alpha}\sqrt{d_{j^*}}\sqrt{\zeta}}\right),$$

which is indeed the case. Turning our attention to the term involving U_3 we have, that by some simple algebra it suffices to verify that,

$$\sum_{i \in S_{j^*}} \Delta_i^2 \geq \left(\frac{144 b_{j^*}}{n\zeta}\right).$$

Using the lower bound in Equation (B.16) we obtain that it is sufficient to ensure,

$$\frac{\epsilon_n^2}{64k^2 d_{j^*}} \geq \left(\frac{144 b_{j^*}}{n\zeta}\right),$$

and with the observation that $d_{j^*} b_{j^*} \leq 2 d_{j^*}^{3/2} a_{j^*} \leq 2V_{\epsilon_n/8}(p_0)$ we obtain,

$$\epsilon_n^2 \geq \left(\frac{18432 k^2 V_{\epsilon_n/8}(p_0)}{n\zeta}\right),$$

which is indeed the case. Finally, we turn our attention to the term involving U_4. In this case we have that it suffices to show that,

$$n^{1/2} \sum_{i \in S_{j^*}} \Delta_i^2 \geq 16\sqrt{\frac{\sum_{i \in S_{j^*}} \Delta_i^3}{\zeta}},$$

by the monotonicity of the ℓ_p norm it suffices then to show that,

$$n^{1/2} \sum_{i \in S_{j^*}} \Delta_i^2 \geq 16\sqrt{\frac{\left[\sum_{i \in S_{j^*}} \Delta_i^2\right]^{3/2}}{\zeta}},$$

and after some simple algebra this yields that it suffices to have,

$$\sum_{i \in S_{j^*}} \Delta_i^2 \geq \frac{(16)^4}{\zeta^2 n^2},$$

and this follows from an essentially identical argument to the one handling the term involving U_2. This completes the proof.

APPENDIX C: PROOFS FOR EXAMPLES OF HÖLDER TESTING

In this Section we provide proofs of the claims in Section 4.1. For convenience, we restate all the claims in the following lemma.

LEMMA C.1.

- *Suppose that p_0 is a standard one-dimensional Gaussian, with mean μ, and variance ν^2, then we have that:*

$$T_0(p_0) = (8\pi)^{1/2}\nu. \tag{C.1}$$

- *Suppose that p_0 is a Beta distribution with parameters α, β. Then we have,*

$$T_0(p_0) = \left(\int_0^1 \sqrt{p_0(x)}dx\right)^2 = \frac{B^2((\alpha+1)/2, (\beta+1)/2)}{B(\alpha, \beta)}, \tag{C.2}$$

where $B : \mathbb{R}^2 \mapsto \mathbb{R}$ is the Beta function. Furthermore, if we take $\alpha = \beta = t \geq 1$, then we have that,

$$\frac{\pi^2}{4e^4} t^{-1/2} \leq T_0(p_0) \leq \frac{e^4}{4} t^{-1/2}. \tag{C.3}$$

- *Suppose that p_0 is Cauchy with parameter α, then we have that,*

$$T_0(p_0) = \infty. \tag{C.4}$$

Furthermore, if $0 \leq \sigma \leq 0.5$ then,

$$\frac{4\alpha}{\pi}\left[\ln^2\left(\frac{1}{\sigma}\right)\right] \leq T_\sigma(p_0) \leq \frac{4\alpha}{\pi}\left[\ln^2\left(\frac{2e}{\pi\sigma}\right)\right]. \tag{C.5}$$

- *Suppose that p_0 has a Pareto distribution with parameter α then we have that,*

$$T_0(p_0) = \infty, \tag{C.6}$$

while the truncated T-functional satisfies:

$$\frac{4\alpha x_0}{(1-\alpha)^2}\left(\sigma^{-\frac{1-\alpha}{2\alpha}} - 1\right)^2 = T_\sigma(p_0) \leq \frac{4\alpha x_0}{(1-\alpha)^2}\sigma^{-\frac{1-\alpha}{\alpha}}. \tag{C.7}$$

PROOF. Notice that Claims (C.4) and (C.6) follow by taking $\sigma \to 0$ in Claims (C.5) and (C.7), respectively. We prove the remaining claims in turn.
Proof of Claim (C.1): Observe that,

$$\begin{aligned} T_0(p_0) &= \frac{1}{\sqrt{2\pi}\nu}\left(\int_{-\infty}^{\infty} \exp(-(x-\mu)^2/(4\nu^2))dx\right)^2 \\ &= \frac{1}{\sqrt{2\pi}\nu} 4\pi\nu^2 \\ &= \sqrt{8\pi}\nu. \end{aligned}$$

Proof of Claim (C.2): The Beta density can be written as:

$$p_0(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}x^{\beta-1} = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}x^{\beta-1},$$

where $\Gamma : \mathbb{R} \mapsto \mathbb{R}$ denotes the Gamma function. Some simple algebra yields that the T-functional is simply:

$$T_0(p_0) = \int_0^1 \sqrt{p_0(x)}dx = \frac{B((\alpha+1)/2, (\beta+1)/2)}{\sqrt{B(\alpha,\beta)}}. \tag{C.8}$$

Proof of Claim (C.3): We now take $\alpha = \beta = t \geq 1$ in the above expression. To prove the claim we use standard approximations to the Beta function derived using Stirling's formula. Recall, that by Stirling's formula we have that:

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \leq n! \leq e\sqrt{n}\left(\frac{n}{e}\right)^n.$$

We begin by upper bounding the Beta function for integers $\alpha, \beta \geq 0$:

$$\begin{aligned} B(\alpha,\beta) &= \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!} = \frac{\alpha!\beta!}{(\alpha+\beta)!}\frac{\alpha+\beta}{\alpha\beta} \\ &\leq \frac{e^2}{\sqrt{2\pi}}\frac{\alpha+\beta}{\alpha\beta}\frac{\sqrt{\alpha\beta}\alpha^\alpha\beta^\beta\exp(\alpha+\beta)}{\sqrt{\alpha+\beta}(\alpha+\beta)^{\alpha+\beta}\exp(\alpha+\beta)} \\ &= \frac{e^2}{\sqrt{2\pi}}\sqrt{\frac{\alpha+\beta}{\alpha\beta}}\frac{\alpha^\alpha\beta^\beta}{(\alpha+\beta)^{\alpha+\beta}}. \end{aligned}$$

Now, setting $\alpha = \beta = t \geq 1$, we obtain:

$$B(t,t) \leq \frac{e^2}{\sqrt{\pi}}\frac{2^{-2t}}{\sqrt{t}}.$$

We can similarly lower bound the Beta function as:

$$B(t,t) \geq \frac{2\sqrt{2}\pi}{e}\frac{2^{-2t}}{\sqrt{t}}.$$

We also need to bound the Beta function at certain non-integer values. In particular, we observe that,

$$B(t+1,t+1) \leq B(t+1/2,t+1/2) \leq B(t,t),$$

so that we can similarly sandwich the Beta function at these non-integer values as:

$$\frac{2\pi}{4e}\frac{2^{-2t}}{\sqrt{t}}. \leq B(t+1/2,t+1/2) \leq \frac{e^2}{\sqrt{\pi}}\frac{2^{-2t}}{\sqrt{t}}.$$

With these bounds in place we can now upper and lower bound the T-functional in (C.8). We can upper bound this expression by considering the cases when t is odd and t is even separately, and taking the worse of these two bounds to obtain:

$$T(p_0) \leq \frac{e^2}{2}t^{-1/4}.$$

Similarly, using the above results we can lower bound the T-functional as:

$$T(p_0) \geq \frac{\pi}{2e^2}t^{-1/4},$$

and this yields the claim.

Proof of Claim (C.5): We are interested in the truncated T-functional. The set B_σ of probability content $1-\sigma$, takes the form $[-\alpha,\alpha]$, where

$$\alpha = \gamma\tan\left(\frac{\pi}{2}(1-\sigma)\right) = \gamma\cot\left(\frac{\pi\sigma}{2}\right).$$

Using the inequality that $\cot(x) \leq \frac{1}{x}$, we can upper bound α as:

$$\alpha \leq \frac{2\gamma}{\pi\sigma}.$$

Similarly, we can (numerically) lower bound α by noting that for $0 \leq \sigma \leq 0.5$ we have that,

$$\alpha \geq \frac{\gamma}{4\sigma}.$$

With these bounds in place, we can now proceed to upper and lower bound the truncated T functional. Concretely,

$$\begin{aligned}
T_\sigma(p_0) &\leq \frac{\gamma}{\sqrt{\pi\gamma}} \int_{-\frac{2\gamma}{\pi\sigma}}^{\frac{2\gamma}{\pi\sigma}} \frac{1}{\sqrt{x^2+\gamma^2}} dx \leq \frac{2\gamma}{\sqrt{\pi\gamma}} \left[\int_0^\gamma \frac{1}{\gamma} dx + \int_\gamma^{\frac{2\gamma}{\pi\sigma}} \frac{1}{x} dx \right] \\
&\leq \frac{2\gamma}{\sqrt{\pi\gamma}} \left[1 + \ln\left(\frac{2}{\pi\sigma}\right) \right] \\
&= 2\sqrt{\frac{\gamma}{\pi}} \left[\ln\left(\frac{2e}{\pi\sigma}\right) \right].
\end{aligned}$$

In a similar fashion, we can lower bound the functional as:

$$T_\sigma(p_0) \geq 2\sqrt{\frac{\gamma}{\pi}} \left[\ln\left(\frac{1}{\sigma}\right) \right].$$

Taken together these bounds give the desired claim.

Proof of Claim (C.7): We treat x_0 as a fixed constant. The CDF for the Pareto family of distributions takes the simple form:

$$F(x) = 1 - \left(\frac{x_0}{x}\right)^\alpha, \quad \text{for} \quad x \geq x_0,$$

we obtain that the set B_σ takes the form $[x_0, x_0\sigma^{-1/\alpha}]$. So that the truncated functional is simply:

$$\begin{aligned}
T_\sigma(p_0) &= \int_{x_0}^{x_0\sigma^{-1/\alpha}} \sqrt{p_0(x; x_0, \alpha)} dx \\
&= \frac{2\sqrt{\alpha x_0}}{1-\alpha} \left(\sigma^{-\frac{1-\alpha}{2\alpha}} - 1\right),
\end{aligned}$$

which yields the desired claim. □

APPENDIX D: PROPERTIES OF THE T-FUNCTIONAL

The rate for Hölder testing is largely dependent on the truncated T-functional of the null hypothesis. In this section we establish several properties of the T-functional, and its stability with respect to perturbations. There are two notions of stability of the truncated T-functional that are of interest: its stability with respect to perturbation of the truncation parameter, and its stability with respect to perturbations of the density p_0. In particular, the truncation stability determines the discrepancy between the upper and lower bounds in Theorem 4.1.

Our interest is in the difference between $T_{\sigma_1}(p_0)$ and $T_{\sigma_2}(p_0)$ (where without loss of generality we take $\sigma_1 \leq \sigma_2$). We show that if the support of the density is stable with respect to the truncation parameter then so is the T-functional. Intuitively, the discrepancy can be large only if the density has a long σ_1-tail but a relatively small σ_2-tail. Returning to the definition of the T-functional in (4.2), we let B_{σ_1} and B_{σ_2} denote the sets that achieve the infimum for T_{σ_1} and T_{σ_2}, respectively. These are not typically well-defined for two reasons: the set may not be unique and the infimum might not be attained. The second problem can be easily dealt with by introducing a small amount of slack. To deal with the non-uniqueness we simply choose the sets that have maximal overlap in Lebesgue measure, i.e. we define B_{σ_1} and B_{σ_2} to be two sets that have maximal Lebesgue overlap such that,

$$\Big(\int_{B_{\sigma_1}} \sqrt{p_0(x)}dx\Big)^2 \geq T_{\sigma_1}(p_0) - \xi,$$
$$\Big(\int_{B_{\sigma_2}} \sqrt{p_0(x)}dx\Big)^2 \geq T_{\sigma_2}(p_0) - \xi,$$

for an arbitrary small $\xi > 0$. The quantity ξ may be taken as small as we like and has no effect when chosen small enough so we ignore it in what follows. We define, $S = B_{\sigma_1} \backslash B_{\sigma_2}$ which measures the stability of the support with respect to changes in the truncation parameter, i.e. if the Lebesgue measure $\mu(S)$ is small then the support is stable. With these definitions in place we have the following lemma:

LEMMA D.1. *For any two truncation levels $\sigma_1 \leq \sigma_2$, we have that,*

$$T^{\gamma}_{\sigma_1}(p_0) - T^{\gamma}_{\sigma_2}(p_0) \leq (\sigma_1 - \sigma_2)^{\gamma} \mu(S)^{1-\gamma}.$$

Remarks:

- Since $\gamma < 1$, this result asserts that if the support of the density is stable with respect to the truncation parameter then so is the truncated T-functional. This is the case in all the examples we considered in Section 4.1.
- If we restrict attention to compactly supported densities then we can upper bound $\mu(S)$ by the Lebesgue measure of the support indicating that in these cases the truncated T-functional is somewhat stable.
- On the other hand this result also gives insight into when the truncated functional is not stable. In particular, it is straightforward to construct examples of densities p_0 which have a very long σ_1-tail but a light σ_2-tail, in which case this discrepancy can be arbitrarily large. Noting

however that in our bounds the regime of interest is when the truncation parameter is not fixed, i.e. when $\sigma \to 0$, in which case this discrepancy can be large only for carefully constructed pathological densities.

PROOF. The result follows using Hölder's inequality:

$$\begin{aligned} T^{\gamma}_{\sigma_1}(p_0) - T^{\gamma}_{\sigma_2}(p_0) &= \int_S p_0^{\gamma}(x)dx \\ &= \mu(S) \int_S \frac{p_0^{\gamma}(x)}{\mu(S)} dx \\ &\le \mu(S) \left(\int_S \frac{p_0(x)}{\mu(S)} dx \right)^{\gamma} \\ &= \mu(S)^{1-\gamma} (\sigma_1 - \sigma_2)^{\gamma}. \end{aligned}$$

□

In order to understand the stability of the T-functional with respect to perturbations of p_0 it is natural to consider a form of the modulus of continuity. We restrict our attention to densities p_0 which have support contained in a fixed set S, and denote these densities by $\mathcal{L}(L_n, S)$, and only consider the case when $d = 1$ and hence $\gamma = 1/2$.

Focussing on the case when the truncation parameter is fixed (say to 0) we define:

$$s(p_0, \tau, S) = \sup_{p, p_0 \in \mathcal{L}(L_n, S), \|p - p_0\|_1 \le \tau} |T_0^{\gamma}(p) - T_0^{\gamma}(p_0)|.$$

With these definitions in place, we have the following result:

LEMMA D.2. *For any p_0, the modulus of continuity of the T-functional is upper bounded as:*

$$s(p_0, \tau, S) \le \sqrt{\tau \mu(S)}.$$

Remark:

- This result guarantees that for densities that are close in ℓ_1, their corresponding T-functionals are close, provided that we restrict attention to compactly supported densities.
- On the other hand, an inspection of the proof below reveals that if we eliminate the restriction of compact support, then for any density p_0, we can construct a density p that is close in ℓ_1 but has an arbitrarily large discrepancy in the T-functional, i.e. the T-functional can be highly unstable to perturbations of p_0 if we allow densities with arbitrary support.

PROOF. Notice that,

$$\begin{aligned}
T_0^\gamma(p) - T_0^\gamma(p_0) &= \int_S (\sqrt{p(x)} - \sqrt{p_0(x)})dx \\
&= \mu(S) \int_S (\sqrt{p(x)} - \sqrt{p_0(x)}) \frac{1}{\mu(S)} dx \\
&\leq \mu(S) \sqrt{\int_S (\sqrt{p(x)} - \sqrt{p_0(x)})^2 \frac{1}{\mu(S)} dx} \\
&\overset{\text{(i)}}{\leq} \sqrt{\mu(S) \|p - p_0\|_1} \leq \sqrt{\tau \mu(S)},
\end{aligned}$$

where (i) uses the fact that the Hellinger distance is upper bounded by the ℓ_1 distance. □

D.1. Proof of Claim (4.6). This claim is a straightforward consequence of Hölder's inequality. We have that,

$$T_\sigma(p_0) = \inf_{B \in \mathcal{B}_\sigma} \left(\int_B p_0^\gamma(x) dx \right)^{1/\gamma}.$$

We restrict our attention to densities with support contained in a fixed set S. We let B_σ denote an arbitrary set in $\mathcal{B}_\sigma$ that minimizes the above integral (dealing with non-uniqueness as before). Then,

$$\begin{aligned}
T_\sigma^\gamma(p_0) &= \mu(B_\sigma) \int_{B_\sigma} \frac{p_0^\gamma(x)}{\mu(B_\sigma)} dx \\
&\overset{\text{(i)}}{\leq} \mu(B_\sigma) \left(\int_{B_\sigma} \frac{p_0(x)}{\mu(B_\sigma)} dx \right)^\gamma \\
&= \mu(B_\sigma)^{1-\gamma} (1-\sigma)^\gamma \leq \mu(S)^{1-\gamma} (1-\sigma)^\gamma
\end{aligned}$$

where (i) uses Hölder's inequality. This yields the claim. For the uniform distribution u on the set S we have that for any set B_σ of mass $1-\sigma$,

$$\begin{aligned}
T_\sigma^\gamma(u) &= \int_{B_\sigma} \frac{1}{\mu^\gamma(S)} dx \\
&= \mu(S)^{1-\gamma}(1-\sigma),
\end{aligned}$$

which matches the result of (4.6) up to constant factors involving σ and γ. In particular, our interest is in the regime when $\sigma \to 0$, and γ is a constant, in which case the two quantities are equal.

APPENDIX E: TECHNICAL RESULTS FOR HÖLDER TESTING

In this section we provide the remaining technical proofs related the Theorem 4.1. We begin with the preliminary Lemmas 4.3 and 4.4.

E.1. Preliminaries.

E.1.1. *Proof of Lemma 4.1.* We first prove the upper bound. Note that,

$$\epsilon_n = \int \min\left\{\frac{p_0(x)}{x}, \frac{\epsilon_n p_0(x)^\gamma}{\mu(x)}\right\} dx \leq \int_{B_{\epsilon_n/\delta}} \frac{\epsilon_n p_0(x)^\gamma}{\mu(x)} + \int_{B^c_{\epsilon_n/\delta}} \frac{p_0(x)}{x} \leq \int_{B_{\epsilon_n/\delta}} \frac{\epsilon_n p_0(x)^\gamma}{\mu(x)} + \frac{\epsilon_n}{\delta x}.$$

which yields that,

$$\mu(x) \leq \left(1 - \frac{1}{\delta x}\right)^{-1} T_{\epsilon_n/\delta}.$$

Choosing $\delta = 2/x$ we obtain the upper bound. In order to prove the lower bound, we first define:

$$G = \left\{x : \frac{\epsilon_n p_0(x)^\gamma}{\mu(x)} < \frac{p_0(x)}{x}\right\}.$$

Note that,

$$\epsilon_n \geq \int_{G^c} \frac{p_0(x)}{x},$$

so we obtain that $\mathbb{P}(G) \geq 1 - x\epsilon_n$. Also, we have that,

$$\epsilon_n \geq \int_G \frac{\epsilon_n p_0(x)^\gamma}{\mu(x)} dx,$$

which in turn yields that,

$$\mu(x) \geq \int_G p_0(x)^\gamma dx \geq T^\gamma_{x\epsilon_n},$$

where the final inequality uses the fact that $T_{x\epsilon_n}$ is defined to be the infimizer of the integral of $p_0(x)^\gamma$ over all sets of measure at least $1 - x\epsilon_n$ while G is one such set.

E.1.2. *Proof of Lemma 4.3.* Let $\mathcal{A}$ be all sets A such that $P_0^n(A) \leq \alpha$. Now

$$\begin{aligned}\zeta_n(\mathcal{P}) &\geq \inf_{\phi} Q(\phi = 0) \geq 1 - \alpha - \sup_{A\in\mathcal{A}} |Q(A) - P_0^n(A)| \\ &\geq 1 - \alpha - \sup_{A} |Q(A) - P_0^n(A)| \\ &= 1 - \alpha - \frac{1}{2}\|Q - P_0^n\|_1.\end{aligned}$$

Note that

$$\|Q - P_0^n\|_1 = \mathbb{E}_0|W_n(Z_1, \ldots, Z_n) - 1| \leq \sqrt{\mathbb{E}_0[W_n^2(Z_1, \ldots, Z_n)] - 1}.$$

The result then follows from (4.20).

E.1.3. *Proof of Lemma 4.4.* We divide the proof into several claims.
Claim 1: Each p_η is a density function. Note that

$$\int p_\eta(x)dx = 1.$$

Now we show it is non-negative. Let $x \in A_j$. Then

$$\begin{aligned}p_\eta(x) &= p_0(x) + \rho_j\eta_j\psi_j(x) \geq p_0(x) - \rho_j\psi_j(x) \\ &\geq p_0(x) - \frac{\rho_j}{c_j^{d/2}h_j^{d/2}}\|\psi\|_\infty.\end{aligned}$$

Now, we observe that for each piece of our partition we have that,

$$p_0(x) \geq \frac{p_0(x_j)}{2} \geq \frac{(\sqrt{d}h_j)^s}{2\theta_1} = L_n(\sqrt{d}h_j)^s,$$

where we use the fact that $\theta_1 = 1/(3L_n)$. We then obtain that it suffices to choose,

$$\rho_j \leq \frac{L_n c_j^{d/2}}{\|\psi\|_\infty} h_j^{d/2+s},$$

which is ensured by the condition in (4.22).

Claim 2: Each $p_\eta \in \mathcal{L}(L_n)$. Let x, y be two points, and that $x \in A_j, y \in A_k$. We consider two cases: when neither of j, k are ∞, and when at least one of them is. Noting that we do not perturb A_∞ the second case follows from a similar argument to that of the first case. In the first case, we have that:

$$
\begin{aligned}
|p_\eta(y) - p_\eta(x)| &\leq |p_0(x) - p_0(y)| + \left| \frac{\rho_k \eta_k}{c_k^{d/2} h_k^{d/2}} \psi\left(\frac{y - x_k}{c_k h_k}\right) - \frac{\rho_j \eta_j}{c_j^{d/2} h_j^{d/2}} \psi\left(\frac{x - x_j}{c_j h_j}\right) \right| \\
&\leq c_{\mathrm{int}} L_n \|x - y\|_2^s + \left| \frac{\rho_k \eta_k}{c_k^{d/2} h_k^{d/2}} \psi\left(\frac{y - x_k}{c_k h_k}\right) - \frac{\rho_k \eta_k}{c_k^{d/2} h_k^{d/2}} \psi\left(\frac{x - x_k}{c_k h_k}\right) \right| \\
&+ \left| \frac{\rho_j \eta_j}{c_j^{d/2} h_j^{d/2}} \psi\left(\frac{y - x_j}{c_j h_j}\right) - \frac{\rho_j \eta_j}{c_j^{d/2} h_j^{d/2}} \psi\left(\frac{x - x_j}{c_j h_j}\right) \right| \\
&\leq c_{\mathrm{int}} L_n \|x - y\|_2^s + \frac{\rho_j}{c_j^{d/2} h_j^{d/2}} \min\left\{2\|\psi\|_\infty, \frac{\|\psi'\|_\infty \|x - y\|}{c_j h_j}\right\} \\
&+ \frac{\rho_k}{c_k^{d/2} h_k^{d/2}} \min\left\{2\|\psi\|_\infty, \frac{\|\psi'\|_\infty \|x - y\|}{c_k h_k}\right\} \\
&\leq c_{\mathrm{int}} L_n \|x - y\|_2^s + \frac{\rho_j}{c_j^{d/2} h_j^{d/2}} \max\{2\|\psi\|_\infty, \|\psi'\|_\infty\} \min\left\{1, \frac{\|x - y\|}{c_j h_j}\right\} \\
&+ \frac{\rho_k}{c_k^{d/2} h_k^{d/2}} \max\{2\|\psi\|_\infty, \|\psi'\|_\infty\} \min\left\{1, \frac{\|x - y\|_2}{c_k h_k}\right\} \\
&\leq c_{\mathrm{int}} L_n \|x - y\|_2^s + \frac{\rho_k \max\{2\|\psi\|_\infty, \|\psi'\|_\infty\} \|x - y\|_2^s}{c_k^{d/2+s} h_k^{d/2+s}} \\
&+ \frac{\rho_j \max\{2\|\psi\|_\infty, \|\psi'\|_\infty\} \|x - y\|_2^s}{c_j^{d/2+s} h_j^{d/2+s}},
\end{aligned}
$$

so that it suffices to ensure that for $i \in \{1, \ldots, N\}$,

$$
\rho_i \leq \frac{(1 - c_{\mathrm{int}}) L_n c_i^{d/2+s} h_i^{d/2+s}}{2 \max\{2\|\psi\|_\infty, \|\psi'\|_\infty\}},
$$

which, noting that $1 \geq c_j \geq 1/4$, is ensured by the condition in (4.22).

Claim 3. $\int |p_0 - p_\eta| \geq \epsilon_n$**.** We have

$$\begin{aligned}\int |p_0 - p_\eta| &= \sum_j \int_{A_j} |p_0 - p_\eta| = \sum_j \int_{A_j} |\rho_j \eta_j \psi_j| \\ &= \sum_j \rho_j \int_{A_j} |\psi_j| = \sum_j \rho_j \int_{A_j} \frac{1}{c_j^{d/2} h_j^{d/2}} \left| \psi\left(\frac{x - x_j}{c_j h_j}\right)\right| \\ &= \sum_j \rho_j c_j^{d/2} h_j^{d/2} \int_{[-1/2,1/2]^d} |\psi| = \omega_2 \sum_j \rho_j c_j^{d/2} h_j^{d/2} \geq \epsilon_n,\end{aligned}$$

where we use the condition in (4.23). Taken together claims 1, 2 and 3 show that $p_\eta \in \mathcal{L}(L_n)$ and that $\|p_\eta - p_0\|_1 \geq \epsilon_n$.

Claim 4: Likelihood ratio bound. For observations $\{Z_1, \ldots, Z_n\}$ the likelihood ratio is given as

$$W_n(Z_1, \ldots, Z_n) = \frac{1}{2^N} \sum_{\eta \in \{-1,1\}^N} \prod_i \frac{p_\eta(Z_i)}{p_0(Z_i)}$$

and

$$\begin{aligned}W_n^2(Z_1, \ldots, Z_n) &= \frac{1}{2^{2N}} \sum_{\eta \in \{-1,1\}^N} \sum_{\nu \in \{-1,1\}^N} \prod_i \frac{p_\eta(Z_i) p_\nu(Z_i)}{p_0(Z_i) p_0(Z_i)} \\ &= \frac{1}{2^{2N}} \sum_{\eta \in \{-1,1\}^N} \sum_{\nu \in \{-1,1\}^N} \prod_i \left(1 + \frac{\sum_{j=1}^N \rho_j \eta_j \psi_j(Z_i)}{p_0(Z_i)}\right)\left(1 + \frac{\sum_{j=1}^N \rho_j \nu_j \psi_j(Z_i)}{p_0(Z_i)}\right).\end{aligned}$$

Taking the expected value over $Z_1, \ldots, Z_n$, and using the fact that the ψ_js have disjoint support we obtain

$$\begin{aligned}E_0[W_n^2(Z_1, \ldots, Z_n)] &= \frac{1}{2^{2N}} \sum_{\eta \in \{-1,1\}^N} \sum_{\nu \in \{-1,1\}^N} \left(1 + \sum_{j=1}^N \rho_j^2 \eta_j \nu_j a_j\right)^n \\ &\leq \frac{1}{2^{2N}} \sum_{\eta \in \{-1,1\}^N} \sum_{\nu \in \{-1,1\}^N} \exp\left(n \sum_j \rho_j^2 \eta_j \nu_j a_j\right)\end{aligned}$$

where

$$\begin{aligned}a_j &= \int_{A_j} \frac{\psi_j^2(z)}{p_0(z)} dz = \frac{1}{p_0(z_j)} \int_{A_j} \psi_j^2(z) \frac{p_0(z_j)}{p_0(z)} dz \\ &\leq \frac{2}{p_0(z_j)}.\end{aligned}$$

Thus $E_0[W_n^2(Z_1,\ldots,Z_n)] \le E_{\eta,\nu} e^{n\langle \eta,\nu\rangle}$ where we use the weighted inner product defined as:

$$\langle \eta,\nu\rangle := \sum_j \rho_j^2 \eta_j \nu_j a_j.$$

Hence,

$$\begin{aligned} E_0[W_n^2(Z_1,\ldots,Z_n)] &\le E_{\eta,\nu} e^{n\langle \eta,\nu\rangle} = \prod_j E e^{n\eta_j\nu_j} \\ &= \prod_j \cosh(n\rho_j^2 a_j) \le \prod_j (1 + n^2\rho_j^4 a_j^2) \le \prod_j \exp(n^2\rho_j^4 a_j^2) \\ &= \exp\left\{\sum_j n^2\rho_j^4 a_j^2\right\} \le \exp\left\{4n^2 \sum_j \frac{\rho_j^4}{p_0^2(x_j)}\right\} \le C_0, \end{aligned}$$

where the final inequality uses the condition in (4.24). From Lemma 4.3, it follows that the Type II error of any test is at least δ.

E.2. Further technical preliminaries. Our analysis of the pruning in Algorithm 2 uses various results that we provide in this section.

LEMMA E.1. *Let P be a distribution with density p and let $\gamma \in [0,1)$. Let*

$$A = \{x : \; p(x) \ge t\}$$

for some t. Define $\theta = P(A)$. Finally, let $\mathcal{B} = \{B : \; P(B) \ge \theta\}$. Then, for every $B \in \mathcal{B}$,

$$\int_A p^\gamma(x)dx \le \int_B p^\gamma(x)dx.$$

PROOF. Let

$$S_1 = A \bigcap B^c, \quad S_2 = A^c \bigcap B.$$

Then

$$\int_A p^\gamma(x)dx - \int_B p^\gamma(x)dx = \int_{S_1} p^\gamma(x)dx - \int_{S_2} p^\gamma(x)dx.$$

So it suffices to show that $\int_{S_1} p^\gamma(x)dx \le \int_{S_2} p^\gamma(x)dx$. Note that:

1. S_1 and S_2 are disjoint,
2. $\inf_{y\in S_1} p(y) \ge \sup_{y \in S_2} p(y)$ and
3. $\int_{S_1} p(x)dx \le \int_{S_2} p(x)dx$.

where the last fact follows since $\int_A p(x)dx \leq \int_B p(x)dx$. Thus, letting $g(x) = 1/p^{1-\gamma}(x)$, we have that

$$g(x) \leq g(y)$$

for all $x \in S_1$ and $y \in S_2$. So

$$\begin{aligned}
\int_{S_1} p^{\gamma}(x)dx &= \int_{S_1} p(x)g(x)dx \leq \sup_{x\in S_1} g(x) \int_{x\in S_1} p(x)dx \\
&\leq \sup_{x\in S_1} g(x) \int_{x\in S_2} p(x)dx \leq \inf_{x\in S_2} g(x) \int_{x\in S_2} p(x)dx \\
&\leq \int_{S_2} p(x)g(x)dx = \int_{S_2} p^{\gamma}(x)dx.
\end{aligned}$$

□

The following lemma concerns the optimal truncation of a piecewise constant function. Suppose we have a piecewise constant positive function f, which is constant on the partition $\{A_1, \ldots, A_N\}$. Without loss of generality suppose that $A_1, \ldots, A_N$ are arranged in decreasing order of the value of f on the cell A_i. The lemma follows from lemma E.1.

Lemma E.2. *With the notation introduced above suppose that we construct a set $A = \bigcup_{i=1}^{t} A_i$ and let $\theta = \int_A f(x)$ then we have that for any $\gamma \leq 1$*

$$\int_A f^{\gamma}(x)dx \leq \inf_{B,\ \int_B f(x)\geq\theta} \int_B f^{\gamma}(x)dx.$$

The following result is the discrete analogue of the one above. Suppose that we have a sequence $\{p_1, \ldots, p_d\}$ of positive numbers sorted as $p_1 \geq p_2 \geq \ldots p_d$. By replacing Lebesgue measure in Lemma E.1 by the counting measure we get:

Lemma E.3. *Suppose we construct a set of indices $A = \{1, \ldots, t\}$ and let $\theta = \sum_{i=1}^{t} p_i$, then we have that,*

$$\sum_{i=1}^{t} p_i^{2/3} \leq \min_{\mathcal{J}} \sum_{j\in\mathcal{J},\ \sum_{k\in\mathcal{J}} p_k \geq \theta} p_j^{2/3}.$$

E.3. Proof of Lemma 4.2. We divide the proof into two steps: the first step analyzes the output of Algorithm 1, and the second step analyzes the pruning of Algorithm 2.

E.3.1. *Analysis of Algorithm 1.* We analyze Algorithm 1, with the parameters: $\theta_1 = 1/(3L_n)$ and $a, b = \epsilon_n/1024$. We allow $\theta_2 > 0$ to be arbitrary.

Before turning our attention to the main properties, we verify that the partition created by Algorithm 1 is indeed finite. It is immediate to check that the partition $\mathcal{P}^\dagger = \{A_1, \ldots, A_{\bar{N}}, A_\infty\}$ has the property that $P_0(A_\infty) \leq a + b$, which yields the upper bound of property (4.16). We claim that no cell A_i has very small diameter. Recall that Algorithm 1 is run on S_a a set of probability content $1 - a$ (centered around the mean of p_0). Define,

$$p_{\min} = \frac{b}{\text{vol}(S_a)},$$

Suppose that,

$$[\text{diam}(A_i)]^s < \frac{1}{4} \min \left\{\theta_1 p_{\min}, \theta_2 p_{\min}^\gamma\right\}, \tag{E.1}$$

then let us denote the parent cell of A_i by U_i and its centroid by y_i. The parent cell U_i, satisfies the condition that:

$$[\text{diam}(U_i)]^s < \frac{1}{2} \min \left\{\theta_1 p_{\min}, \theta_2 p_{\min}^\gamma\right\}.$$

Since this cell was split, we must have that neither stopping rule (4.10) nor (4.11) was satisfied. We claim that if the second stopping rule was not satisfied it must be the case that,

$$p_0(y_i) \leq \frac{p_{\min}}{2}.$$

Indeed, if the second rule is not satisfied we obtain that:

$$\min \{\theta_1 p_0(y_i), \theta_2 p_0^\gamma(y_i)\} \leq \frac{1}{2} \min \left\{\theta_1 p_{\min}, \theta_2 p_{\min}^\gamma\right\},$$

which via some simple case analysis of the min's, together with the fact that $\gamma < 1$ yields the desired claim. Now using the Hölder property and the fact that $\theta_1 = 1/(3L_n)$, we have that:

$$\sup_{x \in U_i} p_0(x) \leq p_0(y_i) + L_n [\text{diam}(U_i)]^s < \frac{p_{\min}}{2} + \frac{p_{\min}}{4} < p_{\min}.$$

This means that the first stopping rule was in fact satisfied and we could not have split U_i. This in turn means that every cell in our partition (excluding A_∞) has diameter at least:

$$[\text{diam}(A_i)]^s > \frac{1}{4} \min \left\{\theta_1 p_{\min}, \theta_2 p_{\min}^\gamma\right\}.$$

This yields that our produced partition is finite and in turn that algorithm terminates in a finite number of steps.

Proof of Claim (4.14)**:** We now show that the partition satisfies the condition that,

$$\frac{1}{4}\min\left\{\theta_1 p_0(x_i), \theta_2 p_0^\gamma(x_i)\right\} \leq [\mathrm{diam}(A_i)]^s \leq \min\left\{\theta_1 p_0(x_i), \theta_2 p_0^\gamma(x_i)\right\}.$$

The upper bound is straightforward since it is enforced by our stopping rule. To observe that the lower bound is always satisfied we note that if

$$[\mathrm{diam}(A_i)]^s < \frac{1}{4}\min\left\{\theta_1 p_0(x_i), \theta_2 p_0^\gamma(x_i)\right\},$$

then denoting the parent cell of A_i to be U_i (with centroid y_i) we obtain that,

$$[\mathrm{diam}(U_i)]^s < \frac{1}{2}\min\left\{\theta_1 p_0(x_i), \theta_2 p_0^\gamma(x_i)\right\}.$$

Using this we obtain that,

$$p_0(y_i) \geq p_0(x_i) - L_n\,[\mathrm{diam}(U_i)]^s \geq \frac{3}{4}p_0(x_i).$$

This yields that,

$$[\mathrm{diam}(U_i)]^s < \frac{1}{2(3/4)^\gamma}\min\left\{\theta_1 p_0(y_i), \theta_2 p_0^\gamma(y_i)\right\} < \min\left\{\theta_1 p_0(y_i), \theta_2 p_0^\gamma(y_i)\right\},$$

where in our final step we use the fact that $\gamma < 1$. This results in a contradiction since this means that U_i satisfies our stopping rule and would not have been split.

Proof of Claim (4.15)**:** This is a straightforward consequence of the previous property. In particular, we have that $[\mathrm{diam}(A_i)]^s \leq \theta_1 p_0(x_i)$, with $\theta_1 = 1/(3L_n)$ so that,

$$\sup_{x\in A_i} p_0(x) \leq p_0(x_i) + L_n\frac{\theta_1 p_0(x_i)}{2} \leq \frac{5}{4}p_0(x_i).$$

Similarly,

$$\inf_{x\in A_i} p_0(x) \leq p_0(x_i) - L_n\frac{\theta_1 p_0(x_i)}{2} \leq \frac{3}{4}p_0(x_i),$$

which yields the desired claim.

E.3.2. *Analysis of Algorithm 2.* We now turn our attention to studying the properties of the pruned partition $\mathcal{P} = \{A_1, \ldots, A_N, A_\infty\}$. For this algorithm, we choose $\theta_2 = \epsilon_n/(8L_n\mu(1))$ and take $c = \epsilon_n/512$.

Proof of Claim (4.14)**:** The pruning algorithm completely eliminates some cells, adding them to A_∞. In the case when $\mathcal{Q}(j^*) \leq c/5$ we change the diameter of the final cell A_N, shrinking it by a $1-\alpha$ factor. By definition $\alpha \leq 1/5$, and this yields Claim (4.14).

Proof of Claim (4.15)**:** Since the pruning step either eliminates cells, adding them to A_∞, or reduces their diameter this claim follows directly from the fact that this property holds for $\mathcal{P}^\dagger$.

Proof of Claim (4.16)**:** The pruning eliminates cells of total additional mass at most c so we obtain that, $P_0(A_\infty) \leq a + b + c \leq \epsilon_n/256$ verifying the upper bound in (4.16). To verify the lower bound, we claim that the difference in the probability mass of the unpruned partition, $\{A_1, \ldots, A_{\tilde{N}}\}$ and the pruned partition $\{A_1, \ldots, A_N\}$ is at least $c/5$, i.e.

$$P_0\Big(\bigcup_{j=1}^{\tilde{N}} A_j\Big) - P_0\Big(\bigcup_{j=1}^{N} A_j\Big) \geq c/5.$$

In the case when $\mathcal{Q}(j^*) \geq c/5$ the claim is direct. When this is not the case then the cell A_N was too large, so that $\mathcal{Q}(j^*) + P_0(A_N) \geq c$, which implies that, $P_0(A_N) \geq 4c/5$. Let x_N be the center of A_N. Using property (4.15) and the fact that $(1-\alpha)^d \leq (1-\alpha)$ verify that,

$$P_0(D_1) \leq 4(1-\alpha)P_0(A_N).$$

Using the definition of α we obtain that $P_0(D_2) \geq c/5$ as desired.

Proof of Claim (4.17)**:** We claim that the partition satisfies the property that,

$$L_n \sum_{i=1}^{N} [\mathrm{diam}(A_i)]^s \,\mathrm{vol}(A_i) \leq \frac{\epsilon_n}{4}. \tag{E.2}$$

Taking this claim as given we verify the property (4.17). We divide the proof into two cases:

1. $P(A_\infty) \geq \epsilon_n/4$: In this case we obtain that,

$$\sum_{i=1}^{N} |P_0(A_i) - P(A_i)| + |P_0(A_\infty) - P(A_\infty)| \geq |P_0(A_\infty) - P(A_\infty)| \geq \epsilon_n/8,$$

using the upper bound in property (4.16).

2. $P(A_\infty) \leq \epsilon_n/4$: In this case we observe that,

$$\int_{A_\infty} |p_0(x) - p(x)|dx \leq \int_{A_\infty} p_0(x)dx + \int_{A_\infty} p(x)dx \leq \frac{3\epsilon_n}{8},$$

and this yields that,

$$\int_{\mathbb{R}^d \setminus A_\infty} |p_0(x) - p(x)|dx \geq \epsilon_n(1 - 3/8).$$

Now denoting by $\bar{p}$ the approximation of p by a density equal to the average of p on each cell of the partition we have that,

$$\begin{aligned}\int_{\mathbb{R}^d \setminus A_\infty} |p_0(x) - p(x)|dx \leq \int_{\mathbb{R}^d \setminus A_\infty} |p_0(x) - \bar{p_0}(x)|dx + \int_{\mathbb{R}^d \setminus A_\infty} |p(x) - \bar{p}(x)|dx \\ + \sum_{i=1}^N |p_0(A_i) - p(A_i)|dx.\end{aligned}$$

For any L_n-Hölder density we have that,

$$\int_{\mathbb{R}^d \setminus A_\infty} |p(x) - \bar{p}(x)|dx \leq L_n \sum_{i=1}^N [\text{diam}(A_i)]^s \operatorname{vol}(A_i) \leq \frac{\epsilon_n}{4},$$

using claim (E.2). This yields that,

$$\sum_{i=1}^N |p_0(A_i) - p(A_i)| + |p_0(A_\infty) - p(A_\infty)| \geq \sum_{i=1}^N |p_0(A_i) - p(A_i)| \geq \epsilon_n(1 - 7/8) = \epsilon_n/8,$$

as desired.

It remains to prove claim (E.2). Notice that,

$$\begin{aligned}L_n \sum_{i=1}^N [\text{diam}(A_i)]^s \operatorname{vol}(A_i) &\leq L_n \sum_{i=1}^N \min\{\theta_1 p_0(x_i), \theta_2 p_0^\gamma(x_i)\} \operatorname{vol}(A_i) \overset{\text{(i)}}{\leq} 2L_n \int_{\mathbb{R}^d} \min\{\theta_1 p_0(x), \theta_2 p_0^\gamma(x)\} dx \\ &= 2L_n \int_{\mathbb{R}^d} \min\left\{\frac{p_0(x)}{3L_n}, \frac{\epsilon_n p_0^\gamma(x)}{8L_n\mu(3/8)}\right\} dx \\ &\overset{\text{(ii)}}{=} \frac{\epsilon_n}{4},\end{aligned}$$

where step (i) uses property (4.15) and (ii) uses the definition of μ in (4.12).
Proof of Claim (4.18)**:** Recall that we have chosen $c = \epsilon_n/512$. In order to prove this claim we need to use properties of the pruning step. Let us

define $\widetilde{p}_0(x)$ as the piecewise constant function formed by replacing $p_0(x)$ by its maximum value over the cell containing x, and 0 outside the support of $\{A_1, \ldots, A_{\bar{N}}\}$. We note that,

$$\int_K p_0^\gamma(x)dx \leq \int_K \widetilde{p}_0^\gamma(x)dx. \tag{E.3}$$

Now, abusing notation slightly and ignoring the set A_∞ we denote the original partition as $\{A_1, \ldots, A_{\bar{N}}\}$, which we take as sorted by the values $p_0(A_i)$, and the pruned partition as $\{A_1, \ldots, A_N\}$, noting that we might potentially have split the last cell A_N into two cells. We let $A = \bigcup_{i=1}^{\bar{N}} A_i$. Let us denote,

$$\mathcal{B} = \left\{B : B \subset A, \int_{B^c} \widetilde{p}_0(x)dx \leq \int_{K^c} \widetilde{p}_0(x)dx\right\}.$$

Using Lemma E.2 we obtain that,

$$\int_K \widetilde{p}_0^\gamma(x)dx \leq \inf_{B\in\mathcal{B}} \int_B \widetilde{p}_0^\gamma(x)dx. \tag{E.4}$$

Noting, that

$$\int_{K^c} \widetilde{p}_0(x)dx \geq \int_{K^c} p_0(x)dx \geq \frac{c}{5},$$

and defining,

$$\mathcal{C} = \left\{C : C \subset A, \int_{C^c} \widetilde{p}_0(x)dx \leq \frac{c}{5}\right\}.$$

we obtain that,

$$\inf_{B\in\mathcal{B}} \int_B \widetilde{p}_0^\gamma(x)dx \leq \inf_{C\in\mathcal{C}} \int_C \widetilde{p}_0^\gamma(x)dx. \tag{E.5}$$

Defining,

$$\mathcal{D} = \left\{D : D \subset A, \int_{D^c} p_0(x)dx \leq \frac{c}{10}\right\},$$

we see that

$$\mathcal{D} \subset \mathcal{C} \subset \mathcal{B}$$

so that

$$\inf_{C\in\mathcal{C}} \int_C \widetilde{p}_0^\gamma(x)dx \leq 2^\gamma \inf_{D\in\mathcal{D}} \int_D p_0^\gamma(x)dx \leq 2^\gamma T_{c/10}^\gamma. \tag{E.6}$$

Putting together Equations (E.3), (E.4), (E.5) and (E.6) we obtain the desired result.

Proof of Claim (4.19): In order to lower bound the density over the pruned partition we will show that our pruning step is approximately a level set truncation. We have that for any point x that is removed and any point y that is retained it must be the case that,

$$p_0(x) \leq 2p_0(y).$$

where we used property (4.15). Let K denote the set of points retained by the pruning. The above observation yields that, there exists some $t \geq 0$ such that,

$$\{p_0 \geq t\} \subseteq K \subseteq \{p_0 \geq t/2\}.$$

We know that $\int_K p_0(x)dx \leq 1 - c/10$. Consider, the set

$$G(u) = \{x : p_0(x) \geq u\}.$$

Suppose that for some u we can show that,

$$\mathbb{P}(K) \leq \mathbb{P}(G(u)),$$

then we can conclude that $t \geq u$, and further that the density on K is at least $u/2$. It thus only remains to find a value u such that $\mathbb{P}(G(u)) \geq 1-c/10$. Suppose we choose $u = \left(\frac{c}{10\mu(c/(10\epsilon_n))}\right)^{1/(1-\gamma)}$, and recall that,

$$\epsilon_n = \int \min\left\{\frac{p_0(x)}{c/(10\epsilon_n)}, \frac{\epsilon_n p_0^{\gamma}(x)}{\mu(c/(10\epsilon_n))}\right\} dx.$$

Over the set G^c the minimizer is always the first term above which yields,

$$\epsilon_n \geq \int_{G^c} \frac{p_0(x)}{c/(10\epsilon_n)} dx,$$

i.e. that $\mathbb{P}(G^c) \leq c/10$, as desired. This in turn yields the claim.

E.4. Proof of Lemma 4.5. To show this, it suffices to show that more mass is truncated from q than is truncated from p, i.e. letting $\{t+1, \dots, N+1\}$ denote the $\epsilon_n/128$ tail of q we need to show that,

$$\sum_{t+1}^{N+1} q_i \geq P_0(A_\infty), \tag{E.7}$$

and then we apply Lemma E.3. To show (E.7) we proceed as follows. Note that $P_0(A_\infty) = q_a$ for some a. If $t \leq a$ then $\sum_{t+1}^{N+1} q_i \geq P_0(A_\infty)$ follows immediately. Now suppose that $t > a$. From the definition of t we know that $q_t + \sum_{t+1}^{N+1} q_i \geq \epsilon_n/128$ so that $\sum_{t+1}^{N+1} q_i \geq \epsilon_n/128 - q_t$. Since $t > a$, $q_t \leq P_0(A_\infty) \leq \epsilon_n/256$ and so $\sum_{t+1}^{N+1} q_i \geq \epsilon_n/256 \geq P_0(A_\infty)$ so that $\sum_{t+1}^{N+1} q_i \geq P_0(A_\infty)$ as required. Thus (E.7) holds.

APPENDIX F: ADAPTING TO UNKNOWN PARAMETERS

In this section, we consider ways to choose the parameter σ for the max test, and for the test in [4], and then consider tests that are adaptive to the typically unknown smoothness parameter L_n.

F.1. Choice of σ. The max test and the test from [4] require choosing the truncation parameter $\sigma = \epsilon_n/8$. In typical settings, we do not assume that ϵ_n is known. We consider the case of the test from [4] though our ideas generalize to the max test in a straightforward way.

Perhaps the most natural way to choose the parameter σ is to solve the critical equation and choose σ accordingly, i.e. we find $\widetilde{\sigma}$ that satisfies:

$$\widetilde{\sigma} = \max\left\{\frac{1}{n}, \sqrt{\frac{V_{\widetilde{\sigma}/16}(p_0)}{n}}\right\}, \tag{F.1}$$

and then we choose the tuning parameter $\sigma := C\max\{1/\alpha, 1/\zeta\}\widetilde{\sigma}$, for a sufficiently large constant $C \geq 1$.

When the unknown $\epsilon_n \geq 8C\max\{1/\alpha, 1/\zeta\}\widetilde{\sigma}$, then it is clear that our choice guarantees that the tuning parameter σ is chosen sufficiently small, i.e. $\sigma \leq \epsilon_n/8$ as desired. It is also clear that the test has size at most α. It remains to understand the Type II error. Inverting the above relationship we see that,

$$n = \max\left\{\frac{1}{\widetilde{\sigma}}, \frac{V_{\widetilde{\sigma}/16}(p_0)}{\widetilde{\sigma}^2}\right\}.$$

Noting that $\sigma/2 \geq \widetilde{\sigma}/16$, and that $C\max\{1/\alpha, 1/\zeta\} \geq 1$ we obtain that,

$$n \geq C\max\{1/\alpha, 1/\zeta\}\max\left\{\frac{1}{\sigma}, \frac{V_{\sigma/2}(p_0)}{\sigma^2}\right\} \geq C\max\{1/\alpha, 1/\zeta\}\max\left\{\frac{1}{\sigma}, \frac{V_{\sigma/2}(p_0)}{\epsilon_n^2}\right\}.$$

An application of Lemma B.1 shows that the Type II error of the test is at most ζ as desired. Thus, we see that this test provides the same result as the test in Theorem 3.3 without knowledge of ϵ_n.

Although adequate from a theoretical perspective, the previous choice of the parameter depends on an unknown (albeit universal) constant. An alternative is to consider a range of possible values for the parameter σ and appropriately adjust the threshold α via a Bonferroni correction. One natural range is to consider scalings of the parameter $\widetilde{\sigma}$ in (F.1). More generally, if we considered $\Sigma = \{\sigma_1, \ldots, \sigma_K\}$, a natural goal would be to compare the risk of the Bonferroni corrected test to the oracle test which minimizes the risk over the set Σ of possible tuning parameters. We leave a more detailed analysis of this test to future work.

F.2. Adapting to unknown L_n. Our tests for Hölder testing, in addition to assuming knowledge of ϵ_n use knowledge of the Hölder constant L_n in constructing the binning. Constructing tests which are adaptive to unknown smoothness parameters is a problem which has received much attention in classical works. The techniques from the previous section can be used to construct tests without knowledge of ϵ_n. We focus in this section on adapting to L_n but note that similar ideas are useful in constructing tests which are adaptive to the parameter s. We take $s = 1$ to simplify notation. We focus only on establishing upper bounds. Some lower bounds follow from standard arguments and we highlight important open questions in the sequel.

In order to define precisely the notion of an adaptive test, we follow the prescription of Spokoiny [3] (see also [1, 2]). As in (4.3) define a sequence of critical radii $w_n(p_0, L)$ as the solutions to the critical equations:

$$w_n(p_0, L) = \left(\frac{L^{d/2} T_{cw_n(p_0,L)}(p_0)}{n} \right)^{2/(4+d)}$$

for a sufficiently small constant $c > 0$. We now define the adaptive upper critical radii as the solutions to the critical equations:

$$w_n^a(p_0, L) = \left(\frac{L^{d/2} \log\log(n) T_{cw_n^a(p_0,L)}(p_0)}{n} \right)^{2/(4+d)}. \tag{F.2}$$

We can upper bound the ratio:

$$\frac{w_n^a(p_0, L)}{w_n(p_0, L)} \leq (\log\log(n))^{2/(4+d)}.$$

This ratio upper bounds the price for adaptivity. It will be necessary to distinguish the (known) smoothness parameter of the null from the possibly unknown parameter L_n in (2.5). We will denote the smoothness parameter of p_0 by L_0. We note that in the setting where L_n was known, we assumed that both $p_0, p \in \mathcal{L}(L_n)$ and this in turn requires that $L_n \geq L_0$.

We take $\alpha, \zeta > 0$ to be fixed constants. For a sufficiently large constant $C > 0$ (which depends on both $\alpha, \zeta > 0$) we define the class of densities:

$$\mathcal{L}(L_n, w_n^a) = \{p : p \in \mathcal{L}(L_n), \|p - p_0\|_1 \geq C w_n^a(p_0, L_n)\}.$$

For some $p_0 \in \mathcal{L}(L_0)$, consider the hypothesis testing problem of distinguishing:

$$H_0 : p = p_0, p_0 \in \mathcal{L}(L_0) \quad \text{versus} \quad H_1 : p \in \bigcup_{L_n \geq L_0} \mathcal{L}(L_n, w_n^a). \tag{F.3}$$

In order to precisely define our testing procedure we first show that there are natural upper bounds on L_n. In particular, we claim that when $L_n \gg n^{2/d} L_0$ then the critical radius remains lower bounded by a constant.

We have the following lemma. We let $C_\ell, c > 0$ denote universal constants.

LEMMA F.1. *If $L_n \geq C_\ell n^{2/d} L_0$, then*

$$\epsilon_n(p_0, L_n) \geq c.$$

Thus we restrict our attention to the regime where $L_n \in [L_0, C n^{2/d} L_0]$, for a sufficiently large constant $C > 0$. A natural strategy is then to consider a discretization of the set of possible values for L_n,

$$\mathsf{L} = \{L_0, 2L_0, \ldots, 2^{\log_2(C n^{2/d})} L_0\}.$$

The multinomial tests we build on (in Theorem 3.3) have critical radii that scale with $\max\{1/\alpha, 1/\zeta\}$ in order to control the Type I and Type II error at α and ζ, respectively. It is possible to improve the dependence on these parameters via a simple sample-splitting scheme. In more detail, to control the Type I and Type II errors at α and ζ we split the sample into roughly $t = \log \max\{1/\alpha, 1/\zeta\}$ groups of equal size, and run the multinomial test with parameters $\widetilde{\alpha}$ and $\widetilde{\zeta}$, each equal to $1/4$ say, on each of the groups. Now, the overall test rejects the null hypothesis if more than $1/2$ of the t group tests reject the null hypothesis. Using a standard Hoeffding bound it is straightforward to verify that this overall test, controls the Type I error at α and Type II error at ζ as desired.

Our adaptive test then simply performs the sample-split version binning test described in Theorem 4.1 for each choice of $L_n \in \mathsf{L}$, with the threshold α reduced by a factor of $\lceil \log_2(C n^{2/d}) + 1 \rceil$. We refer to this test as the adaptive Hölder test. We have the following result:

THEOREM F.1. *Consider the testing problem in* (F.3)*. The adaptive Hölder test has Type I error at most α, and has Type II error at most ζ.*

Remarks:

- Comparing the non-adaptive critical radii in (4.3) and the adaptive critical radii in (F.2) we see that we lose a factor of $(\log\log(n))^{2/(4+d)}$. A natural question is whether such a loss is necessary.
- Classical results [2] consider adapting to an unknown Hölder exponent s and show that for testing uniformity (with deviations in the ℓ_2 metric) a loss of a factor $(\sqrt{\log\log(n)})^{2s/(4s+d)}$ is necessary and sufficient. In our setting, the loss is of a $\log\log$ factor instead of a $\sqrt{\log\log}$ factor and this is a consequence of using sample-splitting to reduce the Type I and Type II errors of our test. We hope to develop a more precise understanding of this situation in future work.

PROOF. The proof follows almost directly from our previous analysis of Theorem 4.1 so we only provide a brief sketch. It is straightforward to check that the Bonferroni correction controls the size of the adaptive Hölder test at α. Let j^* denote the smallest integer such that, $2^{j^*} L_0 \geq L_n$. In order to bound the Type II error, it is sufficient to show that under the alternate, the test corresponding to the index j^* rejects the null hypothesis with probability at least $1 - \zeta$. Noting that the ratio $2^{j^*} L_0 / L_n \leq 2$ this follows directly from the proof of Theorem 4.1. $\square$

F.2.1. *Proof of Lemma F.1.* In order to establish this claim, it suffices to show that the lower bound on the critical radius in (4.3) is at least a constant. By the monotonicity of the critical equation, it suffices to show that for some small constant $c > 0$ we have that,

$$c \leq \left(\frac{L_n^{d/2} T_{Cc}(p_0)}{n} \right)^{2/(4+d)},$$

where $C > 0$ is the universal constant in (4.3). We choose $c < 1/(2C)$ so we obtain that it suffices to show,

$$c \leq \left(\frac{L_n^{d/2} T_{1/2}(p_0)}{n} \right)^{2/(4+d)}.$$

We claim that for any $p_0 \in \mathcal{L}(L_0)$ there is a universal constant $C_1 > 0$ such that,

$$T_{1/2}(p_0) \geq \frac{C_1}{L_0^{d/2}}. \tag{F.4}$$

Taking this claim as given for now we see that,

$$\left(\frac{L_n^{d/2} T_{1/2}(p_0)}{n}\right)^{2/(4+d)} \geq \left(\frac{C_1 L_n^{d/2}}{L_0^{d/2} n}\right)^{2/(4+d)} \geq \left(\frac{C_1}{C_\ell^{d/2}}\right)^{2/(4+d)} \geq c,$$

as desired.

Proof of Claim (F.4): As a preliminary we first produce an upper bound on any Hölder density. We claim that, there exists a constant $C > 0$ depending only on the dimension such that any L_0-Hölder density p_0 is upper bounded as $\|p_0\|_\infty \leq C L_0^{d/(d+1)}$.

Without loss of generality let us suppose the density p_0 is maximized at $x = 0$. The density p_0 is then lower bounded by the function,

$$g_0(x) = (\|p_0\|_\infty - L_0\|x\|_2)\,\mathbb{I}(\|p_0\|_\infty - L_0\|x\|_2 \geq 0).$$

The integral of this function is straightforward to compute, and since p_0 must integrate to 1 we obtain that,

$$1 = \int_x p_0(x)dx \geq \int_x g_0(x)dx = \frac{v_d}{d+1}\frac{\|p\|_\infty^{d+1}}{L_0^d},$$

where v_d denotes the volume of the d-dimensional unit ball. This in turn yields the upper bound,

$$\|p\|_\infty \leq \left(\frac{d+1}{v_d}\right)^{1/(d+1)} L_0^{d/(d+1)},$$

as desired. With this result in place we can lower bound the truncated T-functional. In particular, letting B_σ denote a set of probability content $1-\sigma$ that (nearly) minimizes the truncated T-functional we have that,

$$T_\sigma^\gamma(p_0) = \int_{B_\sigma} p_0^\gamma(x)dx \geq \int_{B_\sigma} \frac{p_0(x)}{\|p\|_\infty^{1-\gamma}}dx \geq \left(\frac{v_d}{d+1}\right)^{1/(3+d)} \frac{1-\sigma}{L_0^{d/(3+d)}},$$

which gives the bound,

$$T_\sigma(p_0) \geq \left(\frac{v_d}{d+1}\right)^{1/2} \frac{(1-\sigma)^{(3+d)/2}}{L_0^{d/2}},$$

as desired. Taking $\sigma = 1/2$ yields the desired claim.

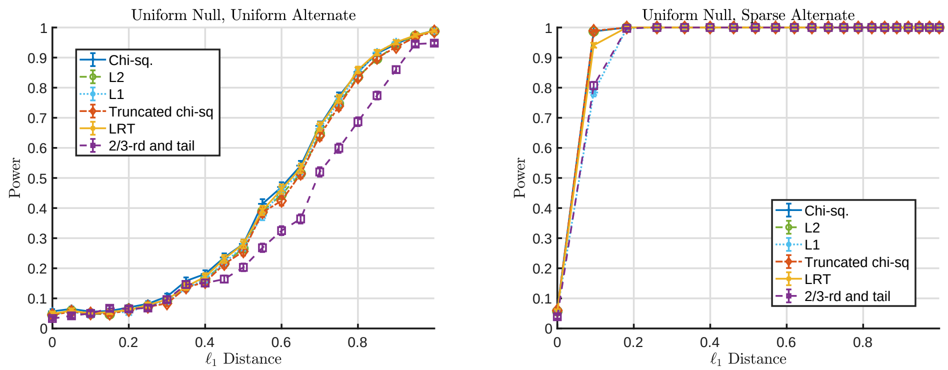

Fig 5: A comparison between the truncated χ^2 test, the 2/3-rd + tail test [4], the χ^2-test, the likelihood ratio test, the ℓ_1 test and the ℓ_2 test. The null is chosen to be uniform, and the alternate is either a dense or sparse perturbation of the null. The power of the tests are plotted against the ℓ_1 distance between the null and alternate. Each point in the graph is an average over 1000 trials. Despite the high-dimensionality (i.e. $n = 300, d = 2000$) the tests have high-power, and perform comparably.

APPENDIX G: ADDITIONAL SIMULATIONS

In this section we re-visit the simulations for multinomials. In addition to the alternatives that are created by dense and sparse perturbations of the null we also consider two other perturbations: one where we perturb each coordinate of the null by an amount proportional to the entry $p_0(i)$, and one where we perturb each coordinate by $p_0(i)^{2/3}$, in magnitude with a Rademacher sign. The latter perturbation is close to the worst-case perturbation considered by [4] in their proof of local minimax lower bounds. We take $n = 300, d = 2000$ and each point in the graph is an average over 1000 trials.

Once again we observe that the truncated χ^2 test we propose, and the 2/3-rd + tail test from [4] are remarkably robust. All tests are comparable when the null is uniform, while distinctions are clearer for the power law null. The ℓ_2 test appears to have high-power against sparse alternatives suggesting potential avenues for future investigation.

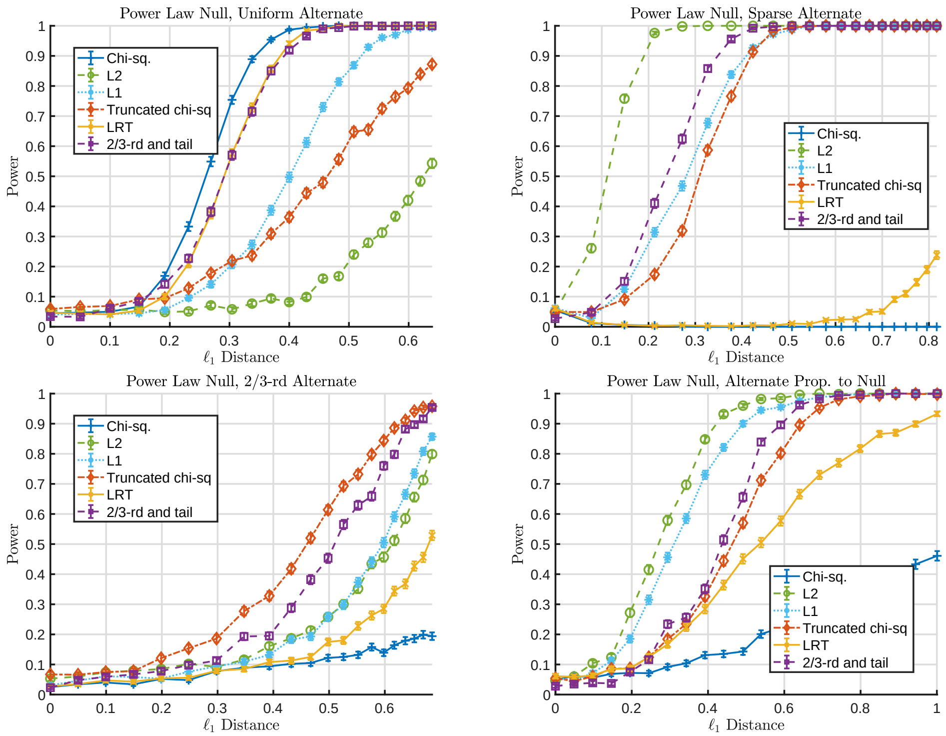

Fig 6: A comparison between the truncated χ^2 test, the 2/3-rd + tail test [4], the χ^2-test, the likelihood ratio test, the ℓ_1 test and the ℓ_2 test. The null is chosen to be a power law with $p_0(i) \propto 1/i$. The alternatives are uniform, sparse (only perturbing the first two coordinates), perturbing each co-ordinate proportional to $p_0(i)^{2/3}$ and perturbing each coordinate proportional to $p_0(i)$.

REFERENCES

[1] Giné, E. and Nickl, R. (2015). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press.

[2] Ingster, Y. I. (1997). Adaptive chi-square tests. *Zapiski Nauchnykh Seminarov POMI* **244** 150–166.

[3] Spokoiny, V. G. (1996). Adaptive hypothesis testing using wavelets. *Ann. Statist.* **24** 2477–2498.

[4] Valiant, G. and Valiant, P. (2014). An Automatic Inequality Prover and Instance Optimal Identity Testing. *2014 IEEE 55th Annual Symposium on Foundations of Computer Science (FOCS)* 51-60.

Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213.
E-mail: siva@stat.cmu.edu
larry@stat.cmu.edu