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:

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

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 $Z_1, \ldots, Z_n \in \mathcal{X}$, 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:

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

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:

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

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 $Z_1, \ldots, Z_n \sim P$ define the counts $X = (X_1, \ldots, X_d)$ 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

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

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 $(X_1, \ldots, X_d)$ 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.* $P_0(\phi_{\text{trunc}} = 1) \le \alpha$. *Furthermore, there is a universal constant* $C > 0$ *such that if for any* $0 < \zeta \le 1$ *we have that*

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

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

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* 2/3*rd* + *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 2/3rd + 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

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

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

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.

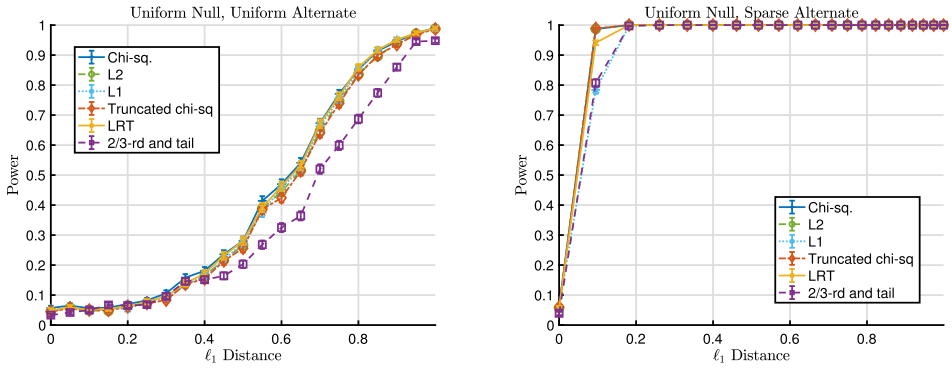

FIG. 1. *A comparison between the truncated χ^2 test, the 2/3rd + tail test* [37], *the χ^2-test and the likelihood ratio test. The null is chosen to be uniform, and the alternate is either a dense or sparse perturbation of the null. The power of the tests are plotted against the ℓ_1 distance between 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-

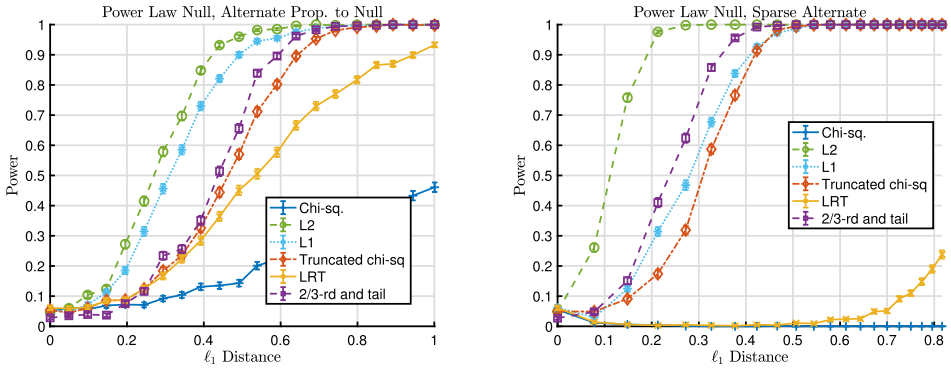

FIG. 2. *A comparison between the truncated χ^2 test, the 2/3rd + tail test* [37], *the χ^2-test and the likelihood ratio test. The null is chosen to be a power law, and the alternate is either a dense or sparse perturbation of the null. The power of the tests are plotted against the ℓ_1 distance between 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 $0 \leq c_{int} < 1$ *we have that,* $p_0 \in \mathcal{L}_s(c_{int}L_n)$*, then the critical radius is lower bounded as*

$$cv_n(p_0) \leq \varepsilon_n\big(p_0, \mathcal{L}_s(L_n)\big). \tag{4.5}$$

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 $\sigma > 0$,

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

  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 $t \geq 1$,

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

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

$$p_0(x) = \begin{cases} L_n x, & 0 \le x \le \dfrac{1}{\sqrt{L_n}}, \\ 2\sqrt{L_n} - L_n x, & \dfrac{1}{\sqrt{L_n}} \le x \le \dfrac{2}{\sqrt{L_n}}, \\ 0, & \text{otherwise}, \end{cases}$$

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

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

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 $i \in \{1, \ldots, \widetilde{N}\}$, 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

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

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

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

   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 $\mathcal{P} = \{A_1, \ldots, A_{\widetilde{N}}, A_\infty\}$.

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

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

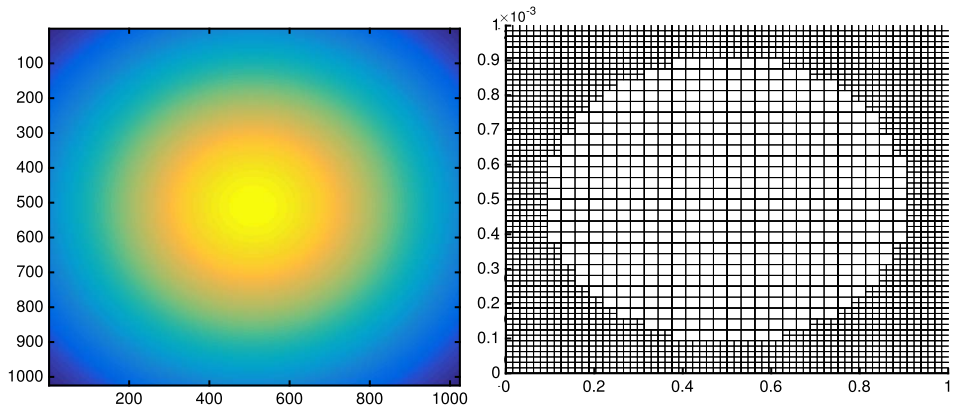

FIG. 3. (a) *A density* p_0 *on* $[0,1]^2$ *evaluated on a* 1000×1000 *grid.* (b) *The corresponding spatially adaptive partition* $\mathcal{P}$ *produced by Algorithm* 1. *Cells of the partition are larger in regions where the density* p_0 *is higher.*

Algorithm 2 Prune Partition

1. *Input*: Unpruned partition $\mathcal{P} = \{A_1, \ldots, A_{\widetilde{N}}, A_\infty\}$ 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 $j \in \{1, \ldots, \widetilde{N}\}$, 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*: $\mathcal{P}^\dagger = \{A_1, \ldots, A_N, A_\infty\}$.

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*

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

2. [*Multiplicative control.*] *The density is multiplicatively controlled on each cell, that is, for* $i \in \{1, \ldots, N\}$ *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 $P_1, \ldots, P_N$ then

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

Given n observations, we define the likelihood ratio

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

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

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

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

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

where $\theta_1 = 1/(3L_n)$ and we let $\theta_2 > 0$ be arbitrary. Let $\eta = (\eta_1, \eta_2, \ldots, \eta_N)$ 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 $j \in \{1, \ldots, N\}$, 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 $\mathcal{P} = \{A_1, \ldots, A_N, A_\infty\}$. 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 $\{\rho_1, \ldots, \rho_N\}$, 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 $j \in \{1, \ldots, N\}$,

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

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

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

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

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*

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

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

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

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

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

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

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

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

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

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

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

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

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

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.

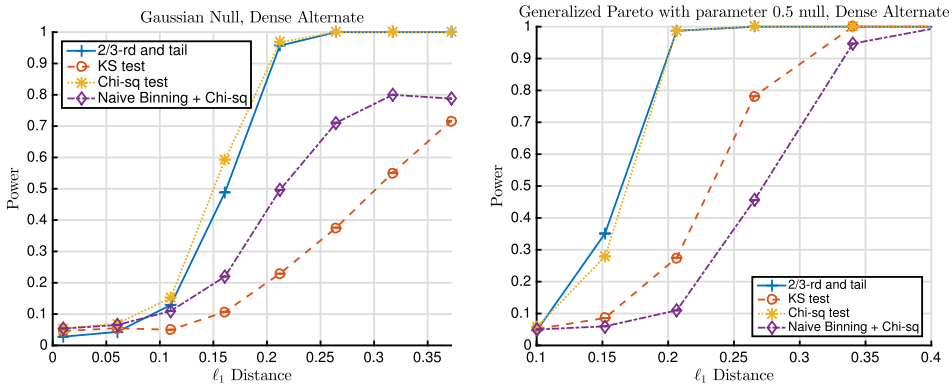

FIG. 4. *A comparison between the KS test, multinomial tests on an adaptive binning and multinomial tests on a fixed bin-width binning. In the figure on the left, we choose the null to be standard Gaussian and on the right we choose the null to be Pareto. The alternate is chosen to be a dense near worst-case, 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 55*th 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