The Annals of Applied Statistics
2018, Vol. 12, No. 2, 727–749
https://doi.org/10.1214/18-AOAS1155SF

HYPOTHESIS TESTING FOR HIGH-DIMENSIONAL MULTINOMIALS: A SELECTIVE REVIEW[1]

BY SIVARAMAN BALAKRISHNAN AND LARRY WASSERMAN

Carnegie Mellon University

In memory of Stephen E. Fienberg

The statistical analysis of discrete data has been the subject of extensive statistical research dating back to the work of Pearson. In this survey we review some recently developed methods for testing hypotheses about high-dimensional multinomials. Traditional tests like the χ^2-test and the likelihood ratio test can have poor power in the high-dimensional setting. Much of the research in this area has focused on finding tests with asymptotically normal limits and developing (stringent) conditions under which tests have normal limits. We argue that this perspective suffers from a significant deficiency: it can exclude many high-dimensional cases when—despite having non-normal null distributions—carefully designed tests can have high power. Finally, we illustrate that taking a minimax perspective and considering refinements of this perspective can lead naturally to powerful and practical tests.

1. Introduction.

Steve Fienberg was a pioneer in the development of theory and methods for discrete data. His textbook [Bishop, Fienberg and Holland (1977)] remains one of the main references for the topic. Our focus in this review is on high-dimensional multinomial models where the number of categories d can grow with, and possibly exceed the sample-size n. Steve's paper [Fienberg and Holland (1973)], written with Paul Holland, was one of the first to consider multinomial data in the high-dimensional case. In Fienberg (1980), Steve provided strong motivation for considering the high-dimensional setting:

> "The fact remains … that with the extensive questionnaires of modern-day sample-surveys, and the detailed and painstaking inventory of variables measured by biological and social-scientists, the statistician is often faced with large sparse arrays full of 0's and 1's, in need of careful analysis."

In this review we focus on hypothesis testing for high-dimensional multinomials. In the context of hypothesis testing, several works [see for instance Read and Cressie (1988), Holst (1972) and references therein] have considered the high-dimensional setting. Hoeffding (1965) (building on an unpublished result of Stein) showed that for testing goodness-of-fit, in sharp contrast to the fixed-d setting, in

Received December 2017; revised February 2018.

[1] This research was supported by NSF Grant DMS-1713003.

Key words and phrases. Hypothesis testing, high-dimensional multinomials.

the high-dimensional setting the likelihood ratio test can be dominated by the χ^2 test. In traditional asymptotic testing theory, the power of tests are often investigated at local alternatives which approach the null as the sample-size grows. In the high-dimensional setting, considering local alternatives, Ivčenko and Medvedev (1980) showed that neither the χ^2 nor the likelihood ratio test are uniformly optimal. These results show some of the difficulties of using classical theory to identify optimal tests in the high-dimensional regime.

Morris (1975) studied the limiting distribution of a wide-range of multinomial test statistics and gave relatively stringent conditions under which these statistics have asymptotically normal limiting distributions. Related investigations appear in the works [Koehler and Larntz (1980), Haberman (1977), Koehler (1986)] and are reviewed in the work of Fienberg (1979). In general, as we illustrate in our simulations, carefully designed tests can have high power under much weaker conditions, even when the null distribution of the test statistics are not Gaussian or χ^2. In many cases, understanding the limiting distribution of the test statistic under the null is important to properly set the test threshold, and indeed this often leads to practical tests. However, in several problems of interest, including goodness-of-fit, two-sample testing and independence testing we can determine practical (nonconservative) thresholds by simulation. In the high-dimensional setting, rather than rely on asymptotic theory and local alternatives, we advocate for using the minimax perspective and developing refinements of this perspective to identify and study optimal tests.

We emphasize that the minimax perspective on testing is not new to the discipline of Statistics. In particular, the work of Ermakov, Ingster, Lepski, Spokoiny, Suslina and co-authors [for instance, Ermakov (1991), Ingster and Suslina (2003), Spokoiny (1996), Lepski and Spokoiny (1999)] laid the foundations of the minimax framework for testing. Due to their work minimax rates, for signal detection in the Gaussian white noise model and more broadly for testing with nonparametric alternatives are relatively well understood [Giné and Nickl (2016)]. More recently, work on minimax hypothesis testing has focused on testing problems with a combinatorial flavor [Addario-Berry et al. (2010), Arias-Castro, Candès and Durand (2011)], and on testing problems in high-dimensional statistics with sparsity constraints [Berthet and Rigollet (2013), Ingster, Tsybakov and Verzelen (2010), Donoho and Jin (2004)].

For high-dimensional *discrete* problems a minimax perspective has been recently developed in a series of works in different fields including statistics, information theory and theoretical computer science and we provide an overview of some important results from these different communities in this paper.[2]

[2] Theoretical computer science centric surveys of a subset of these results include the papers Rubinfeld (2012), Goldreich (2017) and Canonne (2018).

2. Background. Suppose that we have data of the form $Z_1, \dots, Z_n \sim P$ where P is a d-dimensional multinomial and $Z_i \in \{1, \dots, d\}$. We denote the probability mass function for P by $p \in \mathbb{R}^d$. The set of all multinomials is then denoted by

$$\mathcal{M} = \left\{ p = \bigl(p(1), \dots, p(d)\bigr) : p(j) \geq 0 \text{ for all } j, \sum_{j=1}^{d} p(j) = 1 \right\}. \tag{2.1}$$

We are interested in the case where d can be large, possibly much larger than n. In this paper, we focus on two hypothesis testing problems, and defer a discussion of other testing problems to Section 6. The problems we consider are:

2. *Goodness-of-fit testing*: In its most basic form, in goodness-of-fit testing we are interested in testing the fit of the data to a fixed distribution P_0. Concretely, we are interested in distinguishing the hypotheses

$$H_0 : P = P_0 \quad \text{versus} \quad H_1 : P \neq P_0.$$

(i) *Two-sample testing*: In two-sample testing we observe

$$Z_1, \dots, Z_{n_1} \sim P, \qquad W_1, \dots, W_{n_2} \sim Q.$$

In this case, the hypotheses of interest are

$$H_0 : P = Q \quad \text{versus} \quad H_1 : P \neq Q.$$

Notation: Throughout this paper, we write $a_n \asymp b_n$ if a_n/b_n is bounded away from both 0 and ∞ for all large n. In certain cases, to improve readability we make slightly imprecise mathematical statements and indicate this with the $\approx$ relational symbol. In each such case, for the interested reader, we provide a reference to the precise statement.

2.1. *Why hypothesis testing?*

As with any statistical problem, there are many inferential tasks related to multinomial models: estimation, constructing confidence sets, Bayesian inference, prediction and hypothesis testing, among others.

Our focus on testing in this paper is not meant to downplay the importance of these other tasks. Indeed, many would argue that hypothesis testing has received too much attention: over-reliance on hypothesis testing is sometimes cited as one of the causes of the reproducibility crisis. However, there is a good reason for studying hypothesis testing. When trying to understand the theoretical behavior of statistical problems in difficult cases—such as in high dimensional models—hypothesis testing provides a very clean, precise framework.

Hypothesis testing is a good starting point for theoretical investigations into difficult statistical models. As an example, in Section 3.2 we will see that the power of goodness-of-fit tests can vary drastically depending on where the null sits in the simplex, a phenomenon that we refer to as *local minimaxity*. This local minimax phenomenon is very clear and easy to precisely capture in the hypothesis testing framework.

2.2. *Minimax and local minimax testing.* In traditional asymptotic testing theory, local measures of performance are often used to assess the performance of various hypothesis tests. In the local approach, one typically examines the power at a sequence of alternatives at $\theta_0 + C/\sqrt{n}$ where θ_0 denotes the null value of the parameter. The local approach is well-suited to well-behaved, low-dimensional models. It permits very precise power comparisons for distributions which are close to the null hypothesis. Generally, the tools for local analysis are tied to ideas like contiguity and asymptotic normality and in the high-dimensional setting these tools often break down. Furthermore, as we discussed earlier, results of Ivčenko and Medvedev (1980) suggest that the local perspective does not provide a clear overall picture in the high-dimensional case.

For these reasons, we will use the minimax perspective. For goodness-of-fit testing, we can refine the minimax perspective to obtain results that also have a local nature, but in a very different sense than the local results described above. Formally, a test is a map from the samples to $\{0, 1\}$. We let Φ_n denote all level α tests, that is, $\phi \in \Phi_n$ if $\phi(Z_1, \dots, Z_n) \in \{0, 1\}$ and

$$\sup_{P \in \mathcal{P}_0} P^n(\phi = 1) \le \alpha,$$

where $\mathcal{P}_0$ denotes the (possibly composite) collection of possible null distributions. Let

$$\mathcal{M}_\varepsilon = \{p : d(\mathcal{P}_0, p) \ge \varepsilon\},$$

where $\mathcal{P}_0$ is the set of null distributions, $d(\mathcal{P}_0, p) = \inf_{q \in \mathcal{P}_0} d(q, p)$ and $d(p, q)$ is some distance. The maximum type II error of a test $\phi \in \Phi_n$ is

$$R_{n,\varepsilon}(\phi, \mathcal{P}_0) = \sup_{P \in \mathcal{M}_\varepsilon} P^n(\phi = 0).$$

The *minimax risk* is

$$R^\dagger_{n,\varepsilon}(\mathcal{P}_0) = \inf_{\phi \in \Phi_n} R_{n,\varepsilon}(\phi, \mathcal{P}_0).$$

A test $\phi \in \Phi_n$ is minimax optimal if $R_{n,\varepsilon}(\phi) = R^\dagger_{n,\varepsilon}(\mathcal{P}_0)$. It is common to study the minimax risk via a coarse lens by studying instead the *minimax separation*, also called the *critical radius*. The minimax separation ε_n is defined by

$$\varepsilon_n(\mathcal{P}_0) = \inf\{\varepsilon : R^\dagger_{n,\varepsilon}(\mathcal{P}_0) \le 1/2\},$$

which is the smallest ε such that the power is nontrivial. The choice of $1/2$ is not important and any sufficiently small, nonzero number will suffice.

We need to choose a distance d. We will focus on the total variation (TV) distance defined by

$$\mathrm{TV}(P, Q) = \max_A |P(A) - Q(A)|,$$

where the maximum is over all events A. The reason we use total variation distance is because it has a clear probabilistic meaning and is invariant to natural transformations [Devroye and Györfi (1985)]: if $\mathrm{TV}(P, Q) = \varepsilon$ then $|P(A) - Q(A)| \leq \varepsilon$ for every event A. The total variation distance is equivalent to the L_1 distance

$$\mathrm{TV}(P, Q) = \frac{1}{2}\sum_{j=1}^{d} |p(j) - q(j)| \equiv \frac{1}{2}\|p - q\|_1,$$

where p and q are the probability functions corresponding to the distributions P and Q. Other distances, such as L_2, Hellinger and Kullback–Leibler can be used too, but in some cases can be less interpretable and in other cases lead to trivial minimax rates. We revisit the choice of metric in Section 3.4.

Typically we characterize the minimax separation by providing upper and lower bounds on it: upper bounds are obtained by analyzing the minimax separation for practical tests, while lower bounds are often obtained via an analysis of the likelihood ratio test for carefully constructed pairs of hypotheses [see for instance the pioneering work of Ingster and co-authors Ingster and Suslina (2003), Ingster (1997)].

The local minimax separation: Focusing on the problem of goodness-of-fit testing with a simple null, we observe that the minimax separation $\varepsilon_n(p_0)$ is a function of the null distribution. Classical work in hypothesis testing [Ingster and Suslina (2003), Ingster (1997)] has focused on characterizing the global minimax separation, that is, in understanding the quantity

$$\varepsilon_n = \inf\Big\{\varepsilon : \sup_{p_0 \in \mathcal{M}} R^{\dagger}_{n,\varepsilon}(p_0) \leq 1/2\Big\}, \tag{2.2}$$

where $\mathcal{M}$ is defined in (2.1). 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.

Recent work by Valiant and Valiant (2017) showed that for goodness-of-fit testing in the TV metric for high-dimensional multinomials, the critical radius can vary considerably as a function of the null distribution p_0. In this case, the local minimax separation, that is, $\varepsilon_n(p_0)$ provides a much more refined notion of the difficulty of the goodness-of-fit problem. Valiant and Valiant (2017) further provided a locally minimax test, that is, a test that is nearly-optimal for *every possible null distribution*.

Developing refinements to the minimax framework in problems beyond goodness-of-fit testing is an active area of research. For the problem of two-sample testing for multinomials, a recent proposal appears in the work of Acharya et al. (2012). Other works developing a local perspective in testing and estimation include Donoho and Johnstone (1994), Goldenshluger and Lepski (2011), Cai and Low (2015), Chatterjee, Guntuboyina and Sen (2015), Wei and Wainwright (2017).

3. Goodness-of-fit. Let $Z_1, \dots, Z_n \sim P$ where $Z_i \in \{1, \dots, d\}$. Define the vector of counts $(X_1, \dots, X_d)$ where $X_j = \sum_{i=1}^n \mathbb{I}(Z_i = j)$. We consider testing the simple null hypothesis:

$$H_0 : P = P_0 \quad \text{versus} \quad H_1 : P \neq P_0.$$

The most commonly used test statistics are the chi-squared statistic,

$$T_{\chi^2} = \sum_{j=1}^d \frac{(X_j - np_0(j))^2}{np_0(j)}, \tag{3.1}$$

and the likelihood ratio test (LRT) statistic,

$$T_{\mathrm{LRT}} = \sum_{j=1}^d \widehat{p}(j) \log\left(\frac{\widehat{p}(j)}{p_0(j)}\right), \tag{3.2}$$

where $\widehat{p}(j) = X_j/n$. See the book of Read and Cressie (1988) for a variety of other popular multinomial goodness-of-fit tests. As we show in Section 5, when d is large, these tests can have poor power. In particular, they are not minimax optimal. Much of the research on tests in the large d setting has focused on establishing conditions under which these statistics have convenient limiting distributions [see for instance Morris (1975)]. Unfortunately, these conditions are generally not testable, and they rule out many interesting cases, when test statistics despite not having a convenient limiting distribution can have high power (see Section 5).

3.1. *Globally minimax optimal tests.* The global minimax separation rate was characterized in the works of Paninski (2008), Valiant and Valiant (2017). In particular, these works show that the global minimax separation rate [see (2.2)] is given by:

$$\epsilon_n \asymp \frac{d^{1/4}}{\sqrt{n}}. \tag{3.3}$$

This implies, surprisingly, that we can have non-negligible power even when $n \ll d$. In the regime when $n \asymp \sqrt{d}$ most categories of the multinomial are unobserved, but we can still distinguish *any* multinomial from alternatives separated in ℓ_1 with high power. In stark contrast, it can be shown that the minimax estimation rate in the ℓ_1 distance, is $\sqrt{d/n}$ which is much slower than the testing rate. This is a common phenomenon: hypothesis testing is often easier than estimation.

An important fact, elucidated by the minimax perspective, is that none of the traditional tests are minimax. A very simple minimax test, from Balakrishnan and Wasserman (2017) is the truncated χ^2 test defined by

$$T_{\mathrm{trunc}} = \sum_j \frac{(X_j - np_0(j))^2 - X_j}{\max\{p_0(j), \frac{1}{d}\}}. \tag{3.4}$$

The α level critical value t_α is defined by

$$t_\alpha(p_0) = \inf\{t : P_0^n(T > t) \leq \alpha\}.$$

Normal approximations or χ^2 approximations cannot be used to find t_α since the asymptotics are not uniformly valid over $\mathcal{M}$. However, the critical value t_α for the test can easily be found by simulating from P_0. In Section 5 we report some simulation studies that illustrate the gain in power by using this test.

3.2. *Locally minimax optimal tests.* In goodness-of-fit testing, some nulls are easier to test than others. For example, when p_0 is uniform the local minimax risk is quite large and scales as in (3.3). However, when p_0 is sparse the problem effectively behaves as a much lower dimensional multinomial testing problem and the minimax separation can be much smaller. This observation has important practical consequences. Substantial gains in power can be achieved, by adapting the test to the shape of the null distribution p_0.

Roughly, Valiant and Valiant (2017) showed that the local minimax rate is given by

$$\varepsilon_n(p_0) \approx \sqrt{\frac{V(p_0)}{n}}$$

for a functional V that depends on p_0 as

$$V(p_0) \approx \|p_0\|_{2/3} = \left(\sum_{j=1}^{d} p_0^{2/3}(j)\right)^{3/2}.$$

We provide a more precise statement in the Appendix. The fact that the local minimax rate depends on the 2/3rd norm[3] is certainly not intuitive and is an example of the surprising nature of the results in the world of high-dimensional multinomials. When p_0 is uniform the 2/3rd norm is maximal and takes the value $\|p_0\|_{2/3} = \sqrt{d}$ whereas when $p_0 = (1, 0, \dots, 0)$ the 2/3rd norm is much smaller, that is, $\|p_0\|_{2/3} = 1$. This means that a test tailored to p_0 can have dramatic gains in power.

Valiant and Valiant (2017) constructed a test that achieves the local minimax bound. To describe the test we need a few definitions. First, without loss of generality, assume that $p_0(1) \geq p_0(2) \geq \cdots \geq p_0(d)$. Let $\sigma \in [0, 1]$ and define the tail and bulk by

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

[3] We use this terminology for convenience despite the fact that ℓ_p "norms" do not satisfy the triangle inequality for $0 < p < 1$.

and

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

The test is $\phi = \phi_1 \vee \phi_2$ where $\phi_1 = I(T_1(\sigma) > t_1)$, $\phi_2 = I(T_2(\sigma) > t_2)$,

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

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

The test may appear to be somewhat complicated but all the quantities are easy to compute. Furthermore, in practice the thresholds are easily computed by simulation. Other near-local minimax tests for testing multinomials appear in Diakonikolas and Kane (2016), Balakrishnan and Wasserman (2017).

A problem with the above test is that there is a tuning parameter σ. Valiant and Valiant (2017) suggested using $\sigma = \varepsilon/8$. While this choice is useful for theoretical analysis it is not useful in practice as it would require knowing how far P is from P_0 if H_0 is false. In Balakrishnan and Wasserman (2017) we propose ways to select the tuning parameter σ in a data-driven fashion. For instance, one might consider a Bonferroni corrected test. More precisely, let $\Sigma = \{\sigma_1, \dots, \sigma_N\}$ be a grid of values for σ. Let ϕ_j be the test using tuning parameter σ_j and significance level α/N. We then use the Bonferroni corrected test $\phi = \max_j\{\phi_j\}$. It can be shown that, if Σ is chosen carefully, there is only a small loss of power from the Bonferroni correction.

3.3. *Implications for continuous data.* Although the focus of this paper is on discrete data, we would like to briefly mention the fact that these results have implications for continuous data. Our discussion is based on Balakrishnan and Wasserman (2017). We also note the works of Diakonikolas, Kane and Nikishkin (2015a, 2015b, 2017) which provide minimax optimal tests for various univariate continuous testing problems with shape constraints.

Suppose that $X_1, \dots, X_n \sim p$ where p is a density on $[0, 1]$. We want to test $H_0 : p = p_0$. As shown in LeCam (1973) and Barron (1989), the power of any test over the set $\{p : \mathrm{TV}(p, p_0) > \varepsilon\}$ is trivial unless we add further assumptions. For example, suppose we restrict attention to densities p that satisfy the Lipschitz constraint

$$|p(y) - p(x)| \le L|x - y|.$$

In this case, Ingster (1997) showed that, when p_0 is the uniform density, the minimax separation rate is $\varepsilon_n \asymp n^{-2/5}$. The optimal rate can be achieved by binning the data and using a χ^2 test. However, if p_0 is not uniform and the Lipschitz constant

L is allowed to grow with n, Balakrishnan and Wasserman (2017) showed that the local minimax rate is

$$\varepsilon_n(p_0) \asymp \left(\frac{\sqrt{L_n}T(p_0)}{n}\right)^{2/5},$$

where $T(p_0) \approx \int \sqrt{p_0(x)}\,dx$ [a more precise statement of this result can be found in our paper Balakrishnan and Wasserman (2017)]. This rate can be achieved by using a very careful, adaptive binning procedure, and then invoking the test from Section 3.2. The proofs make use of some of the tools for high dimensional multinomials described in the previous sections.

The main point here is that the theory for multinomials has implications for continuous data. It is worth noting that Fienberg and Holland (1973) was one of the first papers to explicitly link the high-dimensional multinomial problem to continuous problems.

3.4. *Testing in other metrics.* A natural question is to characterize the dependence of the local minimax separation and the local minimax test on the choice of metric. We have focused thus far on the TV metric. The paper Daskalakis, Kamath and Wright (2018) [see also Diakonikolas and Kane (2016)] considers goodness-of-fit testing in other metrics. They provide results on the global minimax separation for the Hellinger, Kullback–Leibler and χ^2 metric. In particular, they show that while the global minimax separation is identical for Hellinger and TV, this separation is infinite for the Kullback–Leibler and χ^2 distance because these distances can be extremely sensitive to small perturbations of small entries of the multinomial.

In forthcoming work [Balakrishnan and Wasserman (2018)] we characterize the *local* minimax rate in the Hellinger metric for high-dimensional multinomials as well as for continuous distributions. The optimal choice of test, as well as the local minimax separation can be sensitive to the choice of metric and (in the continuous case) to the precise nature of the smoothness assumptions.

Despite this progress, developing a comprehensive theory for minimax testing of high-dimensional multinomials in general metrics, which provides useful practical insights, remains an important open problem.

3.5. *Composite nulls and imprecise nulls.* Now we briefly consider the problem of goodness-of-fit testing for a composite null. Let $\mathcal{P}_0 \subset \mathcal{M}$ be a subset of multinomials and consider testing

$$H_0 : P \in \mathcal{P}_0 \quad \text{verus} \quad H_1 : P \notin \mathcal{P}_0.$$

A complete minimax theory for this case is not yet available, but many special cases have been studied [Batu, Kumar and Rubinfeld (2004), Acharya, Daskalakis and Kamath (2015), Indyk, Levi and Rubinfeld (2012), Canonne et al. (2016)]. In particular, the work of Acharya, Daskalakis and Kamath (2015), provides some results for testing monotonicity, unimodality and log-concavity of a high-dimensional multinomial. Here, we briefly outline a general approach due to Acharya, Daskalakis and Kamath (2015).

A natural approach to hypothesis testing with a composite null is to split the sample, estimate the null distribution using one sample, and then to test goodness-of-fit to the estimated null distribution using the second sample. In the high-dimensional setting, we cannot assume that our estimate of the null distribution is very accurate (particularly in the TV metric). However, as highlighted in Acharya, Daskalakis and Kamath (2015) even in the high-dimensional setting we can often obtain sufficiently accurate estimates of the null distribution in the χ^2 distance. This observation motivates the study of the following two-stage approach. Use half the data to get an estimate $\widehat{p}_0$ assuming H_0 is true. Now use the other half to test the imprecise null of the form

$$H_0 : d_1(p, \widehat{p}_0) \leq \theta_n \quad \text{verus} \quad H_1 : d(p, \widehat{p}_0) \geq \varepsilon_n.$$

Here, d_1 and d are two possibly different metrics (in the case described above d_1 is the χ^2-distance and d is TV distance). The distribution $\widehat{p}_0$ is treated as fixed. As is typical, our interest is in ranges of the two critical radii $(\theta_n, \varepsilon_n)$ for which we have nontrivial power. Acharya, Daskalakis and Kamath (2015) refer to this as "robust testing" but they are not using the word robust in the usual sense. This imprecise null testing problem has been studied for a variety of metric choices in Valiant and Valiant (2011), Acharya, Daskalakis and Kamath (2015), Jiao, Han and Weissman (2017) and Daskalakis, Kamath and Wright (2018).

Towards developing practical tests for general composite nulls, an important open problem is to provide nonconservative methods for determining the rejection threshold for imprecise null hypothesis tests. In the high-dimensional setting we can no longer rely on limiting distribution theory, and it seems challenging to develop simulation based methods in general. Some alternative proposals have been suggested for instance in Berger and Boos (1994), but warrant further study in the high-dimensional setting.

More specific procedures can be constructed based on the structure of $\mathcal{P}_0$, and important special cases of composite null testing such as two-sample testing and independence testing have been studied in the literature. We turn our attention towards two-sample testing next.

4. Two sample testing.

In this case the data are

$$Z_1, \ldots, Z_{n_1} \sim P, \qquad W_1, \ldots, W_{n_2} \sim Q$$

and the hypotheses are

$$H_0 : P = Q \quad \text{versus} \quad H_1 : P \neq Q.$$

Let X and Y be the corresponding vectors of counts. First, suppose that $n_1 = n_2 := n$. In this case, Chan et al. (2014) showed that the minimax rate is

$$\varepsilon_n \asymp \max\left\{\frac{d^{1/2}}{n^{3/4}}, \frac{d^{1/4}}{n^{1/2}}\right\}. \tag{4.1}$$

The second term in the maximum is identical to the goodness-of-fit rate. In the high-dimensional case $d \geq n$, the minimax separation rate is $\frac{d^{1/2}}{n^{3/4}}$, which is strictly slower than the goodness-of-fit rate. This highlights that, in contrast to the low-dimensional setting, there can be a price to pay for testing when the null distribution is not known precisely (i.e., for testing with a composite null).

Chan et al. (2014) also showed that the (centered) χ^2 statistic

$$T = \sum_j \frac{(X_j - Y_j)^2 - X_j - Y_j}{X_j + Y_j} \tag{4.2}$$

is minimax optimal. The critical value t_α can be obtained using the usual permutation procedure [Lehmann and Romano (2005)]. It should be noted, however, that the theoretical results do not apply to the data-based permutation cutoff but, rather, to a cutoff based on a loose upper bound on the variance of T. In the remainder of this section we discuss some extensions:

1. *Unequal sample sizes*: The problem of two-sample testing with unequal sample sizes has been considered in Diakonikolas and Kane (2016) and Bhattacharya and Valiant (2015), and we summarize some of their results here. Without loss of generality, assume that $n_1 \geq n_2$. The minimax rate is

$$\varepsilon_n \asymp \max\left\{\frac{d^{1/2}}{n_1^{1/4} n_2^{1/2}}, \frac{d^{1/4}}{n_2^{1/2}}\right\}. \tag{4.3}$$

Note that this rate is identical to the minimax goodness of fit rate from Section 3 when $n_1 \geq d$. This makes sense since, when n_1 is very large, P can be estimated to high precision and we are essentially back in the simple goodness-of-fit setting.

From a practical perspective, while the papers [Diakonikolas and Kane (2016), Bhattacharya and Valiant (2015)] propose tests that are near-minimax they are not tests in the usual statistical sense. They contain a large number of unspecified constants and it is unclear how to choose the test threshold in such a way that the level of the test is α. We could choose the constants somewhat loosely then use the permutation distribution to get a level α test, but it is not known if the resulting test is still minimax, highlighting an important gap between theory and practice.

2. *Refinements of minimaxity*: In two-sample testing, unlike in goodness-of-fit testing, it is less clear how precisely to define the local minimax separation. Roughly, we would expect that distinguishing whether two samples are drawn from the same distribution or not should be more difficult if the distributions of the samples are nearly uniform, than if the distributions are concentrated on a small

number of categories. Translating this intuition into a satisfactory refined minimax notion is more involved than in the goodness-of-fit problem.

One such refinement appears in the work of Acharya et al. (2012), who restrict attention to so-called "symmetric" test statistics (roughly, these statistics are invariant to re-labelings of the categories of the multinomial). Their proposal is in the spirit of classical notions of adaptivity and oracle inequalities in statistics [see for instance van de Geer (2016), Donoho et al. (1996) and references therein], where they benchmark the performance of a test/estimator against an oracle test/estimator which is provided some side-information about the local structure of the parameter space. Concretely, they compare the separation achieved by χ^2-type tests to the minimax separation achieved by an oracle that knows the distributions P, Q up to a permutation.

3. *Testing continuous distributions*: Finally, we note that results for two-sample testing of high-dimensional multinomials have implications for two-sample testing in the continuous case. The recent paper of Arias-Castro, Pelletier and Saligrama (2018) building on work by Ingster (1997), considers two-sample testing of smooth distributions by reducing the testing problem to an appropriate high-dimensional multinomial testing problem. Somewhat surprisingly, they observe that at least under sufficiently strong smoothness assumptions, the minimax separation rate for two-sample testing matches that of goodness-of-fit testing.

5. Simulations.

In this section, we report some simulation studies performed to illustrate that tests can have high power even when their limiting distributions are not Gaussian and to illustrate the gains from using careful modifications to classical tests for testing high-dimensional multinomials.

5.1. *Limiting distribution of test statistics.*

One of the main messages of our paper is that tests can have high-power even in regimes where their null distributions are not Gaussian (or more generally well-behaved). As a result, restricting attention to regimes where the null distribution of a test statistic is well-behaved can be severely limiting.

As an illustration, we consider goodness-of-fit testing where the null distribution is a power law, that is, we take $p_0(i) \propto 1/i$. We will consider a high-dimensional setting where $d = 1000$ and $n = 400$. In this setting, as we will show in Section 5.2, the χ^2-statistic performs poorly but the truncated statistic in (3.4) has high power. However, as illustrated in Figure 1, in this high-dimensional regime the limiting distributions of both statistics are quite far from Gaussian. We also observe that the classical χ^2 statistic has a huge variance, which explains its poor power and motivates our introduction of the truncated χ^2 statistic. The truncated χ^2 statistic has a high-power, and achieves the minimax rate for goodness-of-fit testing, and as illustrated in this simulation has a much better behaved distribution under the null.

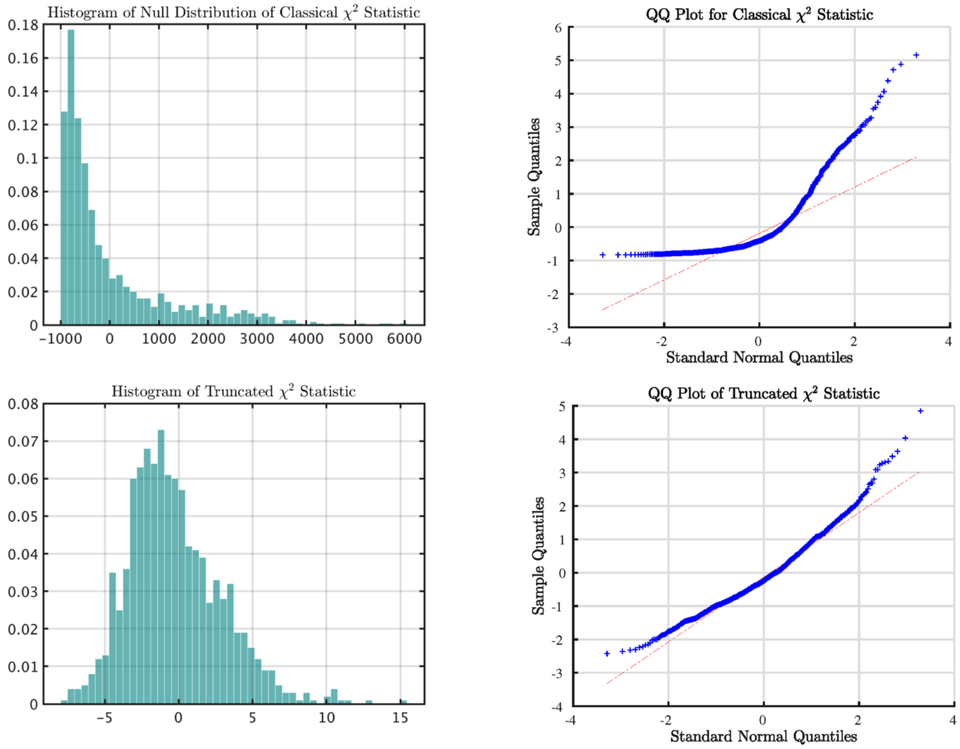

FIG. 1. *A plot of the distribution of the* (*centered*) *classical and truncated* χ^2 *test statistics under a power law null distribution*, *in the high-dimensional setting where* $n = 400$, $d = 1000$, *obtained via simulation*.

5.2. *Testing goodness-of-fit*. In this section, we compare the performance of several goodness-of-fit test statistics. Throughout we take $n = 400$ and $d = 1000$. Closely related simulations appear in our prior work [Balakrishnan and Wasserman (2017)]. In particular, we compare the classical χ^2 statistic in (3.1), the likelihood-ratio test in (3.2), the truncated χ^2 statistic in (3.4) and the two-stage locally minimax 2/3rd and tail test described in Section 3.2, with the ℓ_1 and ℓ_2 test statistics given as

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

We examine the power of these tests under various alternatives:

1. *Minimax alternative*: We perturb each entry by an amount proportional to $p_0(i)^{2/3}$ with a randomly chosen sign. This is close to the worst-case perturbation used in Valiant and Valiant (2017) in their proof of the local-minimax lower bound.

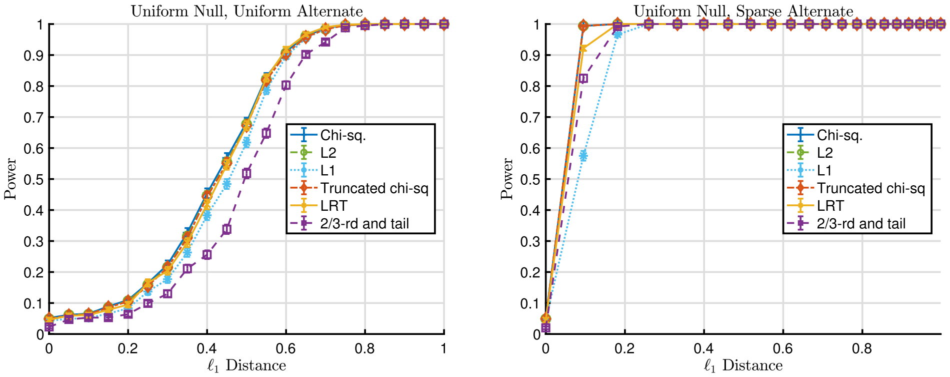

FIG. 2. *A comparison between the truncated* χ^2 *test, the* 2/3*rd* + *tail test* Valiant and Valiant (2017), *the* χ^2*-test, the likelihood ratio test, the* ℓ_1 *test and the* ℓ_2 *test. The null is chosen to be uniform, and the alternate is either a dense or sparse perturbation of the null. The power 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.*, $n = 200, d = 2000$) *the tests have high-power, and perform comparably.*

2. *Uniform dense alternative*: We perturb each entry of the null distribution by a scaled Rademacher random variable. (A Rademacher random variable takes values +1 and −1 with equal probability.)

3. *Sparse alternative*: In this case, we essentially only perturb the first two entries of the null multinomial. We increase the two largest entries of the multinomial and then re-normalize the resulting distribution, this results in a large perturbation to the two largest entries and a relatively small perturbation to the other entries of the multinomial.

4. *Alternative proportional to null*: We perturb each entry of the null distribution by an amount proportional to $p_0(i)$, with a randomly chosen sign.

We observe that the truncated χ^2 test and the 2/3rd + tail test from Valiant and Valiant (2017) are remarkably robust. All tests are comparable when the null is uniform but the two-stage 2/3rd + tail test suffers a slight loss in power due to the Bonferroni correction (see Figure 2). We also note that when the null is uniform the χ^2 test, the truncated χ^2 test and the ℓ_2 test are identical up to centering. The distinctions between the classical tests and the recently proposed modified tests are clearer for the power law null. In particular, from the simulation testing a power-law null against a sparse alternative it is clear that the χ^2 and likelihood ratio test can have very poor power in the high-dimensional setting. The ℓ_2 test appears to have high-power against sparse alternatives but performs poorly against dense alternatives suggesting potential avenues for future investigation.

5.3. *Two-sample testing*. Finally, we turn our attention to the problem of two-sample testing for high-dimensional multinomials. We compare three different test statistics, the two-sample χ^2 statistic (4.2), the ℓ_1 and ℓ_2 statistics:

$$T_{\ell_1} = \sum_{i=1}^{d} \left| \frac{X_i}{n_1} - \frac{Y_i}{n_2} \right|, \qquad T_{\ell_2} = \sum_{i=1}^{d} \left(\frac{X_i}{n_1} - \frac{Y_i}{n_2} \right)^2$$

and an oracle goodness-of-fit statistic. The oracle has access to the distribution P (and ignores the first sample) and tests goodness-of-fit using the second sample. In our simulations, we use the truncated χ^2 test for goodness-of-fit.

Motivated by our simulations in the previous section, we consider the following pairs of distributions in the two-sample setup:

1. *Uniform P, dense perturbation Q*: We take the distribution P to be uniform and the distribution Q to be the distribution where we perturb each entry of P by a scaled Rademacher random variable.
2. *Power-law P, sparse perturbation Q*: Noting the difficulty faced by classical tests for goodness-of-fit testing of a power law versus a sparse perturbation (see Figure 3) we consider a similar setup in the two-sample setting. We take each entry $p(i) \propto 1/i$ and take $q(i)$ to be the sparse perturbation described previously (where the two largest entries are perturbed by a relatively large magnitude).
3. *Minimax P, Q*: This construction is inspired by the work of Batu et al. (2000), and is used in their construction of a minimax lower bound for two-sample testing in the high-dimensional setting (i.e., when $d \gg \max\{n_1, n_2\}$).

For a prescribed separation ε, with probability $n_1/(2d)$ we choose $p(i) = q(i) = 1/n_1$ and with probability $1 - n_1/(2d)$ we choose $p(i) = 1/(2d)$ and $q(i) = 1/(2d) + \varepsilon R_i/(2d)$, where the R_i denote independent Rademacher random variables. Both distributions are then normalized.

Roughly, the two distributions contain a mixture of heavy elements of mass $1/n_1$ and light elements of mass close to $1/d$. The two distributions have ℓ_1 distance close to ε and the insight of Batu et al. (2000) is that in the two-sample setting is quite difficult to distinguish between variations in observed frequencies due to the perturbation of the entries and due to the random mixture of heavy and light entries.

In each case, the cut-off for the tests is determined via the permutation method.

The balanced case: In this case, we set $n_1 = n_2 = 200$ and $d = 400$. We observe in Figure 4 that there is a clear and significant loss in power relative to the goodness-of-fit oracle, indicating the increased complexity of two-sample testing in the high-dimensional setup. We note that, as is clear from the minimax separation rate we do not expect any loss in power in the low-dimensional setting. The loss in power is exacerbated when we consider the minimax two-sample pair (P, Q) described above. We also note that the loss in power is negligible for the power law pair of distributions: due to the rapid decay of their entries these multinomials can be well estimated from a small number of samples.

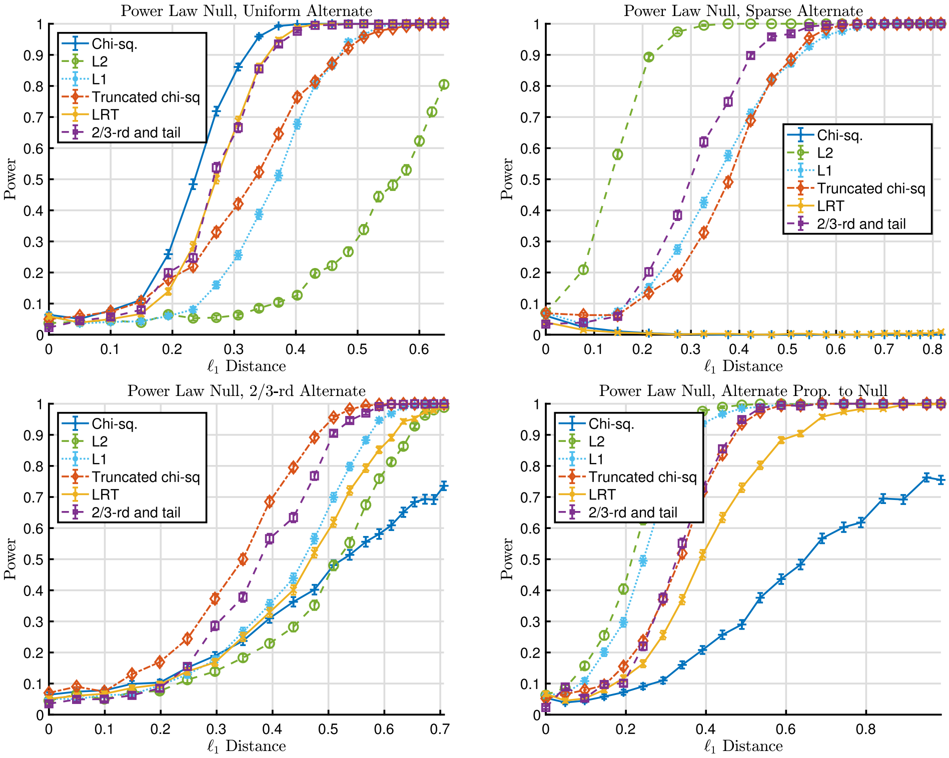

FIG. 3. *A comparison between the truncated χ^2 test, the $2/3$rd + tail test* [Valiant and Valiant (2017)], *the χ^2-test, the likelihood ratio test, the ℓ_1 test and the ℓ_2 test. The null is chosen to be a power law with $p_0(i) \propto 1/i$. We consider four possible alternatives, the first uniformly perturbs the coordinates, the second is a sparse perturbation only perturbing the first two coordinates, the third perturbs each co-ordinate proportional to $p_0(i)^{2/3}$ and the final setting perturbs each coordinate proportional to $p_0(i)$. 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 imbalanced case: In this case, we set $n_1 = 2000, n_2 = 200$ and $d = 400$. The χ^2 statistic for two-sample testing in the imbalanced case is given by

$$T = \sum_{i=1}^{d} \frac{(n_2 X_i - n_1 Y_i)^2 - n_2^2 X_i - n_1^2 Y_i}{X_i + Y_i},$$

where we follow Bhattacharya and Valiant (2015) and use a slightly modified centering of the usual χ^2 statistic. Since the χ^2-statistic is minimax optimal in the balanced case, one might conjecture that this continues to be the case in the imbalanced case. However, as our simulations suggest this is not the case. Somewhat surprisingly, the performance of the χ^2 statistic can degrade when one of the sample sizes is increased (see Figure 5). The ℓ_1 statistic on the other hand appears to be perform as expected, i.e. its power is close to that of the oracle goodness-of-fit

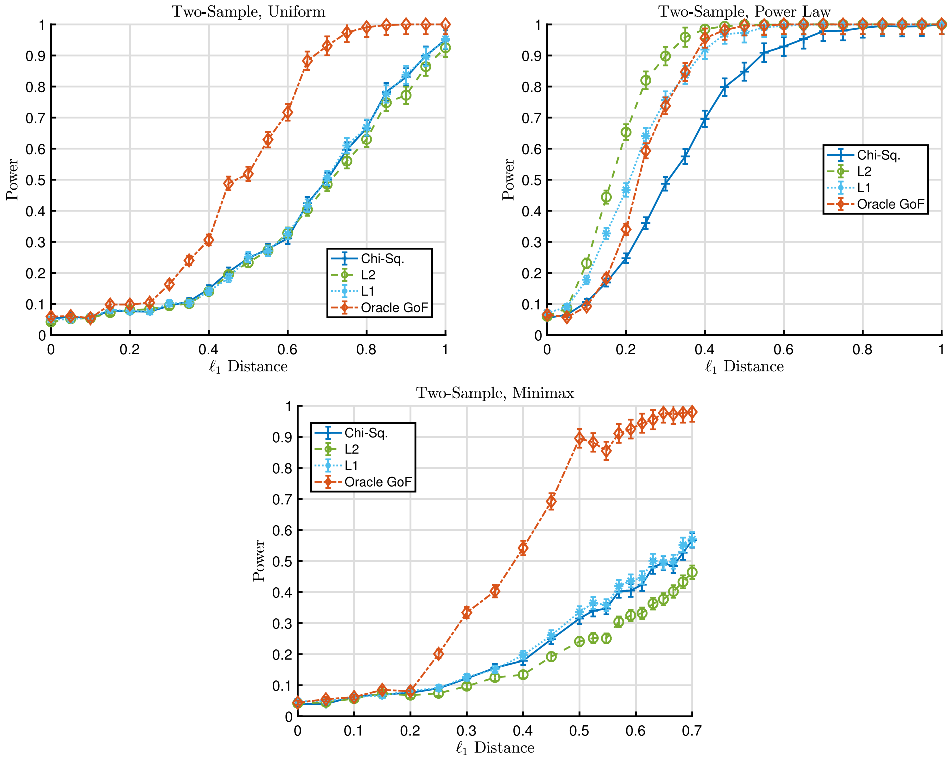

FIG. 4. *A comparison between the* χ^2 *test, the* ℓ_1 *test, the* ℓ_2 *test, and the* (*oracle*) *goodness-of-fit test. In the three settings, the two distributions are chosen as described in the text* (*the distribution* P *is chosen to be either uniform, power-law or minimax*). *The power of the tests are plotted against the* ℓ_1 *distance between* P *and* Q. *The sample sizes from* P *and* Q *are taken to be balanced and are each equal to* 200. *Each point in the graph is an average over* 1000 *trials.*

test in the case when one of the sample sizes is very large, and we believe that this statistic warrants further study.

6. Discussion. Despite the fact that discrete data analysis is an old subject, it is still a vibrant area of research and there is still much that we don't know. Steve Fienberg showed prescience in drawing attention to one of the thorniest issues: understanding high-dimensional multinomials.

Much of the statistical literature has dealt with the high-dimensional case by imposing assumptions on the distribution so that simple limiting distributions can be obtained. Doing so gives up the most appealing property of multinomial inference: it is completely distribution-free. As we have seen in this paper, recent work by a variety of communities has developed new and rather surprising theoretical results. What is often missing in the recent literature is the appreciation that statisticians want tests with precise control of the type I error rate. As a result, there

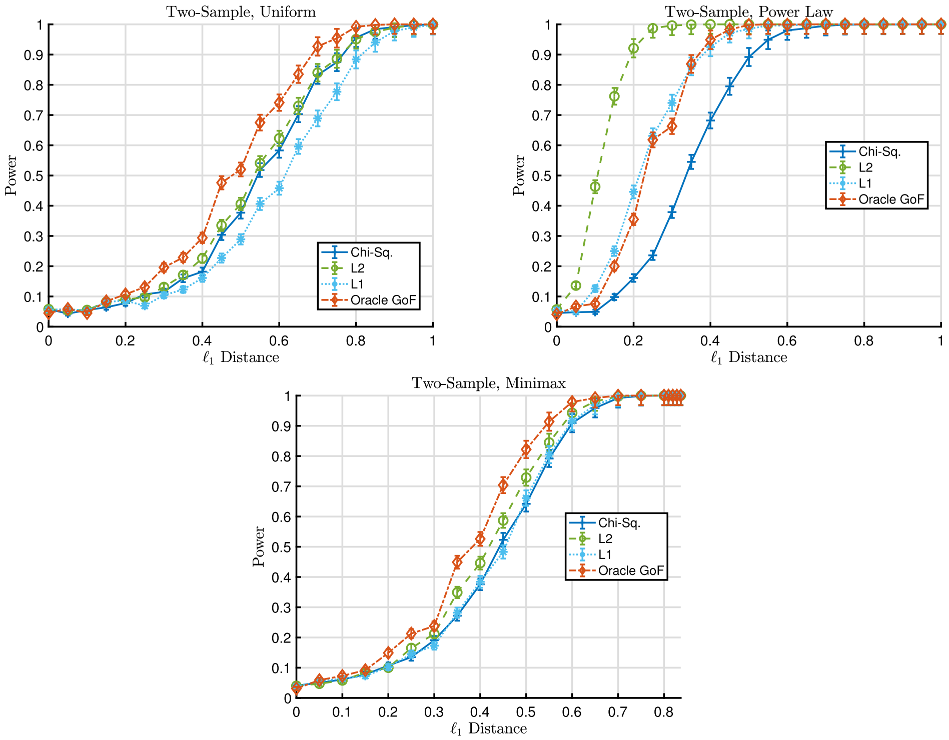

FIG. 5. *A comparison between the* χ^2 *test, the* ℓ_1 *test, the* ℓ_2 *test and the* (*oracle*) *goodness-of-fit test. In the three settings, the two distributions are chosen as described in the text* (*the distribution* P *is chosen to be either uniform, power-law or minimax*). *The power of the tests are plotted against the* ℓ_1 *distance between* P *and* Q. *The sample sizes from* P *and* Q *are taken to be imbalanced, that is, we take* $n_1 = 2000$ *and* $n_2 = 200$. *Each point in the graph is an average over* 1000 *trials.*

remain gaps between theory and practice. This issue is particularly significant in settings with general composite or imprecise null hypotheses where methods based on simulation are not directly applicable.

As alluded to earlier other directions of future research include: developing tractable and interesting refinements of the minimax framework in problems beyond goodness-of-fit testing as well as developing a precise and broad understanding of the role of the geometry of the null and alternate sets of distributions in determining both the minimax rates and well as the optimal tests. From a practical standpoint, we believe it would be fruitful to develop a deeper understanding of the middle-ground between minimax-tests that are designed to have power against a large class of alternatives, and so-called “tailor-made tests” [Bickel, Ritov and Stoker (2006)] which are designed to have much higher power against a narrow class of alternatives.

We have focused on goodness of fit and two sample problems. There is a rich literature on other problems such as independence testing and testing shape constraints [Diakonikolas and Kane (2016), Acharya, Daskalakis and Kamath (2015), Diakonikolas, Kane and Nikishkin (2015a, 2015b, 2017)]. As we discussed earlier, Balakrishnan and Wasserman (2017) showed that these new results for high-dimensional discrete data have implications for continuous data. There is much more to say about this and this is a direction that we are actively pursuing. It would also be interesting to explore the extent to which these results for high-dimensional discrete testing can lead to advancements in other parametric, combinatorial or high-dimensional structured testing problems [Addario-Berry et al. (2010), Arias-Castro, Candès and Durand (2011), Berthet and Rigollet (2013), Ingster, Tsybakov and Verzelen (2010), Donoho and Jin (2004)]. Finally, we have restricted attention in this paper to hypothesis testing. In future work, we will report results on high-dimensional inference using confidence sets and point estimation.

APPENDIX

Here we describe the local minimax results for goodness of fit testing more precisely. 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:

$$\mathcal{Q}_\sigma(p_0) = \left\{ i : \sum_{j=i}^{d} p_0(j) \leq \sigma \right\}. \tag{A.1}$$

The σ-bulk is defined to be

$$\mathcal{B}_\sigma(p_0) = \{ i > 1 : i \notin \mathcal{Q}_\sigma(p_0) \}. \tag{A.2}$$

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

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

Define, ℓ_n and u_n to be the solutions to the equations

$$\begin{aligned} \ell_n(p_0) &= \max\left\{ \frac{1}{n}, \sqrt{\frac{V_{\ell_n(p_0)}(p_0)}{n}} \right\}, \\ u_n(p_0) &= \max\left\{ \frac{1}{n}, \sqrt{\frac{V_{u_n(p_0)/16}(p_0)}{n}} \right\}. \end{aligned} \tag{A.4}$$

With these definitions in place, we are now ready to state the result of Valiant and Valiant (2017). We use $c_1, c_2, C_1, C_2 > 0$ to denote positive universal constants.

THEOREM A.1 [Valiant and Valiant (2017)]. *The local critical radius $\varepsilon_n(p_0, \mathcal{M})$ for multinomial testing is upper and lower bounded as*

$$c_1 \ell_n(p_0) \leq \varepsilon_n(p_0, \mathcal{M}) \leq C_1 u_n(p_0). \tag{A.5}$$

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

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

Acknowledgments. We thank the anonymous reviewers and the editor for their feedback and various suggestions that substantially improved the paper. We are also grateful to Ilias Diakonikolas and Clément Canonne for their suggestions and for pointing out several missing references. All of our simulation results and figures can be reproduced using code made available at (https://github.com/sivab/multinomial).

REFERENCES

ACHARYA, J., DASKALAKIS, C. and KAMATH, G. (2015). Optimal testing for properties of distributions. In *Advances in Neural Information Processing Systems* 3591–3599.

ACHARYA, J., DAS, H., JAFARPOUR, A., ORLITSKY, A., PAN, S. and SURESH, A. (2012). Competitive classification and closeness testing. In *Proceedings of the* 25*th Annual Conference on Learning Theory* **23** 22.1–22.18.

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

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

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

BALAKRISHNAN, S. and WASSERMAN, L. (2017). Hypothesis testing for densities and high-dimensional multinomials: Sharp local minimax rates. Available at arXiv:1706.10003.

BALAKRISHNAN, S. and WASSERMAN, L. (2018). Hypothesis testing for densities and high-dimensional multinomials II: Sharp local minimax rates. Forthcoming.

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

BATU, T., KUMAR, R. and RUBINFELD, R. (2004). Sublinear algorithms for testing monotone and unimodal distributions. In *Proceedings of the* 36*th Annual ACM Symposium on Theory of Computing* 381–390. ACM, New York. MR2121623

BATU, T., FORTNOW, L., RUBINFELD, R., SMITH, W. D. and WHITE, P. (2000). Testing that distributions are close. In 41*st Annual Symposium on Foundations of Computer Science* (*Redondo Beach*, *CA*, 2000) 259–269. IEEE Comput. Soc., Los Alamitos, CA. MR1931824

BERGER, R. L. and BOOS, D. D. (1994). *P* values maximized over a confidence set for the nuisance parameter. *J. Amer. Statist. Assoc.* **89** 1012–1016. MR1294746

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

BHATTACHARYA, B. and VALIANT, G. (2015). Testing closeness with unequal sized samples. In *Advances in Neural Information Processing Systems* 2611–2619.

BICKEL, P. J., RITOV, Y. and STOKER, T. M. (2006). Tailor-made tests for goodness of fit to semiparametric hypotheses. *Ann. Statist.* **34** 721–741. MR2281882

BISHOP, Y. M. M., FIENBERG, S. E. and HOLLAND, P. W. (1977). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA. With the collaboration of Richard J. Light and Frederick Mosteller. Third printing. MR0431514

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

CANONNE, C. L. (2018). A survey on distribution testing: Your data is big. But is it blue? *Theory Comput*. To appear.

CANONNE, C. L., DIAKONIKOLAS, I., GOULEAKIS, T. and RUBINFELD, R. (2016). Testing shape restrictions of discrete distributions. In 33*rd Symposium on Theoretical Aspects of Computer Science*. *LIPIcs. Leibniz Int. Proc. Inform.* **47** Art. No. 25. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern. MR3539122

CHAN, S.-O., DIAKONIKOLAS, I., VALIANT, G. and VALIANT, P. (2014). Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the Twenty-Fifth Annual ACM–SIAM Symposium on Discrete Algorithms* 1193–1203. ACM, New York. MR3376448

CHATTERJEE, S., GUNTUBOYINA, A. and SEN, B. (2015). On risk bounds in isotonic and other shape restricted regression problems. *Ann. Statist.* **43** 1774–1800. MR3357878

DASKALAKIS, C., KAMATH, G. and WRIGHT, J. (2018). Which distribution distances are sublinearly testable? In *Proceedings of the Twenty-Ninth Annual ACM–SIAM Symposium on Discrete Algorithms* 2747–2764. SIAM, Philadelphia, PA.

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

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 Comput. Soc., Los Alamitos, CA. MR3631031

DIAKONIKOLAS, I., KANE, D. M. and NIKISHKIN, V. (2015a). Optimal algorithms and lower bounds for testing closeness of structured distributions. In 2015 *IEEE* 56*th Annual Symposium on Foundations of Computer Science—FOCS* 2015 1183–1202. IEEE Comput. Soc., Los Alamitos, CA. MR3473364

DIAKONIKOLAS, I., KANE, D. M. and NIKISHKIN, V. (2015b). Testing identity of structured distributions. In *Proceedings of the Twenty-Sixth Annual ACM–SIAM Symposium on Discrete Algorithms* 1841–1854. SIAM, Philadelphia, PA. MR3451147

DIAKONIKOLAS, I., KANE, D. M. and NIKISHKIN, V. (2017). Near-optimal closeness testing of discrete histogram distributions. In 44*th International Colloquium on Automata*, *Languages*, *and Programming*. *LIPIcs. Leibniz Int. Proc. Inform.* **80** Art. No. 8. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern. MR3685748

DONOHO, D. and JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962–994. MR2065195

DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455. MR1311089

DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.* **24** 508–539. MR1394974

ERMAKOV, M. S. (1991). Minimax detection of a signal in Gaussian white noise. *Theory Probab. Appl.* **35** 667–679. MR1090496

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

FIENBERG, S. E. (1980). *The Analysis of Cross-Classified Categorical Data*, 2nd ed. MIT Press, Cambridge, MA. MR0623082

FIENBERG, S. E. and HOLLAND, P. W. (1973). Simultaneous estimation of multinomial cell probabilities. *J. Amer. Statist. Assoc.* **68** 683–691. MR0359153

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

GOLDENSHLUGER, A. and LEPSKI, O. (2011). Bandwidth selection in kernel density estimation: Oracle inequalities and adaptive minimax optimality. *Ann. Statist.* **39** 1608–1632. MR2850214

GOLDREICH, O. (2017). *Introduction to Property Testing*. Cambridge Univ. Press, Cambridge.

HABERMAN, S. J. (1977). Log-linear models and frequency tables with small expected cell counts. *Ann. Statist.* **5** 1148–1169. MR0448675

HOEFFDING, W. (1965). Asymptotically optimal tests for multinomial distributions. *Ann. Math. Stat.* **36** 369–408. MR0173322

HOLST, L. (1972). Asymptotic normality and efficiency for certain goodness-of-fit tests. *Biometrika* **59** 137–145. MR0314193

INDYK, P., LEVI, R. and RUBINFELD, R. (2012). Approximating and testing k-histogram distributions in sub-linear time. In *Proceedings of the* 31*st ACM SIGMOD–SIGACT–SIGART Symposium on Principles of Database Systems*, *PODS* 2012 15–22.

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

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

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

IVČENKO, G. I. and MEDVEDEV, JU. I. (1980). Decomposable statistics and hypothesis testing. The case of small samples. *Theory Probab. Appl.* **23** 540–551. MR0516276

JIAO, J., HAN, Y. and WEISSMAN, T. (2017). Minimax estimation of the l_1 distance. Available at arXiv:1705.00807.

KOEHLER, K. J. (1986). Goodness-of-fit tests for log-linear models in sparse contingency tables. *J. Amer. Statist. Assoc.* **81** 483–493. MR0845887

KOEHLER, K. J. and LARNTZ, K. (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials. *J. Amer. Statist. Assoc.* **75** 336–344.

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

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

LEPSKI, O. V. and SPOKOINY, V. G. (1999). Minimax nonparametric hypothesis testing: The case of an inhomogeneous alternative. *Bernoulli* **5** 333–358. MR1681702

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

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

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

RUBINFELD, R. (2012). Taming big probability distributions. *XRDS* **19** 24–28.

SPOKOINY, V. G. (1996). Adaptive hypothesis testing using wavelets. *Ann. Statist.* **24** 2477–2498. MR1425962

VALIANT, G. and VALIANT, P. (2011). Estimating the unseen: An $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs. In *STOC'*11*—Proceedings of the* 43*rd ACM Symposium on Theory of Computing* 685–694. ACM, New York. MR2932019

VALIANT, G. and VALIANT, P. (2017). An automatic inequality prover and instance optimal identity testing. *SIAM J. Comput.* **46** 429–455. MR3614697

VAN DE GEER, S. (2016). *Estimation and Testing Under Sparsity*. *Lecture Notes in Math.* **2159**. Springer, Cham. MR3526202

WEI, Y. and WAINWRIGHT, M. J. (2017). The local geometry of testing in ellipses: Tight control via localized Kolomogorov widths. Available at arXiv:1712.00711.

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