Math. Stat. Learn. 5 (2022), 1–54
DOI 10.4171/MSL/32

© 2022 European Mathematical Society
Published by EMS Press
This work is licensed under a CC BY 4.0 license

Sharp local minimax rates for goodness-of-fit testing in multivariate binomial and Poisson families and in multinomials

Julien Chhor and Alexandra Carpentier

Abstract. We consider the identity testing problem – or goodness-of-fit testing problem – in multivariate binomial families, multivariate Poisson families and multinomial distributions. Given a known distribution p and n i.i.d. samples drawn from an unknown distribution q, we investigate how large $\rho > 0$ should be to distinguish, with high probability, the case $p = q$ from the case $d(p, q) \geq \rho$, where d denotes a specific distance over probability distributions. We answer this question in the case of a family of different distances: $d(p, q) = \|p - q\|_t$ for $t \in [1, 2]$, where $\|\cdot\|_t$ is the entrywise ℓ_t norm. Besides being locally minimax-optimal – i.e. characterizing the detection threshold in dependence of the known matrix p – our tests have simple expressions and are easily implementable.

1. Introduction

We consider the problem of *identity testing* or *goodness-of-fit testing* in multivariate binomial families, multivariate Poisson families and multinomial distributions. At a high level, this problem aims at testing whether or not the data distribution matches a given known distribution. Throughout the paper, we will state the results in the multivariate binomial setting, and will establish the link with multivariate Poisson families and multinomials later on. The problem can be stated as follows: given n i.i.d. realizations of an unknown multivariate Binomial family (see Section 2) with unknown distribution q, and given a known distribution p, we want to test

$$\mathcal{H}_0 \colon p = q \quad \text{vs.} \quad \mathcal{H}_1 \colon d(p, q) \geq \rho,$$

for a given distance d and separation radius ρ.

The difficulty of this testing problem is characterized by the minimal separation radius ρ needed to ensure the existence of a test that is uniformly consistent under both the null and the alternative hypothesis, i.e. a test whose worst-case error is smaller than

2020 Mathematics Subject Classification. Primary 62G10; Secondary 62B10, 62C20.
Keywords. Minimax identity testing, goodness-of-fit testing, multinomial distributions, multivariate poisson families, locality.

a given $\eta > 0$, and to identify such a test. See Section 2 for a precise definition of the setting.

In this paper, we will mostly focus on the following goals:

• We focus on the case where the distance d is the ℓ_t distance, namely, if $p = (p_1, \dots, p_N)$ and $q = (q_1, \dots, q_N)$, then

$$d(p,q) = \left(\sum_{i=1}^{N} |q_i - p_i|^t \right)^{1/t}$$

for any $t \in [1, 2]$. Typically, the case $t = 2$ and $t = 1$ (total variation distance for discrete distributions) are considered, and we interpolate between these two extreme cases.

• Our main objective will be to develop tests – as well as matching lower bounds – for this identity testing problem that are *locally optimal* in that the minimax separation distance ρ should depend tightly on p. Indeed, it is clear that some p will be "easier" to test than others. Consider, for example, the following two extreme cases in the case of discrete (multinomial) distributions over $\{1, \dots, N\}$:

(i) the very "easy" case where p is a Dirac distribution on one of the coordinates, which implies a very low noise, and

(ii) the very "difficult" case where all entries of p are equal to $1/N$, which maximizes the noise.

It is clear that the minimax local separation distance should differ between these two cases and be much smaller in case (i) than in case (ii). We aim at studying the minimax local separation distance for any p, and characterize tightly its shape depending on p.

The existing literature about hypothesis testing [46] is profuse: the goodness-of-fit problem has been thoroughly studied, especially in the case of signal detection in the Gaussian setting, notably by Ingster (see [40]) and has given rise to a vast literature. In parallel to the study of hypothesis testing, there exists a broad literature on the related problem of property testing with seminal papers such as [36, 49].

The identity testing problem in multinomials – i.e. probability distributions over a finite set – has been widely studied in the literature. We refer the reader to [8,18,19] for excellent surveys. When observing n i.i.d. data with unknown discrete distribution q and when fixing a distribution p, the aim is to derive the minimal separation distance ρ so that a uniformly consistent test exists for testing $\mathcal{H}_0 : p = q$ vs. $\mathcal{H}_1(\rho) : d(p,q) \geq \rho$. Note that this problem is also often considered in the dual setting of *sample complexity*, where the goal is to find the minimal number of samples n such that a consistent test exists for a given separation $\rho > 0$. One distinguishes between *global results* which are obtained for the worst case of the distribution p, and *local results*, where the minimax separation distance is required to depend precisely on any given p.

For global results, see e.g. [38] (in Russian), [29, 34, 39, 47], and also in the related two-sample testing problem – where both p, q are unknown and observed through samples – see e.g. [11, 21]. In the present paper, we focus on *local* results. In the case of the ℓ_1 distance, important contributions to local testing have been established in e.g. [28, 53]. Note that these papers provide results in terms of sample complexity, and more recently, the paper [9] has re-considered this problem in terms of minimax separation distance – focusing also on the case of smooth densities. Another quite related work is [14], investigating the rate of goodness-of-fit testing in the multinomial case, in the ℓ_1 and ℓ_2 distances, under privacy constraints. Regarding the related two sample testing problem, see [4, 15, 28, 41]. This multinomial framework proves very useful for a wide range of applications, which include Ising models [26], Bayesian networks [20] or even quantum mechanics [7].

The papers [9,53] are the most related to our present results, due to the equivalence between the multivariate binomial and Poisson distribution settings and the multinomial setting after a Poissonization trick; see Section 3.1 for more details on why our setting encompasses those settings. We postpone a precise discussion between our result and this stream of literature to the core of the paper[1], since it is technical. As high-level comments, we restrict to remarking this stream of literature only considers separation in total variation distance, namely the ℓ_1 distance for discrete distributions.

Note that goodness-of-fit testing for inhomogeneous Erdös–Rényi random graphs (see the definition e.g. in [32]), is a direct an important corollary of our result about multivariate binomial local testing. This result is therefore interesting as only little literature exists about identity testing in random graphs - and to the best of our knowledge, no literature exists about *local* identity testing in the sense described above (see for example [25] for global testing in inhomogeneous random graphs). In recent machine learning and statistical applications, the increasing use of networks has made large random graphs a decisive field of interest. To name a few topics, let us mention community detection, especially in the stochastic block model ([1, 2, 6, 27, 54]), in social networks ([12, 56]), as well as network modeling ([5, 44]), or network dynamics ([13]). The papers [32] and [33] propose an analysis of the two sample case, under sparsity: Given two populations of mutually independent random graphs, each population being drawn respectively from the distributions P and Q, they perform the minimax hypothesis testing $\mathcal{H}_0 : P = Q$ vs. $\mathcal{H}_1 : d(P, Q) \geq \rho$ for a variety of distances d, and identify optimal tests over the classes of sparse graphs that they consider. The paper [48] identifies a computationally efficient algorithm for testing the separability of two hypotheses. Testing between a stochastic block model versus an Erdös–Rényi model has been studied in [30] and [43]. Phase transitions are

[1] We compare with this stream of literature under our upper and lower bounds in Sections 3, and also in the discussion in Section 4.

also known for detecting strongly connected groups or high dimensional geometry in large random graphs ([17]). The paper [51] tests random dot-product graphs in the two sample setting with low-rank adjacency matrices. The paper [31] examines a more general case in which the graphs are not necessarily defined on the same set of vertices. To summarize, only few papers address the construction of efficient tests in random graphs – although this would be valuable in various areas such as social networks [45], brain or 'omics' networks [35, 37], testing chemicals [50] or ecology and evolution [24]. Moreover, and to the best of our knowledge, no paper considers the *local version* of the testing problem, i.e. focuses on obtaining separation distances that depend on the null hypothesis.

The paper is organized as follows: In Section 2, we describe the setting by defining the multivariate binomial model and the minimax framework. In Section 3, state our main theorem, which gives an explicit expression of the minimax separation radius as a function of p and n. In Section 3.1, we establish the equivalence between the binomial, the Poisson and the multinomial settings. In Section 4, we discuss our results, by comparing them with the state of the art, especially with the multinomial setting. In Section 5, we describe our lower bound construction. In Section 6, we describe our tests and state theoretical results guaranteeing their optimality. We finally provide additional comments on our results in Section 7. All proofs are deferred to the appendix.

2. Problem statement

2.1. Setting

We first introduce the Binomial setting. In Section 3.1, we will introduce two other very related settings (the multinomial and the Poisson settings) and prove that the associated minimax rates can be deduced from the Binomial case.

Let $N \in \mathbb{N}$, $N \geq 2$ and define $\mathcal{P}_N = [0,1]^N$. Let $q = (q_1, \dots, q_N) \in \mathcal{P}_N$ be an unknown vector of Bernoulli parameters. Assume that we observe $X_1, \dots, X_n$ i.i.d. such that each X_i can be written as $X_i = (X_i(1), \dots, X_i(N))$ where all of the entries $X_i(1), \dots, X_i(N)$ are mutually independent and $X_i(j) \sim \mathrm{Ber}(q_j)$. We slightly abuse notation and write $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} q$ when $X_1, \dots, X_n$ are generated with this distribution. Assume that n is even: $n = 2k$, for $k \in \mathbb{N}$. This assumption can be made without loss of generality and makes the analysis of the upper bound more convenient by allowing for sample splitting. We denote the total variation distance between two

probability measures by d_{TV} and for any $p \in \mathbb{R}^N$ and for $t > 0$, we define

$$\|p\|_t = \left[\sum_{j=1}^{N} |p_j|^t\right]^{1/t}.$$

This quantity defines a norm whenever $t \geq 1$.

2.2. Minimax testing problem

We now define the testing problem considered in the paper. Let $\eta \in (0,1)$ be a fixed constant and let $t \in [1,2]$. We are given a known vector $p \in \mathcal{P}_N$ and we suppose that the data is generated from an unknown vector q: $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} q$. We are interested in the following testing problem:

$$\mathcal{H}_0^p : q = p \quad \text{vs.} \quad \mathcal{H}_1^{\rho,p,t} : q \in \mathcal{P}_N, \quad \|p - q\|_t \geq \rho. \tag{1}$$

This problem is called "goodness-of-fit testing problem". When no ambiguity arises, we write $\mathcal{H}_0$ and $\mathcal{H}_1$ to denote the null and alternative hypotheses.

A *test* ψ is a measurable function of the observations $X_1, \ldots, X_n$, taking only the values 0 or 1. We measure the quality of any test ψ by its *maximum risk*, defined as

$$\begin{aligned} R(\psi) &:= R_{\rho,p,t,n}(\psi) \\ &= \mathbb{P}_p(\psi = 1) + \sup_{q \text{ s.t. } \|p-q\|_t \geq \rho} \mathbb{P}_q(\psi = 0), \end{aligned} \tag{2}$$

where $R(\psi)$ is the sum of the type-I and the type-II errors.

The *minimax risk* is the risk of the best possible test, if any:

$$\begin{aligned} R^* &:= R^*_{\rho,p,t,n} = \inf_{\psi \text{ test}} R(\psi) \\ &= \inf_{\psi \text{ test}} \Big[\mathbb{P}_p(\psi = 1) + \sup_{Q : \|p-q\|_t \geq \rho} \mathbb{P}_q(\psi = 0)\Big]. \end{aligned}$$

Note that $R^* := R^*_{\rho,p,t,n}$ depends on the choice of the norm indexed by t, the vector p, the separation radius ρ, and the sample size n. Since all quantities depend on p, we say that the testing problem is *local* – around p – as opposed to classical approaches in the minimax testing literature, where one generally only considers a family of vectors p and focuses only on the worst case results over this family, see e.g. [31].

In the following, we fix an absolute constant $\eta \in (0,1)$ and *we are interested in finding the smallest* $\rho^*_{p,t,n}$ *such that* $R^*_{\rho^*_{p,t,n},p,t,n} \leq \eta$:

$$\rho^*_{p,t,n}(\eta) = \inf\{\rho > 0 : R^*_{\rho,p,t,n} \leq \eta\}. \tag{3}$$

We call $\rho^*_{p,t,n}(\eta)$ the η-*minimax separation radius*. Whenever no ambiguity arises, we drop the indexation in n, p, t, η and write simply $\rho^*, R^*_\rho, R_\rho(\psi)$ – but these variables remain important, as will appear later on.

The aim of the paper is to give the explicit expression of $\rho^*_{p,t,n}$ up to constant factors depending only on η and to construct optimal tests, for any $p \in \mathcal{P}_N$ and all $t \in [1, 2]$.

Additional notation. Let $\eta > 0$. For f and g two real-valued functions defined, we say that $f \lesssim_\eta g$ (resp. $f \gtrsim_\eta g$) if there exists a constant $c_\eta > 0$ (resp. $C_\eta > 0$) depending only on η, such that $c_\eta g \leq f$ (resp. $f \geq C_\eta g$). We write $f \asymp_\eta g$ if $g \lesssim_\eta f$ and $f \lesssim_\eta g$. Whenever the constants are absolute, we drop the index η and just write $\lesssim$, $\gtrsim$, $\asymp$. We respectively denote by $x \vee y$ and $x \wedge y$ the maximum and minimum of the two real values x and y.

3. Results

Without loss of generality, assume that $\max_{1\leq j\leq N} p_j \leq \frac{1}{2}$. Otherwise, if for some $j \in \{1, \dots, N\}$, $p_j > \frac{1}{2}$, replace p_j by $1 - p_j$ and replace accordingly $X_i(j)$ by $1 - X_i(j)$ for all $i = 1, \dots, n = 2k$. Without loss of generality, assume that all entries of the known vector p are sorted in decreasing order:

$$p = (p_1 \geq p_2 \geq \cdots \geq p_N).$$

For any index $1 \leq u \leq N$, we define the vectors

$$\begin{cases} p_{\leq u} = (p_1, \dots, p_u, 0, \dots, 0), \\ p_{>u} = (0, \dots, 0, p_{u+1}, \dots, p_N). \end{cases}$$

Let $\eta > 0$. In what follows, we write

$$r = \frac{2t}{4-t} \quad \text{and} \quad b = \frac{4-2t}{4-t}. \tag{4}$$

For p, we also define

$$I = \min\left\{ J : \sum_{i>J} p_i^2 \leq \frac{c_I}{n^2} \right\}, \tag{5}$$

where c_I is a small enough constant depending only on η. We will prove the following theorem.

Theorem 1. *For all $t \in [1, 2]$, the following bound holds, up to a constant depending only on η and t:*

$$\rho^* \asymp_{\eta,t} \sqrt{\frac{\left\| p_{\leq I} \right\|_r}{n}} + \frac{\left\| p_{>I} \right\|_1^{(2-t)/t}}{n^{(2t-2)/t}} + \frac{1}{n},$$

where we recall that $I = I(n, p, t)$.

The lower bounds and the minimax test are given in Section 5 and Section 6.

3.1. Equivalence between the Binomial, the multinomial and the Poisson setting

We now move to the multinomial and Poisson settings. In the following propositions, we state that the multinomial and the multivariate Binomial model are equivalent to the multivariate Poisson setting after using the *Poissonization trick*, and that the results from the binomial setting can be transferred to the other two settings. The Poissonization trick consists in drawing $\widetilde{n} \sim \mathrm{Poi}(n)$ observations instead of n, either from the multinomial or from the multivariate binomial model. The resulting data is exactly distributed as a multivariate Poisson family.

Proposition 1 (Poissonization trick for multinomials). *Let $n \geq 2$ and assume that p, q are probability vectors, i.e. such that*

$$\sum_i p_i = \sum_i q_i = 1.$$

Let $\widetilde{n} \sim \mathrm{Poi}(n)$. Conditional on $\widetilde{n}$, let $Z_1, \dots, Z_{\widetilde{n}} \overset{\text{i.i.d.}}{\sim} \mathcal{M}(q)$. We build the histogram sufficient statistic by defining, for all $j = 1, \dots, N$,

$$H_j = \sum_{i=1}^{\widetilde{n}} \mathbb{1}\{Z_i = j\}.$$

Then for all j, $H_j \sim \mathrm{Poi}(nq_j)$ and $H_1, \dots, H_N$ are mutually independent.

Proposition 2 (Poissonization trick for binomial families). *Let $n \geq 2$ and $\widetilde{n} \sim \mathrm{Poi}(n)$. Conditionally on $\widetilde{n}$, let $X_1, \dots, X_{\widetilde{n}} \overset{\text{i.i.d.}}{\sim} \bigotimes_{j=1}^{N} \mathrm{Ber}(p_j)$. Then*

$$\sum_{i=1}^{\widetilde{n}} X_i \sim \bigotimes_{j=1}^{N} \mathrm{Poi}(np_j).$$

These two propositions are classical and follow from basic properties of the Poisson, multinomial, and Binomial distributions. We rewrite them here only to provide some context on the following equivalences.

Without loss of generality, assume that $p_1 \geq \cdots \geq p_N$. We consider the following settings:

Binomial case. This is the setting considered above. We define

$$\mathcal{P}^{(\mathrm{Bin})} = \{\mathrm{Ber}(p) \ ; \ p \in \mathbb{R}_+^N\},$$

where by convention,

$$\mathrm{Ber}(p) := \bigotimes_{j=1}^{N} \mathrm{Ber}(p_j).$$

We fix $p \in \mathcal{P}^{(\mathrm{Bin})}$ and suppose we observe $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(q)$ for $q \in \mathcal{P}^{(\mathrm{Bin})}$ unknown. We consider the *binomial* testing problem:

$$H_0^{(\mathrm{Bin})}: q = p \quad \text{vs.} \quad H_1^{(\mathrm{Bin})}: \begin{cases} q \in \mathcal{P}^{(\mathrm{Bin})}, \\ \|q - p\|_t \geq \rho. \end{cases}$$

Poisson case. We define

$$\mathcal{P}^{(\mathrm{Poi})} = \{\mathrm{Poi}(p) \ ; \ p \in \mathbb{R}_+^N\},$$

where by convention,

$$\mathrm{Poi}(p) := \bigotimes_{j=1}^{N} \mathrm{Poi}(p_j).$$

We fix $p \in \mathcal{P}^{(\mathrm{Poi})}$ and suppose we observe $Y_1, \dots, Y_n \overset{\text{i.i.d.}}{\sim} \mathrm{Poi}(q)$ for $q \in \mathcal{P}^{(\mathrm{Poi})}$ unknown. We consider the *Poisson* testing problem:

$$H_0^{(\mathrm{Poi})}: q = p \quad \text{vs.} \quad H_1^{(\mathrm{Poi})}: \begin{cases} q \in \mathcal{P}^{(\mathrm{Poi})}, \\ \|q - p\|_t \geq \rho. \end{cases}$$

Multinomial case. We define

$$\mathcal{P}^{(\mathrm{Mult})} = \left\{ \mathcal{M}(p) \mid p \in \mathbb{R}_+^N, \ \sum_{j=1}^{N} p_j = 1 \right\},$$

where $\mathcal{M}(p)$ denotes the multinomial distribution over $\{1, \dots, N\}$. We fix $p \in \mathcal{P}^{(\mathrm{Mult})}$ and suppose we observe $Z_1, \dots, Z_n \overset{\text{i.i.d.}}{\sim} \mathcal{M}(q)$ for $q \in \mathcal{P}^{(\mathrm{Mult})}$ unknown. We consider the *multinomial* testing problem:

$$H_0^{(\mathrm{Mult})}: q = p \quad \text{vs.} \quad H_1^{(\mathrm{Mult})}: \begin{cases} q \in \mathcal{P}^{(\mathrm{Mult})}, \\ \|q - p\|_{\mathcal{M},t} \geq \rho, \end{cases}$$

where for $x = (x_1, \dots, x_N)$:

$$\|x\|_{\mathcal{M},t} = \left[\sum_{j=2}^{N} |x_j|^t \right]^{1/t}$$

is the multinomial norm, defined without taking the first coordinate into account. Indeed, because of the shape constraint $\sum p_j = 1$, the first coordinate does not bring any information and can be deduced from the $N - 1$ coordinates.

For these three testing problems, we define respectively

$$\rho^*_{\mathrm{Bin}}(n, p, t, \eta), \quad \rho^*_{\mathrm{Poi}}(n, p, t, \eta), \quad \rho^*_{\mathrm{Mult}}(n, p, t, \eta)$$

for the minimax separation distances in the sense of equation (3), for each of the testing problems.

We state the following statement regarding the equivalence between all models.

Lemma 1 (Equivalence between the Binomial and Poisson settings). *Let $t \in [1, 2]$. There exist two absolute constants c_{BP}, $C_{\mathrm{BP}} > 0$ depending on η such that $\forall p \in [0, 1]^N$, $\forall n \geq 2\eta > 0$, we have*

$$c_{\mathrm{BP}} \rho^*_{\mathrm{Bin}}(n, p, t, \eta) \leq \rho^*_{\mathrm{Poi}}(n, p) \leq C_{\mathrm{BP}}\, \rho^*_{\mathrm{Bin}}(n, p, t, \eta).$$

Lemma 2 (Equivalence between multinomial and Poisson settings). *Let $t \in [1, 2]$. It holds that $\forall p \in [0, 1]^N$, $\forall n \geq 2\ \eta > 0$, if $\sum_{i=1}^N p_i = 1$, we have*

$$\rho^*_{\mathrm{Mult}}(n, p, t, \eta) \lesssim_\eta \rho^*_{\mathrm{Poi}}(n, p^{-\max}) \lesssim_\eta \rho^*_{\mathrm{Mult}}(n, p, t, \eta),$$

where $p^{-\max} := (p_2, \ldots, p_N)$.

This entails the following corollary regarding the minimax rates of testing in the multinomial model:

Corollary 1. *Let $t \in [1, 2]$. The minimax separation radii in the Poisson and multinomial cases are respectively given by*

$$\rho^*_{\mathrm{Poi}}(n, p, t, \eta) \asymp_\eta \sqrt{\frac{\|p_{\leq I}\|_r}{n}} + \frac{\|p_{>I}\|_1^{(2-t)/t}}{n^{(2t-2)/t}} + \frac{1}{n} \qquad \text{for } p \in \mathcal{P}^{(\mathrm{Poi})},$$

$$\rho^*_{\mathrm{Mult}}(n, p, t, \eta) \asymp_\eta \sqrt{\frac{\|p^{-\max}_{\leq I}\|_r}{n}} + \frac{\|p_{>I}\|_1^{(2-t)/t}}{n^{(2t-2)/t}} + \frac{1}{n} \quad \text{for } p \in \mathcal{P}^{(\mathrm{Mult})},$$

where we recall that $I = I(n, p, t)$.

Note that the upper bounds in the Poisson model are obtained using our tests on the Poisson vector, and the upper bounds in the multinomial model are obtained using our tests on the last $N - 1$ coordinates of the estimates of probabilities of each categories.

4. Discussion

In this entire section, we mostly discuss the multinomial setting – whose rates are given in Corollary 1 – which is the most studied setting in the literature. To alleviate notations, we will write $\rho^*(n, p)$ for the minimax separation distance in the multinomial model, dropping the dependence on η.

4.1. Locality of the results

In the present paper, we derive sharp local minimax rates of testing in the binomial, Poisson and multinomial settings. The locality property is a major aspect of the results: for each fixed p we identify the detection threshold *associated to* p, where p is allowed to be any distribution in the class. For related local results in the case of the ℓ_1 or ℓ_2 norm, see e.g. [9, 14, 28, 53]. This approach is less standard than the usual *global* approach, which consists in finding the largest detection threshold in the class, i.e. for the *worst case* of p; see e.g. [38] (in Russian), [29, 34, 39, 47]. Yet, local results can substantially improve global results: for instance, in the multinomial case and for the ℓ_2 norm, the global separation radius for an N-dimensional multinomial is classically $N^{-1/4}/\sqrt{n}$, and is reached in the case where p is uniform distribution. However, if $p = (1, 0, \dots, 0)$ is a Dirac multinomial, then from our results the rate of testing in ℓ_2 norm is $\frac{1}{n}$, hence much faster than the global rate. Even for fixed N, one can actually find a sequence of null distributions $p^{(n)}$ whose associated separation distance $\rho^*_{\mathrm{Mult}}(n, p^{(n)}, 2, \eta)$ reaches any rate $1/n^\alpha$ for any $1/2 \leq \alpha \leq 1$. This consequently improves the global rate even for less extreme discrete distributions than Dirac multinomials. To give an example, consider an exponentially decreasing multinomial distribution

$$p^{(n)} = \left(\frac{c}{n^{(2\alpha-1)j}}\right)_{j=1}^{N}$$

for the renormalizing constant

$$c = n^{2\alpha-1}\frac{1 - 1/n^{2\alpha-1}}{1 - 1/n^{(2\alpha-1)N}} \asymp n^{2\alpha-1}.$$

Then, evaluating the local rate in ℓ_2 (allowing us to consider the whole set of coefficients as the bulk, see Section 7.1 below), we get

$$\rho^*_{\mathrm{Mult}}(n, p^{(n)}, 2, \eta) \asymp_\eta \sqrt{\frac{\|p^{-\max}\|_2}{n}} + \frac{1}{n} \asymp_\eta \frac{1}{n^\alpha}.$$

4.2. Comparison with existing literature in the multinomial case

Our results are quite related to those of [53], which examines the multinomial testing problem for the ℓ_1 distance and in terms of sample complexity. More precisely, for a fixed N-dimensional multinomial distribution p, and for a fixed separation ρ, this work investigates the smallest number $n^*(p, \rho)$ of samples $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{M}(p)$ needed to ensure that the multinomial testing problem introduced in Section 3.1 has a minimax risk less than $2/3$, for a fixed separation distance $\rho > 0$. Formally this is defined as

$$n^*(p, \rho) = \min\left\{n \in \mathbb{N} : R^*_{\rho,p,t,n} \leq 2/3\right\},$$

where $R^*_{\rho,p,t,n}$ denotes here the minimax risk for the multinomial problem[2]. Note that the quantities n^* and ρ^* are dual, for $\eta = 2/3$.

The authors of [53] prove the following bounds to characterize the optimal sample complexity $n^*(p,\varepsilon)$ when given a fixed $\varepsilon > 0$:

$$\frac{1}{\varepsilon} + \frac{\|p_{-\varepsilon}^{-\max}\|_{2/3}}{\varepsilon} \lesssim n^*(p,\varepsilon) \lesssim \frac{1}{\varepsilon} + \frac{\|p_{-\varepsilon/16}^{-\max}\|_{2/3}}{\varepsilon}.$$

In the above bound, $p = (p_1, \dots, p_N)$, where

$$p_1 \geq \cdots \geq p_N \geq 0 \quad \text{and} \quad \sum_{i=1}^{N} p_i = 1.$$

For $\varepsilon > 0$, let J be the smallest index such that $\sum_{i>J} p_i \leq \varepsilon$. The notation $p_{-\varepsilon}^{-\max}$ denotes $(p_2, \dots, p_J)$.

We generalize the result in several respects:

- We consider the whole range of ℓ_t distances for t in the segment $[1,2]$ and characterize the *local* rates of testing in each case,
- We generalize the multinomial case to the graph case (binomial case) and to the Poisson setting, through the Poissonization trick.

In Appendix D, we justify that the upper and lower bounds from [53], when translated in terms of separation radius as in [9] actually match in the multinomial case, although claimed otherwise by the authors of [9] themselves. It was therefore unclear in the literature so far that matching upper and lower bounds on the critical radius were actually known in the case $t = 1$. All of these cases involve the following ideas. The distribution can be split into bulk (set of large coefficients, with a subgaussian phenomenon) and tail (set of small coefficients, with a subpoissonian phenomenon). To the best of our knowledge, the way we define the tail is new. It allows us to establish a clear cut-off between these two optimal sets, fundamentally differing through the behavior of the second order moment of p.

The present paper can be linked with [16], which considers instance optimal identity testing. Specifically, [16] obtains a different characterization of the sample complexity for the case $t = 1$, in terms of a fundamental quantity in the theory of interpolation of Banach spaces, known as Peetre's K-functional. This functional is defined for all $u > 0$ as

$$\kappa_p(u) = \inf_{p'+p''=p} \|p'\|_1 + u\|p''\|_2.$$

[2] See equation (2) for the definition of this quantity in the graph problem.

This paper proves that for fixed $\varepsilon \in (0,1)$, any test for testing identity to p needs at least $C\kappa_p^{-1}(1-2\varepsilon)$ samples in order to have a risk less than η, where $C > 0$ is a constant depending only on η. In Section 6.3, especially equation (14), this paper discusses the non-tightness of [53]. Note that their bound is not optimal either, but is incomparable to [53]. This paper also provides a testing algorithm considering separately tail and heavy elements of the distribution, as well as a lower bound that uses interpolation theory to divide the problem into two types of elements – the ℓ_1 contribution (heavy elements) and the ℓ_2 ones (uniform-like).

Building on this work, [3, Appendix D] provides a general reduction scheme showing how to perform instance-optimal one-sample testing, given a "regular" (non-instance optimal) one-sample testing algorithm (even only for uniformity testing). This applies in particular to local privacy, or testing under communication constraints, or even without constraints at all.

5. Lower bounds

We recall the definitions of r and b in equation (4). In what follows, index A is defined as

$$A = A_{p,t,n}(\eta) := \max\left\{a \le I : p_a^{b/2} \ge \frac{c_A}{\sqrt{n}\left(\sum_{i\le I} p_i^r\right)^{1/4}}\right\}, \tag{6}$$

where $c_A > 0$ is a small enough constant depending only on η. We adopt the convention that $\max \emptyset = -\infty$ and that $p_{\le -\infty} = \emptyset$ and $p_{>-\infty} = p$. We start by presenting the lower bound part of Theorem 1. We divide the analysis into two parts: a lower bound for the large coefficients of p (bulk) and a lower bound for the small coefficients of p (tail). The bulk will be defined as the set $p_{\le A}$ and the tail as $p_{>A}$.

5.1. Lower bound for the bulk

To prove the lower bound, we identify a radius ρ such that, if the ℓ_t distance between $\mathcal{H}_0$ and $\mathcal{H}_1$ is less than ρ, then any test has risk at least η. Therefore, by definition of ρ^*, ρ is necessarily a lower bound on ρ^*.

Proposition 3. *Let $t \in [1,2]$. There exists a constant $\underline{c}'_\eta > 0$ depending only on η, as well as a distribution q such that for any test ψ, we have*

$$\|(q-p)_{\le A}\|_t \ge \underline{c}'_\eta\left(\frac{\|p_{\le A}\|_r^{r/t}}{\sqrt{n}\|p_{\le I}\|_r^{r/4}} + \frac{1}{n}\right),$$

and

$$\mathbb{P}_p(\psi = 1) + \mathbb{P}_q(\psi = 0) \ge \eta.$$

This implies that

$$\rho = \frac{\| p_{\leq A} \|_r^{r/t}}{\sqrt{n} \| p_{\leq I} \|_r^{r/4}} + \frac{1}{n}$$

is a lower bound on the minimax separation radius ρ^*.

Note that the lower bound in $\frac{1}{n}$ is trivial since changing any entry of p by $\frac{1}{n}$ is not detectable with high probability. Now let us examine the first part of the rate. To prove this lower bound, we use Le Cam's two points method by defining a prior distribution over a discrete subset of $\mathcal{P}_N$ satisfying $\mathcal{H}_1$. More precisely, for all $(\delta_1, \ldots, \delta_A) \in \{\pm 1\}^A$, we define the distribution q_δ such that

$$(q_\delta)_j = \begin{cases} p_j + \delta_i \gamma_j & \text{if } j \leq A, \\ p_j & \text{otherwise,} \end{cases} \tag{7}$$

where, for some small enough constant $c_\gamma > 0$ depending only on η:

$$\gamma_i = \frac{c_\gamma \, p_i^{2/(4-t)}}{\sqrt{n} \left(\sum_{i \leq I} p_i^r \right)^{1/4}}. \tag{8}$$

The mixture

$$\overline{\mathbb{P}}_{\text{bulk}} = \frac{1}{2^A} \sum_{\delta \in \{\pm 1\}^A} q_\delta^{\otimes n}$$

defines a probability distribution over the set of observations $X_1, \ldots, X_n$, such that, conditional on $\delta \in \{\pm 1\}^A$, the observations are i.i.d. with probability distribution q_δ.

The core of the proof is to prove that observations $X_1, \ldots, X_n$ drawn from this mixture distribution $\overline{\mathbb{P}}_{\text{bulk}}$ are so difficult to distinguish from observations $X'_1, \ldots, X'_n$ drawn from $\mathbb{P}_p$, that the risk of any test is necessarily larger than η. This brings us to the conclusion of our proposition since any distribution q_δ is separated away from p by an ℓ_t distance equal to

$$\left(\sum_{i=1}^{A} \gamma_i^t \right)^{1/t} \asymp \frac{\| p_{\leq A} \|_r^{r/t}}{\sqrt{n} \| p_{\leq I} \|_r^{r/4}}.$$

Therefore,

$$\frac{\| p_{\leq A} \|_r^{r/t}}{\sqrt{n} \| p_{\leq I} \|_r^{r/4}}$$

is necessarily a lower bound on the separation radius ρ^* This lower bound is an extension to the case where $t \in [1, 2]$ of the lower bound in [53] which is given for the case $t = 1$, up to some issues that are discussed in details in Section 4.2.

5.1.1. Lower bound for the tail. We now derive a lower bound for the tail $p_{>A}$, containing the smallest coefficients of p. The tail lower bound involves very different phenomena compared to the above bulk lower bound. The reason is that the definition of A implies that on the tail, with high probability, no same coordinate is observed twice or more among the n data.

Proposition 4. *Let $t \in [1,2]$, and consider any test ψ. There exists a constant $\underline{c}'_\eta > 0$ depending only on η and a distribution Q such that*

$$\|(q-p)_{>A}\|_t \geq \underline{c}'_\eta \frac{\|p_{>I}\|_1^{(2-t)/t}}{n^{(2t-2)/t}},$$

and

$$\mathbb{P}_p(\psi = 1) + \mathbb{P}_q(\psi = 0) \geq \eta.$$

To prove this lower bound, we once more use Le Cam's two points method with a *sparse* prior distribution. Define the smallest index $U > I$ such that

$$n^2 p_U \|P_{\geq U}\|_1 \leq c_u < 1,$$

where $c_u > 0$ is a small constant defined in the appendix. We define

$$\overline{\pi} = \frac{c_u}{n^2 \|p_{\geq U}\|_1} \quad \text{and} \quad \pi_i = \frac{p_i}{\overline{\pi}}.$$

Index U has no further meaning than to guarantee that for all $i \geq U$, we have $\pi_i \in [0,1]$. In particular, π_i is a Bernoulli parameter. Now, we define the following prior on q. For any $i < U$, we set $q_i = p_i$. Otherwise, for $i \geq U$, we set $b_i \sim \mathrm{Ber}(\pi_i)$ mutually independent, and

$$q_b(i) = b_i \overline{\pi}, \tag{9}$$

We now consider the mixture of the probability distributions q_b:

$$\overline{\mathbb{P}}_{\text{tail}} = \sum_{b \in \{0,1\}^{\{U+1,\dots,N\}}} \left(\prod_{j>U} \pi_j^{b_j} (1-\pi_j)^{1-b_j} \right) q_b^{\otimes n}.$$

As above, we prove that the data $X_1, \dots, X_n$ drawn from this mixture $\overline{\mathbb{P}}_{\text{tail}}$ is difficult to distinguish from the data $X'_1, \dots, X'_n$ drawn from $\mathbb{P}_p$. Moreover, we show that with high probability, the ℓ_t distance between $\overline{\mathbb{P}}_{\text{tail}}$ and p, is larger, up to an absolute constant than

$$\frac{\|p_{\geq U}\|_1^{(2-t)/t}}{n^{2(t-1)/t}}.$$

Finally, to conclude the proof, we show in Lemma 8 that

$$\frac{\|p_{\geq U}\|_1^{(2-t)/t}}{n^{2(t-1)/t}} + \frac{1}{n} \asymp_\eta \frac{\|p_{>I}\|_1^{(2-t)/t}}{n^{2(t-1)/t}} + \frac{1}{n}$$

in words, that we can replace U by I. This lower bound departs significantly from the one in [53] in the case $t = 1$, which is significantly simpler than for $t > 1$ for the tail coefficients.

5.1.2. Combination of both lower bounds. By combining Propositions 3 and 4, we obtain the following theorem.

Theorem 2. *Let $t \in [1, 2]$, and consider any test ψ. There exists a constant $\underline{c}'_\eta > 0$ depending only on η and a distribution q such that*

$$\|q - p\|_t \geq \underline{c}'_\eta \left(\sqrt{\frac{\|p_{\leq I}\|_r}{n}} + \frac{\|p_{>I}\|_1^{(2-t)/t}}{n^{(2t-2)/t}} + \frac{1}{n} \right),$$

and

$$\mathbb{P}_p(\psi = 1) + \mathbb{P}_q(\psi = 0) \geq \eta.$$

This theorem implies that

$$\rho^* \gtrsim_\eta \sqrt{\frac{\|p_{\leq I}\|_r}{n}} + \frac{\|p_{>I}\|_1^{(2-t)/t}}{n^{(2t-2)/t}} + \frac{1}{n},$$

which is a lower bound on the separation radius ρ^*, up to a positive constant depending only on η.

Note that when combining Propositions 3 and 4, we do not get exactly the expression in Theorem 2. We actually obtain:

We therefore need to show that this expression is equivalent to that in Theorem 2. This is done by using Lemma 9, which states that we can replace

$$\frac{\|p_{\leq A}\|_r^{r/t}}{\sqrt{n}\|p_{\leq I}\|_r^{r/4}} \quad \text{by} \quad \sqrt{\frac{\|p_{\leq I}\|_r}{n}}$$

without changing the rate, i.e.

$$\frac{\|p_{\leq A}\|_r^{r/t}}{\sqrt{n}\|p_{\leq I}\|_r^{r/4}} + \frac{\|p_{>I}\|_1^{(2-t)/t}}{n^{(2t-2)/t}} + \frac{1}{n} \quad \asymp_\eta \quad \sqrt{\frac{\|p_{\leq I}\|_r}{n}} + \frac{\|p_{>I}\|_1^{(2-t)/t}}{n^{(2t-2)/t}} + \frac{1}{n}.$$

Remark on index A. As explained in (7), the optimal prior is of the form $p_i \pm \gamma_i$ where γ_i is proportional to $p_i^{2/(4-t)}$, according to equation (8). Since $2/(4-t) \leq 1$, we can have $\gamma_i > p_i$ if p_i is too small, so that it is impossible to set the optimal prior $p_i \pm \gamma_i$, since $p_i - \gamma_i$ has to be a Bernoulli parameter. The index A is just the last index ensuring $p_A \geq \gamma_A$ so that our lower bound construction is well-defined.

Remark on index I. Index I defines the largest set of coefficients $p_{>I}$ such that under $\mathcal{H}_0$, with high probability, no coordinate $j > I$ is observed twice or more. This is exactly the interpretation of the relation

$$\sum_{j>I} n^2 p_j^2 \leq c_I$$

for a small constant c_I. As shown in Lemma 13, it is important that the definition of A also implies that

$$\sum_{j>A} n^2 p_j^2 \leq c_I + c_A^4,$$

which leads us to tune the constants c_I and c_A such that this sum is small. Therefore, on the actual tail $(p_{>A})$, no same coordinate will be observed twice with high probability under $\mathcal{H}_0$. This is the reason why the phenomena involved are different on the bulk and on the tail. On the bulk, many coordinates are observed at least twice, which allows us to build an estimator based on the *dispersion* of the data around its mean, namely the renormalized χ^2 estimator which is a modified estimator of the variance. Like in the classical Gaussian signal detection setting, the optimal procedure for detecting whether or not the data is drawn from p is to estimate the dispersion of the data.

On the tail, however, each coordinate is observed at most once under $\mathcal{H}_0$, so that the *dispersion* of the data cannot be estimated. On this set, we rather design a prior distribution which mimics the behavior of the null distribution, while being as separated from p as possible. More precisely, we impose that with high probability, no coordinate is observed twice, and such that coordinate-wise, the expected number of observations is equal to that under the null hypothesis p. This prior is therefore designed such that its first order moment is equal to that under the null and its second order moment is unobserved with high probability. Under both of these constraints, we maximize the ℓ_t distance between the null hypothesis p and the possible distributions composing the prior. When $t > 1$, the result of this process is a prior that needs to be relatively sparse - which is significantly more involved than the case $t = 1$ treated in [53].

Remark on the lower bounds. The bulk lower bound is close to that of [53]. The tail lower bound relies on a sparse prior that is an existing technique (for example, in sparse testing, see [10, 23, 42]) and is very different from the construction in [53]. Handling the indices I, A and U require careful manipulations that we believe are new techniques.

6. Upper bounds

Recalling that $n = 2k$, we use sample splitting to define

$$S = \sum_{i=1}^{k} X_i \quad \text{and} \quad S' = \sum_{i=k+1}^{n} X_i.$$

We also write

$$b = \frac{4-2t}{4-t}.$$

6.1. Test for the bulk

We now introduce the following test statistic on the bulk coefficients, i.e. the coefficients with index smaller than A:

$$T_{\text{bulk}} = \sum_{i \le A} \frac{1}{p_i^b} \left(\frac{S_i}{k} - p_i \right) \left(\frac{S_i'}{k} - p_i \right), \tag{10}$$

which is a weighted χ^2 statistic. We now define the test

$$\psi_{\text{bulk}} = \mathbf{1}\left\{ T_{\text{bulk}} > \frac{c_\eta}{n} \| p_{\le A} \|_r^{r/2} \right\},$$

where $c_\eta = 4/\sqrt{\eta}$ is a large enough constant, depending only on η. We prove the following proposition regarding this statistic and the bulk of the vector p.

Proposition 5. *There exists $c'_\eta > 0$, such that the following holds:*

- *Type-I error is bounded:*

$$\mathbb{P}_p(\psi_{\text{bulk}} = 1) \le \eta/2.$$

- *Type-II error is bounded: for any q, such that*

$$\| q_{\le A} \|_t \ge c'_\eta \left(\sqrt{\frac{\| p_{\le I} \|_r}{n}} + \frac{1}{n} \right),$$

it holds that

$$\mathbb{P}_q(\psi_{\text{bulk}} = 0) \le \eta/2.$$

For $t = 1$, we get $r = \frac{2}{3}$, which is the norm identified in [53]. However, our setting is slightly different for three reasons. First, we consider multivariate binomial families rather than multinomials. Second, we consider separation distance for a fixed n instead of sample complexity. Third, our result holds for any $t \in [1, 2]$. However, in Section 3.1, we prove that multivariate binomial and multinomial settings are related and that the rates can be transferred from our setting to the multinomial case.

Note that our cut-off is defined differently from that in [53]. In [53], the cut-off I' is the smallest index such that, for a fixed ε, we have

$$\sum_{i>I'} p_i \leq \varepsilon.$$

This definition therefore only involves the first order moment of the null distribution. In our setting, conversely, we define index I using the second order moment of the null distribution, as the smallest index such that

$$\sum_{i>I} p_i^2 \leq \frac{c_I}{n^2}.$$

The above result also generalizes the bound identified in [53], by characterizing the testing rate for all $t \in [1, 2]$ and sheds light on a duality between the ℓ_t and ℓ_r norms when $r = 2t/(4-t)$.

6.2. Test for the tail coefficients

The tail test is a combination of two tests. We define the histogram of the data which is a sufficient statistic:

$$\forall j > A, \quad N_j := \sum_{i=1}^{n} \mathbb{1}\{X_i = j\}.$$

We first define the test ψ_2, which rejects $\mathcal{H}_0$ whenever one tail coordinate is observed twice:

$$\psi_2 = \mathbb{1}\big\{\exists j > A : N_j \geq 2\big\}. \tag{11}$$

We also define a statistic counting the number of observations on the tail, and the associated test, recalling that $\underline{c}_\eta = 4/\sqrt{\eta}$:

$$T_1 = \sum_{i>A} \frac{N_i}{n} - p_i, \quad \psi_1 = \mathbf{1}\left\{|T_1| > \underline{c}_\eta \sqrt{\frac{\sum_{i>A} p_i}{n}}\right\}. \tag{12}$$

We prove the following proposition regarding this statistic.

Proposition 6. *There exists $\underline{c}'_\eta > 0$, such that the following holds.*

- *Type-I error is bounded:*

$$\mathbb{P}_p(\psi_1 \vee \psi_2 = 1) \leq \eta/2.$$

- *Type-II error is bounded: for any q such that*

$$\|q_{>A}\|_t \geq \underline{c}'_\eta \left(\frac{\|p_{>A}\|_1^{(2-t)/t}}{n^{(2t-2)/t}} + \frac{1}{n} \right),$$

it holds that

$$\mathbb{P}_q(\psi_1 \vee \psi_2 = 0) \leq \eta/2.$$

Recall that the tail is defined such that, with high probability under $\mathcal{H}_0$, no same coordinate is observed at least twice. We therefore combine two tests: The test ψ_2 rejects $\mathcal{H}_0$ if one of the coordinates is observed at least twice, while the test ψ_1 rejects $\mathcal{H}_0$ if the total mass of observed coordinates differs substantially from its expectation under the null. Proposition 6 proves that this combination of tests reaches the optimal rate.

In [53], the tail test only involves the first order moment, which is sufficient in the case of the ℓ_1 norm. Moreover, in the proof of Proposition 6, it becomes clear that for $t = 1$ we only need the test ψ_1 and for $t = 2$ we only need the test ψ_2. However, in the case of the ℓ_t for $t \in (1, 2)$, the combination of both ψ_1 and ψ_2 is necessary.

6.3. Aggregated test

We now combine the above results to define the aggregated test. We define our test as

$$\psi = \psi_{\text{bulk}} \vee \psi_1 \vee \psi_2.$$

This is the test rejecting the null whenever one of the three tests does. Denote by

$$\overline{\rho} = \sqrt{\frac{\|p_{\leq I}\|_r}{n}} + \frac{\|p_{>I}\|_1^{(2-t)/t}}{n^{(2t-2)/t}} + \frac{1}{n}.$$

The following theorem states that this test reaches the rate $\overline{\rho}$, which is the minimax rate ρ^* given in Theorem 1. In other words, it guarantees that, whenever the two hypotheses are $\overline{\rho}$-separated in ℓ_t distance, this test has type-I and type-II errors upper bounded by $\eta/2$, ensuring that its risk is less than η. Since the minimax separation radius ρ^* is the smallest radius ensuring the existence of a test satisfying this condition, we can conclude that $\rho^* \lesssim \overline{\rho}$.

Theorem 3. *There exists $\underline{c}'_\eta > 0$, such that the following holds.*

- *The type-I error is bounded:*

$$\mathbb{P}_p(\psi = 1) \leq \eta/2.$$

- *The type-II error is bounded: for any q such that*

$$\|p - q\|_t \geq \underline{c}'_\eta \left(\sqrt{\frac{\|p_{\leq I}\|_r}{n}} + \frac{\|p_{>I}\|_1^{(2-t)/t}}{n^{(2t-2)/t}} + \frac{1}{n} \right),$$

it holds that

$$\mathbb{P}_q(\psi = 0) \leq \eta/2.$$

6.4. Remarks on the tests

In the bulk tests, we propose test statistics based on sample splitting, whose variance is easier to express. However, those tests could be defined slightly differently without sample splitting, allowing also for the analysis of the case $n = 1$. Denoting by H the histogram of the data, we could define

$$\widetilde{T}_{\text{Bulk}} = \sum_{j \leq A} \frac{1}{p_j^b} \left[\left(\frac{H_j}{n} - p_j \right)^2 - H_j \right]$$

and the associated test

$$\widetilde{\psi}_{\text{bulk}} = \mathbf{1}\left\{ \widetilde{T}_{\text{bulk}} > \frac{\underline{c}_\eta}{n} \| p_{\leq A} \|_r^{r/2} \right\}.$$

This test attains the same upper bound in terms of separation distance – up to multiplicative constants depending on η – as the bulk test we define in equation (10), and is therefore also optimal in the bulk regime.

To understand the interpolation between the extreme cases $t = 1$ and $t = 2$, an important remark is that the tail tests ψ_1 and ψ_2 do not capture the same signals. Under the alternative hypothesis, the test ψ_1 checks that the total mass of the tail coefficients $\|q_{>A}\|_1$ is not to far away from $\|p_{>A}\|_1$. As to test ψ_2, *on the tail*, that is, on a set for which

$$\sum_{j>A}^{N} n^2 p_j^2 \ll 1,$$

it is actually equivalent to using a test for the second order moment. In other words, the test ψ_2 is equivalent to $\widetilde{\psi}_2 = \mathbf{1}\{|T_2| > \frac{\underline{c}_\eta}{n} \|p_{>A}\|_2\}$ for a small constant $\underline{c}_\eta$, where

$$T_2 = \sum_{i>A} \left(\frac{S_i}{k} - p_i \right) \left(\frac{S_i'}{k} - p_i \right).$$

Therefore, the test ψ_2 checks that the second order moment of the tail of distribution $q_{>A}$ is not too different from that of $p_{>A}$, in other words, that it does not contain much greater coefficients than the corresponding values of $p_{>A}$.

7. Further remarks on the results

7.1. Influence of the ℓ_t norm

In this paper, we consider the separation distance in all ℓ_t norms for $t \in [1, 2]$. The choice of t influences the minimax separation distance. The effect of the L^t separation distance for $t \in [1, 2]$ has also been investigated in the paper [22] in the case of goodness-of-fit testing for Hölder continuous densities.

In the extreme case $t = 2$, the minimax separation distance reduces to

$$\rho^* \asymp_\eta \sqrt{\frac{\|p_{\le I}\|_2}{n}} + \frac{1}{n},$$

which can be further simplified as

$$\rho^* \asymp_\eta \sqrt{\frac{\|p\|_2}{n}} + \frac{1}{n}.$$

Indeed, by definition of I, we have $\|p_{>I}\|_2 \lesssim_\eta \frac{1}{n}$. This case has already been solved in [21]. In this case, as discussed earlier, a simple χ^2 test would suffice for reaching this separation distance, and p would only appear in the definition of the threshold of this test. Here we therefore do not need to combine a bulk with a tail test. A single χ^2 test, applied on both the bulk and the tail (i.e. setting $A = N$), would suffice.

We now consider the opposite extreme case $t = 1$. In this case

$$\rho^* \asymp_\eta \sqrt{\frac{\|p_{\le I}\|_{2/3}}{n}} + \|p_{>A}\|_1 + \frac{1}{n}.$$

In the minimax separation distance, the contribution of the Bulk coefficients involves the $\ell_{2/3}$ quasi-norm, as in [53]. In terms of test statistic, this is reflected by the fact that the optimal Bulk test is based on a re-weighted χ^2 test statistic whose weights depend on p. For each entry j, the optimal weight is larger when p_j is small: indeed, for small p_j, coordinate j has smaller variance. This re-weighting differs from the extreme case $t = 2$, since, compared to the ℓ_2 norm, the ℓ_1 norm lays more emphasis on smaller entries of the perturbation $p - q$. As to the tail coefficients, however, the big picture is simpler as the minimax rate with respect to the tail coefficients is $\|p_{>A}\|_1$, which is very large. This rate implies in particular that only the total mass of the perturbations of the tail coefficients matters. We therefore do not need to use the test ψ_2, which is tailored to detect extreme values of the perturbations, and can only restrict to using ψ_1 when it comes to the tail coefficients.

Between the two extreme cases, that is, for $t \in (1, 2)$, we have an interpolation between the two extreme scenarios. When it comes to the bulk, we need to re-weight the test statistics by weights that increase with p_i for entry i as in the case $t = 1$. But the larger t, the milder the reweighting – as the ℓ_t norm puts more weight on large coefficients – until it vanishes for $t = 2$. As for the tail, both tests ψ_1 and ψ_2 are required in this intermediate regime. Indeed, we need to control both the mass of the tail perturbations like for $t = 1$, but also their extreme values like for $t = 2$. Note that [55] had already considered the global problem of ℓ_t testing for discrete distributions and identified (non-matching) upper and lower bounds.

For $t > 2$, the underlying phenomenon is fundamentally different. In this case, the ℓ_t norm emphasizes so much the large deviations that re-weighted χ^2 tests – that

are related to re-weighted second order moment estimation – seem to be sub-optimal for testing. We leave the case $t > 2$ as an open problem.

In the minimax separation distance in ℓ_t norm, the bulk part $\sqrt{\|p_{\le I}\|_r/n}$ involves a duality between the norms ℓ_t and ℓ_r for $r = 2t/(4-t)$, as was also the case for $t = 1$ in [53]. This phenomenon comes from a combination of Hölder's inequality and information theory. Define $\gamma = (\gamma_1, \dots, \gamma_A) \in [0,1]^A$, and define the random vector

$$q = (p_1 + \delta_1\gamma_1, \dots, p_A + \delta_A\gamma_A)$$

for $\delta_i \overset{\text{i.i.d.}}{\sim} \text{Rad}(\frac{1}{2})$ like in (7), except that this time, we *do not* impose that $(\gamma_i)_i$ is defined as in (8). Introduce

$$\Gamma := \left\{(\gamma_1, \dots, \gamma_A) \in [0,1]^A : \sum_{i=1}^A \frac{\gamma_i^4}{p_i^2} \le \frac{C_\gamma}{n^2};\ p_i - \gamma_i \in [0,1],\ p_i + \gamma_i \in [0,1]\right\},$$

where C_γ is a small enough constant depending only on η. Then by Lemma 4 in the appendix, whenever $\gamma \in \Gamma$, the n samples[3] generated from the random vector q have a probability distribution indistinguishable from the null hypothesis p. The largest $\gamma \in \Gamma$, when measured in ℓ_t, therefore provides a lower bound on the minimax separation radius. It is found by solving

$$\max_{\gamma\in\Gamma} \sum_{i=1}^A \gamma_i^t,$$

which can be done using Hölder's inequality

$$\sum_{i=1}^A \gamma_i^t = \sum_{i=1}^A \left(\frac{\gamma_i^4}{p_i^2}\right)^{t/4} p_i^{t/2} \underset{\text{Hölder}}{\le} \left(\sum_{i=1}^A \frac{\gamma_i^4}{p_i^2}\right)^{t/4} \left(\sum_{i=1}^A p_i^r\right)^{(4-t)/4} \le \left(\frac{C_\gamma}{n^2}\right)^{t/4} \|p\|_r^{1/2t},$$

where we have used Hölder's inequality with $a = 4/t$ and $b = 4/(4-t)$. Setting γ^* the vector on the frontier of Γ reaching the equality case in Hölder's inequality, for fixed n, we obtain $\|\gamma^*\|_t \propto \|p\|_r^{1/2}$.

As to the contribution of the tail, we refer the reader to the remarks below Proposition 4.

7.2. Asymptotics as $n \to \infty$

Consider now p as being a *fixed* multinomial distribution, or a fixed vector of Poisson parameters. Then by the definitions of A and I, there exists an integer n_0 such that

[3] Although the proof is written for graph samples, it is argued in Section 3.1 that it can be transposed to the multinomial or the Poisson settings.

for all $n \geq n_0$, we have $I = A = N$. In words, we eventually no longer need to split the distribution into bulk and tail and we can define the bulk as the whole set of coefficients. For n large enough ($n \geq n_0$), the local minimax rate therefore rewrites

$$\rho^*(p,n) \underset{n\to\infty}{\asymp} \begin{cases} \sqrt{\|p^{-\max}\|_r/n} + \frac{1}{n} & \text{in the multinomial case,} \\ \sqrt{\|p\|_r/n} + \frac{1}{n} & \text{in the binomial or Poisson case.} \end{cases}$$

On the other hand the fast rate $\frac{1}{n}$ asymptotically dominates if p is close to a Dirac multinomial distribution in the multinomial setting, or if e.g. $p = 0$ in the binomial and Poisson setting.

A. Lower bound

Let $p \in \mathcal{P}_N$. For $\mathcal{P}_1 := \mathcal{P}_1(\rho)$ a particular collection of elements of $\mathcal{P}_N$ satisfying $\mathcal{H}_{1,\rho}$, we denote by $\mathcal{U}(\mathcal{P}_1)$ the uniform distribution over $\mathcal{P}_1$.

Let $\mathcal{G} = (\{0,1\}^N)^n$ be the set of all possible observations $(X_1, \dots, X_n)$, where $X_i = (X_i(1), \dots, X_i(N))$. The following lemma gives a way to derive a lower bound on ρ^* by giving a sufficient condition, for a fixed ρ, that $R^*(\rho) \geq \eta$.

Lemma 3. *If*

$$\frac{1}{|\mathcal{G}|} \sum_{\mathbf{X}\in\mathcal{G}} \frac{(\mathbb{E}_{q\sim\mathcal{U}(\mathcal{P}_1)}\mathbb{P}_q(X))^2}{\mathbb{P}_p(X)} \leq 1 + 4(1-\eta)^2,$$

then $R^*(\rho) \geq \eta$.

Proof of Lemma 3. We have that

$$\begin{aligned}
R^*(\rho) &\geq \inf_{\psi \text{ test}} \mathbb{P}_p(\psi = 1) + \sup_{q\in\mathcal{P}_1} \mathbb{P}_q(\psi = 0) \qquad \text{(all elements of } \mathcal{P}_1 \text{ satisfy } \mathcal{H}_1) \\
&\geq \inf_{\psi \text{ test}} \mathbb{P}_p(\psi = 1) + \mathbb{E}_{q\sim\mathcal{U}(\mathcal{P}_1)}\mathbb{P}_q(\psi = 0) \\
&\qquad\qquad \text{(the supremum is greater than the integral)} \\
&= 1 + \inf_{\psi \text{ test}} \mathbb{P}_p(\psi = 1) - \mathbb{E}_{q\sim\mathcal{U}(\mathcal{P}_1)}\mathbb{P}_q(\psi = 1) \\
&= 1 - \sup_{\psi \text{ test}} |\mathbb{P}_p(\psi = 1) - \mathbb{E}_{q\sim\mathcal{U}(\mathcal{P}_1)}\mathbb{P}_q(\psi = 1)| \\
&= 1 - d_{TV}(\mathbb{P}_p, \mathbb{E}_{q\sim\mathcal{U}(\mathcal{P}_1)}\mathbb{P}_q) \\
&\geq 1 - \frac{1}{2}\sqrt{\chi^2(\mathbb{E}_{q\sim\mathcal{U}(\mathcal{P}_1)}\mathbb{P}_q || \mathbb{P}_p)},
\end{aligned}$$

where the definition of the χ^2 divergence can be found in [52], as well as the proof for the inequality $d_{TV} \leq \frac{1}{2}\sqrt{\chi^2}$. Therefore,

$$\begin{aligned} R^*(\rho) &\geq 1 - \frac{1}{2}\sqrt{\chi^2(\mathbb{E}_{q\sim\mathcal{U}(\mathcal{P}_1)}\mathbb{P}_q||\mathbb{P}_p)} \\ &= 1 - \frac{1}{2}\sqrt{\frac{1}{|\mathcal{G}|}\sum_{X\in\mathcal{G}}\frac{(\mathbb{E}_{q\sim\mathcal{U}(\mathcal{P}_1)}\mathbb{P}_q(X))^2}{\mathbb{P}_p(X)} - 1}. \end{aligned}$$

Therefore, to have $R^*(\rho) \geq \eta$ it suffices that

$$\frac{1}{|\mathcal{G}|}\sum_{X\in\mathcal{G}}\frac{(\mathbb{E}_{q\sim\mathcal{U}(\mathcal{P}_1)}\mathbb{P}_q(X))^2}{\mathbb{P}_p(X)} \leq 1 + 4(1-\eta)^2. \qquad \blacksquare$$

For all $i = 1, \ldots, N$, let $\gamma_i \in [0, p_i]$ and let $\gamma = (\gamma_i)_i$. We now apply the previous lemma with

$$\mathcal{P}_1 = \{p + (\delta_i\gamma_i)_{i\leq N} \mid \delta \in \{\pm 1\}^N\}.$$

Lemma 4. *There exists a sufficiently small absolute constant c_4 such that, if*

$$\sum_{i=1}^N \frac{\gamma_i^4}{p_i^2} \leq \frac{c_4}{n^2},$$

then for all $\rho \leq \|\gamma\|_t$, we have $R^*(\rho) \geq \eta$.

Proof. We will use Lemma 3 with p and $\mathcal{P}_1$ defined as above.

- We first compute $\mathbb{P}_q(X)$ for some realization $X \in \mathcal{G}$. Let

$$S = \sum_{i=1}^n X_i \in \{0, \ldots, n\}^N$$

and write $S = (s_1, \ldots, s_N)$. We have that

$$\mathbb{P}_p(X) = \prod_{i=1}^N p_i^{s_i}(1-p_i)^{n-s_i}.$$

- We now compute $\mathbb{E}_{q\sim\mathcal{U}(\mathcal{P}_1)}\mathbb{P}_q(X)$: for any $(\delta_i)_i \in \{\pm 1\}^N$, we define

$$q_\delta = p + (\delta_i\gamma_i)_{1\leq i\leq N}.$$

Then we have

$$\mathbb{P}_{q_\delta}(X) = \prod_{i=1}^N (p_i + \delta_i\gamma_i)^{s_i}(1 - p_i - \delta_i\gamma_i)^{n-s_i}.$$

Therefore, we have

$$\begin{aligned}
&\frac{1}{|\mathcal{G}|}\sum_{X\in\mathcal{G}}\frac{(\mathbb{E}_{q\sim\mathcal{U}(\mathcal{P}_1)}\mathbb{P}_q(X))^2}{\mathbb{P}_p(X)}\\
&\quad=\frac{1}{|\mathcal{G}|}\sum_{X\in\mathcal{G}}\sum_{\delta,\delta'}\prod_{i=1}^{N}\frac{(p_i+\delta_i\gamma_i)^{s_i}(1-p_i-\delta_i\gamma_i)^{n-s_i}}{p_i^{s_i}(1-p_i)^{n-s_i}}\\
&\qquad\times(p_i+\delta_i'\gamma_i)^{s_i}(1-p_i-\delta_i'\gamma_i)^{n-s_i}\\
&\quad=\frac{1}{|\mathcal{G}|}\sum_{\delta,\delta'}\prod_{i=1}^{N}\sum_{l=0}^{n}\binom{n}{l}\left(p_i+(\delta_i+\delta_i')\gamma_i+\frac{\delta_i\delta_i'\gamma_i^2}{p_i}\right)^{l}\\
&\qquad\times\left(1-p_i-(\delta_i+\delta_i')\gamma_i+\frac{\delta_i\delta_i'\gamma_i^2}{1-p_i}\right)^{n-l}\\
&\quad=\frac{1}{|\mathcal{G}|}\sum_{\delta,\delta'}\prod_{i=1}^{N}\left(1+\frac{\delta_i\delta_i'\gamma_i^2}{p_i(1-p_i)}\right)^{n}=\prod_{i=1}^{N}\left[\frac{1}{4}\sum_{\delta_i,\delta_i'\in\{\pm1\}}\left(1+\frac{\delta_i\delta_i'\gamma_i^2}{p_i(1-p_i)}\right)^{n}\right]\\
&\quad=\prod_{i=1}^{N}\left[\frac{1}{2}\left(1+\frac{\gamma_i^2}{p_i(1-p_i)}\right)^{n}+\frac{1}{2}\left(1-\frac{\gamma_i^2}{p_i(1-p_i)}\right)^{n}\right]\\
&\quad\le\prod_{i=1}^{N}\left[\frac{1}{2}\exp\left(\frac{n\gamma_i^2}{p_i(1-p_i)}\right)+\frac{1}{2}\exp\left(\frac{-n\gamma_i^2}{p_i(1-p_i)}\right)\right]\\
&\quad=\prod_{i=1}^{N}\cosh\left(\frac{n\gamma_i^2}{p_i(1-p_i)}\right)\le\exp\left(\sum_{i=1}^{N}\frac{n^2\gamma_i^4}{2p_i^2(1-p_i)^2}\right).
\end{aligned}$$

Note that

$$\begin{aligned}
\exp\left(\sum_{i=1}^{N}\frac{n^2\gamma_i^4}{2p_i^2(1-p_i)^2}\right)&\le 1+4(1-\eta)^2\\
&\Leftrightarrow\quad\sum_{i=1}^{N}\frac{\gamma_i^4}{p_i^2(1-p_i)^2}\le\frac{2c_A^4}{n^2}\\
&\Leftarrow\quad\sum_{i=1}^{N}\frac{\gamma_i^4}{p_i^2}\le\frac{c_A^4}{2n^2},
\end{aligned}\tag{13}$$

where $c_A^4:=\log(1+4(1-\eta)^2)$ and since $\forall\, i$, we have $p_i\le\frac{1}{2}$. The result follows by Lemma 3. ∎

This means the following: let $\gamma:=(\gamma_i)_i$ satisfying (13) and let $\rho=\|\gamma\|_t$. Then all points $p+(\delta_i\gamma_i)_{1\le i\le|\mathcal{G}|}$ are located at a distance ρ from p in terms of ℓ_t norm, so that the corresponding adjacency matrices are at a distance ρ from each other in ℓ_t norm.

Moreover, we proved that for the uniform prior on this set of points $\mathcal{P}_1$, we have $R^*(\rho) \geq \eta$, which yields $\rho^* \geq \rho$.

We now prove the lower bound by combining Lemmas 5–10.

Lemma 5. *It holds that*

$$\rho_t^* \gtrsim_\eta \rho_1 := \frac{\|p_{\leq A}\|_r^{r/t}}{\sqrt{n}\|p_{\leq I}\|_r^{r/4}}.$$

Proof of Lemma 5. For a small enough constant c_A depending only on η, we define the quantity

$$a = \frac{c_A}{\sqrt{n}\left(\sum_{i\leq I} p_i^r\right)^{1/4}}. \tag{14}$$

For all $\delta \in \{\pm 1\}^A$, let $q_\delta = ((q_\delta)_i)_{i=1,\dots,N}$ such that

- $\forall i \leq A, (q_\delta)_i = p_i + a\delta_i p_i^{2/(4-t)}$, where a is defined in (14);
- $\forall i > A, (q_\delta)_i = p_i$.

Let $\mathcal{P}_1 = \{q_\delta \mid \delta \in \{\pm 1\}^A\}$. We set a uniform prior on $\mathcal{P}_1$. With the notation of Lemma 4, we just set $\gamma_i = a p_i^{2/(4-t)}$ if $i \leq A$ and 0 otherwise. In terms of $\|\cdot\|_t$ norm, any distribution where this prior puts mass is separated from p with a distance ρ such that

$$\begin{aligned}\rho &= a\left\|\left(p_i^{2/(4-t)}\right)_{i=1,\dots,A}\right\|_t \\ &= \frac{c_A}{\sqrt{n}\left(\sum_{i\leq I} p_i^r\right)^{1/4}}\left(\sum_{i\leq A} p_i^r\right)^{1/t} \asymp_\eta \frac{\|p_{\leq A}\|_r^{r/t}}{\sqrt{n}\|p_{\leq I}\|_r^{r/4}} = \rho_1.\end{aligned}$$

According to Lemma 4, taking $c_A^4 \leq c_4$ this prior gives a minimax risk greater than η since

$$\sum_{i\leq A}\frac{\gamma_i^4}{p_i^2} \leq a^4 \sum_{i\leq A} p_i^{(8/(4-t))-2} = \frac{c_A^4}{n^2} \leq \frac{c_4}{n^2}.$$ ∎

Lemma 6. *Assume that* $\|p_{>I}\|_1 \geq \frac{1}{n}$. *Then it holds that*

$$\rho_t^* \gtrsim_\eta \rho_2 := \frac{\|p_{\geq I}\|_1^{(2-t)/t}}{n^{(2t-2)/t}}.$$

Proof of Lemma 6. We divide the proof in two steps. In the first step, we prove that the prior concentrates with high probability on a zone located at

$$\frac{\|p_{\geq U}\|_1^{(2-t)/t}}{n^{(2t-2)/t}} + \frac{1}{n},$$

up to a multiplicative constant. In the second step, we prove that the prior is indistinguishable from the null hypothesis p, by proving that the total variation between p and this prior is small.

First step. We prove that the prior concentrates with high probability on a zone located at

$$\frac{\|p_{\geq U}\|_1^{(2-t)/t}}{n^{(2t-2)/t}} + \frac{1}{n},$$

up to a multiplicative constant. By assumption we have $\|p_{>I}\|_1 \geq \frac{1}{n}$.

Let U be the smallest index greater than or equal to A such that

$$n^2 p_U \|p_{\geq U}\|_1 \leq c_u,$$

where $c_u = \frac{\eta}{10} \wedge \frac{1}{2}(1-\eta)^2$.

Let

$$\overline{\pi} = \frac{c_u}{n^2 \|p_{\geq U}\|_1} \quad \text{and} \quad \pi_i = \frac{p_i}{\overline{\pi}}.$$

We set the following sparse prior: for all $i < U$, we set $q_i = p_i$ and for all $i \geq U$, we draw $b_i \sim \mathcal{B}(\pi_i)$ mutually independent, and we define $q_i = b_i \overline{\pi}$. We write $q = (q_i)_i$ for the corresponding distribution parameter and $\mathcal{Q}$ for the prior distribution.

Before showing that the data distribution coming from this prior – namely $\mathbb{E}_{q\sim\mathcal{Q}}\mathbb{P}_q$ – is close enough to $\mathbb{P}_\pi$ in total variation, we first prove that $q \sim \mathcal{Q}$ is such that $\|q-p\|_t$ is with high probability larger – up to a positive multiplicative constant that depends only on u – than ρ_2. We have

$$\begin{aligned}
\mathbb{E}_{q\sim\mathcal{Q}}\big[\|p-q\|_t^t\big] &= \mathbb{E}_{(b_i)_i\sim\otimes\mathcal{B}(\pi_i)}\left[\sum_{i\geq U}|b_i\overline{\pi}-p_i|^t\right] \\
&= \overline{\pi}^t\,\mathbb{E}_{(b_i)_i\sim\otimes\mathcal{B}(\pi_i)}\left[\sum_{i\geq U}|b_i-\pi_i|^t\right] \\
&= \overline{\pi}^t \sum_{i\geq U} \pi_i(1-\pi_i)^t + (1-\pi_i)\pi_i^t \\
&\geq 4^{-1}\overline{\pi}^t \sum_{i\geq U}\pi_i + \pi_i^t \geq 4^{-1}\overline{\pi}^t \sum_{i\geq U}\pi_i,
\end{aligned}$$

since $\forall i \geq U$, it holds that $\pi_i \leq c_u \leq \frac{1}{2}$, and

$$\begin{aligned}
\mathbb{V}_{q\sim\mathcal{Q}}\big[\|p-q\|_t^t\big] &= \overline{\pi}^{2t}\sum_{i\geq U}\mathbb{V}_{b_i\sim\mathcal{B}(\pi_i)}|b_i-\pi_i|^t \\
&= \overline{\pi}^{2t}\sum_{i\geq U}\pi_i(1-\pi_i)\big[(1-\pi_i)^t-\pi_i^t\big]^2 \leq \overline{\pi}^{2t}\sum_{i\geq U}\pi_i.
\end{aligned}$$

We now show that

$$\big[\mathbb{E}_{q\sim\mathcal{Q}}\big[\|p-q\|_t^t\big]\big]^2 \gg \mathbb{V}_{q\sim\mathcal{Q}}\big[\|p-q\|_t^t\big].$$

This is equivalent to proving $\sum_{i\geq U}\pi_i \gg 1$, or equivalently, $n^2\|p_{\geq U}\|_1^2 \gg c_u$.

By Lemma 8, we are necessarily in the case $\|p_{\geq U}\|_1 \geq \frac{1}{3}\|p_{>I}\|_1$. Indeed, suppose that $\|p_{\geq U}\|_1 < \frac{1}{3}\|p_{>I}\|_1$, then by Lemma 8, we would have

$$\|p_{>I}\|_1 \leq \|p_{\geq U}\|_1 + \frac{\sqrt{c_I}}{n} \leq \frac{1}{3}\|p_{>I}\|_1 + \frac{\sqrt{c_I}}{n},$$

hence $\|p_{>I}\|_1 \leq \frac{3}{2}\frac{\sqrt{c_I}}{n}$, which is excluded because we assume $\|p_{>I}\|_1 \geq \frac{1}{n}$.

Therefore,

$$\|p_{\geq U}\|_1^2 n^2 \geq \frac{1}{9} \gg c_u.$$

We conclude using Chebyshev's inequality. Therefore, this prior is indeed separated away from the null distribution by a distance greater than $\overline{\pi}\sum_{i\geq U}\pi_i$ up to a constant, or equivalently, greater than

$$\frac{\|p_{\geq U}\|_1^{(2-t)/t}}{n^{2(t-1)/t}}.$$

Second step. We now show that this prior is indistinguishable from p, i.e. has a Bayes risk strictly greater than η. We denote by $\overline{\mathbb{P}}_{\text{tail}} = \mathbb{E}_{q\sim\mathcal{Q}}[\mathbb{P}_q]$, the prior distribution used to lower bound the minimax risk. We always have

$$R^* \geq 1 - d_{TV}\big(\mathbb{P}_p, \overline{\mathbb{P}}_{\text{tail}}\big).$$

Moreover, we recall that for any realization $X = (X_1, \dots, X_n)$, we write

$$S = \sum_{i=1}^{n} X_i.$$

We have

$$\begin{aligned} d_{TV}\big(\mathbb{P}_p, \overline{\mathbb{P}}_{\text{tail}}\big) &= \frac{1}{2}\sum_{X\in\mathcal{G}} |\mathbb{P}_p(X) - \overline{\mathbb{P}}_{\text{tail}}(X)| \\ &= \frac{1}{2}\sum_{X\in\mathcal{G}:\, \forall i\geq U,\, s_i\leq 1} |\mathbb{P}_p(X) - \overline{\mathbb{P}}_{\text{tail}}(X)| + \frac{1}{2}\sum_{X\in\mathcal{G}:\, \exists i\geq U,\, \text{s.t. } s_i\geq 2} |\mathbb{P}_p(X) - \overline{\mathbb{P}}_{\text{tail}}(X)|. \end{aligned}$$

This allows us to split the total variation into two terms: The first one will be the principal term, while the second one will be negligible. We first prove the negligibility of the second term.

Since s is a sufficient statistic, we have

$$\begin{aligned} &\sum_{X\in\mathcal{G}:\, \exists i\geq U,\, \text{s.t. } s_i\geq 2} |\mathbb{P}_p(X) - \overline{\mathbb{P}}_{\text{tail}}(X)| \\ &\quad\leq \big[\mathbb{P}_p(\exists i \geq U;\, s_i \geq 2) + \overline{\mathbb{P}}_{\text{tail}}(\exists i \geq U;\, s_i \geq 2)\big] \\ &\quad\leq \sum_{i=U}^{|\mathcal{G}|}\big[1 - \mathbb{P}_p(s_i = 0) - \mathbb{P}_p(s_i = 1) + 1 - \overline{\mathbb{P}}_{\text{tail}}(s_i = 0) - \overline{\mathbb{P}}_{\text{tail}}(s_i = 1)\big]. \end{aligned}$$

Let us fix $i \in \{U, \dots, N\}$. We will use the following inequalities which hold for all $n \in \mathbb{N}$, $x \in [0, 1]$:

$$(1-x)^n \geq 1 - nx, \quad (1-x)^n \geq 1 - nx + \frac{n}{4}x^2, \quad (1-x)^n \leq 1 - nx + \frac{n^2}{2}x^2.$$

First term in the sum: $\sum_{i=U}^{N}[1 - \mathbb{P}_p(s_i = 0) - \mathbb{P}_p(s_i = 1)]$. We recall that by the definition of U, since $U > I$, it holds that $\forall i \geq U, np_i \leq c_I$, so that for any $i \geq U$, we have

$$\begin{aligned} 1 - \mathbb{P}_p(s_i = 0) - \mathbb{P}_p(s_i = 1) &= 1 - (1-p_i)^n - np_i(1-p_i)^{n-1} \\ &\leq 1 - \Big[1 - np_i + \frac{n}{4}p_i^2\Big] - np_i\big[1 - (n-1)p_i\big] \\ &\leq n^2 p_i^2. \end{aligned}$$

Summing over all $i = U, \dots, N$ yields that

$$\sum_{i=U}^{N}\big[1 - \mathbb{P}_p(s_i = 0) - \mathbb{P}_p(s_i = 1)\big] \leq c_I.$$

Second term in the sum: $\sum_{i=U}^{N}[1 - \overline{\mathbb{P}}_{\text{tail}}(s_i = 0) - \overline{\mathbb{P}}_{\text{tail}}(s_i = 1)]$. We recall that by the definition of U, since $U > I$, it holds that $\forall i \geq U, np_i \leq c_I$, so that for any $i \geq U$, we have

$$\begin{aligned} &1 - \overline{\mathbb{P}}_{\text{tail}}(s_i = 0) - \overline{\mathbb{P}}_{\text{tail}}(s_i = 1) \\ &\quad = 1 - \big[1 - \pi_i + \pi_i(1-\overline{\pi})^n\big] - \pi_i n\overline{\pi}(1-\overline{\pi})^{n-1} \\ &\quad = \pi_i - \pi_i(1-\overline{\pi})^n - \pi_i n\overline{\pi}(1-\overline{\pi})^{n-1} \\ &\quad \leq \pi_i - \pi_i(1 - n\overline{\pi}) - \pi_i n\overline{\pi}(1-(n-1)\overline{\pi}) \\ &\quad = n(n-1)\pi_i\overline{\pi}^2 = n(n-1)p_i\overline{\pi} \leq n^2 c_u \frac{p_i}{n^2\|p_{\geq U}\|_1} = c_u\frac{p_i}{\|p_{\geq U}\|_1}. \end{aligned}$$

Summing over all $i = U, \dots, N$ yields that

$$\sum_{i=U}^{N}\big[1 - \overline{\mathbb{P}}_{\text{tail}}(s_i = 0) - \overline{\mathbb{P}}_{\text{tail}}(s_i = 1)\big] \leq c_u\frac{\|p_{\geq U}\|_1}{\|p_{\geq U}\|_1} = c_u.$$

Therefore,

$$d_{TV}\big(\mathbb{P}_p, \overline{\mathbb{P}}_{\text{tail}}\big) = \underbrace{\frac{1}{2}\sum_{X \in \mathcal{G}:\, \forall i \geq U,\, s_i \leq 1} |\mathbb{P}_p(X) - \overline{\mathbb{P}}_{\text{tail}}(X)|}_{\text{principal term}} + c_I + c_u. \tag{15}$$

Now, we can upper bound the total variation by the χ^2 divergence on the high probability event that we only observe 0 or 1 for each coordinate $i \geq U$ corresponding to the principal term. Since s is a sufficient statistic, we have

$$\begin{aligned}
&\sum_{X \in \mathcal{G}: \forall i \geq U, s_i \leq 1} |\mathbb{P}_p(X) - \overline{\mathbb{P}}_{\text{tail}}(X)| \\
&\quad \leq \sqrt{\sum_{X \in \mathcal{G}:\, \forall i \geq U,\, s_i \leq 1} \frac{\left(\mathbb{P}_p(X) - \overline{\mathbb{P}}_{\text{tail}}(X)\right)^2}{\mathbb{P}_p(X)}} \sqrt{\underbrace{\sum_{X \in \mathcal{G}:\, \forall i \geq U,\, s_i \leq 1} \mathbb{P}_p(X)}_{\leq 1}} \\
&\quad \leq \sqrt{\sum_{X \in \mathcal{G}:\, \forall i \geq U,\, s_i \leq 1} \frac{\overline{\mathbb{P}}_{\text{tail}}(X)^2}{\mathbb{P}_p(X)} - 1 + 2c_u} \\
&\quad = \sqrt{\prod_{i=U}^{N} \left(\sum_{j=0}^{1} \frac{\overline{\mathbb{P}}_{\text{tail}}(s_i = j)^2}{\mathbb{P}_p(s_i = j)} \right) - 1 + 2c_u}. \qquad (16)
\end{aligned}$$

Computation of $\sum_{k=0}^{1} \overline{\mathbb{P}}_{\text{tail}}(s_i = k)^2 / \mathbb{P}(s_i = k)$. We have

$$\sum_{k=0}^{1} \frac{\overline{\mathbb{P}}_{\text{tail}}(s_i = k)^2}{\mathbb{P}_p(s_i = k)} = \frac{\left[1 - \pi_i + \pi_i(1 - \overline{\pi})^n\right]^2}{(1 - p_i)^n} + \frac{\left[\pi_i n \overline{\pi}(1 - \overline{\pi})^{n-1}\right]^2}{n p_i (1 - p_i)^{n-1}}.$$

The first term becomes

$$\begin{aligned}
\frac{\left[1 - \pi_i + \pi_i(1 - \overline{\pi})^n\right]^2}{(1 - p_i)^n} &\leq \frac{\left[1 - \pi_i + \pi_i\left(1 - n\overline{\pi} + (n^2/2)\overline{\pi}^2\right)\right]^2}{1 - n p_i} \\
&= 1 - n p_i + n^2 p_i \overline{\pi} + \frac{\left((n^2/2) p_i \overline{\pi}\right)^2}{1 - n p_i} \\
&\leq 1 - n p_i + n^2 p_i \overline{\pi} + \frac{n^4 p_i^2 \overline{\pi}^2}{4(1 - c_I)} \\
&\leq 1 - n p_i + n^2 p_i \overline{\pi} + \frac{c_u^2}{4(1 - c_I)}.
\end{aligned}$$

The second term becomes

$$\frac{\left[\pi_i n \overline{\pi}(1 - \overline{\pi})^{n-1}\right]^2}{n p_i (1 - p_i)^{n-1}} = n p_i \frac{(1 - \overline{\pi})^{2n-2}}{(1 - p_i)^{n-1}} \leq n p_i \quad \text{since } \overline{\pi} \geq p_i.$$

We can now sum the two terms as follows:

$$\sum_{k=0}^{1} \frac{\overline{\mathbb{P}}_{\text{tail}}(s_i = k)^2}{\mathbb{P}_p(s_i = k)} = 1 + n^2 p_i \overline{\pi} + \frac{c_u^2}{4(1 - c_I)}.$$

So that

$$\prod_{i=U}^{N}\left(\sum_{k=0}^{1}\frac{\overline{\mathbb{P}}_{\text{tail}}(s_i=k)^2}{\mathbb{P}_p(s_i=k)}\right)=\prod_{k=U}^{N}\left(1+n^2p_i\overline{\pi}+\frac{c_u^2}{4(1-c_I)}\right)$$
$$\leq \exp\left(c_u+\frac{c_u^2}{1-c_I}\right)$$
$$\leq \exp\frac{3}{2}c_u \leq 1+3c_u \quad \text{since } \frac{3}{2}c_u\leq 1.$$

Now, using (15) and (16), we have

$$d_{TV}\left(\mathbb{P}_p,\overline{\mathbb{P}}_{\text{tail}}\right)\leq\frac{1}{2}\sqrt{5c_u}+c_I+c_u\leq 1-\eta$$

by the definition of c_u, c_I. This concludes the proof. ∎

Lemma 7. *Assume that* $\|p_{\geq I}\|_1\leq\frac{1}{n}$. *Then it holds that*

$$\rho_t^*\gtrsim\rho_3:=\frac{1}{n}.$$

Proof of Lemma 7. We introduce q such that $q_1=p_1+(1-\eta)/n$ and $q_j=p_j$ for all $j\geq 2$. We then have

$$R^*\geq\inf_{\psi\text{ test}}\mathbb{P}_p(\psi=1)+\mathbb{P}_q(\psi=0)=1-d_{TV}(\mathbb{P}_p,\mathbb{P}_q)$$
$$=1-n\,d_{TV}\left(\bigotimes_{i<j}\mathcal{B}(p_i),\bigotimes_{i<j}\mathcal{B}(q_i)\right)$$
$$=1-n\,d_{TV}\left(\mathcal{B}(p_1),\ \mathcal{B}(q_1)\right)$$
$$=1-n\,|p_1-q_1|=1-n\,\frac{1-\eta}{n}=\eta.$$

This concludes the proof. ∎

Lemma 8. *It holds that*

$$\|p_{\geq U}\|_1+\frac{1}{n}\asymp\|p_{>I}\|_1+\frac{1}{n}.$$

Moreover, we either have $\|p_{\geq U}\|_1\geq\frac{1}{3}\|p_{>I}\|_1$ *or* $\|p_{>I}\|_1\leq\|p_{\geq U}\|_1+\sqrt{c_I}/n$.

Proof of lemma 8. If

$$\|p_{\geq U}\|_1\geq\frac{1}{3}\|p_{>I}\|_1,$$

then the result is clear. Now, suppose

$$\|p_{\geq U}\|_1<\frac{1}{3}\|p_{>I}\|_1.$$

We have $\|p_{\geq U}\|_1 < \frac{1}{2}\|P_{I\to U}\|$, where $P_{I\to U} = (p_{I+1}, \dots, p_{U-1})$. We then have

$$\begin{aligned} p_{U-1}^2 + \frac{c_I}{2n^2} &\geq p_{U-1}^2 + \frac{1}{2}\sum_{i=I+1}^{U-1} p_i^2 \geq p_{U-1}\left(p_{U-1} + \frac{1}{2}\sum_{i=I+1}^{U-1} p_i\right) \\ &> p_{U-1}\left(p_{U-1} + \sum_{i\geq U} p_i\right) \geq p_{U-1}\sum_{i\geq U-1} p_i \\ &= p_{U-1}\|P_{\geq U-1}\|_1 > \frac{c_u}{n^2} \end{aligned}$$

by the definition of U. Therefore,

$$p_{U-1}^2 > \frac{2c_u - c_I}{2n^2} \Rightarrow \forall I < i < U, \quad p_i^2 > \frac{c_I}{2n^2} \quad \text{since } c_u \geq c_I.$$

Moreover,

$$\frac{c_I}{n^2} \geq \sum_{I<i<U} p_i^2 > (I - U - 1)p_{U-1}^2 > (I - U - 1)\frac{c_I}{2n^2}.$$

So that

$$I - U - 1 < 2, \quad \text{i.e. } I - U - 1 \leq 1.$$

Thus,

$$\begin{aligned} \|p_{>I}\|_1 &\leq \|P_{I\to U}\|_1 + \|p_{\geq U}\|_1 \\ &\leq (I - U - 1)p_{I+1} + \|p_{\geq U}\|_1 \\ &\leq \frac{\sqrt{c_I}}{n} + \|p_{\geq U}\|_1 \lesssim \|p_{\geq U}\|_1 + \frac{1}{n}. \end{aligned}$$

Hence the result. ∎

Lemma 9. *Let ρ_1 and ρ_2 be defined as in Lemmas 5 and 6. We have*

$$\rho_1 + \rho_2 \asymp \sqrt{\frac{\|p_{\leq I}\|_r}{n}} + \rho_2.$$

Proof of Lemma 9. Clearly,

$$\rho_1 + \rho_2 \leq \sqrt{\frac{\|p_{\leq I}\|_r}{n}} + \rho_2.$$

To prove

$$\rho_1 + \rho_2 \gtrsim_\eta \sqrt{\frac{\|p_{\leq I}\|_r}{n}} + \rho_2,$$

there are two cases:

(i) If $A = I$, then the result is clear.

(ii) Otherwise, $I > A$. Note that by setting $p'_i := np_i$ for all $i = 1, \dots, N$, the result to be shown can be rewritten as

$$\frac{\|p'_{\le A}\|_r^{r/t}}{\|p'_{\le I}\|_r^{r/4}} + \|p'_{\ge I}\|_1^{2-t} \asymp \sqrt{\|p'_{\le I}\|_r} + \|p'_{\ge I}\|_1^{2-t}. \tag{17}$$

By definition of A and I, we have

$$p_I'^{2-r}\left(\sum_{i\ge I} p'_i\right)^{2-r} = \left(\sum_{i\ge I} p'_I p'_i\right)^{2-r} \ge \left(\sum_{i\ge I} p_i'^2\right)^{2-r} \gtrsim_\eta 1$$

and

$$p_I'^{2b}\sum_{i\le I} p_i'^r \le p_{A+1}'^{2b}\sum_{i\le I} p_i'^r \le c_A^4 \asymp 1 \quad \text{by definition of } A.$$

Hence, by noticing that $2b = 2 - r$, we have

$$\left(\sum_{i\ge I} p'_i\right)^{2-r} \gtrsim_\eta \sum_{i\le I} p_i'^r,$$

which yields

$$\|p'_{\ge I}\|_1^{2-t} \ge \sqrt{\|p'_{\le I}\|_r} \ge \frac{\|p'_{\le A}\|_r^{r/t}}{\|p'_{\le I}\|_r^{r/4}}$$

by raising to the power $\frac{1}{2r}$. This condition yields the result of the lemma, by replacing p' by np. ∎

Lemma 10. *It holds that*

$$\|p_{>I}\|_1 + \frac{1}{n} \asymp \|p_{>A}\|_1 + \frac{1}{n}.$$

Proof of Lemma 10. If $A = I$, then the result is clear. Now, suppose that $A < I$. By the definition of A, we have

$$\frac{c_A^4}{n^2} > p_{A+1}^{2b}\sum_{i\le I} p_i^r \ge \sum_{i=A+1}^{I} p_i^2 \ge p_I \sum_{i=A+1}^{I} p_i \Rightarrow \frac{c_A^4}{n^2\sum_{i=A+1}^{I} p_i} \ge p_I.$$

Moreover, if $I < N$, then

$$\frac{c_I}{n^2} \le \sum_{i>I} p_i^2 \le p_{I+1}\sum_{i>I} p_i \Rightarrow p_{I+1} \ge \frac{c_I}{n^2\sum_{i>I} p_i}.$$

So that

$$\sum_{i>I} p_i \geq \frac{c_I}{c_A^4} \sum_{i=A+1}^{I} p_i,$$

and consequently $\|p_{>I}\|_1 \gtrsim \|p_{>A}\|_1$ if we impose moreover that $c_A^4 \gtrsim c_I$, which can be done without loss of generality.

Now if $I = N$, we have $\|p_{>I}\|_1 = 0$ and $p_N > \frac{\sqrt{c_I}}{n}$ and

$$\begin{aligned} p_{A+1}^{2b} < \frac{c_A^4}{n^2 \sum_{i=1}^N p_i^r} &\Rightarrow \sum_{j=A+1}^{N} p_{A+1}^{2b} p_j^r \leq \frac{c_A^4}{n^2} \\ &\Rightarrow \sum_{j=A+1}^{N} p_j^2 \leq \frac{c_A^4}{n^2} \\ &\Rightarrow p_N \|p_{>A}\|_1 \leq \frac{c_A^4}{n^2} \\ &\Rightarrow \frac{\sqrt{c_I}}{n} \|p_{>A}\|_1 \leq \frac{c_A^4}{n^2}, \end{aligned}$$

and hence $\|p_{>A}\|_1 \lesssim \frac{1}{n}$, so that $\|p_{>A}\|_1 + \frac{1}{n} \asymp \|p_{>I}\|_1 + \frac{1}{n} \asymp \frac{1}{n}$. ∎

B. Upper bound

Define $\Delta = q - p$. In the following, $c > 0$ denotes an absolute constant, depending only on η. We call

$$\rho = \sqrt{\frac{\|p_{\leq I}\|_r}{n}} + \frac{\|p_{\geq A}\|_1^{(2-t)/t}}{n^{(2-2t)/t}} + \frac{1}{n},$$

and we prove $\rho^* \lesssim_\eta \rho$.

We start with the three following lemmas which control the expectation and variance of the statistics T_{bulk}, T_1, T_2. We recall that $k = \frac{n}{2}$.

Lemma 11 (Bounds on expectation and variance of T_{bulk}). *Let T_{bulk} be defined as in equation (10). The expectation and variance of T_{bulk} satisfy*

$$\mathbb{E}[T_{\text{bulk}}] = \sum_{i \leq A} \frac{\Delta_i^2}{p_i^b}, \quad \mathbb{V}[T_{\text{bulk}}] \leq \sum_{i \leq A} \frac{1}{p_i^{2b}} \left(\frac{q_i^2}{k^2} + \frac{2}{k} q_i \Delta_i^2 \right).$$

Lemma 12 (Bounds on expectation and variance of T_1). *Let T_1 be defined as in equation (12). The expectation and variance of T_1 satisfy*

$$\mathbb{E}[T_1] = \sum_{i>A} q_i - p_i, \quad \mathbb{V}[T_1] \leq \sum_{i>A} \frac{q_i}{n}.$$

We then study the null and alternative hypotheses in the following subsection, bounding the probability of error of the test ψ.

B.1. Under the null hypothesis $\mathcal{H}_0$

We start by assuming that $p = q$. We recall that $\underline{c}_\eta = \frac{4}{\sqrt{\eta}}$.

Test ψ_{bulk}. Moreover, for the bulk, since $p = q$, we have by Lemma 11 that

$$\mathbb{E}[T_{\text{bulk}}] = 0 \quad \text{and} \quad \mathbb{V}[T_{\text{bulk}}] = \sum_{i \le A} \frac{p_i^r}{n^2}.$$

Therefore, by Chebyshev's inequality,

$$\mathbb{P}\left(T_{\text{bulk}} > \underline{c}_\eta \sqrt{\sum_{i \le A} \frac{p_i^r}{n^2}} \right) \le \frac{\eta}{16},$$

so that

$$\mathbb{P}(\psi_{\text{bulk}} = 1) \le \frac{\eta}{16}. \tag{18}$$

Test ψ_1. Since $p = q$, we have by Lemma 12 that

$$\mathbb{E}(T_1) = 0 \quad \text{and} \quad \mathbb{V}(T_1) \le \sqrt{\frac{\sum_{i>A} p_i}{n}}.$$

By the same argument, ψ_1's type-I error is upper bounded as

$$\mathbb{P}_p\big(\psi_1 = 1\big) = \mathbb{P}_p\left(T_1 > \underline{c}_\eta \sqrt{\frac{\sum_{i>A} p_i}{n}} \right) \le \frac{1}{\underline{c}_\eta^2} = \frac{\eta}{16},$$

so that by definition of ψ_1, we have

$$\mathbb{P}_p\big(\psi_1 = 1\big) \le \frac{\eta}{16}. \tag{19}$$

Test ψ_2. By Lemmas 13 and 14, we have

$$\mathbb{P}(\psi_2 = 1) \le c_I + c_A^4 \le \frac{\eta}{16}, \tag{20}$$

by choosing the constants c_I and c_A depending only on η sufficiently small.

Conclusion. Putting together equations (19), (18) and (20), we get that the type-I error of $\psi = \psi_{\text{bulk}} \vee \psi_1 \vee \psi_2$ is upper bounded as

$$\mathbb{P}\big(\psi = 1\big) \le \sum_{i \in \{\text{bulk},1,2\}} \mathbb{P}\big(\psi_i = 1\big) \le \frac{3\eta}{16} < \eta/2.$$

B.2. Under the alternative hypothesis $\mathscr{H}_1(\rho)$

Suppose that for some constant $\overline{c}_\eta > 0$, we have $\|\Delta\|_t \geq 2\overline{c}_\eta \rho$. By the triangle inequality, there are two cases:

(i) Either $\|\Delta_{\leq A}\|_t \geq \overline{c}_\eta \rho$, or

(ii) $\|\Delta_{>A}\|_t \geq \overline{c}_\eta \rho$.

Proposition 7 (Study in case (i)). *There exists a large enough constant $\overline{c}_\eta^{(\text{bulk})} > 0$ such that if $\|\Delta_{\leq A}\|_t \geq \overline{c}_\eta^{(\text{bulk})} \rho$, then*

$$\mathbb{P}(\psi_{\text{bulk}} = 1) \geq 1 - \eta/6.$$

Proposition 8 (Study in case (ii)). *If $\|\Delta_{>A}\|_t \geq c\rho$, then*

$$\mathbb{P}(\psi_1 \vee \psi_2 = 1) \geq 1 - \frac{2\eta}{3}.$$

Proof of Proposition 7. Suppose $\|\Delta_{\leq A}\|_t \geq c\rho$ for some constant c. We show that if c is large enough, then the test ψ_{Bulk} will detect it. To do so, we compute a constant c' depending on c such that if $\|\Delta_{\leq A}\|_t \geq c\rho$, then $\mathbb{V}(T_{\text{Bulk}}) \leq c'\,\mathbb{E}(T_{\text{Bulk}})^2$ and such that $\lim_{c\to+\infty} c' = 0$.

By definition of ρ, we have in particular that

$$\|\Delta_{\leq A}\|_t \geq c\sqrt{\frac{\|p_{\leq I}\|_r}{n}} \vee \frac{c}{n},$$

and hence

$$\frac{1}{n^2} \leq \frac{1}{c^4}\frac{\|\Delta_{\leq A}\|_t^4}{\|p_{\leq I}\|_r^2} \wedge \frac{\|\Delta_{\leq A}\|_t^2}{c^2} \tag{21}$$

Using Lemma 11 we split $\mathbb{V}[T_{\text{bulk}}]$ into four terms as follows:

$$\begin{aligned}\mathbb{V}[T_{\text{bulk}}] &\leq \sum_{i\leq A}\frac{1}{p_i^{2b}}\left(\frac{(p_i+\Delta_i)^2}{n^2} + \frac{2}{n}(p_i+\Delta_i)\Delta_i^2\right)\\ &\leq \underbrace{\frac{2}{n^2}\sum_{i\leq A} p_i^r}_{①} + \underbrace{\frac{2}{n^2}\sum_{i\leq A}\frac{\Delta_i^2}{p_i^{2b}}}_{②} + \underbrace{\frac{2}{n}\sum_{i\leq A} p_i^{1-2b}\Delta_i^2}_{③} + \underbrace{\frac{2}{n}\sum_{i\leq A}\frac{\Delta_i^3}{p_i^{2b}}}_{④}.\end{aligned}$$

Now we show that each of the four terms is less than $\mathbb{E}[T_{\text{bulk}}]^2$, up to a constant.

Term ①. By Hölder's inequality, we have

$$\sum_{i\le A}\Delta_i^t \le \left[\sum_{i\le A}\left(\frac{\Delta_i^t}{p_i^{bt/2}}\right)^{2/t}\right]^{t/2}\left[\sum_{i\le A}\left(p_i^{bt/2}\right)^{2/(2-t)}\right]^{(2-t)/2}$$
$$= \left(\sum_{i\le A}\frac{\Delta_i^2}{p_i^b}\right)^{t/2}\left(\sum_{i\le A}p_i^r\right)^{1-t/2}.$$

Hence,

$$\|\Delta_{\le A}\|_t \le \left(\sum_{i\le A}\frac{\Delta_i^2}{p_i^b}\right)^{1/2}\left(\sum_{i\le A}p_i^r\right)^{(2-t)/2t}. \tag{22}$$

Moreover, we have

$$\frac{1}{n^2} \le \frac{\|\Delta_{\le A}\|_t^4}{c^4\|p_{\le I}\|_r^2},$$

so that term ① becomes

$$\frac{2}{n^2}\sum_{i\le A}p_i^r \le 2\sum_{i\le A}p_i^r\left(\sum_{i\le A}\Delta_i^t\right)^{4/t}\frac{1}{c^4\left(\sum_{i\le I}p_i^r\right)^{2/r}}$$
$$\le \frac{2}{c^4}\left(\sum_{i\le A}p_i^r\right)^{1-2/r}\left(\sum_{i\le A}\frac{\Delta_i^2}{p_i^b}\right)^2\left(\sum_{i\le A}p_i^r\right)^{(4-2t)/t} \quad \text{(by (22))}$$
$$= \frac{2}{c^4}\left(\sum_{i\le A}\frac{\Delta_i^2}{p_i^b}\right)^2 = \frac{2}{c^4}\mathbb{E}[T_{\text{bulk}}]^2. \tag{23}$$

Term ②. By definition of index A, we have

$$p_A^b \ge \frac{c_A^2}{\left(\sum_{j\le I}p_j^r\right)^{1/2}n} =: \widetilde{c}\frac{1}{\left(\sum_{j\le I}p_j^r\right)^{1/2}n}.$$

Using this condition, term ② becomes

$$\sum_{i\le A}\frac{1}{p_i^{2b}}\frac{\Delta_i^2}{n^2} \le \frac{1}{n^2}\frac{1}{p_A^b}\sum_{i\le A}\frac{\Delta_i^2}{p_i^b} \le \widetilde{c}^{-1}\frac{1}{n}\left(\sum_{j\le I}p_j^r\right)^{1/2}\left(\sum_{i\le A}\frac{\Delta_i^2}{p_i^b}\right). \tag{24}$$

Moreover, since

$$\sqrt{\frac{\|p_{\le I}\|_r}{n}} \le \rho \le \frac{1}{c}\|\Delta_{\le A}\|_t,$$

and using (22), we have

$$\frac{1}{n}\left(\sum_{j\le I} p_j^r\right)^{1/2} = \frac{1}{n^b}\left(\sqrt{\frac{\|p_{\le I}\|_r}{n}}\right)^r \le \frac{1}{n^b c^r}\left(\sum_{i\le A}\frac{\Delta_i^2}{p_i^b}\right)^{r/2}\left(\sum_{i\le A} p_i^r\right)^{b/2} \le \frac{1}{c^2}\sum_{i\le A}\frac{\Delta_i^2}{p_i^b}. \tag{25}$$

In the last inequality, we used the fact proved in case number ① that

$$\frac{1}{n^b}\left(\sum_{i\le A} p_i^r\right)^{b/2} \lesssim \frac{1}{c^{2b}}\mathbb{E}[T_{\text{bulk}}]^b$$

and the relation $\frac{r}{2} + b = 1$.

Plugging in (24) yields that the second term ② is bounded by $\mathbb{E}[T_{\text{bulk}}]^2$.

Term ③. This term becomes

$$\begin{aligned}
\frac{1}{n}\sum_{i\le A} p_i^{1-2b}\Delta_i^2 &\le \frac{\|\Delta_{\le A}\|_t^2}{c^2\left(\sum_{i\le I} p_i^r\right)^{1/r}}\sum_{i\le A} p_i^{1-2b}\Delta_i^2 \\
&\le \frac{1}{c^2}\left(\sum_{i\le A}\frac{\Delta_i^2}{p_i^b}\right)\left(\sum_{i\le A} p_i^r\right)^{(4-2t)/2t-1/r}\sum_{i\le A} p_i^{1-2b}\Delta_i^2 \qquad \text{(using (22))} \\
&\le \frac{1}{c^2}\left(\sum_{i\le A}\frac{\Delta_i^2}{p_i^b}\right)\left(\sum_{i\le A} p_i^r\right)^{-1/2}\left(\sum_{i\le A} p_i^{(2/3)(1-2b)}\Delta_i^{4/3}\right)^{3/2} \qquad \text{(since } \|\cdot\|_1 \le \|\cdot\|_{2/3}\text{)}.
\end{aligned}$$

Moreover, by Hölder's inequality with $\frac{1}{3/2} + \frac{1}{3} = 1$, we have

$$\begin{aligned}
\sum_{i\le A} p_i^{(2/3)(1-2b)}\Delta_i^{4/3} &\le \left(\sum_{i\le A}\left(\frac{p_i^{(2/3)(1-2b)}\Delta_i^{4/3}}{p_i^{(2/3)t/(4-t)}}\right)^{3/2}\right)^{2/3}\left(\sum_{i\le A}\left(p_i^{(2/3)t/(4-t)}\right)^3\right)^{1/3} \\
&\le \left(\sum_{i\le A}\frac{\Delta_i^2}{p_i^b}\right)^{2/3}\left(\sum_{i\le A} p_i^r\right)^{1/3}.
\end{aligned}$$

So that

$$\left(\sum_{i\le A} p_i^{(2/3)(1-2b)}\Delta_i^{4/3}\right)^{3/2} \le \left(\sum_{i\le A}\frac{\Delta_i^2}{p_i^b}\right)\left(\sum_{i\le A} p_i^r\right)^{1/2},$$

i.e.

$$\left(\sum_{i\le I} p_i^r\right)^{-1/2}\left(\sum_{i\le A} p_i^{(2/3)(1-2b)}\Delta_i^{4/3}\right)^{3/2} \le \left(\sum_{i\le A}\frac{\Delta_i^2}{p_i^b}\right).$$

This yields that the third term satisfies:

$$\frac{1}{n}\sum_{i\le A} p_i^{1-2b}\Delta_i^2 \le \frac{1}{c^2}\left(\sum_{i\le A}\frac{\Delta_i^2}{p_i^b}\right)^2 = \frac{1}{c^2}\mathbb{E}[T_{\mathrm{bulk}}]^2.$$

Term ④. The fourth term becomes

$$\begin{aligned}\frac{1}{n}\left\|\left(\frac{|\Delta_i|}{p_i^{2b/3}}\right)_{i\le A}\right\|_3^3 &\le \frac{1}{n}\left\|\left(\frac{|\Delta_i|}{p_i^{2b/3}}\right)_{i\le A}\right\|_2^3 \\ &= \frac{1}{n}\left(\sum_{i\le A}\frac{\Delta_i^2}{p_i^{4b/3}}\right)^{3/2} \le \frac{1}{n^{1/2}}\left(\sum_{i\le A}\frac{\Delta_i^2}{p_i^b}\right)^{3/2}\left(\sum_{i\le I} p_i^r\right)^{1/4},\end{aligned}$$

where in the last step we have used the fact that

$$p_i^{b/3} \ge \frac{1}{\left(\sum_{i\le I} p_i^r\right)^{1/6} n^{1/3}}.$$

Then using (24), we have

$$\frac{1}{\sqrt{n}}\left(\sum_{i\le I} p_i^r\right)^{1/4} \lesssim \left(\sum_{i\le A}\frac{\Delta_i^2}{p_i^b}\right)^{1/2}.$$

So the term ④ is upper-bounded by $\frac{1}{c^2}\mathbb{E}[T_{\mathrm{bulk}}]^2$.

Conclusion. By Chebyshev's inequality, the type-II error of ψ_{Bulk} is bounded as

$$\begin{aligned}\mathbb{P}(\psi_{\mathrm{Bulk}}=0) &= \mathbb{P}\left(T_{\mathrm{Bulk}} \le \frac{c_\eta}{n}\|p_{\le A}\|_r^{r/2}\right) \\ &= \mathbb{P}\left(\mathbb{E}(T_{\mathrm{Bulk}}) - T_{\mathrm{Bulk}} \ge \mathbb{E}(T_{\mathrm{Bulk}}) - \frac{c_\eta}{n}\|p_{\le A}\|_r^{r/2}\right) \\ &\le \mathbb{P}\left(|\mathbb{E}(T_{\mathrm{Bulk}}) - T_{\mathrm{Bulk}}| \ge \mathbb{E}(T_{\mathrm{Bulk}}) - \frac{c_\eta}{n}\|p_{\le A}\|_r^{r/2}\right) \\ &\le \frac{\mathbb{V}(T_{\mathrm{Bulk}})}{\left(\mathbb{E}(T_{\mathrm{Bulk}}) - \frac{c_\eta}{n}\|p_{\le A}\|_r^{r/2}\right)^2} \quad \text{(by Chebyshev's inequality)} \\ &\le \frac{c'\mathbb{E}(T_{\mathrm{Bulk}})^2}{\left(\mathbb{E}(T_{\mathrm{Bulk}}) - \frac{c_\eta}{n}\|p_{\le A}\|_r^{r/2}\right)^2}.\end{aligned}$$

Moreover, using (23), for c large enough, we have that

$$\mathbb{E}(T_{\mathrm{Bulk}}) \ge \frac{c}{n}\|p_{\le A}\|_r^{r/2} \ge 2\frac{c_\eta}{n}\|p_{\le A}\|_r^{r/2},$$

so that the denominator is well-defined. Finally, since $\lim_{c\to+\infty} c' = 0$, the type-II error of this test goes to 0 as c goes to infinity, so for c large enough, the type-II error is upper-bounded by $\eta/6$. ∎

We now move to the proof of Proposition 8.

Proof of Proposition 8. We will need the following two lemmas.

Lemma 13. *It holds by definition of A that $\|p_{>A}\|_2^2 \leq \frac{C_A}{n^2}$ for $C_A = c_A^4 + c_I$.*

Proof of Lemma 13. If $A = I$, then the result is clear by definition of I. Otherwise, by definition of A, we have

$$p_{A+1}^{2b} \sum_{i \leq I} p_i^r < \frac{c_A^4}{n^2} \Rightarrow p_{A+1}^{2b} \sum_{i=A+1}^{I} p_i^r < \frac{c_A^4}{n^2}$$
$$\Rightarrow \sum_{i=A+1}^{I} p_i^2 < \frac{c_A^4}{n^2} \Rightarrow \sum_{i>A} p_i^2 < \frac{c_A^4 + c_I}{n^2}. \qquad \blacksquare$$

Lemma 14. *For fixed $j > A$, the probability that coordinate j is observed at least twice is upper-bounded by $n^2 p_j^2$.*

Proof of Lemma 14. The probability that coordinate j is observed at least twice is

$$1 - (1-p_j)^n - np_j(1-p_j)^{n-1} \leq 1 - (1-np_j) - np_j\big[1-(n-1)p_j\big]$$
$$\leq n^2 p_j^2. \qquad \blacksquare$$

Under H_0. We upper bound the type-I error of tests ψ_1 and ψ_2. For ψ_2: by Lemma 13, we have

$$\mathbb{P}(\psi_2 = 1) \leq \sum_{j>A} n^2 p_j^2 \leq C_A \leq \frac{\eta}{4}.$$

As to the test ψ_1: we have

$$\mathbb{P}(\psi_1 = 1) = \mathbb{P}\left(|T_1| > \underline{c}_\eta \sqrt{\frac{\sum_{i>A} p_i}{n}}\right) \leq \frac{\eta}{4}$$

by Chebyshev's inequality. By union bound, the type-I error of $\psi_1 \vee \psi_2$ is less than $\eta/2$.

Under H_1. If $\|\Delta_{>A}\|_t \geq c\rho$, we now show that either ψ_1 or ψ_2 will detect it.

Note. From now until the end of the proof, we drop the index "$> A$" and write only e.g. $\|p\|_2$, $\|\Delta\|_2$ instead of $\|p_{>A}\|_2$, $\|\Delta_{>A}\|_2$.

By Hölder's inequality, we have

$$\|\Delta\|_2^{2(t-1)} \|\Delta\|_1^{2-t} \geq \|\Delta\|_t^t \geq C\left(\frac{\|p\|_1^{2-t}}{n^{2t-2}} + \frac{1}{n^t}\right) = C\,\frac{1}{n^{2t-2}}\left(\|p\|_1^{2-t} + \frac{1}{n^{2-t}}\right)$$

for $C = C_1 C_2$, where

$$C_1 = \left(\left(\frac{20}{\eta}(\underline{c}_\eta + 1) + 1\right)\right)^{2-t}, \quad C_2 = \left(\frac{1}{4}\big(\log(4/\eta)^2 \vee 9/100\big) + c_I\right)^{(t-1)/2},$$

so that one of the following two relations must hold:

$$\|\Delta\|_2^{2(t-1)} \geq C_2 \frac{1}{n^{2t-2}} \quad \text{or} \quad \|\Delta\|_1^{2-t} \geq C_1\left(\|p\|_1^{2-t} + \frac{1}{n^{2-t}}\right).$$

First case: $\|\Delta\|_2^{2(t-1)} \geq C_2/n^{2t-2}$. Then $\|\Delta\|_2 \geq C_2^{1/2(t-1)}/n$, so that

$$\|q\|_2 \geq C_2^{1/(t-1)}/n - \|p\|_2 \geq \frac{1}{n}\big(C_2^{1/(t-1)} - c_I\big).$$

The test ψ_2 accepts if, and only if, all coordinates are observed at most once. This probability corresponds to

$$\begin{aligned} q(\forall j > A,\, N_j = 0 \text{ or } N_j = 1) &= \prod_{j>A}\big[(1-q_j)^n + n q_j (1-q_j)^{n-1}\big] \\ &= \prod_{j>A}(1-q_j)^{n-1}\big(1 + (n-1)q_j\big) \\ &= \prod_{j>A}(1-q_j)^{n'}(1+n'q_j), \quad \text{writing } n' = n-1. \end{aligned}$$

Let $I_- = \{j > A : n q_j \leq \frac{1}{2}\}$ and $I_+ = \{j > A : n q_j > \frac{1}{2}\}$. Recall that for $x \in (0, 1/2]$, it holds that $\log(1+x) \leq x - x^2/3$. Then, for $j \in I_-$, we have

$$\begin{aligned} (1-q_j)^{n'}(1+n'q_j) &= \exp\big\{n' \log(1-q_j) + \log(1+n'q_j)\big\} \\ &\leq \exp\left\{-n'q_j + n'q_j - \frac{n'^2 q_j^2}{3}\right\} \\ &= \exp\left(-\frac{n'^2 q_j^2}{3}\right). \end{aligned}$$

Now, for $j \in I_+$, we have

$$n' \log(1-q_j) + \log(1+n'q_j) \leq -n'q_j + \log(1+n'q_j) \leq -\frac{1}{10} n'q_j$$

using the inequality $-0.9x + \log(1+x) \leq 0$ true for all $x \geq \frac{1}{2}$. Therefore, we have upper bounded the type-II error of ψ_2 by

$$q(\psi = 0) \leq \exp\left(-\frac{1}{3}\sum_{j\in I_-} n'^2 q_j^2 - \frac{1}{10}\sum_{j\in I_+} n'q_j\right)$$

$$\leq \exp\left(-\frac{1}{3}\sum_{j\in I_-} n'^2 q_j^2 - \frac{1}{10}\left(\sum_{j\in I_+} n'^2 q_j^2\right)^{1/2}\right)$$
$$= \exp\left(-\frac{1}{3}(S - S_+) - \frac{1}{10}(S_+)^{1/2}\right)$$

for

$$S = \sum_{j>A} n'^2 q_j^2 \quad \text{and} \quad S_+ = \sum_{j\in I_+} n'^2 q_j^2.$$

Now, $S_+ \mapsto -\frac{S}{3} + \frac{1}{3}S_+ - \sqrt{S_+}/10$ is convex over $[0, S]$ so its maximum is reached on the boundaries of the domain and is therefore equal to

$$\left(-\frac{\sqrt{S}}{10}\right) \vee -\frac{S}{3} = -\frac{\sqrt{S}}{10}$$

for $S \geq 9/100$. Now, since

$$\|q\|_2^2 \geq \frac{C_2^{2/(t-1)}}{n^2} \geq 4\frac{C_2^{2/(t-1)}}{n'^2},$$

we have $S = n'^2\|q\|_2^2 \geq \log(4/\eta)^2 \vee 9/100$, which ensures $q(\psi_2 = 0) \leq \eta/4$.

Second case: $\|\Delta\|_1^{2-t} \geq C_1(\|p\|_1^{2-t} + 1/n^{2-t})$. Then

$$\|\Delta\|_1 \geq C_1^{1/(2-t)}\left(\|p\|_1 \vee \frac{1}{n}\right) \geq \frac{C_1^{1/(2-t)}}{2}\left(\|p\|_1 + \frac{1}{n}\right).$$

We will need the following lemma.

Lemma 15. *If $\sum_{j>A}\Delta_j \geq 3\sum_{j>A} p_j$, then $|\sum_{j>A}\Delta_j| \geq \frac{1}{2}\|\Delta\|_1$.*

Proof. Define $J_+ = \{j > A : q_j \geq p_j\}$ and $J_- = \{q_j < p_j\}$. Define also

$$s = \frac{\sum_{j>A}\Delta_j}{\sum_{j>A} p_j}, \quad s_+ = \frac{\sum_{j\in J_+}\Delta_j}{\sum_{j>A} p_j}, \quad s_- = -\frac{\sum_{j\in J_-}\Delta_j}{\sum_{j>A} p_j}.$$

Then by assumption $s_+ - s_- = s \geq 3$. Moreover,

$$s_- = \frac{\sum_{j\in J_-} p_j - q_j}{\sum_{j>A} p_j} \leq 1.$$

Thus, $s_+ \geq 3 \geq 3s_-$, so that $2(s_+ - s_-) \geq s_+ + s_-$, which yields the result. ∎

Note that by definition of the second case, we have for some constant C that

$$C\|p\|_1 \leq \|\Delta\|_1 \leq \|q\|_1 + \|p\|_1,$$

hence that $\|q\|_1 \geq (C-1)\|p\|_1$, and therefore taking $C \geq 5$ ensures that the assumption of Lemma 15 are met.

We can now upper bound the type-II error of ψ_1:

$$
\begin{aligned}
q(\psi_1 = 0) &= q\left(\Big|\sum_{j>A} \frac{N_j}{n} - p_j\Big| \leq \underline{c}_\eta \sqrt{\frac{\|p\|_1}{n}}\right) \\
&\leq q\left(\Big|\sum_{j>A} q_j - p_j\Big| - \Big|\sum_{j>A} \frac{N_j}{n} - q_j\Big| \leq \underline{c}_\eta \sqrt{\frac{\|p\|_1}{n}}\right) && \text{(by the triangular inequality)} \\
&\leq q\left(\frac{1}{2}\|q-p\|_1 - \underline{c}_\eta \sqrt{\frac{\|p\|_1}{n}} \leq \Big|\sum_{j>A} \frac{N_j}{n} - q_j\Big|\right) && \text{(by Lemma 15)} \\
&\leq \frac{\frac{1}{n}\sum_{j>A} q_j}{\left(\frac{1}{2}\|q-p\|_1 - \underline{c}_\eta \sqrt{\|p\|_1/n}\right)^2} && \text{(by Chebyshev's inequality)} \\
&\leq \frac{\|q\|_1/n}{\left(\frac{1}{2}\|q\|_1 - \frac{1}{2}\|p\|_1 - \underline{c}_\eta \sqrt{\|p\|_1/n}\right)^2} && \text{(by triangular inequality)} \\
&\leq \frac{\|q\|_1/n}{\left(\frac{1}{2}\|q\|_1 - \frac{1}{2}\|p\|_1 - \underline{c}_\eta\big(\|p\|_1 + 1/n\big)\right)^2} && \text{(using } \sqrt{xy} \leq x+y\text{)} \\
&\leq \frac{\|q\|_1/n}{\left(\frac{1}{2}\|q\|_1 - (\underline{c}_\eta + 1)\big(\|p\|_1 + 1/n\big)\right)^2}.
\end{aligned}
$$

Now set $z = (\underline{c}_\eta + 1)(\|p\|_1 + 1/n)$. The function

$$f : x \mapsto \frac{x}{n(x/2 - z)^2}$$

is decreasing. Moreover, for $x \geq 20z/\eta$, we have

$$f(x) \leq \frac{20z/\eta}{n(10z/\eta - z)^2} = \frac{20\eta}{nz(10-\eta)^2} \underset{nz \leq 1}{\leq} \frac{20\eta}{81} \leq \eta/4,$$

which proves that, whenever

$$\|q\|_1 \geq \frac{20}{\eta}(\underline{c}_\eta + 1)\big(\|p\|_1 + 1/n\big),$$

we have $q(\psi_1 = 0) \leq \eta/4$. This condition is guaranteed when

$$\|\Delta\|_1 \geq \left(\frac{20}{\eta}(\underline{c}_\eta + 1) + 1\right)\big(\|p\|_1 + 1/n\big) = C_1^{1/(2-t)}\big(\|p\|_1 + 1/n\big). \qquad \blacksquare$$

Proof of Lemma 11.

Expectation. We have that

$$\begin{aligned}\mathbb{E}[T_{\text{bulk}}] &= \sum_{i\le A}\frac{1}{p_i^b}\left(\mathbb{E}\left[\frac{S_i}{k}-p_i\right]\mathbb{E}\left[\frac{S_i'}{k}-p_i\right]\right)\\ &= \sum_{i\le A}\frac{1}{p_i^b}(p_i-q_i)^2.\end{aligned}$$

Variance. We have that

$$\begin{aligned}\mathbb{V}(T_{\text{bulk}}) &= \sum_{i\le A}\frac{1}{p_i^{2b}}\left(\mathbb{E}\left[\left(\frac{S_i}{k}-p_i\right)^2\left(\frac{S_i'}{k}-p_i\right)^2\right]-\mathbb{E}\left[\left(\frac{S_i}{k}-p_i\right)\left(\frac{S_i'}{k}-p_i\right)\right]^2\right)\\ &= \sum_{i\le A}\frac{1}{p_i^{2b}}\left(\mathbb{E}\left[\left(\frac{S_i}{k}-p_i\right)^2\right]^2-(p_i-q_i)^4\right),\end{aligned}$$

since the $(S_i, S_i')_i$ are independent. And so by a bias-variance decomposition, and since $S_i, S_i' \sim \mathcal{B}(k, q_i)$, we obtain

$$\begin{aligned}\mathbb{V}(T_{\text{bulk}}) &= \sum_{i\le A}\frac{1}{p_i^{2b}}\left(\left[\mathbb{V}\left(\frac{S_i}{k}\right)+\mathbb{E}\left[\left(\frac{S_i}{k}-p_i\right)\right]^2\right]^2-(p_i-q_i)^4\right)\\ &= \sum_{i\le A}\frac{1}{p_i^{2b}}\left(\left[\frac{q_i(1-q_i)}{k}+(p_i-q_i)^2\right]^2-(p_i-q_i)^4\right)\\ &= \sum_{i\le A}\frac{1}{p_i^{2b}}\left(\frac{q_i^2(1-q_i)^2}{k^2}+\frac{2}{k}q_i(1-q_i)(p_i-q_i)^2\right)\\ &\le \sum_{i\le A}\frac{1}{p_i^{2b}}\left(\frac{q_i^2}{k^2}+\frac{2}{k}q_i(p_i-q_i)^2\right).\end{aligned}$$

■

Proof of Lemma 12. We therefore have

$$\mathbb{E}[T_1] = \mathbb{E}\left[\sum_{i>A}\frac{S_i+S_i'}{n}-p_i\right] = \sum_{i>A}q_i-p_i,$$

and

$$\begin{aligned}\mathbb{V}[T_1] &= \mathbb{V}\left[\sum_{i>A}\frac{S_i+S_i'}{n}\right] = \sum_{i>A}\frac{\mathbb{V}[S_i]+\mathbb{V}[S_i']}{n^2} \quad \text{(by independence of } (S_i, S_i')_i)\\ &= \sum_{i>A}\frac{q_i(1-q_i)}{n} \le \sum_{i>A}\frac{q_i}{n},\end{aligned}$$

which completes the proof. ■

C. Equivalence between the Binomial, Poisson and multinomial settings

We now prove that the rates for goodness-of-fit testing in the Binomial, Poisson and multinomial cases are equivalent.

Proof of Lemma 1. We first prove $\rho^*_{\mathrm{Poi}}(n, p) \le C_{\mathrm{BP}}\, \rho^*_{\mathrm{Bin}}(n, p)$. Let $n \ge 2$, and let $Y_1, \dots, Y_n \overset{\text{i.i.d.}}{\sim} \mathrm{Poi}(q)$. We consider a random function ϕ such that for any Poisson family $Y_1, \dots, Y_n \overset{\text{i.i.d.}}{\sim} \mathrm{Poi}(q)$, we have

$$\begin{cases} \phi(Y_1, \dots, Y_n) = (X_1, \dots, X_{\widetilde{n}}) \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(q), \quad \text{where } \widetilde{n} \sim \mathrm{Poi}(n) \perp\!\!\!\perp (Y_i)_i, \\ \sum_{i=1}^{\widetilde{n}} X_i = \sum_{i=1}^{n} Y_i. \end{cases}$$

In words, ϕ is a function which takes n Poisson random variables (or equivalently one Poisson random variable $\mathrm{Poi}(nq)$) and decomposes them into $\widetilde{n} \sim \mathrm{Poi}(n)$ Bernoulli i.i.d. random variables whose sum is $\sum_{i=1}^{n} Y_i$.

Let $\widetilde{n} \sim \mathrm{Poi}(n)$ be the random length of $\phi(Y_1, \dots, Y_n)$. We can choose a small constant $c = c(\eta)$ such that the event

$$\mathcal{A}_1 := \{\widetilde{n} \ge cn\}$$

has probability larger than $1 - \eta/4$. Moreover, for $m \ge cn$ we can define the function

$$\pi(x_1, \dots, x_m) = (x_1, \dots, x_{\lfloor cn \rfloor}).$$

Let ψ_{Bin} be the test associated to the *binomial* testing problem

$$H_0 : q = p \quad \text{vs.} \quad H_1 : \|p - q\|_t \ge \rho_{\mathrm{Bin}}\Big(cn, p, \frac{\eta}{2}\Big).$$

In particular, $R(\psi_{\mathrm{Bin}}) \le \eta/2$. Now, we define the test

$$\psi = \begin{cases} \psi_{\mathrm{Bin}} \circ \pi \circ \phi & \text{if } \mathcal{A}_1, \\ 0 & \text{otherwise} \end{cases}$$

and we show that, when associated to the *Poissonian* testing problem

$$H_0 : q = p \quad \text{vs.} \quad H_1 : \|p - q\|_t \ge \rho$$

with $\rho = \rho_{\mathrm{Bin}}(cn, p, \frac{\eta}{2})$, it has a risk less than η. We first analyze its type-I error:

$$\begin{aligned} \mathbb{P}_{H_0}\big(\psi(Y_1^n) = 1\big) &\le \mathbb{P}_{H_0}\big(\mathcal{A}_1 \cap \psi(Y_1^n) = 1\big) + \mathbb{P}_{H_0}(\overline{\mathcal{A}_1}) \\ &\le \mathbb{P}_{H_0}\big(\psi(Y_1^n) = 1 | \mathcal{A}_1\big) + \frac{\eta}{4} \\ &\le \mathbb{P}_{H_0}\big(\psi_{\mathrm{Bin}}(X_1, \dots, X_{\lfloor cn \rfloor}) = 1 | \mathcal{A}_1\big) + \frac{\eta}{4} \\ &= \mathbb{P}_{X_1^{\lfloor cn \rfloor} \sim \mathrm{Ber}(p)^{\otimes \lfloor cn \rfloor}}\big(\psi_{\mathrm{Bin}}(X_1, \dots, X_{\lfloor cn \rfloor}) = 1\big) + \frac{\eta}{4}. \end{aligned}$$

For the type-II error, the same steps show that for any vector q:

$$\mathbb{P}_q\big(\psi(Y_1^n)=0\big) \le \mathbb{P}_{X_1^{\lfloor cn\rfloor}\sim \mathrm{Ber}(q)^{\otimes \lfloor cn\rfloor}}\big(\psi_{\mathrm{Bin}}(X_1,\dots,X_{\lfloor cn\rfloor})=0\big)+\frac{\eta}{4}.$$

We can now compute the risk of ψ when $\rho=\rho_{\mathrm{Bin}}(cn,p,\frac{\eta}{2})$:

$$\begin{aligned}
R(\psi) &= \mathbb{P}_{H_0}\big(\psi(Y_1^n)=1\big)+\sup_{\|p-q\|_t\ge\rho}\mathbb{P}_q\big(\psi(Y_1^n)=0\big)\\
&\le \frac{\eta}{2}+\mathbb{P}_{X_1^{\lfloor cn\rfloor}\sim \mathrm{Ber}(p)^{\otimes \lfloor cn\rfloor}}\big(\psi_{\mathrm{Bin}}(X_1,\dots,X_{\lfloor cn\rfloor})=1\big)\\
&\quad+\sup_{\|p-q\|_t\ge\rho}\mathbb{P}_{X_1^{\lfloor cn\rfloor}\sim \mathrm{Ber}(q)^{\otimes \lfloor cn\rfloor}}\big(\psi_{\mathrm{Bin}}(X_1,\dots,X_{\lfloor cn\rfloor})=0\big)\\
&= \frac{\eta}{2}+R(\psi_{\mathrm{Bin}})\\
&= \frac{\eta}{2}+\frac{\eta}{2}=\eta.
\end{aligned}$$

This proves $\rho^*_{\mathrm{Poi}}(n,p)\le\rho^*_{\mathrm{Bin}}(cn,p,\frac{\eta}{2})\asymp\rho^*_{\mathrm{Bin}}(n,p,\eta)$.

We now show $\rho^*_{\mathrm{Poi}}(n,p)\ge c_{\mathrm{BP}}\,\rho^*_{\mathrm{Bin}}(n,p)$. Let $X_1,\dots,X_n\sim\mathrm{Ber}(q)$ i.i.d. For some small constant $\overline{c}>0$, let $\widetilde{n}\sim\mathrm{Poi}(\lfloor\overline{c}n\rfloor)$. We choose $\overline{c}>0$ such that

$$\mathcal{A}_2=\{\widetilde{n}\le n\} \tag{26}$$

has probability larger than $1-\frac{\eta}{4}$. Consider the extended sequence of multivariate Bernoulli random variables $(\widetilde{X}_i)_i$ such that

$$\begin{cases}\widetilde{X}_i=X_i & \text{if } i\le n,\\ \widetilde{X}_i\sim\mathrm{Ber}(q) & \text{otherwise,}\end{cases}$$

and such that $(\widetilde{X}_i)_i$ are mutually independent. Let $Y=\sum_{i=1}^{\widetilde{n}}X_i\sim\mathrm{Poi}(\lfloor\overline{c}n\rfloor q)$. The sum is a sufficient statistic of the parameter q for Poisson random variables so we can define a function

$$\overline{\phi}(Y)=(Y_1,\dots,Y_{\lfloor\overline{c}n\rfloor})$$

such that $Y_i\overset{\text{i.i.d}}{\sim}\mathrm{Poi}(q)$ and $\sum_{i=1}^{\lfloor\overline{c}n\rfloor}Y_i=\sum_{i=1}^{\widetilde{n}}X_i$. Moreover, for $m\le n$, we set

$$\overline{\pi}(y_1,\dots,y_n,m)=(y_1,\dots,y_m).$$

On $\mathcal{A}_2$, we do not even need to extend the sequence of observations. We call ψ_{Poi} the test associated to the *Poisson* testing problem:

$$H_0\colon q=p\quad\text{vs.}\quad H_1\colon\|p-q\|_t\ge\rho_{\mathrm{Poi}}\Big(\lfloor\overline{c}n\rfloor,p,\frac{\eta}{2}\Big).$$

We define the randomized test

$$\bar{\psi} = \begin{cases} \psi_{\mathrm{Poi}} \circ \bar{\pi} \circ \bar{\phi}(Y) & \text{if } \mathcal{A}_2, \\ 0 & \text{otherwise.} \end{cases} \tag{27}$$

We show that this test has a risk less than η. For the type-I error:

$$\begin{aligned} \mathbb{P}_{H_0}\big(\bar{\psi}(Y) = 1\big) &\le \mathbb{P}_{H_0}\big(\mathcal{A}_2 \cap \bar{\psi}(Y) = 1\ \big) + \mathbb{P}_{H_0}(\bar{\mathcal{A}}_2) \\ &\le \mathbb{P}_{H_0}\big(\bar{\psi}(Y) = 1\ |\mathcal{A}_2\big) + \frac{\eta}{4} \\ &\le \mathbb{P}_{H_0}\big(\psi_{\mathrm{Poi}}(Y_1, \dots, Y_{\lfloor \bar{c}n \rfloor}) = 1 | \mathcal{A}_2\big) + \frac{\eta}{4} \\ &= \mathbb{P}_{Y_1^{\lfloor \bar{c}n \rfloor} \sim \mathrm{Poi}(p)^{\otimes \lfloor \bar{c}n \rfloor}}\big(\psi_{\mathrm{Poi}}(Y_1, \dots, Y_{\lfloor \bar{c}n \rfloor}) = 1\big) + \frac{\eta}{4}. \end{aligned}$$

For the type-II error, the same steps show that for any vector q:

$$\mathbb{P}_q\big(\bar{\psi}(Y) = 0\big) \le \mathbb{P}_{Y_1^{\lfloor \bar{c}n \rfloor} \sim \mathrm{Poi}(q)^{\otimes \lfloor \bar{c}n \rfloor}}\big(\psi_{\mathrm{Poi}}(Y_1, \dots, Y_{\lfloor \bar{c}n \rfloor}) = 0\big) + \frac{\eta}{4}.$$

We can now compute the risk of $\bar{\psi}$ when $\rho = \rho_{\mathrm{Poi}}(\bar{c}n, p, \frac{\eta}{2})$:

$$\begin{aligned} R(\psi) &= \mathbb{P}_{H_0}\big(\bar{\psi}(Y) = 1\big) + \sup_{\|p-q\|_t \ge \rho} \mathbb{P}_q\big(\psi(Y) = 0\big) \\ &\le \frac{\eta}{2} + \mathbb{P}_{Y_1^{\lfloor \bar{c}n \rfloor} \sim \mathrm{Poi}(p)^{\otimes \lfloor \bar{c}n \rfloor}}\big(\psi_{\mathrm{Poi}}(Y_1, \dots, Y_{\lfloor \bar{c}n \rfloor}) = 1\big) \\ &\quad + \sup_{\|p-q\|_t \ge \rho} \mathbb{P}_{Y_1^{\lfloor \bar{c}n \rfloor} \sim \mathrm{Poi}(q)^{\otimes \lfloor \bar{c}n \rfloor}}\big(\psi_{\mathrm{Poi}}(Y_1, \dots, Y_{\lfloor \bar{c}n \rfloor}) = 0\big) \\ &= \frac{\eta}{2} + R(\psi_{\mathrm{Poi}}) \\ &= \frac{\eta}{2} + \frac{\eta}{2} = \eta. \end{aligned}$$

This proves $\rho^*_{\mathrm{Bin}}(n, p) \le \rho^*_{\mathrm{Poi}}(\bar{c}n, p, \frac{\eta}{2}) \asymp \rho^*_{\mathrm{Poi}}(n, p, \eta)$. ∎

Proof of Lemma 2. We first prove that $\rho^*_{\mathrm{Mult}}(n, p) \lesssim \rho^*_{\mathrm{Poi}}(n, p^{-\max})$ when $\sum p_i = 1$ by following the same steps as for proving $\rho_{\mathrm{Bin}} \lesssim \rho_{\mathrm{Poi}}$: we draw $\tilde{n} \sim \mathrm{Poi}(\bar{c}n)$ and $Z_1, \dots, Z_{\tilde{n}} \overset{\text{i.i.d.}}{\sim} \mathcal{M}(q)$. Then the histogram is a sufficient statistic of $Z_1, \dots, Z_{\tilde{n}}$ for q. It is defined as

$$\begin{pmatrix} N_1 \\ \vdots \\ N_d \end{pmatrix} := \begin{pmatrix} \sum_{i=1}^{\tilde{n}} \mathbb{1}\{Z_i = 1\} \\ \vdots \\ \sum_{i=1}^{\tilde{n}} \mathbb{1}\{Z_i = d\} \end{pmatrix} \sim \mathrm{Poi}(nq),$$

where we recall that for any vector $v = (v_1, \dots, v_\ell)$ with nonnegative entries, we denote by $\mathrm{Poi}(v)$ the distribution $\otimes_{j=1}^{\ell} \mathrm{Poi}(v_j)$. On $\mathcal{A}_2$, defined in (26), we have

$$\begin{pmatrix} N_2 \\ \vdots \\ N_d \end{pmatrix} \sim \mathrm{Poi}(n(q_2, \dots, q_d)),$$

so we can just apply the exact same steps to prove that, if $q = p$ then the test $\overline{\psi}$ from (27) has type-I error less than $\frac{\eta}{2}$ and if $\|q - p\|_{\mathcal{M},t} \geq \rho_{\mathrm{Poi}}(\overline{c}n, p, \frac{\eta}{2})$, its type-II error is less than $\frac{\eta}{2}$.

We now prove the converse bound: $\rho^*_{\mathrm{Poi}}(n, p^{-\max}, \eta) \lesssim_\eta \rho^*_{\mathrm{Mult}}(n, p, \eta)$. Note that the constants denoted by C and depending on η, are allowed to vary from line to line. Let $p = (p_1, \dots, p_d)$ be a probability vector and $q = (q_2, \dots, q_N)$ and assume that we observe

$$(X_2, \dots X_N) \sim \bigotimes_{j=2}^{N} \mathrm{Poi}(nq_j) = \mathrm{Poi}(nq).$$

We consider the testing problem

$$H_0 : q = p^{-\max} \quad \text{vs.} \quad H_1 : \|q - p^{-\max}\|_t \geq \rho. \tag{28}$$

We exhibit a test ψ and a constant $C > 0$ such that if $\rho \geq C\rho^*_{\mathrm{Mult}}(n, p, \eta)$, then its risk for problem (29) is at most η. For any $m \in \mathbb{N}^*$, let ψ_m be a test such that, if $Y_1, \dots, Y_m$ are *multinomial* observations drawn with discrete distribution $q' = (q'_1, \dots, q'_d)$ such that $\sum_j q'_j = 1$, then its risk for the following testing problem is at most η:

$$H_0 : q' = p \quad \text{vs.} \quad H_1 : \|q' - p\|_{\mathcal{M},t} \geq \rho^*_{\mathrm{Mult}}(p, m, \eta). \tag{29}$$

Now, draw $X_1 \sim \mathrm{Poi}(np_1)$ independently on $(X_2, \dots, X_N)$, so that $(X_1, \dots, X_N) \sim \mathrm{Poi}(n\overline{q})$, where $\overline{q} = (p_1, q_2, \dots, q_d)$. For some large enough constants C, C' depending only on η, let also

$$\psi_0(X_1, X_2, \dots, X_N) = \mathbb{1}\left\{ \left| \sum_{j=1}^{N} X_j - n \right| \geq C\sqrt{n} \right\},$$

where

$$\begin{cases} \mathbb{P}\left(|\mathrm{Poi}(n) - n| \geq C\sqrt{n} \right) \leq \frac{\eta}{100}, \\ \mathbb{P}\left(|\mathrm{Poi}(\lambda) - n| < C\sqrt{n} \right) \leq \frac{\eta}{100}, \quad \text{whenever } |\lambda - n| \geq C'\sqrt{n}. \end{cases}$$

We define the randomized test ψ such that, conditional on $m := \sum_{j=1}^{N} X_j$:

$$\psi(X_2, \dots, X_N) \mid m = \psi_0(X_1, \dots, X_N) \vee \psi_m(X_1, \dots, X_N).$$

First, if $|\|\overline{q}\| - 1| > C'/\sqrt{n}$, then with probability at least $1 - \eta/100$, ψ_0 will detect it. From now on, assume that $|\|\overline{q}\| - 1| \le C'/\sqrt{n}$. We now prove that for some large enough constant C, C', if $\|\overline{q} - p\|_{\mathcal{M},t} \ge C\rho^*_{\mathrm{Mult}}(p, n, \eta)$, then

$$\left\| \frac{\overline{q}}{\|\overline{q}\|_1} - p \right\|_{\mathcal{M},t} \ge C'\rho^*_{\mathrm{Mult}}(p, n, \eta).$$

Indeed,

$$\begin{aligned}
\left\| \frac{\overline{q}}{\|\overline{q}\|_1} - p \right\|_{\mathcal{M},t} &\ge \|\overline{q} - p\|_{\mathcal{M},t} - \left\| \frac{\overline{q}}{\|\overline{q}\|_1} - \overline{q} \right\|_{\mathcal{M},t} \\
&\ge C\rho^*_{\mathrm{Mult}}(p, n, \eta) - \|q\|_t \left| \frac{1 - \|\overline{q}\|_1}{\|\overline{q}\|_1} \right| \\
&\ge C\rho^*_{\mathrm{Mult}}(p, n, \eta) - \left[\|p\|_{\mathcal{M},t} + \|p - q\|_{\mathcal{M},t} \right] \frac{C'}{\sqrt{n}} \\
&\ge C\rho^*_{\mathrm{Mult}}(p, n, \eta) - \|p\|_{\mathcal{M},t} \frac{C'}{\sqrt{n}}.
\end{aligned}$$

Now, since $\|p\|_1 \le 1$ and $r = 2t/(4 - t) \le t$, we have $\|\cdot\|_r \ge \|\cdot\|_t$ so that

- for some small enough $c > 0$,

$$\frac{C'}{\sqrt{n}} \|p_{\le A}\|_{\mathcal{M},t} \le \frac{C}{\sqrt{n}} \sqrt{\|p^{-\max}\|_r} \le c\rho^*_{\mathrm{Mult}}(p, n, \eta)$$

  provided that n is greater than a suitable constant depending on η.

- By Hölder's inequality, we get

$$\begin{aligned}
\frac{C'}{\sqrt{n}} \|p_{>A}\|^t_{\mathcal{M},t} &\le \frac{C'}{\sqrt{n}} \|p_{>A}\|_1^{2-t} \|p_{>A}\|_2^{(t-1)} \\
&\le \frac{C'}{\sqrt{n}} \|p_{>A}\|_1^{2-t} \cdot \left(\frac{1}{n^2} \right)^{(t-1)} \le c\rho^*_{\mathrm{Mult}}(p, n, \eta).
\end{aligned}$$

Therefore, we get

$$\left\| \frac{\overline{q}}{\|\overline{q}\|_1} - p \right\|_{\mathcal{M},t} \ge C\rho^*_{\mathrm{Mult}}(p, n, \eta). \tag{30}$$

Now, choose n larger than a suitable constant depending only on η such that

$$\mathbb{P}\left(\mathrm{Poi}(n) \ge \frac{n}{2} \right) \ge 1 - \frac{\eta}{100}.$$

Conditional on $m = \sum_{j=1}^N X_j$, the observations $(X_1, \ldots, X_N)$ follow a multinomial distribution $\mathcal{M}(m, \overline{q}/\|\overline{q}\|_1)$. Hence, with probability at least $1 - \eta/2$, the test ψ_m will conclude in favor of H_1 in view of (30) whenever $m \ge \frac{n}{2}$, since

$$\rho^*_{\mathrm{Mult}}(p, n, \eta) \ge C\rho^*_{\mathrm{Mult}}\left(p, \frac{n}{2}, \frac{\eta}{2} \right).$$

We now prove that the risk of ψ for problem (29) is at most η.

On the other hand, if $\overline{q} = \overline{p}$, then with probability $\geq 1 - \eta/100$:

$$\psi_0(X_1, \dots, X_N) = 0,$$

and whenever $m \geq n/2$, we have $\psi_m(X_1, \dots, X_N) = 0$ with probability at least $1 - \eta/4$ by definition of ψ_m, since

$$\rho^*_{\text{Mult}}(p, n, \eta) \geq C\rho^*_{\text{Mult}}\Big(p, \frac{n}{2}, \frac{\eta}{4}\Big).$$

To conclude, we can explicitly bound from above the risk of test ψ as

$$\begin{aligned}
&\mathbb{P}_p(\psi = 1) + \sup_{\|p-q\|_{\mathcal{M},t} \geq C\rho^*(p,n,\eta)} (\psi = 0) \\
&\leq 2\mathbb{P}\Big(m < \frac{n}{2}\Big) + \mathbb{P}_p\Big(\psi = 1 | m \geq \frac{n}{2}\Big) + \sup_{\|p-q\|_{\mathcal{M},t} \geq C\rho^*(p,n,\eta)} \mathbb{P}_q\Big(\psi = 0 | m \geq \frac{n}{2}\Big) \\
&\leq \frac{2\eta}{100} + \frac{\eta}{4} + \frac{\eta}{100} + \eta/2 \leq \eta,
\end{aligned}$$

which proves that $\rho^*_{\text{Poi}} \lesssim \rho^*_{\text{Mult}}$. ∎

D. Tightness of Balakrishnan and Wasserman (2019) in the multinomial case

For fixed n and for two absolute constants $C, c > 0$, define ε_+ as the largest quantity satisfying

$$\varepsilon_+ \leq C\sqrt{\frac{\|p^{-\max}_{-\varepsilon_+/16}\|_{2/3}}{n}} + \frac{C}{n}$$

and ε_- as the smallest quantity satisfying

$$\varepsilon_- \geq c\sqrt{\frac{\|p^{-\max}_{-\varepsilon_-}\|_{2/3}}{n}} + \frac{c}{n}.$$

By [9], the critical radius ρ^* satisfies $\varepsilon_- \lesssim \rho^* \lesssim \varepsilon_+$.

(1) First case: If $\varepsilon_+ \leq 16\varepsilon_-$, then the bounds match.

(2) Second case: Otherwise,

$$\varepsilon_+ \leq C\sqrt{\frac{\|p^{-\max}_{-\varepsilon_+/16}\|_{2/3}}{n}} + \frac{C}{n} \leq C\sqrt{\frac{\|p^{-\max}_{-\varepsilon_-}\|_{2/3}}{n}} + \frac{C}{n} \leq \frac{C}{c}\varepsilon_-,$$

so that the bounds also match.

Acknowledgments. Both authors acknowledge fruitful discussions with Alexandre Tsybakov, Cristina Butucea and Rajarshi Mukherjee.

Funding. A. Carpentier is partially supported by the Deutsche Forschungsgemeinschaft (DFG) Emmy Noether grant MuSyAD (CA 1488/1-1), by the DFG-314838170, GRK 2297 MathCoRe, by the FG DFG, by the DFG CRC 1294 "Data Assimilation", Project A03, by the Forschungsgruppe FOR 5381 "Mathematical Statistics in the Information Age – Statistical Efficiency and Computational Tractability", Project TP 02, by the Agence Nationale de la Recherche (ANR) and the DFG on the French-German PRCI ANR ASCAI CA 1488/4-1 "Aktive und Batch-Segmentierung, Clustering und Seriation: Grundlagen der KI" and by the UFA-DFH through the French-German Doktorandenkolleg CDFA 01-18 and by the SFI Sachsen-Anhalt for the project RE-BCI.

References

[1] E. Abbe, Community detection and stochastic block models: recent developments. *J. Mach. Learn. Res.* **18** (2017), Art. ID 177 MR 3827065

[2] E. Abbe and C. Sandon, Achieving the ks threshold in the general stochastic block model with linearized acyclic belief propagation. In *NIPS 2016 – Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 1342–1350, Curran Associates Inc., 2016

[3] J. Acharya, C. L. Canonne, and H. Tyagi, Inference under information constraints. II. Communication constraints and shared randomness. *IEEE Trans. Inform. Theory* **66** (2020), no. 12, 7856–7877 Zbl 1457.62092 MR 4191031

[4] J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, S. Pan, and A. Suresh, Competitive classification and closeness testing. In *COLT 2012 – 25th Annual Conference on Learning Theory*, pp. 22.1–22.18, Proceedings of Machine Learning Research 23, PMLR, 2012

[5] R. Albert and A.-L. Barabási, Statistical mechanics of complex networks. *Rev. Modern Phys.* **74** (2002), no. 1, 47–97 Zbl 1205.82086 MR 1895096

[6] E. Arias-Castro and N. Verzelen, Community detection in dense random networks. *Ann. Statist.* **42** (2014), no. 3, 940–969 Zbl 1305.62035 MR 3210992

[7] C. Bădescu, R. O'Donnell, and J. Wright, Quantum state certification. In *STOC 2019 – Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pp. 503–514, ACM, New York, 2019 Zbl 1433.68146 MR 4003359

[8] S. Balakrishnan and L. Wasserman, Hypothesis testing for high-dimensional multinomials: a selective review. *Ann. Appl. Stat.* **12** (2018), no. 2, 727–749 Zbl 1405.62061 MR 3834283

[9] S. Balakrishnan and L. Wasserman, Hypothesis testing for densities and high-dimensional multinomials: sharp local minimax rates. *Ann. Statist.* **47** (2019), no. 4, 1893–1927 Zbl 1466.62307 MR 3953439

[10] Y. Baraud, Non-asymptotic minimax rates of testing in signal detection. *Bernoulli* **8** (2002), no. 5, 577–606 Zbl 1007.62042 MR 1935648

[11] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White, Testing that distributions are close. In *41st Annual Symposium on Foundations of Computer Science (Redondo Beach, CA, 2000)*, pp. 259–269, IEEE Comput. Soc. Press, Los Alamitos, CA, 2000 Zbl 1281.68227 MR 1931824

[12] P. Bedi and C. Sharma, Community detection in social networks. *Wiley Interdiscip. Rev. Data Mining and Knowledge Discovery* **6** (2016), no. 3, 115–135

[13] N. Berger, C. Borgs, J. T. Chayes, and A. Saberi, On the spread of viruses on the internet. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 301–310, ACM, New York, 2005 Zbl 1297.68029 MR 2298278

[14] T. Berrett and C. Butucea, Locally private non-asymptotic testing of discrete distributions is faster using interactive mechanisms. *NeurIPS 2020 – Advances in Neural Information Processing Systems 33*, pp. 3164–3173, Curran Associates, Inc., 2020

[15] B. Bhattacharya and G. Valiant, Testing closeness with unequal sized samples. In *NeurIPS 2015 – Advances in Neural Information Processing Systems 28*, pp. 2611–2619, Curran Associates, Inc., 2015

[16] E. Blais, C. L. Canonne, and T. Gur, Distribution testing lower bounds via reductions from communication complexity. *ACM Trans. Comput. Theory* **11** (2019), no. 2, Art. ID 6 Zbl 1440.68323 MR 3940784

[17] S. Bubeck, J. Ding, R. Eldan, and M. Z. Rácz, Testing for high-dimensional geometry in random graphs. *Random Structures Algorithms* **49** (2016), no. 3, 503–532 Zbl 1349.05315 MR 3545825

[18] C. L. Canonne, A survey on distribution testing: Your data is big. But is it blue? *Theory Comput.* **9** (2020), 1–100

[19] C. L. Canonne, Topics and techniques in distribution testing: A biased but representative sample.

[20] C. L. Canonne, I. Diakonikolas, D. M. Kane, and A. Stewart, Testing Bayesian networks. *IEEE Trans. Inform. Theory* **66** (2020), no. 5, 3132–3170 Zbl 1448.62081 MR 4089774

[21] S.-O. Chan, I. Diakonikolas, G. Valiant, and P. Valiant, Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1193–1203, ACM, New York, 2014 Zbl 1421.68184 MR 3376448

[22] J. Chhor and A. Carpentier, Goodness-of-fit testing for Hölder continuous densities: Sharp local minimax rates. 2021, arXiv:2109.04346

[23] O. Collier, L. Comminges, and A. B. Tsybakov, Minimax estimation of linear and quadratic functionals on sparsity classes. *Ann. Statist.* **45** (2017), no. 3, 923–958 Zbl 1368.62191 MR 3662444

[24] D. P. Croft, J. R. Madden, D. W. Franks, and R. James, Hypothesis testing in animal social networks. *Trends in Ecology & Evolution* **26** (2011), no. 10, 502–507

[25] S. Dan and B. B. Bhattacharya, Goodness-of-fit tests for inhomogeneous random graphs. In *ICML 2020 – Proceedings of the 37th International Conference on Machine Learning*, pp. 2335–2344, Proceedings of Machine Learning Research 119, PMLR, 2020

[26] C. Daskalakis, N. Dikkala, and G. Kamath, Testing Ising models. *IEEE Trans. Inform. Theory* **65** (2019), no. 11, 6829–6852 Zbl 1433.62146 MR 4030862

[27] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E* **84** (2011) no. 6, Art. ID 066106

[28] I. Diakonikolas and D. M. Kane, A new approach for testing properties of discrete distributions. In *FOCS 2016 – 57th Annual IEEE Symposium on Foundations of Computer Science*, pp. 685–694, IEEE Computer Soc., Los Alamitos, CA, 2016 MR 3631031

[29] M. S. Ermakov, Minimax nonparametric testing of hypotheses on a distribution density. *Theory Probab. Its Appl.* **39** (1995), no. 3, 396–416 MR 1347182

[30] C. Gao and J. Lafferty, Testing network structure using relations between small subgraph probabilities. 2017, arXiv:1704.06742

[31] D. Ghoshdastidar, M. Gutzeit, A. Carpentier, and U. von Luxburg, Two-sample tests for large random graphs using network statistics. 2017, arXiv:1705.06168

[32] D. Ghoshdastidar, M. Gutzeit, A. Carpentier, and U. von Luxburg, Two-sample hypothesis testing for inhomogeneous random graphs. *Ann. Statist.* **48** (2020), no. 4, 2208–2229 Zbl 1456.62108 MR 4134792

[33] D. Ghoshdastidar and U. von Luxburg, Practical methods for graph two-sample testing. *NeurIPS 2018 – Advances in Neural Information Processing Systems 31*, pp. 3019–3028, Curran Associates, Inc., 2018

[34] E. Giné and R. Nickl, *Mathematical foundations of infinite-dimensional statistical models*. Camb. Ser. Stat. Probab. Math. 40, Cambridge Univ. Press, New York, 2016 Zbl 1358.62014 MR 3588285

[35] C. E. Ginestet, J. Li, P. Balachandran, S. Rosenberg, and E. D. Kolaczyk, Hypothesis testing for network data in functional neuroimaging. *Ann. Appl. Stat.* **11** (2017), no. 2, 725–750 Zbl 1391.62217 MR 3693544

[36] O. Goldreich, S. Goldwasser, and D. Ron, Property testing and its connection to learning and approximation. *J. ACM* **45** (1998), no. 4, 653–750 Zbl 1065.68575 MR 1675099

[37] D. R. Hyduke, N. E. Lewis, and B. Ø. Palsson, Analysis of omics data with genome-scale models of metabolism. *Molecular BioSystems* **9** (2013), no. 2, 167–174

[38] Y. I. Ingster, Asymptotically minimax testing of nonparametric hypotheses on the density of the distribution of an independent sample. *Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI)* **136** (1984), 74–96 Zbl 0561.62006 MR 758477

[39] Y. I. Ingster, A minimax test of nonparametric hypotheses on the density of a distribution in L_p metrics. *Teor. Veroyatnost. i Primenen.* **31** (1986), no. 2, 384–389 MR 851000

[40] Y. I. Ingster and I. A. Suslina, *Nonparametric goodness-of-fit testing under Gaussian models*. Lect. Notes Stat. 169, Springer, New York, 2003 Zbl 05280099 MR 1991446

[41] I. Kim, S. Balakrishnan, and L. Wasserman, Robust multivariate nonparametric tests via projection averaging. *Ann. Statist.* **48** (2020), no. 6, 3417–3441 Zbl 1460.62087 MR 4185814

[42] S. Kotekal and C. Gao, Minimax rates for sparse signal detection under correlation. 2021, arXiv:2110.12966

[43] J. Lei, A goodness-of-fit test for stochastic block models. *Ann. Statist.* **44** (2016), no. 1, 401–424 Zbl 1331.62283 MR 3449773

[44] L. Lovász, *Large networks and graph limits*. Amer. Math. Soc. Colloq. Publ. 60, Amer. Math. Soc. Providence, RI, 2012 Zbl 1292.05001 MR 3012035

[45] S. Moreno and J. Neville, Network hypothesis testing using mixed kronecker product graph models. In *2013 IEEE 13th International Conference on Data Mining*, pp. 1163–1168, IEEE, 2013

[46] J. Neyman and E. S. Pearson, On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. R. Soc. Lond. A* **231** (1933), 289–337 Zbl 0006.26804

[47] L. Paninski, A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Trans. Inform. Theory* **54** (2008), no. 10, 4750–4755 Zbl 1322.62082 MR 2591136

[48] J. Qian and V. Saligrama, Efficient minimax signal detection on graphs. *NeurIPS 2014 – Advances in Neural Information Processing Systems 27*, pp. 2708–2716, Curran Associates, Inc., 2014

[49] R. Rubinfeld and M. Sudan, Robust characterizations of polynomials with applications to program testing. *SIAM J. Comput.* **25** (1996), no. 2, 252–271 Zbl 0844.68062 MR 1379300

[50] N. Shervashidze, P. Schweitzer, E. J. van Leeuwen, K. Mehlhorn, and K. M. Borgwardt, Weisfeiler–Lehman graph kernels. *J. Mach. Learn. Res.* **12** (2011), 2539–2561 Zbl 1280.68194 MR 2845672

[51] M. Tang, A. Athreya, D. L. Sussman, V. Lyzinski, Y. Park, and C. E. Priebe, A semiparametric two-sample hypothesis testing problem for random graphs. *J. Comput. Graph. Statist.* **26** (2017), no. 2, 344–354 Zbl 1450.62040 MR 3640191

[52] A. B. Tsybakov, *Introduction to nonparametric estimation*. Springer Ser. Statist., Springer, New York, 2009 Zbl 1176.62032 MR 2724359

[53] G. Valiant and P. Valiant, An automatic inequality prover and instance optimal identity testing. *SIAM J. Comput.* **46** (2017), no. 1, 429–455 Zbl 1362.62107 MR 3614697

[54] N. Verzelen and E. Arias-Castro, Community detection in sparse random networks. *Ann. Appl. Probab.* **25** (2015), no. 6, 3465–3510 Zbl 1326.05145 MR 3404642

[55] B. Waggoner, ℓ_p testing and learning of discrete distributions. In *ITCS 2015 – Proceedings of the 6th Innovations in Theoretical Computer Science*, pp. 347–356, ACM, New York, 2015 Zbl 1364.68364 MR 3419028

[56] M. Wang, C. Wang, J. X. Yu, and J. Zhang, Community detection in social networks: an in-depth benchmarking study with a procedure-oriented framework. *Proceedings of the VLDB Endowment* **8** (2015), no. 10, 998–1009

Received 26 January 2021; revised 22 April 2022.

Julien Chhor
CREST-ENSAE, 5, Avenue Le Chatelier, 91120 Palaiseau, France; julien.chhor@ensae.fr

Alexandra Carpentier
Institut für Mathematik, Universität Potsdam, Campus Golm, Haus 9,
Karl-Liebknecht-Straße 24–25, 14476 Potsdam, Germany; carpentier@uni-potsdam.de