Topics in high-dimensional and non-parametric inference

Julien Chhor

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2022IPPAG005

Thèse de doctorat

Topics in high-dimensional and non-parametric inference

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École Nationale de la Statistique et de l'Administration Économique

École doctorale n°574 École Doctorale de Mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées

Thèse présentée et soutenue à Palaiseau, le 30/08/2022, par

JULIEN CHHOR

Composition du Jury :

Cristina Butucea Professeur, CREST-ENSAE	Présidente
Chao Gao Assistant Professor, University of Chicago	Rapporteur
Béatrice Laurent-Bonneau Professeur, INSA Toulouse	Rapportrice
Yannick Baraud Full Professor, University of Luxemburg	Examinateur
Richard Samworth Professor, University of Cambridge	Examinateur
Alexandre Tsybakov Professeur, CREST-ENSAE	Directeur de thèse

Remerciements

Mes plus sincères remerciements vont tout d'abord à mon directeur de thèse, Alexandre Tsybakov. Tout au long de ces trois années, ton implication, tes encouragements et tes conseils m'ont été extrêmement précieux. Très disponible, tu as su me guider avec beaucoup d'enthousiasme et de bienveillance, toujours heureux de partager tes connaissances encyclopédiques et en ayant profondément à coeur la réussite de tes étudiants. Pour toutes ces raisons, je mesure la chance que j'ai eue de débuter en recherche avec toi et j'espère que notre collaboration se poursuivra bien au-delà de ma thèse.

I would also like to warmly thank Chao Gao and Béatrice Laurent, who generously accepted to review this thesis and provided me with very detailed and positive reports as well as lots of constructive remarks. I am deeply honored and I couldn't be more thankful. I would also like to express my gratitude to Cristina Butucea, Richard Samworth and Yannick Baraud for agreeing to be members of my Jury.

Je souhaite également remercier du fond du coeur toutes celles et ceux avec qui j'ai pu travailler au cours de ma thèse. En premier lieu, un immense merci à mon encadrante de stage, Alexandra, qui en plus d'être une mathématicienne particulièrement douée, est une personne remarquable par sa gentillesse et son dévouement. Je tiens également à remercier Olga, avec qui c'est un vrai plaisir de travailler. Merci également à Jaouad, qui m'a énormément appris.

Cette aventure aurait été bien différente si je n'avais pas rencontré de camarades aussi formidables. Merci à Flore pour ton enthousiasme communicatif et pour toutes les heures passées à s'entraider. Suzanne, merci pour ta gentillesse et pour tous les moments musicaux dont je me souviendrai bien longtemps. Merci à Yannis de toujours veiller à la bonne ambiance du labo et de sensibiliser les gens à l'écologie. Un grand merci également à Alexandre, Amir, Arshak, Arya, Aurélien, Avo, Badr, Clara, Côme, Corentin, Dang, Davit, Etienne, Evgenii, Fabien, François-Pierre, Gabriel, Geoffrey, Hugo, Jérémy, Jules, Lionel, Lucas, Lucie, Maria, Martin, Meyer, Nayel, Nicolas, Sasila, Simo, Solenne, Théo, Xiao, Yann, Yannick, Younès, Zong. Merci aux membres permanents du CREST, Arnak, Nicolas, Anna, Victor-Emmanuel, Jaouad, Cristina, Matthieu, Guillaume, Vianney pour toutes ces discussions pleines de bons conseils.

Merci aussi à mes amis, sur qui je sais que je pourrai toujours compter. Une liste loin d'être exhaustive: Timothée, Guillaume, Pauline, Pierre-Etienne, Clément, Olivier, Ruihua, Quentin, Mathieu, Cyril, Tarek, Chenzhang, Pierre, Tom, Amélie, Galatée, Vincent, Antoine, Angelica.

Merci à Thomas, qui continue de me suivre même depuis Singapour, et à Léo, pour toutes ces belles aventures. Merci enfin à mes parents et à ma petite soeur, qui m'ont toujours soutenu et ont toujours su m'aider à faire les bons choix.

Contents

Chapter 1

Introduction

1.1 Introduction

This thesis explores some topics in minimax testing (Part I, Chapters 2 and 3), estimation with learning constraints (Part II, Chapter 4) and benign overfitting in non-parametric regression (Part III, Chapter 5).

- The first part (Chapters 2 and 3) is devoted to minimax testing. Given n i.i.d. observations with distribution p, the goodness-of-fit testing problem aims at testing equality to a given probability distribution p_0, against an alternative composed of distributions separated from p_0 with respect to some distance over the considered class of distributions. We consider two goodness-of-fit testing problems. The first one concerns the discrete case (Chapter 2), where the observations can be multivariate Binomial or Poisson families or follow multinomial distributions. The second one (Chapter 3) is an extension of the first to the continuous case, where the observations are i.i.d. with some probability density over $\mathbb{R}^d$ that belongs to the Hölder class of function with known parameters $\alpha, L > 0$. In both cases, we are interested in the *local* version of the problem (see Subsection 1.2.3)

- The second part proposes to study an estimation problem under learning constraints. Namely, in Chapter 4, we study the interactions between robustness to adversarial contamination and local differential privacy in the context of learning discrete distributions.

- In the third part (Chapter 5), we consider the setting of non-parametric regression. We propose an estimator of the regression function that is minimax optimal adaptively to the unknown smoothness, and that continuously interpolates the data points with high probability - a phenomenon called "benign overfitting".

The notation may change from chapter to chapter.

1.2 Minimax testing

Let $\mathcal{P} = \{\mathbb{P}_\theta \mid \theta \in \Theta\}$ denote a family of probability distributions over a measurable space $(\mathcal{X}, \mathcal{U})$, where Θ is a set of parameters, not necessarily finite-dimensional. Let $\Theta_0, \Theta_1 \subset \Theta$ be two disjoint

subsets of Θ : $\Theta_0 \cap \Theta_1 = \emptyset$. We observe $n \in \mathbb{N}^*$ i.i.d. datapoints $X_1, \dots, X_n \overset{iid}{\sim} \mathbb{P}_\theta$ for some unknown $\theta \in \Theta$. Given $X_1, \dots, X_n$, we consider the testing problem:

$$H_0 : \theta \in \Theta_0 \qquad \text{against} \qquad H_1 : \theta \in \Theta_1. \tag{1.1}$$

Definition 1.1. *A **test** is a measurable function of the observations taking its values in* $\{0, 1\}$*:*

$$\psi : \mathcal{X}^n \longrightarrow \{0, 1\}.$$

In this work we focus on constructing minimax optimal tests. There are essentially three paradigms for studying the quality and optimality of a test.

1. **Neyman-Pearson's approach**: Neyman-Pearson's approach [1] is well adapted if H_0 is the hypothesis that one wishes to believe by default, unless the data provides strong evidence against it. The Type-I error probability of a test ψ is defined as the worst-case probability of deciding in favor of H_1 if H_0 holds: $\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\psi = 1)$ while the Type-II error function is the function $\Theta_1 \longrightarrow [0, 1]$; $\theta \mapsto \mathbb{P}_\theta(\psi = 0)$. This paradigm posits that among the two possible errors (deciding wrongly in favor of H_1 or deciding wrongly in favor of H_0), deciding wrongly in favor of H_1 is the most dangerous one. An ideal test would never wrongly reject H_0 – however, in usual cases, the only test having 0 Type-I error probability is the trivial test $\psi \equiv 0$. It is therefore necessary to allow the tests to have a small Type-I error probability, constrained to be at most α for some $\alpha \in [0, 1]$ chosen beforehand. Under this constraint, the *optimal* test ψ_{NP} would be the test performing uniformly in the best way at each point of Θ_1, hence having uniformly the smallest Type-II error function. Mathematically, the optimal test ψ_{NP} would solve the following problem: $\forall \theta \in \Theta_1, \mathbb{P}_\theta(\psi_{NP} = 0) = \inf_{\psi \in \Psi_\alpha} \mathbb{P}_\theta(\psi = 0)$ where Ψ_α denotes the set of all tests with Type-I error probability at most α. Such a test, called a *uniformly most powerful test* is not guaranteed to exist. Therefore, the Neyman-Pearson approach does not allow for a universal notion of optimality.

2. **Bayesian approach:** In the Bayesian approach, one defines a prior probability distribution π over $\Theta_0 \cup \Theta_1$ and looks for a test ψ minimizing the *Bayes risk* defined as $\mathbb{E}_{\theta \sim \pi}\left[\mathbb{P}_\theta\left(\psi = \mathbb{1}_{\theta \in \Theta_0}\right)\right]$, if such a test exists. Here $\mathbb{E}_{\theta \sim \pi}$ denotes the expectation with respect to π and $\mathbb{1}$ denotes the indicator function. In this approach, it is always guaranteed that at least one test ψ^* has a Bayes risk satisfying $\mathbb{E}_{\theta \sim \pi}\left[\mathbb{P}_\theta\left(\psi^* = \mathbb{1}_{\theta \in \Theta_0}\right)\right] = \inf_\psi \mathbb{E}_{\theta \sim \pi}\left[\mathbb{P}_\theta\left(\psi = \mathbb{1}_{\theta \in \Theta_0}\right)\right]$. Therefore, the notion of optimality is well-defined. The drawback of this approach is that the choice of π is subjective and different choices of π lead, in general, to different solutions.

3. **Minimax approach:** The minimax paradigm has been widely studied since Wald (1949) and more recently developed in the context of tests in the works of Yuri Ingster, see [183] for a detailed account. The Type-I error of a test ψ is defined as $\max_{\theta \in \Theta_0} \mathbb{P}_\theta(\psi = 1)$ and the Type-II error of ψ as $\max_{\theta \in \Theta_1} \mathbb{P}_\theta(\psi = 0)$. One measures the quality of a test through its *risk*, defined as the sum of its Type-I and Type-II errors: $R_{\Theta_0, \Theta_1}(\psi) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\psi = 1) + \sup_{\theta \in \Theta_1} \mathbb{P}_\theta(\psi = 0)$. In this paradigm, H_0 and H_1 therefore play equally important roles and one aims at constructing tests that have both small Type-I and Type-II errors *in the worst case*. As we will see later on, this approach always ensures that there exist optimal tests in a sense defined below.

In what follows, we place ourselves in the minimax setting. We first give some definitions that will be used in the thesis. The **minimax risk** is defined as the risk of the best test, if any.

Definition 1.2. *The minimax risk associated with Problem (1.1) is defined as*

$$\begin{aligned} R^*_{\Theta_0,\Theta_1} &= \inf_{\psi} R_{\Theta_0,\Theta_1}(\psi) \\ &= \inf_{\psi} \left\{ \sup_{\theta\in\Theta_0} \mathbb{P}_\theta\left(\psi = 1\right) + \sup_{\theta\in\Theta_1} \mathbb{P}_\theta\left(\psi = 0\right) \right\}, \end{aligned}$$

where the infimum is taken over all tests ψ.

Note that if $\Theta_0, \Theta_1 \subseteq \Theta$ are such that $R^*_{\Theta_0,\Theta_1} = 1$, then random guessing is optimal. Indeed, define the test $\widetilde{\Delta}$ that takes values 1 and 0 with probability $\frac{1}{2}$ independently of the observed data. Its risk is equal to

$$R_{\Theta_0,\Theta_1}(\widetilde{\Delta}) = \sup_{\theta\in\Theta_0} \mathbb{P}_\theta\Big(\widetilde{\Delta} = 1\Big) + \sup_{\theta\in\Theta_1} \mathbb{P}_\theta\Big(\widetilde{\Delta} = 0\Big) = \frac{1}{2} + \frac{1}{2} = 1.$$

Thus, if $R^*_{\Theta_0,\Theta_1} = 1$, then $\widetilde{\Delta}$ is optimal, so that the problem has trivial solution. It is therefore natural to consider only the testing problems where $R^*_{\Theta_0,\Theta_1}$ is smaller than 1, and we will assume from now on that $R_{\Theta_0,\Theta_1}(\widetilde{\Delta}) \leq \eta$ where $\eta < 1$ is chosen in advance. Minimax testing with composite null hypothesis has been studied in a wide variety of settings, see for instance [155, 135, 71, 173, 166, 40] to cite but a few. However, in the present thesis, we will focus on the particular case where H_0 is simple.

1.2.1 Goodness-of-fit testing problem

The case of a simple null hypothesis is referred to as the *goodness-of-fit testing problem*. Assume that Θ is a metric space equipped with the distance **dist** and fix $\eta \in (0,1)$ as well as $\theta_0 \in \Theta$. The total variation between two probability measures $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$ is defined as $TV(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \sup_{U\in\mathcal{U}} \left|\mathbb{P}_\theta(U) - \mathbb{P}_{\theta'}(U)\right|$. We assume that the total variation distance is continuous with respect to **dist**. For $\rho > 0$, the **goodness-of-fit** testing problem, also called the **identity testing problem**, is defined as follows:

$$H_0 : \ \theta = \theta_0 \qquad \text{against} \qquad H_1^{(\rho)} : \ \begin{cases} \theta \in \Theta, \\ \mathbf{dist}(\theta, \theta_0) \geq \rho. \end{cases} \tag{$\star$}$$

One could wonder why the testing problem is not defined as

$$H_0' : \ \theta = \theta_0 \qquad \text{against} \qquad H_1' : \ \begin{cases} \theta \in \Theta, \\ \theta \neq \theta_0. \end{cases} \tag{1.2}$$

To understand why, denote by R^* the minimax risk of problem (1.2). In this case, for any $\theta_1 \neq \theta_0$, we would have

$$1 \geq \inf_{\psi} \left\{ \mathbb{P}_{\theta_0}(\psi = 1) + \mathbb{P}_{\theta_1}(\psi = 0) \right\}$$

$$
\begin{aligned}
&= 1 + \inf_{\psi} \{\mathbb{P}_{\theta_1}(\psi = 0) - \mathbb{P}_{\theta_0}(\psi = 0)\} \\
&= 1 - \sup_{U \in \mathcal{U}} \mathbb{P}_{\theta_1}(U) - \mathbb{P}_{\theta_0}(U) \\
&= 1 - \sup_{U \in \mathcal{U}} |\mathbb{P}_{\theta_1}(U) - \mathbb{P}_{\theta_0}(U)| \\
&= 1 - TV(\mathbb{P}_{\theta_0}, \mathbb{P}_{\theta_1}) \longrightarrow 1 \quad \text{when } \mathbf{dist}(\theta_1, \theta_0) \to 0,
\end{aligned}
$$

since the total variation distance is continuous with respect to **dist**. For testing problem (1.2), we therefore have $R^*_{\Theta_0,\Theta_1} = 1$. In other words, random guessing is optimal and the problem is trivial. To ensure that the minimax risk is less than $\eta \in (0,1)$, we therefore define the goodness-of-fit testing problem as in ($\star$). To do this, we introduce the notation

$$
R^*(\rho) = \inf_{\psi} \left[\mathbb{P}_{\theta_0}(\psi = 1) + \sup \left\{ \mathbb{P}_{\theta}(\psi = 0) \;\middle|\; \theta \in \Theta,\ \mathbf{dist}(\theta, \theta_0) \geq \rho \right\} \right] \tag{1.3}
$$

to denote the minimax risk associated with problem ($\star$). Noting that $\rho \mapsto R^*(\rho)$ is a non-increasing function, we aim at finding the smallest separation distance $\rho > 0$ ensuring that $R^*(\rho) \leq \eta$.

Definition 1.3. *[Minimax separation radius] The minimax separation radius of problem ($\star$) is defined as*

$$
\rho^*(n, \theta_0, \Theta, \boldsymbol{dist}, \eta) = \inf \left\{ \rho > 0 \;\middle|\; R^*(\rho) \leq \eta \right\}.
$$

Moreover, we denote the risk of any test ψ by

$$
R(\rho, \psi) := \mathbb{P}_{\theta_0}(\psi = 1) + \sup \left\{ \mathbb{P}_{\theta}(\psi = 0) \,\middle|\, \theta \in \Theta,\ \mathbf{dist}(\theta, \theta_0) \geq \rho \right\}.
$$

The aim of the goodness-of-fit testing problem is two-fold:

1. Derive ρ^* up to multiplicative constants.
2. Find a test ψ^* and a constant $C > 0$ such that $R(C\rho^*, \psi) \leq \eta$. Such a test is called a *minimax optimal test*.

Remark: In the literature, the goodness-of-fit testing problem is sometimes formulated in an alternate way. In the above definition, we fixed the number of observations n and aimed at deriving the corresponding minimax separation radius ρ^*. However, the testing problem can also be expressed in terms of *sample complexity*: For a fixed precision $\epsilon > 0$, derive how many observations n^* are necessary and sufficient, up to a multiplicative constant depending only on η, **dist** and Θ, in order for the testing problem

$$
H_0 : \theta = \theta_0 \quad \text{against} \quad H_1(\epsilon) : \begin{cases} \theta \in \Theta, \\ \mathbf{dist}(\theta, \theta_0) \geq \epsilon, \end{cases}
$$

to have a minimax risk at most η. This problem statement can be understood as a dual version of the statement based on the fixed number of observations and Definition 1.3 and often leads to very

similar results. This convention is the most popular in the computer science community (see for instance [71, 95, 81, 100, 67]), whereas the formulation based on the fixed number of observations and Definition 1.3 is standard in the statistics community.

1.2.2 Classical rates for signal detection

In the framework described above, the problem of signal detection is defined as the special case where $\Theta_0 = \{0\}$. One of the earliest papers on the subject is [13] where the observations are assumed to follow the Gaussian white noise model, which we define below. The series of subsequent papers [22] is regarded as a landmark in non-parametric signal detection. The case where Θ is an ellipsoid is considered in [20], and the case where Θ is a Sobolev or Besov ball is considered in [26], [182], [31] to cite just a few. The above references all deal with the asymptotic regime. On the contrary, the present thesis will focus on non-asymptotic rates, which are less common in the literature (see for instance [36], [59], [87], [162], [44], [174], [51]). We do not give a comprehensive survey of the literature and we refer the reader to [183] for an excellent overview. We will limit ourselves to reviewing the classical facts that are helpful to be compared with the results developed in the thesis.

Gaussian setting

The spherical Gaussian framework represents one of the most classical problems studied in signal detection. For $d \geq 1$, assume that we observe $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(\theta, I_d)$ where I_d denotes the identity matrix of size d and $\theta \in \mathbb{R}^d$. The testing problem can be written as

$$H_0 : \theta = 0 \quad \text{against} \quad H_1(\rho) : \begin{cases} \theta \in \mathbb{R}^d, \\ \|\theta\|_2 \geq \rho, \end{cases} \tag{1.4}$$

where $\rho > 0$ and $\|\cdot\|_2$ denotes the Euclidean norm. In this classical case, the minimax separation radius is known to be $\rho^* \asymp d^{\frac{1}{4}}/\sqrt{n}$ and the minimax optimal test can be written as $\mathbb{1}\left\{\|\bar{X}_n\|_2^2 \geq d + c\right\}$ where $c > 0$ is a constant and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ (see for example [162], [87]). It follows that, for any precision $\epsilon > 0$, testing equality to 0 against the alternative $\|\theta\|_2^2 \geq \epsilon^2$ is only possible if $n \gtrsim \frac{\sqrt{d}}{\epsilon^2}$.

The interest of this result is best understood when compared with the corresponding estimation problem. Given $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(\theta, I_d)$, the minimax estimation risk over $\mathbb{R}^d$ (defined as $\inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}^d} \mathbb{E}\|\hat{\theta} - \theta\|^2$) is known to be equal to $\frac{d}{n}$, and can be achieved by the empirical mean $\hat{\theta} = \bar{X}_n$. In other words, for any precision $\epsilon > 0$, building an estimator $\hat{\theta}$ of θ with an estimation risk $\sup_{\theta \in \mathbb{R}^d} \mathbb{E}\|\hat{\theta} - \theta\|^2 \leq \epsilon^2$ is only possible when $n \geq \frac{d}{\epsilon^2}$.

Therefore, an interesting phenomenon arises when $\frac{\sqrt{d}}{n} \lesssim \|\theta\|_2^2 \ll \frac{d}{n}$. On the one hand, detecting from the data that $\theta \neq 0$ is possible with high probability. But on the other hand, no estimator can outperform the trivial estimator equal to 0. This point illustrates an important advantage of testing over estimation. Although it only provides a binary piece of information (thus being more limited than what estimation offers), testing achieves faster rates than estimation. In high dimen-

sions, this improvement can help considerably reduce the sample size needed for drawing statistical conclusions.

Consequently, goodness-of-fit testing has received considerable attention over the past decades, and the Gaussian setting with L^2 separation has been extensively studied see for example [183]). Beyond the L^2 case, some works such as [111] investigated the effect of the separation distance on the minimax separation radius as well as on the minimax tests. Namely, given $X \sim \mathcal{N}(\theta, \sigma^2 I_d)$, the work [111] considers the following testing problem:

$$H_0 : \theta = 0 \quad \text{against} \quad H_1 : \theta \in \mathbb{R}^d,\ \|\theta\|_t \geq \rho, \tag{1.5}$$

for $t \geq 1$, and shows that the minimax separation radius scales as $\rho^* \asymp \sigma d^{\frac{1}{t}-\frac{1}{4}}$ when $t \in [1, 2]$ and as $\rho^* \asymp \sigma d^{\frac{1}{2t}}$ otherwise.

Nonparametric setting

A non-parametric variant of the spherical Gaussian setting described above is the Gaussian white noise model, where one observes the random process X_t defined via $dX_t = f(t)dt + \sigma dW_t$, $t \in [0, 1]$, and where $(W_t)_t$ denotes a standard Wiener process. Assume that the unknown signal f belongs to a Sobolev class $W(\beta, L)$ (see [188] for the definition) for some $\beta, L > 0$. In this nonparametric setting, the signal detection problem is defined as follows:

$$H_0 : f = 0 \quad \text{against} \quad H_1(\rho) : \begin{cases} \|f\|_2 \geq \rho, \\ f \in W(\beta, L). \end{cases} \tag{1.6}$$

For this problem, the minimax separation radius is of the order of $\sigma^{\frac{4\beta}{4\beta+1}}$ (see [183]). This nonparametric model is closely related to the previous one in the periodic case: Namely, one can show that by projecting onto the Fourier basis, the Gaussian white noise model is equivalent to the Gaussian sequence model

$$Y_j = \theta_j + \sigma\xi_j, \quad j \in \mathbb{N},$$

for $\xi_j \overset{iid}{\sim} \mathcal{N}(0, 1)$ and $\theta \in Q(\beta, L)$ where $Q(\beta, L) = \left\{\theta \in \ell^2(\mathbb{N}) : \sum_{j\in\mathbb{N}} \theta_j^2 j^2 \leq \frac{L^2}{\pi^{2\beta}}\right\}$, cf. [188]. By the Parseval identity, problem (1.6) can be equivalently re-written as

$$H_0 : \theta = 0 \quad \text{against} \quad H_1(\rho) : \begin{cases} \|\theta\|_2 \geq \rho, \\ \theta \in Q(\beta, L), \end{cases} \tag{1.7}$$

which is closely related to Problem (1.4).

We can now introduce the most closely related problem to the results developed in this thesis. This problem aims at testing equality to the uniform distribution over $[0, 1]^d$ against an alternative composed of Hölder-continuous densities over $[0, 1]^d$ and separated from the null hypothesis in L^2 distance [179]. Denoting by p_0 the uniform density over $[0, 1]^d$ and given n i.i.d. observations

$X_1, \dots, X_n$ with Hölder-smooth density p over $[0,1]^d$ and known smoothness parameter $\alpha > 0$, this problem consists in testing

$$H_0 : p = p_0 \quad \text{against} \quad H_1(\rho) : \begin{cases} \|p - p_0\|_2 \geq \rho, \\ p \in H(\alpha). \end{cases} \tag{1.8}$$

Here $H(\alpha)$ denotes the class of α-Hölder continuous functions over $[0,1]^d$ with smoothness parameter α and Lipschitz constant normalized to 1. The asymptotic minimax separation radius as $n \to \infty$ is known to be $\rho^* \asymp n^{-\frac{2\alpha}{4\alpha+d}}$, see e.g. [22]. Again, we can compare this rate with the minimax estimation rate of a Hölder-smooth density in $H(\alpha)$ which is known to be $n^{-\frac{\alpha}{2\alpha+d}}$ (see [188]), always slower than the non-parametric testing rate.

Density testing problems or estimation of the quadratic functional have also been considered in [24, 44, 42, 27, 33].

1.2.3 Local testing problem

In this subsection, we introduce a further distinction between *local* and *global* testing problems. We assume that Θ, **dist** and η are fixed. The *local* version of problem ($\star$) aims to determine how ρ^* precisely depends on θ_0 and n, see e.g. [104], [95]. Conversely, the *global* testing problem only establishes the worst-case of all separation radii $\rho^*(n, \theta_0)$ for θ_0 in the class Θ: $\rho^*_{\text{global}}(n) := \sup_{\theta_0 \in \Theta} \rho^*_{\text{local}}(n, \theta_0)$.

A local dependency of ρ^* on θ naturally arises when the variance of the observations depends on the parameter θ involved in the testing problem. For example, assume that we observe $X \sim \mathcal{N}(\theta, I_d)$ and that we fix $\theta_0 \in \mathbb{R}^d$. Consider the following Gaussian testing problem

$$H_0 : \theta = \theta_0 \quad \text{against} \quad H_1 : \begin{cases} \theta \in \Theta, \\ \|\theta - \theta_0\|_2 \geq \rho. \end{cases}$$

In this testing problem local and global minimax separation radii do not differ, as the covariance matrix remains equal to I_d independently of θ_0. Therefore, in the spherical Gaussian setting, testing equality to any $\theta_0 \in \mathbb{R}^d$ is equally difficult. This is no longer the case in the Bernoulli setting. Indeed, for some fixed $p_0 \in [0,1]$ and for observations $X_1, \dots, X_n \overset{iid}{\sim} \text{Ber}(p)$ distributed as Bernoulli random variables with parameter $p \in [0,1]$, consider the following goodness-of-fit testing problem:

$$H_0 : p = p_0 \quad \text{against} \quad H_1(\rho) : \begin{cases} p \in [0,1], \\ |p - p_0| \geq \rho. \end{cases}$$

In this model, a sufficient statistic is the sum $\sum_{i=1}^n X_i$, whose variance $\sqrt{np(1-p)}$ depends on the parameter p involved in the testing problem. Here, $\rho^*(n, p_0)$ exhibits a very different behavior from the global separation radius when p_0 varies in $[0,1]$. For $p_0 = \frac{1}{2}$, the separation radius scales as $\rho^*(n, \frac{1}{2}) \asymp 1/\sqrt{n}$ and attains the global separation radius, while for $p_0 = 0$, it reduces to $\rho^*(n, 0) \asymp 1/n$ (these claims follow from Theorem 2.1 proved below). This very simple example

therefore illustrates that local rates can considerably refine the global ones. Locality will represent one of the major points of focus in the present thesis.

1.3 Inference with learning constraints

The second part (Chapter 4) of this thesis is devoted to estimating discrete distributions under learning constraints. This chapter aims at studying the interactions between local differential privacy and robustness to outliers in this estimation problem. Here we briefly introduce the settings of local differential privacy and of robust statistics.

1.3.1 Local differential privacy

One of the major challenges posed by today's data is to protect the user's privacy when collecting potentially sensitive information. For instance, a large amount of the collected data, such as medical data, can be *sensitive*, meaning that they contain information that should not be disclosed to the statistician. In order to ensure confidentiality, many solutions have been considered.

- *Anonymize the data*: A first naive solution consists in removing the individuals' names without further modifying the dataset. While this method can seem to preserve confidentiality, the resulting dataset is in fact extremely vulnerable. Indeed, it has experimentally been shown in [34] that 87% of the US population could be indirectly identified using only the following pieces of information: {date of birth, gender, ZIP}. By combining different datasets, an attacker could therefore be able to recover some sensitive information concerning the majority of the users.

- *Cryptography*: In order to protect user's privacy, cryptography can be employed to collect data in an encrypted way so that individual information cannot be deduced from its encoded version. A typical example is when a user needs to access the computational power of a cloud computing service, but refuses to disclose the sensitive information to this entity. In this case, such cryptography methods as homomorphic encryption or Secure multiparty computation can be employed. However, many issues can arise when using cryptography. First, its computational cost can be large. Second, such procedures may not be statistically robust if some of the collected data are contaminated, for example according to a Huber contamination model (see the definition in Section 1.3.2). Third, an attacker that would hack the encryption key could recover all the sensitive information that this method aims at protecting.

- *Differential privacy*: To address cryptography's latter drawback, the Differential Privacy technique proposes to randomly perturb the data, often by adding noise to it. The added noise should be sufficiently strong to ensure that none of the individual data can be recovered with sufficient precision from its noisy counterpart. At the same time, the noise should be weak enough to ensure that characteristics of the global population can be inferred when observing sufficiently many individuals. As opposed to cryptography methods, this technique guarantees a fundamental impossibility of revealing sensitive information even under attacks. Though using this technique often leads to computationally tractable methods, their statistical

cost can be heavy or even prohibitive, as we will see later on. It is therefore of primary interest to understand under which conditions differential privacy can be efficiently used, which is one of the themes addressed in this thesis.

Below, we consider the problem of estimating discrete distributions under differential privacy. We briefly formalize this notion. Let $X = (X_1, \dots, X_n)$ be a random vector in the measurable space $(\mathcal{X}^n, \mathcal{A}^n, \mathbb{P}^n)$. We would like to produce a new random variable Z, called the *privatized* version of X. Formally, the random variable Z takes its values on a second measurable space $(\mathcal{Z}, \mathcal{B})$ and is generated using the mechanism $Z|X = x \sim Q(\cdot|x)$ where $Q(\cdot\,|\cdot)$ is a Markov Kernel also called the "Privacy Mechanism". More precisely, for all $x \in \mathcal{X}^n$, we assume that $Q(\cdot|x)$ is a probability distribution, and that $Q(A\,|\,\cdot\,)$ is a measurable function for all $A \in \mathcal{A}^n$. Let $\alpha \in (0, 1)$. There are two main approaches to define the differential privacy constraint.

1. **Central (or global) differential privacy:** A privacy mechanism Q is said to be globally differentially private [46, 43] if for all $A \in \mathcal{B}$ and for all $x, x' \in \mathcal{X}^n$ such that $\sum_{i=1}^n \mathbb{1}_{\{x_i \neq x'_i\}} = 1$, we have
$$\frac{Q(A|x)}{Q(A|x')} \leq e^{\alpha}.$$
The statistician should never access $(X_1, \dots, X_n)$ which is the reason why we produce Z. However, this approach unfortunately requires that some central unit be trusted to access the whole dataset $(X_1, \dots X_n)$ in order to produce the privatized data Z. This setting is therefore very vulnerable to attacks if the central unit is hacked and the whole sensitive dataset $(X_1, \dots, X_n)$ is revealed.

2. **Local differential privacy:** To address this issue, a second formalism was proposed in [54]. The idea is to generate the privatized datapoint Z_i when actually collecting the data. Each user is responsible for sending their privatized datapoint $Z_i \sim Q(\cdot|X_i)$ to the statistician without ever revealing the true value X_i. Formally, a privacy mechanism Q satisfying $Q(dz|x) = Q(dz_1|x_1)Q(dz_2|x_2, z_1) \dots Q(dz_n|x_n, z_1, \dots, z_{n-1})$ is said to be *locally differentially private* if for any $A \in \mathcal{B}$, any $i \in [n]$, any $z_1, \dots z_{i-1} \in \mathcal{Z}^{i-1}$, and any $x, x' \in \mathcal{X}$, we have
$$\frac{Q(A|x, z_1, \dots, z_{i-1})}{Q(A|x', z_1, \dots, z_{i-1})} \leq e^{\alpha}. \tag{1.9}$$

This method never assumes the existence of a central unit, as each privatized data Z_i is generated using only X_i and all of the publicly available data $Z_1, \dots, Z_{i-1}$. A privacy mechanism satisfying (1.9) is said to be an *interactive mechanism*. A more restricted class of privacy mechanisms can be defined as follows. A privacy mechanism Q is said to be *non-interactive* if $Q(dz|x) = \prod_{i=1}^n Q_i(dz_i|x_i)$ where each Markov Kernel Q_i satisfies
$$\frac{Q_i(A|x)}{Q_i(A|x')} \leq e^{\alpha}, \ \forall x, x' \in \mathcal{X}. \tag{1.10}$$

This is the class of all local differential privacy mechanisms generating each privatized data Z_i using X_i only. This class of mechanisms allows the statistician to collect the data in parallel, each individual being privatized independently. In contrast, general interactive mechanisms defined in (1.9)

require the data to be collected sequentially, which appears a major drawback. However, interactive mechanisms have been shown to achieve better statistical performance than non-interactive ones in certain settings [126, 124]. In such cases, they should be preferred whenever possible, especially given the high statistical cost of local differential privacy.

We now explain why (1.9) is a natural definition for a privacy mechanism. This condition formalizes the fact that, given Z_i, no inference is possible about the original X_i, even with the knowledge of the publicly available data $Z_1, \dots Z_{i-1}$. To appreciate why, let us fix $z_1, \dots, z_{i-1} \in \mathcal{Z}$ and assume that we observe the privatized data Z_i. For any $x \in \mathcal{X}$, define the probability measure $\mathbb{P}_x(\cdot) = Q(\cdot|x, Z_1, \dots, Z_{i-1})$ and consider the family of two-point testing problems

$$H_0 : Z_i|X_i, Z_1, \dots, Z_{i-1} \sim \mathbb{P}_x \quad \text{against} \quad H_1 : Z_i|X_i, Z_1, \dots, Z_{i-1} \sim \mathbb{P}_{x'}, \tag{1.11}$$

for any pair $x, x' \in \mathcal{X}$. Fixing $x, x' \in \mathcal{X}$, the likelihood ratio test $\psi := \mathbb{1}\left\{\frac{d\mathbb{P}_x}{d\mathbb{P}_{x'}}(Z_i) > 1\right\}$ is optimal in the sense that it minimizes the sum of Type-I and Type-II errors over all tests. Let $A = \left\{\frac{d\mathbb{P}_x}{d\mathbb{P}_{x'}}(Z_i) > 1\right\}$, then the risk $R(\psi) =$ Type-I + Type-II error satisfies:

$$\begin{aligned} R(\psi) &= \mathbb{P}_x(A) + 1 - \mathbb{P}_{x'}(A) = 1 + \frac{\mathbb{P}_x(A) - \mathbb{P}_{x'}(A)}{\mathbb{P}_x(A)}\mathbb{P}_x(A) \\ &\geq 1 + (1 - e^{\alpha}) \quad \text{using (1.9)} \\ &\geq 1 - e\alpha \quad \text{for } \alpha \in (0, 1). \end{aligned}$$

We recall that random-guessing has a minimax risk equal to 1 (see Section 1.2). We also recall the formalism of Section 1.2 and fix $\eta \in (0, 1)$. Under the constraint (1.9), we can therefore choose α small enough to make any of the testing problems (1.11) infeasible for any pair $x, x' \in \mathcal{X}$. By "infeasible", we mean that the minimax risk of any such problem will always be greater than η. Therefore, condition (1.11) is akin to imposing that no inference is possible about X_i when observing $Z_1, \dots, Z_i$.

In this thesis, we propose to estimate discrete distributions under local differential privacy and contamination (see Section 1.3.2). Estimating discrete distributions under local differential privacy alone has been considered in [62]. The paper [62] discusses two privacy mechanisms for discrete distributions.

- **The RAPPOR mechanism**, introduced in [62], [82]. Let $d \geq 2$ and $X \in [d]$ be a random variable (we do not specify its distribution). The variable X can be privatized as follows. We define the random vector $Z \in \{0, 1\}^d$ with mutually independent entries conditionally on X, such that

$$\forall j \in [d] : \quad Z(j) = \begin{cases} \mathbb{1}_{X=j} & \text{with probability } 1 - \lambda, \\ 1 - \mathbb{1}_{X=j} & \text{with probability } \lambda, \end{cases}$$

  where $\lambda = \frac{1}{e^{\alpha/2}+1}$. The above channel, denoted by Q, is α-locally differentially private. Indeed, fix any $x, x' \in [d]$ and define $e_x = (\mathbb{1}_{x=j})_{j=1}^d$ and $e_{x'} = (\mathbb{1}_{x'=j})_{j=1}^d$. Then, for any

vector $z \in \{0,1\}^d$ we have y definition of Z:

$$Q(Z = z|x) = \lambda^{1-z_x}(1-\lambda)^{z_x} \prod_{j \neq x} \lambda^{z_j} (1-\lambda)^{1-z_j},$$

and noting that $\frac{\lambda}{1-\lambda} = e^{-\alpha/2}$, we get

$$\frac{Q(Z = z|x)}{Q(Z = z|x')} = \left(\frac{\lambda}{1-\lambda}\right)^{2z_x} \left(\frac{1-\lambda}{\lambda}\right)^{2z_{x'}} = \exp\left(\alpha(z_{x'} - z_x)\right) \in \left[e^{-\alpha}, e^{\alpha}\right].$$

- **The Laplace mechanism.** Let $W_1, \ldots, W_d$ be i.i.d.random variables with distribution $\Lambda(1)$ independent of X. Here $\Lambda(1)$ stands for the standard Laplace distribution. The random vector $Z = (\mathbb{1}_{X=j})_{j=1}^d + \frac{2}{\alpha}(W_1, \ldots, W_d)$ is a privatized version of X and we can check that the corresponding mechanism is α-LDP. Indeed, for any two points $x, x' \in [d]$, defining $e_x = (\mathbb{1}_{x=j})_{j=1}^d$ and $e_{x'} = (\mathbb{1}_{x'=j})_{j=1}^d$ and fixing any vector $z \in \mathbb{R}^d$, we have

$$\frac{Q(Z = z|X = x)}{Q(Z = z|X = x')} = \exp\left(\frac{\alpha}{2}(\|z - e_x\|_1 - \|z - e_{x'}\|_1)\right) \in \left[e^{-\alpha}, e^{\alpha}\right].$$

  For multinomial estimation, this mechanism is slightly better than the RAPPOR mechanism, in the sense that it improves the minimax estimation risk by an absolute constant factor.

Local differential privacy comes with a high statistical cost, as it requires that nearly all the information contained in the original data should be lost when generating the privatized data. For multinomial estimation, [62] observed that under α local differential privacy, the minimax rate with n observations is the same as the rate with $n\alpha^2/d$ observations with no privacy. The effect of privacy therefore amounts to shrinking the number of observations by a factor α^2/d, which, in high dimension, can be prohibitive.

More precisely for discrete distribution over d elements, the minimax estimation rate in total variation under α local differential privacy is $\frac{d}{\alpha\sqrt{n}}$ where n denotes the number of i.i.d. observations [37, 54]. Comparing this rate with the minimax estimation rate without the privacy constraint, known to be $\sqrt{\frac{d}{n}}$ [73], we note that estimating a discrete distribution under α-LDP with n observations is as difficult as estimating the same distribution without privacy but with only $\frac{\alpha^2 n}{d}$ observations.

1.3.2 Robustness to outliers

One of the most classical assumptions in statistics is to consider that the data are i.i.d.However, this assumption may fail to hold in practice, especially when working with large datasets. To weaken this assumption, a very popular approach is to assume that only one part of the datapoints, called 'inliers', are i.i.d. with some true distribution of interest. The remaining part, referred to as 'outliers' or 'contamination', consists of data points which do not follow the target distribution. In this case, we say that the dataset is *contaminated*. Informally, the goal of robust learning is

to build estimators that are not much affected by contamination, while being statistically nearly as good as the optimal estimators in absence of contamination, see e.g. [127, 121, 86]. Of course, contamination often degrades the statistical rates and an interesting question is to quantify to what extent. There are several different ways to define contamination, as summarized in [123].

1. **Huber's contamination [21], [5]:** This is one of the most commonly studied contamination models. It supposes that there exist two unknown probability distributions p and q as well as a level of contamination $\epsilon \in (0, \frac{1}{2})$. The i.i.d. observations $X_1 \dots, X_n$ are assumed to follow the mixture model $(1-\epsilon)p + \epsilon q$ where p is the target distribution, while q is an unknown contamination. Note that the number of outliers is random and follows $Bin(n, \epsilon)$.

2. **Huber's deterministic contamination:** A distribution is said to follow the Huber deterministic contamination model if there exists a set $\mathcal{O} \subset [n]$ of cardinality at most $\lceil n\epsilon \rceil$ and two distributions p, q such that for all $i \notin \mathcal{O}$, $X_i \sim p$ and for all $i \in \mathcal{O}$, $X_i \sim q$ and all of the observations $X_1, \dots, X_n$ are mutually independent.

3. **Oblivious contamination:** This model is similar to the Huber's deterministic contamination, except that the family of outliers follows some joint distribution $Q_{\mathcal{O}}$. Hence, the outliers are not assumed to be i.i.d.

4. **Parameter contamination:** For some set of outliers $\mathcal{O}$ chosen in advance, the outliers $X_i, i \in \mathcal{O}$ are independent from the inliers $X_i, i \notin \mathcal{O}$, and each outlier $X_i, i \in \mathcal{O}$ is drawn from some distribution q_i, belonging to the same class as p.

5. **Adversarial contamination:** Together with the Huber contamination model, this is one of the most popular models, on top of being the most general one. In this model, clean i.i.d. data $X_1, \dots X_n$ are generated from a distribution p. For some $\epsilon \in (0, \frac{1}{2})$, an adversary replaces $\lceil n\epsilon \rceil$ of the data with new data points. Nothing is assumed the outliers: They are allowed to be deterministic or random, or to arbitrarily dependent on one another and on the inliers. Alternatively, another popular approach is to assume that $n - \lceil n\epsilon \rceil$ datapoints are drawn i.i.d. with distribution p and that the remaining $\lceil n\epsilon \rceil$ are arbitrarily chosen by the adversary. The adversary is supposed to have full knowledge of p, of the data $X_1, \dots, X_n$ and of the estimator.

We do not provide an overview of the very rich literature on robust estimation, but we refer the reader to [186], [187], [181], [180] where one can find a good introduction to the subject. The field of robust statistics has been pioneered in the sixties by [2], [3], [8]. It mainly encompasses two different approaches, namely robustness to contamination - which we focus on in this thesis - and robustness to heavy tails, which we now only briefly review. Heavy tailed distributions are distributions for which extremely large values can be observed with a substantial probability. One emblematic problem is the mean estimation problem with heavy tails. The breakthrough paper [58] revolutionized our understanding of this problem, by proposing an estimator of the mean with a sub-Gaussian rate, while merely assuming that the distribution has a second order moment. These results were then extended to high dimensions in [114] with an exponential-time procedure, and further improved by [136, 108], which proposed polynomial-time procedures with comparable statistical property. Further references and techniques concerning robustness to heavy tails can be

found in [113].

In this thesis, we place ourselves in the adversarial setting and consider robust estimation of high-dimensional discrete distributions, under the additional learning constraint of local differential privacy (see Section 1.3.1). Without the privacy constraint, several works considered the problem of robustly learning discrete distributions under adversarial contamination, for example [93], [130], [137], [131], [138], [123]. The paper [161] extended the results to densities.

Discrete distributions being instances of sub-Gaussian distributions, it is beneficial to recall the classical estimation rates for learning high-dimensional normal distributions under adversarial contamination. Let $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \Sigma)$ where $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ is a known positive semi-definite and symmetric matrix. Assume that an adversary knowing the data, the underlying distribution as well as the statistician's estimator can replace $\lceil \epsilon n \rceil$ of the data-points with $\lceil \epsilon n \rceil$ outliers for some $\epsilon \in (0, \frac{1}{2})$. Nothing is assumed on the outliers, in particular, they need not follow a probability distribution and can be chosen arbitrarily far away from μ. We consider the problem of estimating μ in the L^2 norm in this setting. From the theory developed in [80, 101, 123] the minimax estimation rate, up to absolute constant factors, is lower bounded by the the risk in the Huber contamination model, scaling as as

$$\sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}} + \epsilon\sqrt{\|\Sigma\|_{op}}, \tag{1.12}$$

where $\|\cdot\|_{op}$ denotes the operator norm. Among all tractable methods, the best rate achieved so far in the adversarial setting, up to absolute constant factors, is as follows (see [171])

$$\sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}} + \epsilon\sqrt{\log\left(\frac{1}{\epsilon}\right)}\sqrt{\|\Sigma\|_{op}}. \tag{1.13}$$

It has been highlighted that polynomial-time estimators can have worse statistical guarantees than computationally intractable ones [84, 60, 61]. In the present setting, it is conjectured that the extra factor $\sqrt{\log\left(\frac{1}{\epsilon}\right)}$ represents a computational gap between polynomial-time methods and computationally intractable ones, see e.g. [89].

1.3.3 Combining robustness with privacy

Although robust learning and local differential privacy are both widely studied fields of research, combining the two settings is just starting to be explored. The links between robustness and *global* differential privacy have been well studied in [120, 115, 112]. However, in the case of *local* differential privacy, only recent works have considered this interaction: [157, 176], where the authors provide upper and lower bounds for estimating discrete distributions under the two constraints, in a different setting from what we consider. The lower bound was later tightened in [149]. The paper [163] also considers the mean estimation problem under local differential privacy and robustness to outliers.

The above papers consider adversarial contamination as well as Huber contamination. It is important to note that when combining robustness with privacy, the contamination can essentially occur at two different steps.

- **Contamination before privacy:** The adversary can play before privacy and replace some of the non-privatized data with outliers. Then, the privacy channel Q is applied on this corrupted dataset.
- **Contamination after privacy:** The adversary can also play at the second step, after privatization, by replacing some of the privatized data-points with outliers.

In this thesis, we consider the second setting. These two configurations may seem very similar and one could expect them to lead to comparable statistical rates. However, it is not the case, and the two settings actually involve quite different phenomena. Denote by $R^*_{privacy}$ the minimax estimation rate under privacy alone and by R^*_{contam} the minimax estimation rate under contamination alone. In each of the above papers [157, 176, 149], whenever contamination occurs *before* privacy, the minimax rate always scales as $R^*_{privacy} + R^*_{contam}$ (in other words, there is no extra statistical cost due to the interaction of the two constraints). When contamination comes *after* privacy, however, the minimax risk scales as $R^*_{privacy} + \frac{\sqrt{d}}{\alpha} R^*_{contam}$, which is always at least as large as the previous rate. This latter rate reveals an interesting interplay between the two constraints. To the best of our knowledge, there is no unifying result generalizing this phenomenon to an arbitrary setting, which can be a very interesting direction for future work.

1.4 Benign overfitting

Benign overfitting is a counter-intuitive phenomenon that was recently discovered in the deep learning community. It has been experimentally observed that deep neural network can achieve very good generalization performance while perfectly fitting noisy training data [167, 105, 154]. This phenomenon seems to go against the classical bias–variance trade-off argument which assumes a necessary balance between overfitting and underfitting. When plotting the test error of a neural net as a function of the number of its parameters, the paper [105] was the first to experimentally exhibit the so-called "double descent risk curve", that reconciles the U-shaped curve predicted by the bias–variance trade-off with the observation that good prediction accuracy is achievable with overfitting. With the aim of understanding this phenomenon, a series of papers studied benign overfitting in the linear regression model, which is perhaps the simplest case where this phenomenon can occur (see [122], [146], [133], [144], [152], [175] and the references therein). We refer the reader to [153] for an excellent review of the recent field. In this thesis, we only study benign overfitting in the context of nonparametric regression. However, we give here a quick overview of benign overfitting in some other related settings.

1.4.1 Ridge (and ridgeless) regression

In the linear regression model, there exists a deep connection between interpolation and ridgeless regression. Consider the linear regression model

$$y = \mathbb{X}\theta^* + \xi. \tag{1.14}$$

Here, $y = (Y_1, \dots, Y_n)^\top \in \mathbb{R}^n$, $\mathbb{X} = (X_1, \dots, X_n)^\top \in \mathbb{R}^{n \times d}$, and $\xi \in \mathbb{R}^n$. For some $\lambda > 0$, the ridge regression estimator is defined as the unique vector $\hat{\theta}^R$ minimizing $\frac{1}{n}\|\mathbb{X}\theta - y\|^2 + \lambda\|\theta\|^2$. This program is equivalent to minimizing $\|\mathbb{X}\theta - y\|^2$ subject to $\|\theta\| \leq b$ or to minimizing $\|\theta\|$ subject to $\frac{1}{n}\|\mathbb{X}\theta - y\|^2 \leq c$, for some constants b, c. We note that, in the last program, taking $\lambda \to 0$ is equivalent to taking $c \to 0$, so that the minimum-norm interpolating estimator $\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^d} \|\theta\|$ subject to $\|\mathbb{X}\theta - y\| = 0$ is the limiting ridge estimator as $\lambda \to 0$, also called the ridgeless estimator (note that this problem has a solution in the overparametrized regime where $d > n$). In most cases, studying benign overfitting in the linear regression setting makes extensive use of this estimator.

Linear case with Gaussian covariates

Assume that $(y, \mathbb{X})$ are jointly Gaussian. We assume that $\mathbb{E}\mathbb{X} = 0$ and that $\frac{1}{n}\mathbb{E}\mathbb{X}\mathbb{X}^\top = \Sigma \in \mathbb{R}^{d \times d}$ where $d \in \mathbb{N}^*$. In the asymptotic case where $\frac{d}{n} \to \gamma > 0$, the (weighted) ridgeless estimator was studied in [172], [165], [147] for a general known covariance matrix Σ.

In the non-asymptotic case, the paper [122] gives necessary and sufficient conditions on the covariance matrix Σ for benign overfitting to occur, that is, for the ridgeless estimator to be near minimax optimal. The authors show that overparametrization is necessary as well as very specific conditions over the decay of the eigenvalues of Σ. In [146], the authors precisely studied the ridge regression estimator in the overparametrized setting, and established non-asymptotic generalization bounds in the case of a general known covariance matrix Σ, and showed that those bounds are tight for a range of regularization parameter values. These results were then refined in [175]. In the linear model, the main conclusion is that the benign overfitting phenomenon requires *overparametrization*, which in a sense approaches the non-parametric setting, as well as an *unbalanced spectrum* of the design matrix with a specific decay of its eigenvalues.

1.4.2 Kernel ridge(less) regression

Extensions to kernel ridgeless regression in RKHS were considered in [142] when the sample size n and the dimension d were assumed to satisfy $n \asymp d$, and in [143] for a more general case $d \asymp n^\alpha$ for $\alpha \in (0, 1)$. These papers give data-dependent upper bounds on the risk that can be small assuming favorable spectral properties of the data and the kernel matrix. On the other hand, if d is constant (independent of n) then the least-norm interpolating estimator with respect to the Laplace kernel is inconsistent [116]. Kernel ridge regression is one of the usual statistical techniques for estimating functions. This setting is therefore closely related to the setting of non-parametric regression, which is addressed in this thesis.

1.4.3 Non-parametric regression

In the setting of non-parametric regression, one is given n i.i.d. observations $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$ for $i = 1, \dots, n$, where $Y_i = f(X_i) + \epsilon_i$. Here, the ϵ_i's are i.i.d. noise random variables that are independent from X_i, and f is an unknown function that we want to estimate. We assume that f belongs to the Hölder class of functions. In the setting of non-parametric regression with square loss and known Hölder smoothness $\beta \leq 2$, it was shown that there exist interpolating estimators attaining minimax optimal rates [106]. Namely, it is proved in [106] that interpolation with minimax

optimal rates on such Hölder classes can be achieved by Nadaraya-Watson estimator with a singular kernel.

1.5 Summary of the contributions

1.5.1 Chapter 2: Local Goodness-of-fit testing in discrete models

This Chapter is based on the paper: "Sharp Local Minimax Rates for Goodness-of-Fit Testing in multivariate Binomial and Poisson families and in multinomials" [132], by Julien Chhor and Alexandra Carpentier (arXiv:2012.13766), to appear in Mathematical Statistics and Learning.

We consider the local goodness-of-fit testing problem for discrete distributions such as multivariate Binomial or Poisson families and multinomial distributions. For fixed null discrete distribution p_0, we derive the nonasymptotic local minimax separation radius $\rho^*(n, p_0, t)$ up to constants depending only on η, for all ℓ_t separation distances with $t \in [1, 2]$. Furthermore, we establish the tight dependency of ρ^* on p_0. We also give the corresponding local minimax tests. The main idea is to introduce a new way of splitting the null distribution p_0 into bulk and tail parts, and to accurately determine the contribution of the tail. Our approach provides understanding of how very small coefficients of p_0 contribute to the minimax separation radius local testing problems for a variety of separation distances.

1.5.2 Chapter 3: Local Goodness-of-fit testing for Hölder continuous densities

This Chapter is based on the paper "Goodness-of-Fit Testing for Hölder-Continuous Densities: Sharp Local Minimax Rates" [158] by Julien Chhor and Alexandra Carpentier (arXiv:2109.04346).

In the continuous setting, we assume that the observations are i.i.d. with the same unknown density p having α-Hölder smoothness over $\mathbb{R}^d$, for $\alpha > 0$. The null density p_0 is assumed to satisfy the same smoothness conditions. We address the local testing problem in all L_t separation distances, $t \in [1, 2]$, and for all smoothness parameter $\alpha > 0$. We establish nonasymptotic upper and lower bounds on the minimax separation radius $\rho^*(n, p_0, t)$, and prove that the bounds are always matching. We also explicitly construct local minimax tests by introducing novel test statistics which we believe could be of independent interest. For $\alpha > 1$, we need an additional technical assumption on the densities which we believe is quite mild. We also introduce a new way of splitting the domain $\mathbb{R}^d$ into bulk and tail parts. Our analysis of the tail part reveals how very small values of the null density contribute to the minimax separation radius, which, to the best of our knowledge, was not understood in the literature.

1.5.3 Chapter 4: Robust estimation of discrete distributions under local differential privacy

This Chapter is based on the paper "Robust Estimation of Discrete Distributions under Local Differential Privacy" [169], by Julien Chhor and Flore Sentenac (arXiv:2202.06825).

We consider the problem of estimating a d dimensional discrete distribution p under the constraints of local differential privacy and robustness to adversarial contamination. We assume that the non corrupted and non-privatized data are i.i.d. with distribution p, and are grouped in n batches of size $k \geq 1$. Moreover, we assume that each one of these nk datapoints is privatized using an α-LDP mechanism for $\alpha \in (0, 1)$. We further assume that an adversary replaces $\lceil \epsilon n \rceil$ of the privatized *batches* with arbitrarily chosen batches. When $k = 1$, this setting encompasses the classical setting where the dataset consists of n non-corrupted datapoints, that an adversary can contaminate by replacing $\lceil \epsilon n \rceil$ of the data points with outliers.

We propose a polynomial-time algorithm and prove that with high probability, it estimates p with the rate $1 \wedge \left\{ \frac{d}{\alpha\sqrt{nk}} + \frac{\epsilon\sqrt{d}}{\alpha\sqrt{k}}\sqrt{\log(1/\epsilon)} \right\}$ in total variation distance. We show that this rate is tight up to the extra $\sqrt{\log(1/\epsilon)}$ factor, i.e. we prove that for any estimator $\hat{p}$, there exists a distribution p such that with constant probability, we have $TV(\hat{p}, p) \geq c \left(1 \wedge \left\{ \frac{d}{\alpha\sqrt{nk}} + \frac{\epsilon\sqrt{d}}{\alpha\sqrt{k}} \right\} \right)$ for some $c > 0$. Up to constants, this rate can be strictly larger than the sum of the estimation rate under privacy only, which is $\frac{d}{\alpha\sqrt{nk}}$, and of the contamination rate only, which is $\sqrt{\frac{d}{nk}} + \frac{\epsilon}{\sqrt{k}}$. More precisely, our results show that the contamination term gets inflated by a factor $\frac{\sqrt{d}}{\alpha}$.

1.5.4 Chapter 5: Benign overfitting in adaptive non-parametric regression

This chapter is based on the paper "Benign Overfitting and Adaptive Nonparametric Regression" [170] with Suzanne Sigalla and Alexandre Tsybakov (arXiv:2206.13347).

We consider the setting of non-parametric regression with squared L^2 loss. We assume that the observations consist of n i.i.d. pairs (X_i, Y_i) such that $Y_i = f(X_i) + \xi_i$ where f belongs to a subset of the Hölder class $\Sigma(\beta, L)$ for unknown $\beta \in (0, \beta_{\max}]$ and ξ_i are i.i.d. σ_ξ-sub-Gaussian noise. We use local polynomial estimators with singular kernels to construct a minimax optimal estimator $\hat{f}$ that is a continuous function, which is adaptive to unknown $\beta \in [0, \beta_{\max}]$ and that interpolates all of the data points with high probability, i.e. such that $\hat{f}(X_i) = Y_i$ for all $i = 1, \dots, n$.

Introduction en Français

Cette thèse explore différents sujets d'inférence statistique parmi lesquels les tests minimax, l'estimation sous contraintes et l'overfitting bénin dans la régression non-paramétrique.

- La première partie est consacrée aux tests minimax. Étant donné n observations i.i.d. de loi p inconnue, le problème de test d'adéquation vise à tester l'égalité de p à une distribution de probabilité donnée p_0 contre une alternative composée de distributions séparées de p_0 au sens d'une certaine distance sur la classe de distributions considérée. Nous étudions deux problèmes différents. Le premier concerne le cas discret, où les observations peuvent être des familles multivariées binomiales ou de Poisson ou des distributions multinomiales. La seconde est une extension de la première au cas continu, où les observations sont i.i.d. avec une certaine densité de probabilité sur $\mathbb{R}^d$ qui appartient à la classe de Hölder de fonction avec des paramètres connus $\alpha, L > 0$. Dans les deux cas, nous nous intéressons spécifiquement à la version *locale* du problème (voir section 1.6.3)

- La deuxième partie propose un problème d'estimation sous contrainte. Nous étudions les interactions entre la robustesse à la contamination adversariale et la confidentialité différentielle locale pour l'estimation de distributions discrètes.

- La troisième partie porte sur la régression non-paramétrique. Nous construisons un estimateur par polynômes locaux, à la fois optimal au sens minimax, adaptatif à la régularité inconnue de la fonction à estimer, et présentant la propriété d'interpoler continûment tous les points de données avec grande probabilité - un phénomène appelé "overfitting bénin".

Les notations peuvent changer d'un chapitre à l'autre.

1.6 Tests Minimax

Soit $\mathcal{X}$ un ensemble et $\mathcal{P} = \{\mathbb{P}_\theta \mid \theta \in \Theta\}$ une famille de lois de probabilité sur un espace measurable $(\mathcal{X},\mathcal{U})$, où Θ est un ensemble de paramètres, de dimension éventuellement infinie. Soient Θ_0 et $\Theta_1 \subset \Theta$ deux sous ensembles disjoints de Θ : $\Theta_0 \cap \Theta_1 = \emptyset$. On suppose que l'on dispose de $n \in \mathbb{N}^*$ observations $X_1, \cdots, X_n \overset{iid}{\sim} \mathbb{P}_\theta$ pour un certain $\theta \in \Theta$ inconnu. Au vu des observations $X_1, \cdots, X_n$, on souhaite effectuer le problème de test suivant :

$$H_0 : \theta \in \Theta_0 \qquad \text{contre} \qquad H_1 : \theta \in \Theta_1. \tag{1.15}$$

Définition: Un ***test*** ψ est une fonction mesurable des données, prenant seulement les valeurs 0 et 1.

$$\psi : \mathcal{X}^n \longrightarrow \{0,1\}.$$

Dans cette thèse, nous nous intéressons à la notion d'optimalité minimax. L'erreur de type I d'un test ψ est définie comme $\max_{\theta\in\Theta_0} \mathbb{P}_\theta(\psi = 1)$ et l'erreur de type II comme $\max_{\theta\in\Theta_1} \mathbb{P}_\theta(\psi = 0)$. On mesure la qualité d'un test par son *risque*, défini comme la somme de ses erreurs de type I et de type II : $R_{\Theta_0,\Theta_1}(\psi) = \sup_{\theta\in\Theta_0} \mathbb{P}_\theta(\psi = 1) + \sup_{\theta\in\Theta_1} \mathbb{P}_\theta(\psi = 0)$. Dans ce paradigme, H_0 et H_1 sont donc symétriques et le problème consiste à construire des tests ayant simultanément de faibles erreurs de type I et II *dans le pire des cas*. Comme nous le verrons plus tard, ce paradigme garantit toujours l'existence de tests optimaux dans un sens défini ci-dessous.

Le **risque minimax** est défini comme le risque du meilleur test, s'il existe.

Definition 1.4. *(Risque minimax): Le risque minimax associé au problème* (1.15) *est défini comme*

$$\begin{aligned} R^*_{\Theta_0,\Theta_1} &= \inf_\psi R_{\Theta_0,\Theta_1}(\psi) \\ &= \inf_\psi \left\{ \sup_{\theta\in\Theta_0} \mathbb{P}_\theta(\psi = 1) + \sup_{\theta\in\Theta_1} \mathbb{P}_\theta(\psi = 0) \right\}, \end{aligned}$$

où l'infimum est pris sur tous les tests $\psi : \mathcal{X}^n \longrightarrow \{0,1\}$.

Une première remarque est que si $\Theta_0, \Theta_1 \subseteq \Theta$ sont tels que $R^*_{\Theta_0,\Theta_1} = 1$, alors la stratégie de décision optimale est de répondre au hasard. En effet, appelons $\widetilde{\Delta}$ le test randomisé, qui prend les valeurs 0 et 1 avec équiprobabilité de manière indépendante des données. Son risque est égal à 1 : en effet, $\forall\theta \in \Theta_0, \mathbb{P}_\theta(\widetilde{\Delta} = 1) = \frac{1}{2}$ et $\forall\theta \in \Theta_1, \mathbb{P}_\theta(\widetilde{\Delta} = 0) = \frac{1}{2}$, d'où

$$R_{\Theta_0,\Theta_1}(\widetilde{\Delta}) = \sup_{\theta\in\Theta_0} \mathbb{P}_\theta\left(\widetilde{\Delta} = 1\right) + \sup_{\theta\in\Theta_1} \mathbb{P}_\theta\left(\widetilde{\Delta} = 0\right) = \frac{1}{2} + \frac{1}{2} = 1.$$

Si $R^*_{\Theta_0,\Theta_1} = 1$, le test $\widetilde{\Delta}$ est optimal et le problème est trivial. Il est donc naturel de ne considérer que des problèmes de test avec $R^*_{\Theta_0,\Theta_1} < 1$. On supposera donc désormais que $R^*_{\Theta_0,\Theta_1} \leq \eta$ où $\eta \in (0,1)$ est un niveau de risque choisi à l'avance.

1.6.1 Problème du Goodness-of-Fit

Soit **dist** une distance sur Θ et fixons $\eta \in (0,1)$ ainsi que $\theta_0 \in \Theta$. Supposons que la distance de variation totale soit continue par rapport à **dist**. Pour $\rho > 0$, le problème de test de **Goodness-of-fit**, également appelé **problème de test d'identité**, s'écrit comme suit :

$$H_0 : \theta = \theta_0 \qquad \text{contre} \qquad H_1^{(\rho)} : \begin{cases} \theta \in \Theta, \\ \mathbf{dist}(\theta, \theta_0) \geq \rho. \end{cases} \tag{$\star$}$$

On pourrait se demander pourquoi on ne définit pas le problème de la manière suivante

$$H_0' : \ \theta = \theta_0 \qquad \text{contre} \qquad H_1' : \ \begin{cases} \theta \in \Theta, \\ \theta \neq \theta_0. \end{cases} \tag{1.16}$$

La raison est la suivante: désignons par R^* le risque minimax du problème (1.16). Dans ce cas, pour tout $\theta_1 \neq \theta_0$, nous aurions

$$\begin{aligned} 1 &\geq \inf_{\psi} \left\{ \mathbb{P}_{\theta_0}(\psi = 1) + \mathbb{P}_{\theta_1}(\psi = 0) \right\} \\ &= 1 + \inf_{\psi} \left\{ \mathbb{P}_{\theta_1}(\psi = 0) - \mathbb{P}_{\theta_0}(\psi = 0) \right\} \\ &= 1 - \sup_{U \in \mathcal{U}} \mathbb{P}_{\theta_1}(U) - \mathbb{P}_{\theta_0}(U) \\ &= 1 - \sup_{U \in \mathcal{U}} \left| \mathbb{P}_{\theta_1}(U) - \mathbb{P}_{\theta_0}(U) \right| \\ &= 1 - TV(\mathbb{P}_{\theta_0}, \mathbb{P}_{\theta_1}) \longrightarrow 1 \quad \text{lorsque } \mathbf{dist}(\theta_1, \theta_0) \to 0, \end{aligned}$$

puisque la distance de variation totale est continue par rapport à **dist**. Le problème de test (1.16) est donc associé à un risque minimax $R^* = 1$, autrement dit, il serait optimal de répondre au hasard. Pour s'assurer que le risque minimax est inférieur à $\eta \in (0, 1)$, il est donc nécessaire de définir le problème de Goodness-of-Fit comme dans l'équation ($\star$) dans le cadre minimax. Plus précisément, introduisons la notation

$$R^*(\rho) = \inf_{\psi} \mathbb{P}_{\theta_0}(\psi = 1) + \sup \left\{ \mathbb{P}_{\theta}(\psi = 0) \ \middle| \ \theta : \mathbf{dist}(\theta, \theta_0) \geq \rho \right\}, \tag{1.17}$$

pour désigner le risque minimal associé au problème ($\star$). En notant que $\rho \mapsto R^*(\rho)$ est une fonction décroissante, on cherche à trouver la plus petite distance de séparation $\rho > 0$ assurant $R^*(\rho) \leq \eta$.

Definition 1.5 (Minimax separation radius)**.** *Le rayon de séparation minimax du problème* ($\star$) *est défini comme*

$$\rho^*(n, \theta_0, \Theta, \boldsymbol{dist}, \eta) = \inf \left\{ \rho > 0 \ \middle| \ R^*(\rho) \leq \eta \right\}.$$

De plus, nous désignerons le risque de tout test ψ par la quantité

$$R(\rho, \psi) := \mathbb{P}_{\theta_0}(\psi = 1) + \sup \left\{ \mathbb{P}_{\theta}(\psi = 0) \ \middle| \ \theta : \mathbf{dist}(\theta, \theta_0) \geq \rho \right\}.$$

L'objectif du problème de test de Goodness-of-Fit est double :

1. Identifier ρ^* à constantes multiplicatives près.
2. Trouver un test ψ^* et une constante $C > 0$ tels que $R(C\rho^*, \psi) \leq \eta$. Un tel test est appelé un test *minimax-optimal*.

Remarque: Dans la littérature, le problème de Goodness-of-Fit est parfois formulé d'une manière différente. Dans la définition ci-dessus, nous nous sommes intéressés au rayon de séparation minimax

pour un nombre fixé de données n. Cependant, le problème du test d'identité peut également être formulé en termes de *sample complexity.* En d'autres termes, pour une précision fixée $\epsilon > 0$, on s'itéresse dans ce cas à déterminer le nombre d'observations n^* nécessaire et suffisant, à une constante multiplicative près dépendant uniquement de $\eta, \mathbf{dist}, \Theta$, pour que le problème de test

$$H_0 : \theta = \theta_0 \quad \text{contre} \quad H_1(\epsilon) : \begin{cases} \theta \in \Theta, \\ \mathbf{dist}(\theta, \theta_0) \geq \epsilon, \end{cases}$$

ait un risque minimax au plus égal à η. Cette formulation est une version duale du problème de Goodness-of-Fit ($\star$) et conduit souvent à des résultats très similaires. Cette convention est la plus souvent utilisée dans la communauté computer science (voir par exemple [71, 95, 81, 100, 67]), alors que la forumlation ($\star$) est standard dans la communauté statistique.

1.6.2 Vitesses classiques pour la détection de signal

Dans le cadre minimax, le problème de détection de signal correspond au cas particulier où $\Theta_0 = \{0\}$. L'un des premiers articles traitant du problème de détection du signal est [13] dans le modèle de bruit blanc gaussien. La série d'articles [22] est considérée comme une référence dans le domaine de la détection de signal non paramétrique. Le cas où Θ est un ellipsoïde est considéré dans [20], ou par des boules de Sobolev ou de Besov dans [26], [182], [31] pour ne citer que quelques exemples. Notons que les références ci-dessus traitent du régime asymptotique. Dans cette thèse, cependant, nous nous concentrerons sur les vitesses non asymptotiques, qui est un cadre moins standard dans la littérature (voir par exemple [36], [59], [87], [162], [44], [174], [51]). Nous ne donnons pas ici un aperçu exhaustif de la littérature et référons le lecteur à [183] pour un excellent aperçu. Nous nous limiterons à passer en revue les résultats classiques qu'il est utile de comparer avec les résultats développés dans la thèse.

Cas gaussien

Le premier cadre classique de détection de signal est le problème de test de Goodness-of-Fit dans le cadre gaussien. Pour $d \geq 1$, supposons que l'on observe $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(\theta, I_d)$ où I_d désigne la matrice identité de taille d. On considère le problème de test

$$H_0 : \theta = 0 \quad \text{contre} \quad H_1(\rho) : \begin{cases} \theta \in \Theta, \\ \|\theta\|_2 \geq \rho, \end{cases} \tag{1.18}$$

où $\rho > 0$ et $\|\cdot\|_2$ désigne la norme euclidienne. Le rayon minimax de séparation pour le problème (1.18) est $\rho^* \asymp d^{\frac{1}{4}}/\sqrt{n}$ (c.f. [162], [87]) et le test minimax optimal est $\mathbb{1}\left\{\|\bar{X}_n\|_2^2 \geq d + c\right\}$ où $c > 0$ est une constante et $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Il s'ensuit que, pour toute précision $\epsilon > 0$, tester l'égalité à 0 par rapport à l'alternative $\|\theta\|_2^2 \geq \epsilon^2$ n'est possible que si $n \gtrsim \frac{\sqrt{d}}{\epsilon^2}$.

Nous pouvons comparer ceci avec la vitesse classique d'estimation. Si, au vu de l'observation $X \sim \mathcal{N}(\theta, I_d)$, on cherche à estimer θ en perte L^2, l'estimateur optimal minimax de θ sur $\mathbb{R}^d$ est $\hat{\theta} = X$,

et il atteint la vitesse $\sup_{\theta\in\mathbb{R}^d}\mathbb{E}\|\hat{\theta}-\theta\|^2 = d$. En d'autres termes, au vu de $X_1,\dots,X_n \overset{iid}{\sim} \mathcal{N}(\theta, I_d)$, construire un estimateur $\hat{\theta}$ de θ avec une précision d'estimation de $\sup_{\theta\in\mathbb{R}^d}\mathbb{E}\|\hat{\theta}-\theta\|^2 = \epsilon^2$ n'est possible que si $n \geq \frac{d}{\epsilon^2}$.

Un phénomène intéressant se produit donc lorsque $\frac{\sqrt{d}}{n} \lesssim \|\theta\|_2^2 \ll \frac{d}{n}$. En effet, il est d'une part possible de détecter que $\theta \neq 0$, tandis qu'il est d'autre part impossible d'améliorer l'estimateur trivial égal à 0. Il s'agit de l'un des avantages des tests par rapport à l'estimation : bien que les tests ne fournissent qu'une information binaire (donc plus limitée que ce que l'estimation permet d'obtenir), ils permettent d'atteindre des vitesses plus rapides. Cet avantage devient d'autant plus important en grande dimension qu'il permet de restreindre la quantité de données nécessaires pour faire de l'inférence.

Cadre non-paramétrique

Le problème de test non paramétrique le plus proche de nos résultats consiste à tester l'égalité d'une distribution inconnue à la distribution uniforme sur $[0,1]^d$ contre une alternative composée de densités hölderiennes sur $[0,1]^d$ et séparées en distance L^2 [179]. Plus précisément, en définissant $p_0 = 1$ comme la densité uniforme sur $[0,1]^d$ et étant donné n observations i.i.d. $X_1,\dots,X_n$ avec une densité hölderienne p sur $[0,1]^d$ et un paramètre de régularité connu $\alpha > 0$, le problème considéré s'écrit

$$H_0 : p = p_0 \quad \text{contre} \quad H_1(\rho) : \begin{cases} \|p - p_0\|_2 \geq \rho \\ p \in H(\alpha), \end{cases} \tag{1.19}$$

où $H(\alpha)$ est la classe des fonctions hölderiennes sur $[0,1]^d$ de paramètre de régularité α et constante de Lipschitz normalisée à 1. On montre que la vitesse asymptotique minimax lorsque $n \to \infty$ est $\rho^* \asymp n^{-\frac{2\alpha}{4\alpha+d}}$ (voir par exemple [22]). Encore une fois, on peut comparer cette vitesse avec la vitesse minimax d'estimation d'une densité hölderienne dans $H(\alpha)$ qui est $n^{-\frac{\alpha}{2\alpha+d}}$ (voir [188]), toujours plus lente que la vitesse non-paramétrique de test.

1.6.3 Tests locaux

Dans cette sous-section, nous introduisons une distinction supplémentaire en tests *locaux* et *globaux*. Supposons que Θ, **dist** et η sont fixés. Le problème de test local s'intéresse à la dépendance précise de ρ^* par rapport à θ_0 et n. Au contraire, le problème de test *global* vise à étudier le rayon de séparation $\rho^*(n,\theta_0)$ dans le pire cas lorsque θ_0 varie dans la classe Θ : $\rho^*_{\text{global}}(n) := \sup_{\theta_0\in\Theta} \rho^*_{\text{local}}(n,\theta_0)$.

Les vitesses locales apparaissent naturellement lorsque la variance des observations dépend du paramètre θ que l'on veut tester. Par exemple, supposons que l'on observe $X \sim \mathcal{N}(\theta, I_d)$ et que l'on fixe $\theta_0 \in \mathbb{R}^d$. Considérons le problème de test gaussien suivant

$$H_0 : \theta = \theta_0 \quad \text{contre} \quad H_1 : \begin{cases} \theta \in \Theta, \\ \|\theta - \theta_0\|_2 \geq \rho. \end{cases}$$

Dans ce problème de test, les rayons minimax de séparation locaux ne diffèrent pas des rayons globaux, car la matrice de covariance des données reste égale à I_d indépendamment de θ_0. Supposons maintenant que l'on dispose de $X_1, \dots, X_n \overset{iid}{\sim} Ber(p)$ où $p \in [0,1]$ et fixons $p_0 \in [0,1]$. On s'intéresse au problème suivant

$$H_0 : p = p_0 \quad \text{contre} \quad H_1(\rho) : \begin{cases} p \in [0,1], \\ |p - p_0| \geq \rho. \end{cases}$$

Une statistique exhaustive est la somme $\sum_{i=1}^n X_i$, dont la variance $\sqrt{np(1-p)}$ dépend du paramètre p à tester. Dans ce cadre, $\rho^*(n, p_0)$ présente des comportements très différents du rayon de séparation global lorsque p_0 varie dans $[0,1]$. En effet, pour $p_0 = 0$, on a $\rho^*(n,0) \asymp 1/n$, tandis que pour $p_0 = \frac{1}{2}$, on a $\rho^*(n, \frac{1}{2}) \asymp 1/\sqrt{n}$ (ces affirmations découlent du Théorème 2.1 prouvé dans cette thèse). Dans ce cadre très simple, les vitesses locales peuvent considérablement raffiner les vitesses globales, ce qui, en grande dimension, peut s'avérer d'un grand intérêt. Dans cette thèse, une attention particulière se portera sur la construction tests *locaux* et non globaux, et nous nous attacherons à prouver leur optimalité.

1.7 Inférence sous contrainte

Dans la deuxième partie de la thèse (chapitre 4), nous nous intéressons à l'estimation de distributions discrètes sous contraintes de learning. Le but de ce chapitre est d'étudier les intéractions entre la contrainte de *confidentialité locale différentielle* et la contrainte de *robustesse aux outliers*.

1.7.1 Confidentialité locale différentielle

De nouveaux défis ont récemment vu le jour concernant le traitement des données. L'un d'entre eux provient du caractère sensible des données collectées à grande échelle sur internet ou sur nos comportements d'achat. De manière générale, bien que le traitement de ces données puisse être un enjeu important, il devient primordial d'assurer la confidentialité des informations traitées par le statisticien. Pour répondre à ce problème, plusieurs solutions ont été envisagées.

- *Anonymiser les données*: Bien que naturelle, cette méthode peut s'avérer très vulnérable. Il a été montré expérimentalement dans [34] que 87% de la population des Etats-Unis pouvait être indirectement identifiée à partir des informations suivantes: {date de naissance, genre, code postal}.
- *Cryptographie*: Une autre solution naturelle est de crypter les données pour les rendre indéchiffrables, par exemple à l'aide de méthodes comme le chiffrement homomorphe ou le *Secure multiparty computation*. Cependant, cette méthode présente également des défauts : le coût en temps de calcul est élevé, ce type de procédure peut ne pas être robuste au sens statistique du terme, et les données peuvent être reconstituées si un attaquant parvient à accéder à la clé de chiffrement.
- *Confidentialité différentielle*: Cette technique consiste à modifier les données, souvent en leur ajoutant un bruit. Le bruit choisi doit être suffisamment fort pour qu'il soit impossible de

reconstituer les données initiales à partir de leur version bruitée. Il doit également être suffisamment faible pour que, si suffisamment d'individus sont observés, il soit possible d'inférer des caractéristiques globales de la population. Le statisticien n'a jamais accès aux données non-buitées ; en revanche, il est libre de choisir le mécanisme de privatisation des données qui convient le mieux au problème statistique considéré. Contrairement aux méthodes cryptographiques, cette technique garantit donc l'impossibilité de reconstituer des informations sensibles, même en cas d'attaque. Bien que l'utilisation de cette technique conduise souvent à des méthodes calculables en temps raisonnable, leur coût statistique peut être prohibitif, comme nous le verrons par la suite. Il est donc de primordial de comprendre dans quelles conditions la confidentialité différentielle peut être utilisée efficacement, ce qui constitue l'un des thèmes abordés dans cette thèse.

Dans cette thèse, nous nous intéressons à l'estimation de distributions discrètes sous contrainte de confidentialité locale différentielle. Soit $X = (X_1, \dots, X_n)$ un vecteur aléatoire dans l'espace mesurable $(\mathcal{X}^n, \mathcal{A}^n, \mathbb{P}^n)$. On souhaite produire une nouvelle variable aléatoire Z, appelée version *privatisée* de X. Formellement, Z prend ses valeurs sur un second espace mesurable $(\mathcal{Z}, \mathcal{B})$ et est générée selon le mécanisme $Z|X = x \sim Q(\cdot|x)$ où $Q(\cdot\,|\cdot)$ est un noyau de Markov également appelé *Mécanisme de confidentialité*. En d'autres termes, pour tout $x \in \mathcal{X}^n$, nous supposons que $Q(\cdot|x)$ est une distribution de probabilité, et pour tout $A \in \mathcal{A}^n$, $Q(A|\cdot)$ est une fonction mesurable. Soit $\alpha \in (0, 1)$. Il existe deux approches principales pour définir la contrainte de confidentialité locale différentielle.

1. **Confidentialité différentielle centrale (ou globale):** On dit qu'un mécanisme de confidentialité Q vérifie la condition globale de confidentialité différentielle [46, 43] si pour tout $A \in \mathcal{B}$ et pour tout $x, x' \in \mathcal{X}^n$ tel que $\sum_{i=1}^n \mathbb{1}_{\{x_i \neq x'_i\}} = 1$, on a

$$\frac{Q(A|x)}{Q(A|x')} \leq e^{\alpha}.$$

   Le statisticien n'accède jamais à $(X_1, \dots, X_n)$, mais seulement à Z. Cependant, cette approche nécessite malheureusement qu'une unité centrale ait accès à l'ensemble des données $(X_1, \dots X_n)$ afin de produire Z. Si l'unité centrale est piratée, l'ensemble des données sensibles $(X_1, \dots, X_n)$ peut être révélé, ce qui rend cette méthode relativement vulnérable.

2. **Local differential privacy:** Une solution à ce problème est d'introduire un deuxième formalisme comme proposé dans [54]. L'idée est de produire les données privatisées au moment-même où elles sont collectées: chaque utilisateur envoie au statisticien la valeur privatisée $Z_i \sim Q(\cdot|X_i)$ sans jamais révéler la véritable valeur X_i. Formellement, un mécanisme de confidentialité Q satisfaisant $Q(dz|x) = Q(dz_1|x_1)Q(dz_2|x_2, z_1) \dots Q(dz_n|x_n, z_1, \dots, z_{n-1})$ est dit être *localement différentiellement privé* si pour tout $A \in \mathcal{B}$, pour tout $i \in [n]$, pour tout $z_1, \dots z_{i-1} \in \mathcal{Z}^{i-1}$, pour tout $x, x' \in \mathcal{X}$:

$$\frac{Q(A|x, z_1, \dots, z_{i-1})}{Q(A|x', z_1, \dots, z_{i-1})} \leq e^{\alpha}. \tag{1.20}$$

Cette méthode ne nécessite pas l'existence d'une unité centrale car chaque donnée privatisée Z_i est produite uniquement à l'aide de X_i et de toutes les données déjà accessibles publiquement $Z_1, \ldots, Z_{i-1}$. Un mécanisme de confidentialité satisfaisant (1.20) est appelé *mécanisme interactif*. Une classe plus restreinte de mécanismes de confidentialité peut être définie comme suit. Un mécanisme de confidentialité Q est dit *non-interactif* si $Q(dz|x) = \prod_{i=1}^{n} Q_i(dz_i|x_i)$ où chaque noyau de Markov Q_i vérifie

$$\frac{Q_i(A|x)}{Q_i(A|x')} \leq e^{\alpha}, \ \forall x, x' \in \mathcal{X}. \tag{1.21}$$

Ce type de mécanisme permet de collecter les données indépendamment les unes des autres. A l'inverse, les mécanismes interactifs généraux définis comme dans (1.20) nécessitent que les données soient collectées de manière séquentielle. Il s'agit d'un inconvénient majeur des mécanismes interactifs. En contrepartie, ces derniers peuvent dans certains cas atteindre de meilleures vitesses statistiques que les mécanismes non-interactifs [126, 124] et doivent par conséquent être privilégiés dès que possible.

En effet, la confidentialité différentielle locale s'accompagne d'un coût statistique élevé. Elle nécessite que la quasi-totalité de l'information contenue dans $X_1, \ldots, X_n$ soit perdue pour produire $Z_1, \ldots, Z_n$. Il a été observé dans [62] que pour les lois discrètes, la précision optimale d'estimation atteignable avec n observations sous contrainte de α-*local differential privacy* est la même que celle atteignable avec $n\alpha^2/d$ observations sans *local differential privacy*. L'effet de la confidentialité revient donc à réduire le nombre d'observations d'un facteur α^2/d, ce qui, en dimension élevée, peut s'avérer prohibitif.

1.7.2 Robustesse

L'une des approches les plus classiques en statistiques est de faire l'hypothèse que les données sont indépendantes et identiquement distribuées. Cependant, cette hypothèse n'a pas nécessairement de raison d'être vérifiée en pratique. Par exemple, de très grands ensembles de données sont susceptibles de contenir de observations provenant de distributions différentes. Pour assouplir cette condition, un point de vue classique est de supposer que seulement une partie des données, appelées *inliers*, provient d'une distribution d'intérêt. Le reste des données, désignées par le terme *ouliers*, est supposé ne pas provenir de cette distribution cible. Dans ce cas, on dit que le jeu de données est *contaminé*. De manière informelle, le but de l'apprentissage robuste est de construire des estimateurs dont les performances soient affectées par la contamination aussi peu que possible. Naturellement, les outliers dégradent souvent les performances statistiques. Le papier [123] répertorie différents modèles de contamination couramment étudiés.

1. **Contamination de Huber [21], [5]:** Il s'agit de l'un des modèles les plus étudiés. Il suppose l'existence de deux distributions p et q ainsi qu'un taux de contamination $\epsilon \in (0, \frac{1}{2})$ tels que $X_1 \ldots, X_n \overset{iid}{\sim} (1-\epsilon)p + \epsilon q$. Dans ce modèle, p est la distribution cible inconnue, et q représente la contamination. Le nombre d'*outliers* est aléatoire et suit la loi $\mathrm{Bin}(n, \epsilon)$.

2. **Contamination déterministe de Huber:** On dit qu'une distribution suit le modèle de contamination déterministe de Huber s'il existe un ensemble $\mathcal{O} \subset [n]$ de cardinal au plus $\lceil n\epsilon \rceil$

et deux distributions p, q telles que pour tout $i \notin \mathcal{O}$, $X_i \sim p$ et pour tout $i \in \mathcal{O}$, $X_i \sim q$ et toutes les observations $X_1, \ldots, X_n$ sont mutuellement indépendantes.

3. **Olivious Contamination:** Ce modèle est similaire à la contamination déterministe de Huber, à ceci près que la famille d'*outliers* suit une certaine distribution conjointe $Q_{\mathcal{O}}$. Par conséquent, les *outliers* ne sont pas supposés être i.i.d.

4. **Contamination des paramètres:** Pour un ensemble d'*outliers* $\mathcal{O}$ choisi à l'avance, les *outliers* $X_i, i \in \mathcal{O}$ sont indépendants des *inliers* $X_i, i \notin \mathcal{O}$, et chaque *outlier* $X_i, i \in \mathcal{O}$ est distribué selon q_i, appartenant à la même classe que p.

5. **Contamination adversariale:** Avec le modèle de contamination de Huber, il s'agit de l'un des modèles de contamination les plus populaires. Il s'agit également du modèle le plus général. Dans ce modèle, les données i.i.d. propres $X_1, \ldots X_n$ sont générées à partir d'une distribution p. Pour un certain $\epsilon \in (0, 1)$, un adversaire remplace $\lceil n\epsilon \rceil$ des données par de nouvelles données qui peuvent être déterministes ou aléatoires, et pouvant dépendre arbitrairement des inliers. Une autre approche populaire consiste à supposer que $n - \lceil n\epsilon \rceil$ points de données sont tirés de manière i.i.d. avec distribution p et que les $\lceil n\epsilon \rceil$ restants sont choisis arbitrairement par l'adversaire. L'adversaire est supposé avoir une connaissance parfaite de p, des données $X_1, \ldots, X_n$ et de l'estimateur.

Nous ne donnons ici qu'un aperçu succinct de la litérature sur l'estimation robuste, et nous référons le lecteur à [186], [187], [181], [180] qui contiennent d'excellentes introductions au domaine. Le domaine des statistiques robustes a commencé à être étudié dans les années 1960 [2], [3], [8]. Il englobe principalement deux thématiques, à savoir la robustesse à la contamination - qui est étudiée dans cette thèse - et la robustesse aux queues lourdes. L'article [58] prouve qu'il est possible d'estimer la moyenne d'une distribution à queue lourde avec une vitesse sous-gaussienne en supposant simplement que la distribution a un moment d'ordre 2. Ces résultats ont été étendus en dimension quelconque dans [114] avec une procédure en temps exponentiel, et encore améliorés par [136, 108], qui proposent des procédures en temps polynomial avec des propriétés statistiques comparables. D'autres références et techniques concernant la robustesse aux queues lourdes peuvent être trouvées dans [113].

Dans cette thèse, nous nous plaçons dans le cadre adversarial et nous nous intéressons à l'estimation robuste de distributions discrètes, sous la contrainte de confidentialité différentielle locale. Sans la contrainte de confidentialité, plusieurs travaux ont résolu le problème de l'apprentissage robuste de distributions discrètes dans un cadre adversarial, par exemple [93], [130], [137], [131], [138], [123]. L'article [161] étend ces résultats au cas des densités.

1.7.3 Combinaison des deux contraintes

Bien que la robustesse et la confidentialité différentielle soient deux domaines de recherche active, la combinaison de ces deux contraintes est un sujet de recherche extrêmement récent. Les liens entre la robustesse et la confidentialité différentielle globale ont été étudiés dans [120, 115, 112]. Dans le cadre de la confidentialité différentielle locale, seuls des travaux récents ont envisagé cette interaction [157, 176], où les auteurs fournissent des bornes supérieures et inférieures pour l'estimation de

distributions discrètes sous les deux contraintes, dans un cadre différent de celui que nous considérons. L'article [163] étudie également le problème de l'estimation de la moyenne avec contraintes de confidentialité différentielle locale et de robustesse aux outliers. Les articles ci-dessus se placent dans le cadre adversarial et dans le cadre de la contamination de Huber. Il est important de noter que combiner robustesse et confidentialité peut s'effectuer selon deux procédures différentes.

- **Contamination avant la privatisation:** L'adversaire peut jouer avant la privacy et remplacer certaines données non privatisées par des outliers. Le mécanisme de confidentialité Q est seulement appliqué dans un deuxième temps sur cet ensemble de données corrompues.
- **Contamination après la privatisation:** L'adversaire peut également jouer à la deuxième étape, après la privatisation, en remplaçant certains des points de données privatisés par des outliers.

Dans cette thèse, nous nous plaçons dans le second cadre. Ces deux configurations peuvent sembler très similaires et l'on pourrait s'attendre à ce qu'elles conduisent à des vitesses statistiques comparables. Pourtant, il n'en est rien, et les phénomènes en jeu dans les deux cas ci-dessus se révèlent très différents. Ainsi, dans chacun des articles ci-dessus [157, 176, 149], la vitesse minimax est toujours plus rapide en cas de contamination survenant *avant* qu'*après* la privacy. Plus précisément, en désignant par $R^*_{privacy}$ la vitesse d'estimation minimax sous la seule contrainte de confidentialité et par R^*_{contam} la vitesse d'estimation minimax sous la seule contrainte de contamination, la vitesse minimax globale pour les deux contraintes combinées est toujours égale à $R^*_{privacy} + R^*_{contam}$ avec contamination avant la privacy et $R^*_{privacy} + \frac{\sqrt{d}}{\alpha} R^*_{contam}$ avec contamination après la privacy. A notre connaissance, il n'existe aucun résultat étendant ce phénomène à un cadre général, ce qui peut être une future direction de travail très intéressante.

1.8 Overfitting bénin

L'overfitting bénin est un phénomène contre-intuitif récemment découvert dans la communauté du deep learning. Il a été observé expérimentalement que les réseaux de neurones profonds peuvent atteindre de très bonnes performances de généralisation tout en s'adaptant parfaitement aux données d'apprentissage bruitées [167, 105, 154]. Ce phénomène semble aller à l'encontre du compromis classique entre biais et variance, qui suppose un nécessaire équilibre entre overfitting et underfitting. En traçant l'erreur de test d'un réseau de neurones en fonction du nombre de ses paramètres, l'article [105] a été le premier à exhiber expérimentalement la "courbe de risque à double descente". Celle-ci réconcilie la courbe en forme de U prédite par le compromis biais-variance, et l'observation que l'overfitting est compatible avec de bonnes performances de prédiction. Pour comprendre ce phénomène, une série d'articles se sont penchés l'overfitting bénin dans le cadre de la régression linéaire, qui est peut-être le cadre le plus simple où ce phénomène peut se produire (see [122], [146], [133], [144], [152], [175]). Nous renvoyons le lecteur à [153] pour un aperçu plus complet. Dans cette thèse, nous étudions uniquement l'overfitting bénin dans le cadre de la régression non-paramétrique. Cependant, nous donnons ici un rapide aperçu de l'overfitting bénin dans d'autres contextes.

1.8.1 Régression Ridge (et Ridge-less)

Dans le cadre de la régression linéaire, il existe un lien profond entre l'interpolation et la régression ridge. Supposons que l'on dispose de n observations i.i.d. $(X_i, Y_i), i = 1, \dots, n$ où

$$y = \mathbb{X}\theta^* + \xi. \tag{1.22}$$

Ici, $y = (Y_1, \dots, Y_n)^\top$, $\mathbb{X} = (X_1, \dots, X_n)^\top$. Pour un certain $\lambda > 0$, l'estimateur de la régression ridge est défini comme le vecteur unique $\hat{\theta}^R$ minimisant $\frac{1}{n}\|\mathbb{X}\theta - y\|^2 + \lambda\|\theta\|^2$. Ce programme d'optimisation est en fait équivalent à minimser $\|\mathbb{X}\theta - y\|^2$ sous contrainte $\|\theta\| \leq b$ ou encore à minimiser $\|\theta\|$ sous contrainte $\frac{1}{n}\|\mathbb{X}\theta - y\|^2 \leq c$, pour certaines constantes b, c. Notons que, dans le dernier programme, prendre $\lambda \to 0$ est équivalent à prendre $c \to 0$, de sorte que l'estimateur interpolant de norme minimale $\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^d} \|\theta\|$ t.q. $\|\mathbb{X}\theta - y\| = 0$ est la limite de l'estimateur ridge lorsque $\lambda \to 0$, également appelé estimateur ridgeless (notons que ce problème a une solution dans le régime surparamétré où $d > n$). Dans la plupart des cas, l'étude de l'overfitting bénin dans le cadre de la régression linéaire fait couramment usage de cet estimateur.

Cas linéaire avec covariables gaussiennes

Supposons que les (X_i, Y_i) sont conjointement gaussiens. Nous supposons que $\mathbb{E}\mathbb{X} = 0$ et que $\frac{1}{n}\mathbb{E}\mathbb{X}\mathbb{X}^\top = \Sigma \in \mathbb{R}^{d \times d}$ où $d \in \mathbb{N}^*$. Dans le cas asymptotique où $\frac{d}{n} \to \gamma > 0$, l'estimateur ridgeless (pondéré) a été étudié dans [172], [165], [147] pour une matrice de covariance générale connue Σ, et [140] montre qu'avec un design aléatoire et dans le régime surparamétré où $d > n$, il peut être optimal d'avoir un paramètre de régularisation $\lambda \leq 0$.

Dans le cas non-asymptotique, l'article [122] donne des conditions nécessaires et suffisantes sur la matrice de covariance Σ pour avoir le phénomène d'overfitting bénin, c'est-à-dire pour que l'estimateur ridgeless soit proche de l'optimum minimax. Les auteurs montrent que la surparamétrisation ainsi que des conditions très spécifiques sur la décroissance des valeurs propres de Σ sont nécessaires pour que ce phenomène se produise. Le papier [146] étudie précisément l'estimateur de régression ridge dans le cadre surparamétré et donne des bornes non asymptotiques de généralisation pour une matrice de covariance générale Σ. Il montre de plus que ces bornes sont optimales sur une certaine échelle de valeurs des paramètres de régularisation. Ces résultats ont ensuite été raffinés dans [175].

1.8.2 Kernel ridge(less) regression

Des extensions à la régression ridgeless à noyau dans les RKHS ont été étudiées dans [142] lorsque la taille de l'échantillon n et la dimension d sont supposées satisfaire $n \asymp d$, et dans [143] pour un cas plus général $d \asymp n^\alpha$ pour $\alpha \in (0, 1)$. Ces articles donnent des bornes supérieures dépendant des données sur le risque qui peuvent être faibles en supposant des propriétés spectrales sur la matrice de covariance des données et sur le noyau. D'autre part, si d est constant (indépendant de n), alors l'estimateur interpolant de moindre norme par rapport au noyau de Laplace est inconsistant [116]. La régression ridge à noyau est une des méthodes statistiques d'estimation de fonctions. Ce cadre

est donc étroitement lié au cadre de la régression non-paramétrique, que nous étudions dans cette thèse.

1.8.3 Régression non-paramétrique

Dans le cadre de la régression non paramétrique, on dispose de n observations i.i.d. $(X_i, Y_i), i = 1, \ldots, n$ où $Y_i = f(X_i) + \epsilon_i$, $X_i \in \mathbb{R}^d$, $Y_i \in \mathbb{R}$ où les ϵ_i sont des variables aléatoires de bruit i.i.d. indépendantes de X_i et f est une fonction inconnue à estimer. Nous supposons que f appartient à une classe non-paramétrique donnée, comme les classes de fonctions de Hölder, Sobolev ou Besov par exemple. Dans le cadre de la régression non-paramétrique avec régularité de Hölder connue $\beta \leq 2$, il existe des estimateurs interpolants atteignant des vitesses optimales minimax [106]. En particulier, il a été prouvé dans [106] que l'estimateur Nadaraya-Watson avec un noyau singulier peut interpoler les données tout en atteignant des vitesses minimax optimales.

Part I

Minimax testing

Chapter 2

Sharp Local Minimax Rates for Goodness-of-Fit Testing in multivariate Binomial and Poisson families and in multinomials

This Chapter is based on the paper: "Sharp Local Minimax Rates for Goodness-of-Fit Testing in multivariate Binomial and Poisson families and in multinomials" [132], by Julien Chhor and Alexandra Carpentier (arXiv:2012.13766), to appear in Mathematical Statistics and Learning.

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 iid 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.

Keywords: Minimax Identity Testing, Goodness-of-fit Testing, Multinomial Distributions, Multivariate Poisson Families, Locality.

2.1 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 2.2 - with unknwon distribution q, and given a known distribution p, we want to test

$$\mathcal{H}_0 : p = q \qquad vs \qquad \mathcal{H}_1 : d(p,q) \geq \rho,$$

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 a given $\eta > 0$, and to identify such a test. See Section 2.2 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 e.g. 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 [1] 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 [183] - 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 [25, 30].

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 [168], [128], [96] for excellent surveys. When observing n iid 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. [16] (in Russian), [17], [23], [179], [48], and also in the related two-sample testing problem - where both p, q are unknown and observed through samples - see e.g. [32, 66]. 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. [95], [81]. Note that these papers provide results in terms of sample complexity, and more recently, the paper [104] has re-considered this problem in terms of minimax separation distance - focusing also on the case of smooth densities. Another quite related work is [124], 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 [57, 72, 81, 139]. This multinomial framework proves very useful for a wide range of applications, which include Ising models [109], bayesian networks [129] or even quantum mechanics [103].

The papers [95, 104] 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 2.3.1 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*, 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 [135]), 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 [134] 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 ([85], [65], [74], [77], [53]), in social networks ([78], [76]), as well as network modeling ([35], [185]), or network dynamics ([41]). The papers [135] and [102] 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 [70] 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 [90] and [83]. Phase transitions are also known for detecting strongly connected groups or high dimensional geometry in large random graphs ([79]). The paper [94] tests random dot-product graphs in the two sample setting with low-rank adjacency matrices. The paper [91] 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 [64], brain or 'omics' networks [92] [63], testing chemicals [56] or ecology and evolution [52]. Moreover, and to

*We compare with this stream of literature under our upper and lower bounds in Sections 2.3, and also in the discussion in Section 2.4.

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 2.2, we describe the setting by defining the multivariate binomial model and the minimax framework. In Section 2.3, state our main theorem, which gives an explicit expression of the minimax separation radius as a function of p and n. In Section 2.3.1, we establish the equivalence between the binomial, the Poisson and the multinomial settings. In Section 2.4, we discuss our results, by comparing them with the state of the art, especially with the multinomial setting. In Section 2.5, we describe our lower bound construction. In Section 2.6, we describe our tests and state theoretical results guaranteeing their optimality. We finally provide additional comments on our results in Section 2.7. All proofs are deferred to the Appendix.

2.2 Problem statement

2.2.1 Setting

We first introduce the Binomial setting. In Section 2.3.1, 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 $X_1, \dots, X_n$ iid such that each X_i can be written as $X_i = \Big(X_i(1), \dots, X_i(N)\Big)$ where all of the entries $X_i(1), \dots, X_i(N)$ are mutually independent and $X_i(j) \sim Ber(q_j)$. We slightly abuse notation and write $X_1, \dots, X_n \overset{iid}{\sim} q$ 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 *wlog* 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

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

2.2.2 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: $X_1, \dots, X_n \overset{iid}{\sim} q$. We are interested in the following testing problem:

$$\mathcal{H}_0^p : q = p \qquad \text{vs} \qquad \mathcal{H}_1^{\rho,p,t} : q \in \mathcal{P}_N;\ \|p - q\|_t \geq \rho. \tag{2.1}$$

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 $X_1, \dots, X_n$, taking only the values 0 or 1. We measure the quality of any test ψ by its ***maximum risk***, defined as:

$$\begin{aligned} R(\psi) &:= R_{\rho,p,t,n}(\psi) \\ &= \mathbb{P}_p(\psi = 1) + \sup_{\substack{q \text{ s.t.} \\ \|p-q\|_t \geq \rho}} \mathbb{P}_q(\psi = 0). \end{aligned} \tag{2.2}$$

$R(\psi)$ 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:

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

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. [91].

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** $R^*_{\rho^*_{p,t,n},p,t,n} \leq \eta$:

$$\rho^*_{p,t,n}(\eta) = \inf \left\{ \rho > 0 \ : \ R^*_{\rho,p,t,n} \leq \eta \right\}. \tag{2.3}$$

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.

2.3 Results

Without loss of generality, assume that $\max_{1\le j\le N} p_j \le \frac{1}{2}$. 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 $i = 1,\cdots,n = 2k$. *Wlog*, assume that all entries of the known vector p are sorted in decreasing order:

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

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

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

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

$$r = \frac{2t}{4-t} \qquad \text{and} \qquad b = \frac{4-2t}{4-t}. \tag{2.4}$$

for p we also define

$$I = \min\left\{ J : \sum_{i>J} p_i^2 \le \frac{c_I}{n^2} \right\} \tag{2.5}$$

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

Theorem 2.1. *For all $t \in [1,2]$, the following bound holds, up to a constant depending only on η and t:*

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

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

The lower bounds and the minimax test are given in Section 2.5 and Section 2.6.

2.3.1 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 $\widetilde{n} \sim Poi(n)$ 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.

Prop 2.1 (Poissonization trick for multinomials)**.** *Let $n \ge 2$ and assume that p, q are probability vectors, i.e. such that $\sum_i p_i = \sum_i q_i = 1$. Let $\widetilde{n} \sim Poi(n)$. Conditional on $\widetilde{n}$, let $Z_1,\cdots,Z_{\widetilde{n}} \overset{iid}{\sim} \mathcal{M}(q)$. We build the histogram sufficient statistic by defining, for all $j = 1,\cdots,N$, $H_j = \sum_{i=1}^{\widetilde{n}} \mathbb{1}\{Z_i = j\}$. Then for all j, $H_j \sim Poi(nq_j)$ and $H_1,\cdots,H_N$ are mutually independent.*

Prop 2.2 (Poissonization trick for binomial families)**.** *Let $n \geq 2$ and $\widetilde{n} \sim Poi(n)$. Conditional on $\widetilde{n}$, let $X_1, \cdots, X_{\widetilde{n}} \overset{iid}{\sim} \bigotimes_{j=1}^{N} Ber(p_j)$. Then $\sum_{i=1}^{\widetilde{n}} X_i \sim \bigotimes_{j=1}^{N} Poi(np_j)$.*

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 equivalences that follow.

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

1. **Binomial case:** This is the setting considered above. We define $\mathcal{P}^{(Bin)} = \{Ber(p);\ p \in \mathbb{R}_+^N\}$ where by convention, $Ber(p) := \bigotimes_{j=1}^{N} Ber(p_j)$. We fix $p \in \mathcal{P}^{(Bin)}$ and suppose we observe $X_1, \cdots, X_n \overset{iid}{\sim} Ber(q)$ for $q \in \mathcal{P}^{(Bin)}$ unknown. We consider the **binomial** testing problem:

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

2. **Poisson case:** $\mathcal{P}^{(Poi)} = \{Poi(p);\ p \in \mathbb{R}_+^N\}$ where by convention, $Poi(p) := \bigotimes_{j=1}^{N} Poi(p_j)$. We fix $p \in \mathcal{P}^{(Poi)}$ and suppose we observe $Y_1, \cdots, Y_n \overset{iid}{\sim} Poi(q)$ for $q \in \mathcal{P}^{(Poi)}$ unknown. We consider the **Poisson** testing problem:

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

3. **Multinomial case** $\mathcal{P}^{(Mult)} = \left\{ \mathcal{M}(p) \middle|\ p \in \mathbb{R}_+^N, \sum_{j=1}^{N} p_j = 1 \right\}$ where $\mathcal{M}(p)$ denotes the multinomial distribution over $\{1, \ldots, N\}$. We fix $p \in \mathcal{P}^{(Mult)}$ and suppose we observe $Z_1, \cdots, Z_n \overset{iid}{\sim} \mathcal{M}(q)$ for $q \in \mathcal{P}^{(Mult)}$ unknown. We consider the **Multinomial** testing problem:

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

   where for $x = (x_1, \cdots, x_N)$: $\|x\|_{\mathcal{M},t} = \left[\sum_{j=2}^{N} |x_j|^t\right]^{1/t}$ 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 $\rho^*_{Bin}(n, p, t, \eta)$, $\rho^*_{Poi}(n, p, t, \eta), \rho^*_{Mult}(n, p, t, \eta)$ for the minimax separation distances in the sense of Equation (2.3), 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 $c_{BP},\ C_{BP} > 0$ depending on η such that $\forall p \in [0,1]^N,\ \forall n \geq 2\ \eta > 0,$:

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

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 $\sum_{i=1}^N p_i = 1$:

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

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

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

Corollary 2.1. *Let $t \in [1,2]$. The minimax separation radii in the Poisson and multinomial cases are respectively given by:*

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

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

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.

2.4 Discussion

In this entire section, we mostly discuss the Multinomial setting - whose rates are given in Corollary 2.1 - 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.

2.4.1 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. [95], [81], [104] [124]. 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. [16] (in Russian), [17], [23], [179], [48]. 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 $\rho^*_{Mult}(n, p^{(n)}, 2, \eta)$ 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 2.7.1 below), we get:

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

2.4.2 Comparison with existing literature in the multinomial case

Our results are quite related to those of [95], 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 $X_1, \cdots, X_n \overset{iid}{\sim} \mathcal{M}(p)$ needed to ensure that the Multinomial testing problem introduced in Section 2.3.1 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[†]. Note that the quantities n^* and ρ^* are dual, for $\eta = 2/3$.

[95] proves the following bounds to characterize the optimal sample complexity $n^*(p, \epsilon)$ when given a fixed $\epsilon > 0$:

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

In the above bound, $p = (p_1, \cdots, p_N)$ where $p_1 \geq \cdots \geq p_N \geq 0$ and $\sum_{i=1}^N p_i = 1$. For $\epsilon > 0$, let J be the smallest index such that $\sum_{i>J} p_i \leq \epsilon$. The notation $p_{-\epsilon}^{-\max}$ 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.

[†]See Equation (2.2) for the definition of this quantity in the graph problem.

In Appendix 2.D, we justify that the upper and lower bounds from [95], when translated in terms of separation radius as in [104] actually match in the multinomial case, although claimed otherwise by the authors of [104] 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 [107], which considers instance optimal identity testing. Specifically, [107] 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

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

This paper proves that for fixed $\epsilon \in (0, 1)$, any test for testing identity to p needs at least $C\kappa_p^{-1}(1 - 2\epsilon)$ 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 [95]. Note that their bound is not optimal either, but is incomparable to [95]. 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, [118] 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.

2.5 Lower bounds

We recall the definitions of r and b in equation (2.4). In what follows, index A is defined as

$$A = A_{p,t,n}(\eta) := \max\left\{ a \leq I : p_a^{b/2} \geq \frac{c_A}{\sqrt{n}\Big(\sum\limits_{i\leq I} p_i^r\Big)^{\frac{1}{4}}} \right\}, \tag{2.6}$$

where $c_A > 0$ is a small enough constant depending only on η. We adopt the convention that $\max \emptyset = -\infty$ and that $p_{\leq -\infty} = \emptyset$ and $p_{>-\infty} = p$. We start by presenting the lower bound part of Theorem 2.1. 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 $p_{\leq A}$ and the tail as $p_{>A}$.

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 2.1. *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*

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

and

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

This implies that $\rho = \frac{\left\|p_{\le A}\right\|_r^{r/t}}{\sqrt{n}\left\|p_{\le I}\right\|_r^{r/4}} + \frac{1}{n}$ 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 $(\delta_1, \cdots, \delta_A) \in \{\pm 1\}^A$ we define the distribution q_δ such that:

$$(q_\delta)_j = \begin{cases} p_j + \delta_i \gamma_j & \text{if } j \le A \\ p_j & \text{otherwise,} \end{cases} \tag{2.7}$$

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

$$\gamma_i = \frac{c_\gamma\, p_i^{\frac{2}{4-t}}}{\sqrt{n}\left(\sum_{i \le I} p_i^r\right)^{\frac{1}{4}}}. \tag{2.8}$$

The mixture

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

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

The core of the proof is to prove that observations $X_1, \dots, X_n$ drawn from this mixture distribution $\bar{\mathbb{P}}_{\text{bulk}}$ are so difficult to distinguish from observations $X'_1, \dots X'_n$ 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 $\left(\sum_{i=1}^A \gamma_i^t\right)^{\frac{1}{t}} \asymp \frac{\left\|p_{\le A}\right\|_r^{r/t}}{\sqrt{n}\left\|p_{\le I}\right\|_r^{r/4}}$. Therefore, $\frac{\left\|p_{\le A}\right\|_r^{r/t}}{\sqrt{n}\left\|p_{\le I}\right\|_r^{r/4}}$ 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 [95] which is given for the case $t = 1$, up to some issues that are discussed in details in Subsection 2.4.2.

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, *whp*, no same coordinate is observed twice or more among the n data.

Proposition 2.2. *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*

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

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

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

Index U has no further meaning than to guarantee that for all $i \geq U : \pi_i \in [0,1]$. 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 $b_i \sim Ber(\pi_i)$ mutually independent, and

$$q_b(i) = b_i \bar{\pi}, \tag{2.9}$$

We now consider the mixture of the probability distributions q_b:

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

As above, we prove that the data $X_1, \dots, X_n$ drawn from this mixture $\bar{\mathbb{P}}_{\text{tail}}$ 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 $\bar{\mathbb{P}}_{\text{tail}}$ and p, is larger, up to an absolute constant than

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

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

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

in words, that we can replace U by I. This lower bound departs significantly from the one in [95] in the case $t = 1$, which is significantly simpler than for $t > 1$ for the tail coefficients.

Combination of both lower bounds

By combining Propositions 2.1 and 2.2, we obtain the following theorem.

Theorem 2.2. *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*

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

and

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

This theorem implies that

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

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

Note that when combining Propositions 2.1 and 2.2, we do not get exactly the expression in Theorem 2.2. We actually obtain:

$$\rho^* \gtrsim_\eta \frac{\|P_{\leq A}\|_r^{r/t}}{\sqrt{n}\|P_{\leq I}\|_r^{r/4}} + \frac{\|p_{>I}\|_1^{\frac{2-t}{t}}}{n^{\frac{2t-2}{t}}} + \frac{1}{n}.$$

We therefore need to show that this expression is equivalent to that in Theorem 2.2. This is done by using Lemma 9, which states that we can replace $\frac{\|P_{\leq A}\|_r^{\frac{r}{t}}}{\sqrt{n}\|P_{\leq I}\|_r^{\frac{r}{4}}}$ by $\sqrt{\frac{\|p_{\leq I}\|_r}{n}}$ without changing the rate, i.e.

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

Remark on index A: As explained in (2.7), the optimal prior is of the form $p_i \pm \gamma_i$ where γ_i is proportional to $p_i^{\frac{2}{4-t}}$, according to Equation (2.8). Since $\frac{2}{4-t} \leq 1$, 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, *whp*, 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 *whp* under 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, 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 it as possible. More precisely, we impose that *whp*, no coordinate is observed twice, and such that coordinate-wise, the expected number of observations is equal to that under the null hypothesis p. In short, this prior is designed such that its first order moment is equal to that under the null and its second order moment is unobserved *whp*. 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 [95].

Remark on the lower bounds: The bulk lower bound is close to that of [95]. The tail lower bound relies on a sparse prior that is an existing technique (for example in sparse testing, see [36], [87], [162]) and is very different from the construction in [95]. Handling the indices I, A and U require careful manipulations that we believe are new techniques.

2.6 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

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

Test for the bulk coefficients

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

$$T_{\text{bulk}} = \sum_{i \leq A} \frac{1}{p_i^b} \left(\frac{S_i}{k} - p_i \right) \left(\frac{S_i'}{k} - p_i \right), \tag{2.10}$$

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

$$\psi_{\text{bulk}} = \mathbf{1} \left\{ T_{\text{bulk}} > \frac{\underline{c}_\eta}{n} \left\| p_{\leq A} \right\|_r^{\frac{r}{2}} \right\},$$

where $\underline{c}_\eta = 4/\sqrt{\eta}$ 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 2.3. *There exists $\underline{c}'_\eta > 0$, such that the following holds.*

- *Type I error is bounded:*
$$\mathbb{P}_p(\psi_{\text{bulk}} = 1) \leq \eta/2.$$
- *Type II error is bounded: for any q such that*
$$\|q_{\leq A}\|_t \geq \underline{c}'_\eta \left(\sqrt{\frac{\|p_{\leq I}\|_r}{n}} + \frac{1}{n} \right),$$
it holds that
$$\mathbb{P}_q(\psi_{\text{bulk}} = 0) \leq \eta/2.$$

For $t = 1$, we get $r = \frac{2}{3}$, which is the norm identified in [95]. 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 Subsection 2.3.1, 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 [95]. In [95], the cut-off I' is the smallest index such that, for a fixed ϵ: $\sum_{i>I'} p_i \leq \epsilon$. 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 [95], 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 $r = \frac{2t}{4-t}$.

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:

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

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

$$\psi_2 = \mathbb{1}\Big\{\exists j > A : N_j \geq 2\Big\} \tag{2.11}$$

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

$$T_1 = \sum_{i>A} \frac{N_i}{n} - p_i, \quad \psi_1 = \mathbf{1}\left\{|T_1| > \underline{c}_\eta \sqrt{\frac{\sum_{i>A} p_i}{n}}\right\}. \tag{2.12}$$

We prove the following proposition regarding this statistic.

Proposition 2.4. *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*
$$\|q_{>A}\|_t \geq \underline{c}'_\eta \left(\frac{\|p_{>A}\|_1^{\frac{2-t}{t}}}{n^{\frac{2t-2}{t}}} + \frac{1}{n} \right),$$
it holds that
$$\mathbb{P}_q(\psi_1 \vee \psi_2 = 0) \leq \eta/2.$$

Recall that the tail is defined such that, *whp* 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 2.4 proves that this combination of tests reaches the optimal rate.

In [95], 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 2.4, 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.

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

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

The following theorem states that this test reaches the rate $\bar{\rho}$, which is the minimax rate ρ^* given in Theorem 2.1. In other words, it guarantees that, whenever the two hypotheses are $\bar{\rho}$-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 $\rho^* \lesssim \bar{\rho}$.

Theorem 2.3. *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*
$$\|p - q\|_t \geq \underline{c}'_\eta \left(\sqrt{\frac{\|p_{\leq I}\|_r}{n}} + \frac{\|p_{>I}\|_1^{\frac{2-t}{t}}}{n^{\frac{2t-2}{t}}} + \frac{1}{n} \right),$$
it holds that
$$\mathbb{P}_q(\psi = 0) \leq \eta/2.$$

2.6.1 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

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

and the associated test:

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

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 (2.10), 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 $p_{>A}$.

2.7 Further remarks on the results

2.7.1 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.

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: $\|p_{>I}\|_2 \lesssim_\eta \frac{1}{n}$. This case has already been solved in [66]. 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 [95]. 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 [75] 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 $\sqrt{\frac{\|p_{\le I}\|_r}{n}}$ involves a duality between the norms ℓ_t and ℓ_r for $r = \frac{2t}{4-t}$ - as was also the case for $t = 1$ in [95]. 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 $q = (p_1 + \delta_1\gamma_1, \cdots, p_A + \delta_A\gamma_A)$ for $\delta_i \overset{iid}{\sim} Rad(\frac{1}{2})$ like in (2.7), except that this time, we *do not* impose that $(\gamma_i)_i$ is defined as in (2.8). 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[‡] 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 $a = \frac{4}{t}$ and $b = \frac{4}{4-t}$. Setting γ^* the vector on the frontier of Γ reaching the equality case in Hölder's inequality, we obtain for fixed n: $\|\gamma^*\|_t \propto \|p\|_r^{1/2}$.
As to the contribution of the tail, we refer the reader to the remarks below Proposition 2.2.

[‡]Although the proof is written for graph samples, it is argued in Subsection 2.3.1 that it can be transposed to the multinomial or the Poisson settings.

2.7.2 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 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:

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

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.

Appendix

2.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 $\mathcal{G} = \left(\{0,1\}^N\right)^n$ 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*

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

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

Proof of Lemma 3. We have that:

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

where the definition of the χ^2 divergence can be found in [188], as well as the proof for the inequality $d_{TV} \leq \frac{1}{2}\sqrt{\chi^2}$. Therefore:

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

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

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

□

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

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

Lemma 4. *There exists a sufficiently small absolute constant c_4 such that, if $\sum_{i=1}^{N} \frac{\gamma_i^4}{p_i^2} \le \frac{c_4}{n^2}$, then for all $\rho \le \|\gamma\|_t$ we have $R^*(\rho) \ge \eta$.*

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 $S = \sum_{i=1}^n X_i \in \{0, \dots, n\}^N$ and write $S = (s_1, \dots, s_N)$. 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 $q_\delta = p + (\delta_i \gamma_i)_{1 \le i \le N}$. 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:

$$\frac{1}{|\mathcal{G}|} \sum_{X \in \mathcal{G}} \frac{\left(\mathbb{E}_{q \sim \mathcal{U}(\mathcal{P}_1)} \mathbb{P}_q(X) \right)^2}{\mathbb{P}_p(X)} = \frac{1}{|\mathcal{G}|} \sum_{X \in \mathcal{G}} \sum_{\delta, \delta'} \prod_{i=1}^{N} \frac{(p_i + \delta_i \gamma_i)^{s_i} (1 - p_i - \delta_i \gamma_i)^{n - s_i}}{p_i^{s_i} (1 - p_i)^{n - s_i}} \times (p_i + \delta_i' \gamma_i)^{s_i} (1 - p_i - \delta_i' \gamma_i)^{n - s_i}$$

$$= \frac{1}{|\mathcal{G}|} \sum_{\delta, \delta'} \prod_{i=1}^{N} \sum_{l=0}^{n} \binom{n}{l} \left(p_i + (\delta_i + \delta_i') \gamma_i + \frac{\delta_i \delta_i' \gamma_i^2}{p_i} \right)^l \left(1 - p_i - (\delta_i + \delta_i') \gamma_i + \frac{\delta_i \delta_i' \gamma_i^2}{1 - p_i} \right)^{n - l}$$

$$= \frac{1}{|\mathcal{G}|} \sum_{\delta, \delta'} \prod_{i=1}^{N} \left(1 + \frac{\delta_i \delta_i' \gamma_i^2}{p_i (1 - p_i)} \right)^n = \prod_{i=1}^{N} \left[\frac{1}{4} \sum_{\delta_i, \delta_i' \in \{\pm 1\}} \left(1 + \frac{\delta_i \delta_i' \gamma_i^2}{p_i (1 - p_i)} \right)^n \right]$$

$$= \prod_{i=1}^{N} \left[\frac{1}{2} \left(1 + \frac{\gamma_i^2}{p_i (1 - p_i)} \right)^n + \frac{1}{2} \left(1 - \frac{\gamma_i^2}{p_i (1 - p_i)} \right)^n \right]$$

$$\leq \prod_{i=1}^{N} \left[\frac{1}{2} \exp\left(\frac{n\gamma_i^2}{p_i(1-p_i)} \right) + \frac{1}{2} \exp\left(\frac{-n\gamma_i^2}{p_i(1-p_i)} \right) \right]$$

$$= \prod_{i=1}^{N} \cosh\left(\frac{n\gamma_i^2}{p_i(1-p_i)} \right) \leq \exp\left(\sum_{i=1}^{N} \frac{n^2\gamma_i^4}{2p_i^2(1-p_i)^2} \right)$$

Note that

$$\begin{aligned} \exp\left(\sum_{i=1}^{N} \frac{n^2\gamma_i^4}{2p_i^2(1-p_i)^2} \right) &\leq 1 + 4(1-\eta)^2 \\ \iff \quad & \sum_{i=1}^{N} \frac{\gamma_i^4}{p_i^2(1-p_i)^2} \leq \frac{2c_A^4}{n^2} \\ \impliedby \quad & \sum_{i=1}^{N} \frac{\gamma_i^4}{p_i^2} \leq \frac{c_A^4}{2n^2} \end{aligned} \tag{1}$$

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

□

This means the following: let $\gamma := (\gamma_i)_i$ satisfying (1) and let $\rho = \|\gamma\|_t$. Then all points $p + (\delta_i\gamma_i)_{1\leq i\leq|\mathcal{G}|}$ 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*

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

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

$$a = \frac{c_A}{\sqrt{n}\left(\sum_{i\leq I} p_i^r\right)^{\frac{1}{4}}} \tag{2.13}$$

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

- $\forall i \leq A\,, \quad (q_\delta)_i = p_i + a\delta_i p_i^{\frac{2}{4-t}}$ where a is defined in (2.13)
- $\forall i > A\,, \quad (q_\delta)_i = p_i.$

Let $\mathcal{P}_1 = \left\{q_\delta \mid \delta \in \{\pm 1\}^A\right\}$. We set a uniform prior on $\mathcal{P}_1$. With the notation of Lemma 4, we just set $\gamma_i = a p_i^{\frac{2}{4-t}}$ 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:

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

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

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

□

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

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

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

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

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 $q_i = b_i \bar{\pi}$. 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

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

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

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

We now show that $\left[\mathbb{E}_{q\sim\mathcal{Q}}\left[\|p-q\|_t^t\right]\right]^2 \gg \mathbb{V}_{q\sim\mathcal{Q}}\left[\|p-q\|_t^t\right]$. 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

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

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, $\|p_{\geq U}\|_1^2n^2\geq\frac{1}{9}\gg c_u$. We conclude using Chebyshev's inequality. Therefore, this prior is indeed separated away from the null distribution by a distance greater than $\bar{\pi}\sum_{i\geq U}\pi_i$ up to a constant, or equivalently, greater than $\frac{\|p_{\geq U}\|_1^{\frac{2-t}{t}}}{n^{\frac{2(t-1)}{t}}}$.

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

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

Moreover, we recall that for any realization $X = (X_1, \dots, X_n)$ we write $S = \sum_{i=1}^n X_i$. We have

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

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.
We have - since s is a sufficient statistic

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

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

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

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 $U > I$ we have $\forall i \geq U\ np_i \leq c_I$ so that for any $i \geq U$

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

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

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

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

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

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

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

Therefore

$$d_{TV}\left(\mathbb{P}_p, \bar{\mathbb{P}}_{\text{tail}}\right) = \underbrace{\frac{1}{2} \sum_{X \in \mathcal{G}: \forall i \geq U, s_i \leq 1} \left|\mathbb{P}_p(X) - \bar{\mathbb{P}}_{\text{tail}}(X)\right|}_{\text{principal term}} + c_I + c_u \tag{2.14}$$

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. We have - since s is a sufficient statistic

$$\begin{aligned}
&\sum_{X \in \mathcal{G}: \forall i \geq U, s_i \leq 1} \left|\mathbb{P}_p(X) - \bar{\mathbb{P}}_{\text{tail}}(X)\right| \qquad (2.15)\\
\leq &\sqrt{\sum_{X \in \mathcal{G}: \forall i \geq U, s_i \leq 1} \frac{\left(\mathbb{P}_p(X) - \bar{\mathbb{P}}_{\text{tail}}(X)\right)^2}{\mathbb{P}_p(X)}} \sqrt{\underbrace{\sum_{X \in \mathcal{G}: \forall i \geq U, s_i \leq 1} \mathbb{P}_p(X)}_{\leq 1}}\\
\leq &\sqrt{\sum_{X \in \mathcal{G}: \forall i \geq U, s_i \leq 1} \frac{\bar{\mathbb{P}}_{\text{tail}}(X)^2}{\mathbb{P}_p(X)} - 1 + 2c_u} = \sqrt{\prod_{i=U}^{N} \left(\sum_{j=0}^{1} \frac{\bar{\mathbb{P}}_{\text{tail}}(s_i = j)^2}{\mathbb{P}_p(s_i = j)}\right) - 1 + 2c_u}. \qquad (2.16)
\end{aligned}$$

Computation of $\sum_{k=0}^{1} \frac{\bar{\mathbb{P}}_{\text{tail}}(s_i=k)^2}{\mathbb{P}(s_i=k)}$. :

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

The first term writes:

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

The second term writes:

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

We can now sum the two terms:

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

So that

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

Now, using (2.14) and (2.16), we have: $d_{TV}(\mathbb{P}_p, \bar{\mathbb{P}}_{\text{tail}}) \leq \frac{1}{2}\sqrt{5c_u} + c_I + c_u \leq 1 - \eta$ 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 $q_1 = p_1 + \frac{1-\eta}{n}$ and $q_j = p_j$ for all $j \geq 2$.

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

This concludes the proof. □

Lemma 8. *It holds : $\|p_{\geq U}\|_1 + \frac{1}{n} \asymp \|p_{>I}\|_1 + \frac{1}{n}$.*
Moreover, we either have $\|p_{\geq U}\|_1 \geq \frac{1}{3}\|p_{>I}\|_1$ or $\|p_{>I}\|_1 \leq \|p_{\geq U}\|_1 + \frac{\sqrt{c_I}}{n}$

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 $P_{I\to U} = (p_{I+1}, \cdots, p_{U-1})$. We 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,

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

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

$$I - U - 1 < 2 \quad i.e. \quad I - U - 1 \leq 1$$

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.

- If $A = I$ then the result is clear.
- Otherwise, $I > A$. Note that by setting $p'_i := np_i$ for all $i = 1, \cdots, N$, the result to show can be rewritten as:

$$\frac{\|p'_{\leq A}\|_r^{\frac{r}{t}}}{\|p'_{\leq I}\|_r^{\frac{r}{4}}} + \|p'_{\geq I}\|_1^{2-t} \asymp \sqrt{\|p'_{\leq I}\|_r} + \|p'_{\geq I}\|_1^{2-t}. \tag{2.17}$$

We have by definition of A and I:

$$p_I'^{2-r}\left(\sum_{i\geq I} p_i'\right)^{2-r} = \left(\sum_{i\geq I} p_I' p_i'\right)^{2-r} \geq \left(\sum_{i\geq I} p_i'^2\right)^{2-r} \gtrsim_\eta 1 \text{ and}$$

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

Hence, by noticing that $2b = 2-r$ we have $\left(\sum_{i\geq I} p_i'\right)^{2-r} \gtrsim_\eta \sum_{i\leq I} p_i'^r$, which yields $\|p'_{\geq I}\|_1^{2-t} \geq \sqrt{\|p'_{\leq I}\|_r} \geq \frac{\|p'_{\leq A}\|_r^{\frac{r}{t}}}{\|p'_{\leq I}\|_r^{\frac{r}{4}}}$ by raising to the power $\frac{1}{2r}$. This condition yields the result of the lemma, by replacing p' by np.

□

Lemma 10. $\|p_{>I}\|_1 + \frac{1}{n} \asymp \|p_{>A}\|_1 + \frac{1}{n}$.

Proof of lemma 10. If $A = I$ then the result is clear. Now, suppose that $A < I$. We have, by the definition of A:

$$\frac{c_A^4}{n^2} > pA + 1^{2b}\sum_{i\leq I} p_i^r \geq \sum_{i=A+1}^{I} p_i^2 \geq p_I \sum_{i=A+1}^{I} p_i \implies \frac{c_A^4}{n^2\sum_{i=A+1}^{I} p_i} \geq p_I$$

Moreover if $I < N$,

$$\frac{c_I}{n^2} \leq \sum_{i>I} p_i^2 \leq p_{I+1}\sum_{i>I} p_i \implies p_{I+1} \geq \frac{c_I}{n^2\sum_{i>I} p_i}$$

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

$$p_{A+1}^{2b} < \frac{c_A^4}{n^2\sum_{i=1}^{N} p_i^r} \implies \sum_{j=A+1}^{N} p_{A+1}^{2b} p_j^r \leq \frac{c_A^4}{n^2}$$

$$\implies \sum_{j=A+1}^{N} p_j^2 \leq \frac{c_A^4}{n^2}$$

$$\implies P_N\|p_{>A}\|_1 \leq \frac{c_A^4}{n^2}$$

$$\implies \frac{\sqrt{c_I}}{n}\|p_{>A}\|_1 \leq \frac{c_A^4}{n^2}$$

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

□

2.B Upper bound

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

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

and we prove: $\rho^* \lesssim_\eta \rho$.
We start with the three following lemmas which control the expectation and variance of the statistics $T_{\text{bulk}}, T_1, T_2$. 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 (2.10). The expectation and variance of T_{bulk} satisfy:*

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

Lemma 12 (Bounds on expectation and variance of T_1)**.** *Let T_1 be defined as in equation (2.12). The expectation and variance of T_1 satisfy:*

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

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

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

$$\mathbb{P}\left(\psi_{\text{bulk}} = 1\right) \le \frac{\eta}{16}, \tag{2.18}$$

Test ψ_1. Since $p = q$, we have by Lemma 12 that $\mathbb{E}(T_1) = 0$ and $\mathbb{V}(T_1) \le \sqrt{\frac{\sum_{i>A} p_i}{n}}$.
By the same argument ψ_1's type-I error is upper bounded as:

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

so that by definition of ψ_1

$$\mathbb{P}_p\left(\psi_1 = 1\right) \le \frac{\eta}{16}, \tag{2.19}$$

Test ψ_2. We have by Lemmas 13 and 14

$$\mathbb{P}\left(\psi_2 = 1\right) \le c_I + c_A^4 \le \frac{\eta}{16}, \tag{2.20}$$

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

Conclusion : Putting together equations (2.19), (2.18) and (2.20) we get that the type I error of $\psi = \psi_{\text{bulk}} \vee \psi_1 \vee \psi_2$ is upper bounded as

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

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

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

- **First case:** Either $\|\Delta_{\le A}\|_t \ge \bar{c}_\eta \rho$
- **Second case:** Or $\|\Delta_{> A}\|_t \ge \bar{c}_\eta \rho$

Proposition 2.5 (Study in the **First case**). *There exists a large enough constant $\bar{c}_\eta^{(\text{bulk})} > 0$ such that if $\|\Delta_{\le A}\|_t \ge \bar{c}_\eta^{(\text{bulk})} \rho$, then*

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

Proposition 2.6 (Study in the **Second case**). *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 2.5. 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 $\mathbb{V}(T_{Bulk}) \leq c'\,\mathbb{E}(T_{Bulk})^2$ and such that $\lim_{c\to+\infty} c' = 0$.
By definition of ρ, we have in particular: $\|\Delta_{\leq A}\|_t \geq c\sqrt{\frac{\|p_{\leq I}\|_r}{n}} \vee \frac{c}{n}$, hence

$$\frac{1}{n^2} \leq \frac{1}{c^4}\frac{\|\Delta_{\leq A}\|_t^4}{\|p_{\leq I}\|_r^2} \wedge \frac{\|\Delta_{\leq A}\|_t^2}{c^2} \tag{2.21}$$

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

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

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

Term ①: We have by Hölder's inequality:

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

$$\text{Hence}\qquad \|\Delta_{\leq A}\|_t \leq \left(\sum_{i\leq A}\frac{\Delta_i^2}{p_i^b}\right)^{\frac{1}{2}}\left(\sum_{i\leq A} p_i^r\right)^{\frac{2-t}{2t}}. \tag{2.22}$$

Moreover, we have $\frac{1}{n^2} \leq \frac{\|\Delta_{\leq A}\|_t^4}{c^4\|p_{\leq I}\|_r^2}$ so that the term ① writes:

$$\frac{2}{n^2}\sum_{i\leq A} p_i^r \leq 2\sum_{i\leq A} p_i^r\left(\sum_{i\leq A}\Delta_i^t\right)^{\frac{4}{t}}\frac{1}{c^4\left(\sum_{i\leq I} p_i^r\right)^{\frac{2}{r}}}$$

$$\leq \frac{2}{c^4}\left(\sum_{i\leq A} p_i^r\right)^{1-\frac{2}{r}}\left(\sum_{i\leq A}\frac{\Delta_i^2}{p_i^b}\right)^2\left(\sum_{i\leq A} p_i^r\right)^{\frac{4-2t}{t}} \quad \text{by (2.22)}$$

$$= \frac{2}{c^4}\left(\sum_{i\leq A}\frac{\Delta_i^2}{p_i^b}\right)^2 = \frac{2}{c^4}\mathbb{E}[T_{\text{bulk}}]^2. \tag{2.23}$$

Term (2): The condition $a \leq p_A^{\frac{b}{2}}$ ensures that:

$$p_A^b \geq a^2 = \frac{c_A^2}{\sqrt{2}(\sum_{j\leq I} p_j^r)^{1/2}n} =: \widetilde{c}\frac{1}{(\sum_{j\leq I} p_j^r)^{1/2}n}.$$

Using this condition, the term (2) writes:

$$\sum_{i\leq A}\frac{1}{p_i^{2b}}\frac{\Delta_i^2}{n^2} \leq \frac{1}{n^2}\frac{1}{p_A^b}\sum_{i\leq A}\frac{\Delta_i^2}{p_i^b} \leq \widetilde{c}^{-1}\frac{1}{n}\left(\sum_{j\leq I} p_j^r\right)^{\frac{1}{2}}\left(\sum_{i\leq A}\frac{\Delta_i^2}{p_i^b}\right). \tag{2.24}$$

Moreover, since $\sqrt{\frac{\|p_{\leq I}\|_r}{n}} \leq \rho \leq \frac{1}{c}\|\Delta_{\leq A}\|_t$ we have, using (2.22):

$$\frac{1}{n}\left(\sum_{j\leq I} p_j^r\right)^{\frac{1}{2}} = \frac{1}{n^b}\left(\sqrt{\frac{\|p_{\leq I}\|_r}{n}}\right)^r \leq \frac{1}{n^b c^r}\left(\sum_{i\leq A}\frac{\Delta_i^2}{p_i^b}\right)^{\frac{r}{2}}\left(\sum_{i\leq A} p_i^r\right)^{\frac{b}{2}} \leq \frac{1}{c^2}\sum_{i\leq A}\frac{\Delta_i^2}{p_i^b}. \tag{2.25}$$

In the last inequality, we use the fact proved in case number (1) that $\frac{1}{n^b}\left(\sum_{i\leq A} p_i^r\right)^{\frac{b}{2}} \leq \frac{1}{c^{2b}}\mathbb{E}[T_{\text{bulk}}]^b$ and the relation $\frac{r}{2} + b = 1$

Plugging in (2.24) yields that the second term (2) is bounded by $\mathbb{E}[T_{\text{bulk}}]^2$

Term (3): This term writes:

$$\frac{1}{n}\sum_{i\leq A} p_i^{1-2b}\Delta_i^2 \leq \frac{\|\Delta_{\leq A}\|_t^2}{c^2\left(\sum_{i\leq I} p_i^r\right)^{\frac{1}{r}}}\sum_{i\leq A} p_i^{1-2b}\Delta_i^2$$

$$\leq \frac{1}{c^2}\left(\sum_{i\leq A}\frac{\Delta_i^2}{p_i^b}\right)\left(\sum_{i\leq A} p_i^r\right)^{\frac{4-2t}{2t}-\frac{1}{r}}\sum_{i\leq A} p_i^{1-2b}\Delta_i^2 \quad \text{using (2.22)}$$

$$\leq \frac{1}{c^2}\left(\sum_{i\leq A}\frac{\Delta_i^2}{p_i^b}\right)\left(\sum_{i\leq A} p_i^r\right)^{\frac{-1}{2}}\left(\sum_{i\leq A} p_i^{\frac{2}{3}(1-2b)}\Delta_i^{\frac{4}{3}}\right)^{\frac{3}{2}} \quad \text{since } \|\cdot\|_1 \leq \|\cdot\|_{\frac{2}{3}}.$$

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

$$\sum_{i\leq A} p_i^{\frac{2}{3}(1-2b)}\Delta_i^{\frac{4}{3}} \;\leq\; \left(\sum_{i\leq A}\left(\frac{p_i^{\frac{2}{3}(1-2b)}\Delta_i^{\frac{4}{3}}}{p_i^{\frac{2}{3}\frac{t}{4-t}}}\right)^{\frac{3}{2}}\right)^{\frac{2}{3}}\left(\sum_{i\leq A}\left(p_i^{\frac{2}{3}\frac{t}{4-t}}\right)^3\right)^{\frac{1}{3}} \;\leq\; \left(\sum_{i\leq A}\frac{\Delta_i^2}{p_i^b}\right)^{\frac{2}{3}}\left(\sum_{i\leq A} p_i^r\right)^{\frac{1}{3}}.$$

So that

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

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

This yields that the third term satisfies:

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

Term (4): The fourth term writes:

$$\frac{1}{n}\left\|\left(\frac{|\Delta_i|}{p_i^{\frac{2b}{3}}}\right)_{i\leq A}\right\|_3^3 \;\leq\; \frac{1}{n}\left\|\left(\frac{|\Delta_i|}{p_i^{\frac{2b}{3}}}\right)_{i\leq A}\right\|_2^3 \;=\; \frac{1}{n}\left(\sum_{i\leq A}\frac{\Delta_i^2}{p_i^{\frac{4b}{3}}}\right)^{\frac{3}{2}} \;\leq\; \frac{1}{n^{\frac{1}{2}}}\left(\sum_{i\leq A}\frac{\Delta_i^2}{p_i^b}\right)^{\frac{3}{2}}\left(\sum_{i\leq I} p_i^r\right)^{\frac{1}{4}},$$

where in the last step we have used the fact that

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

Then using (2.24):

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

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

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

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

Moreover, using (2.23), we have that for c large enough, $\mathbb{E}(T_{Bulk}) \geq \frac{c}{n}\|p_{\leq A}\|_{\frac{r}{2}}^{\frac{r}{2}} \geq 2\frac{c_\eta}{n}\|p_{\leq A}\|_{\frac{r}{2}}^{\frac{r}{2}}$ 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 2.6

Proof of Proposition 2.6. We will need the two following 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 :

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

□

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

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

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

□

Under H_0: We upper bound the type-I error of tests ψ_1 and ψ_2. For ψ_2: by Lemma 13, $\mathbb{P}(\psi_2 = 1) \leq \sum_{j>A} n^2p_j^2 \leq C_A \leq \frac{\eta}{4}$.

As to test ψ_1: $\mathbb{P}(\psi_1 = 1) = \mathbb{P}(|T_1| > \underline{c}_\eta \sqrt{\frac{\sum_{i>A} p_i}{n}}) \leq \frac{\eta}{4}$ 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. **Until the end of the proof, we drop from now on the indexation "$> A$" and write only e.g. $\|p\|_2, \|\Delta\|_2$ instead of $\|p_{>A}\|_2, \|\Delta_{>A}\|_2$.**

We have by Hölder's inequality:

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

for $C = C_1 C_2$ where $C_1 = \left(\left(\frac{20}{\eta} (\underline{c}_\eta + 1) + 1 \right) \right)^{2-t}$,
$C_2 = \left(\frac{1}{4} \left(\log(4/\eta)^2 \vee 9/100 \right) + c_I \right)^{(t-1)/2}$ so that one of the 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 $\|q\|_2 \geq C_2^{1/(t-1)} / n - \|p\|_2 \geq \frac{1}{n} \left(C_2^{1/(t-1)} - c_I \right)$.

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

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

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

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

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:

$$\begin{aligned} q(\psi = 0) &\leq \exp\left(-\frac{1}{3} \sum_{j \in I_-} {n'}^2 q_j^2 - \frac{1}{10} \sum_{j \in I_+} n' q_j \right) \\ &\leq \exp\left(-\frac{1}{3} \sum_{j \in I_-} {n'}^2 q_j^2 - \frac{1}{10} \left(\sum_{j \in I_+} {n'}^2 q_j^2 \right)^{1/2} \right) \\ &= \exp\left(-\frac{1}{3}(S - S_+) - \frac{1}{10} (S_+)^{1/2} \right) \text{ for } S = \sum_{j > A} {n'}^2 q_j^2 \text{ and } S_+ = \sum_{j \in I_+} {n'}^2 q_j^2. \end{aligned}$$

Now, $S_+ \mapsto -\frac{S}{3} + \frac{1}{3} S_+ - \frac{\sqrt{S_+}}{10}$ is convex over $[0, S]$ so its maximum is reached on the boundaries of the domain and is therefore equal to $(-\frac{\sqrt{S}}{10}) \vee -\frac{S}{3} = -\frac{\sqrt{S}}{10}$ for $S \geq 9/100$. Now, since 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: $\|\Delta\|_1^{2-t} \geq C_1 \left(\|p\|_1^{2-t} + \frac{1}{n^{2-t}} \right)$. 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 $\left| \sum_{j>A} \Delta_j \right| \geq \frac{1}{2} \|\Delta\|_1$*

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

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

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, $s_+ \geq 3 \geq 3 s_-$ 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:

$$q(\psi_1 = 0) = q\left(\left| \sum_{j>A} \frac{N_j}{n} - p_j \right| \leq \underline{c}_\eta \sqrt{\frac{\|p\|_1}{n}} \right)$$

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

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 $\|q\|_1 \geq \frac{20}{\eta}(\underline{c}_\eta + 1)(\|p\|_1 + 1/n)$, we have $q(\psi_1 = 0) \leq \eta/4$. This condition is guaranteed when $\|\Delta\|_1 \geq \Big(\frac{20}{\eta}(\underline{c}_\eta + 1) + 1\Big)(\|p\|_1 + 1/n) = C_1^{1/(2-t)}(\|p\|_1 + 1/n)$ □

Proof of lemma 11.

- Expectation:

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

- Variance:

$$
\mathbb{V}(T_{\text{bulk}}) = \sum_{i\leq A} \frac{1}{p_i^{2b}}\left(\mathbb{E}\left[\left(\frac{S_i}{k} - p_i\right)^2\left(\frac{S_i'}{k} - p_i\right)^2\right] - \mathbb{E}\left[\left(\frac{S_i}{k} - p_i\right)\left(\frac{S_i'}{k} - p_i\right)\right]^2\right)
$$

$$= \sum_{i \leq A} \frac{1}{p_i^{2b}} \left(\mathbb{E}\left[\left(\frac{S_i}{k} - p_i \right)^2 \right]^2 - (p_i - q_i)^4 \right),$$

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

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

□

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

$$\mathbb{V}[T_1] = \mathbb{V}\left[\sum_{i > A} \frac{S_i + S_i'}{n} \right] = \sum_{i > A} \frac{\mathbb{V}\left[S_i\right] + \mathbb{V}\left[S_i'\right]}{n^2} \qquad \text{by independence of the } (S_i, S_i')_i$$

$$= \sum_{i > A} \frac{q_i(1-q_i)}{n} \leq \sum_{i > A} \frac{q_i}{n}$$

□

2.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 case are equivalent.

Proof of Lemma 1. **We first prove** $\rho^*_{Poi}(n,p) \leq C_{BP}\, \rho^*_{Bin}(n,p)$. Let $n \geq 2$, and let $Y_1, \cdots, Y_n \overset{iid}{\sim} Poi(q)$. We consider a random function ϕ such that for any Poisson family $Y_1, \cdots, Y_n \overset{iid}{\sim} Poi(q)$,

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

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

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

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

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

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

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

$$H_0 : q = p \qquad \text{v.s.} \qquad H_1 : \|p - q\|_t \geq \rho_{Bin}(cn, p, \frac{\eta}{2})$$

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

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

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

$$H_0 : q = p \qquad \text{v.s.} \qquad H_1 : \|p - q\|_t \geq \rho$$

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

$$\begin{aligned} \mathbb{P}_{H_0}\left(\psi(Y_1^n) = 1\right) &\leq \mathbb{P}_{H_0}\left(\mathcal{A}_1 \cap \psi(Y_1^n) = 1\right) + \mathbb{P}_{H_0}(\bar{\mathcal{A}}_1) \\ &\leq \mathbb{P}_{H_0}\left(\psi(Y_1^n) = 1 | \mathcal{A}_1\right) + \frac{\eta}{4} \\ &\leq \mathbb{P}_{H_0}\left(\psi_{Bin}(X_1, \cdots, X_{\lfloor cn \rfloor}) = 1 | \mathcal{A}_1\right) + \frac{\eta}{4} \\ &= \mathbb{P}_{X_1^{\lfloor cn \rfloor} \sim Ber(p)^{\bigotimes \lfloor cn \rfloor}}\left(\psi_{Bin}(X_1, \cdots, X_{\lfloor cn \rfloor}) = 1\right) + \frac{\eta}{4} \end{aligned}$$

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

$$\mathbb{P}_q\left(\psi(Y_1^n) = 0\right) \leq \mathbb{P}_{X_1^{\lfloor cn \rfloor} \sim Ber(q)^{\bigotimes \lfloor cn \rfloor}}\left(\psi_{Bin}(X_1, \cdots, X_{\lfloor cn \rfloor}) = 0\right) + \frac{\eta}{4}$$

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

$$\begin{aligned} R(\psi) &= \mathbb{P}_{H_0}\left(\psi(Y_1^n) = 1\right) + \sup_{\|p-q\|_t \geq \rho} \mathbb{P}_q\left(\psi(Y_1^n) = 0\right) \\ &\leq \frac{\eta}{2} + \mathbb{P}_{X_1^{\lfloor cn \rfloor} \sim Ber(p)^{\bigotimes \lfloor cn \rfloor}}\Big(\psi_{Bin}(X_1, \cdots, X_{\lfloor cn \rfloor}) = 1\Big) \\ &\quad + \sup_{\|p-q\|_t \geq \rho} \mathbb{P}_{X_1^{\lfloor cn \rfloor} \sim Ber(q)^{\bigotimes \lfloor cn \rfloor}}\Big(\psi_{Bin}(X_1, \cdots, X_{\lfloor cn \rfloor}) = 0\Big) \\ &= \frac{\eta}{2} + R(\psi_{Bin}) \\ &= \frac{\eta}{2} + \frac{\eta}{2} = \eta \end{aligned}$$

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

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

$$\mathcal{A}_2 = \{\widetilde{n} \leq n\} \tag{2.26}$$

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

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

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

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

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

$$\bar{\pi}(y_1, \cdots, y_n, m) = (y_1, \cdots, y_m)$$

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:

$$H_0 : q = p \qquad \text{v.s.} \qquad H_1 : \|p - q\|_t \geq \rho_{Poi}(\lfloor \overline{c}n \rfloor, p, \frac{\eta}{2})$$

We define the randomized test

$$\bar{\psi} = \begin{cases} \psi_{Poi} \circ \bar{\pi} \circ \bar{\phi}(Y) & \text{if } \mathcal{A}_2 \\ \\ 0 & \text{otherwise.} \end{cases} \tag{2.27}$$

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

$$\begin{aligned}
\mathbb{P}_{H_0}\left(\bar{\psi}(Y)=1\right) &\leq \mathbb{P}_{H_0}\left(\mathcal{A}_2 \cap \bar{\psi}(Y)=1\ \right) + \mathbb{P}_{H_0}(\bar{\mathcal{A}}_2) \\
&\leq \mathbb{P}_{H_0}\left(\bar{\psi}(Y)=1\ |\mathcal{A}_2\right) + \frac{\eta}{4} \\
&\leq \mathbb{P}_{H_0}\left(\psi_{Poi}(Y_1,\cdots,Y_{\lfloor \bar{c}n\rfloor})=1|\mathcal{A}_2\right) + \frac{\eta}{4} \\
&= \mathbb{P}_{Y_1^{\lfloor \bar{c}n\rfloor}\sim Poi(p)\bigotimes^{\lfloor \bar{c}n\rfloor}}\left(\psi_{Poi}(Y_1,\cdots,Y_{\lfloor \bar{c}n\rfloor})=1\right) + \frac{\eta}{4}
\end{aligned}$$

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

$$\mathbb{P}_q\left(\bar{\psi}(Y)=0\right) \leq \mathbb{P}_{Y_1^{\lfloor \bar{c}n\rfloor}\sim Poi(q)\bigotimes^{\lfloor \bar{c}n\rfloor}}\left(\psi_{Poi}(Y_1,\cdots,Y_{\lfloor \bar{c}n\rfloor})=0\right) + \frac{\eta}{4}$$

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

$$\begin{aligned}
R(\psi) &= \mathbb{P}_{H_0}\left(\bar{\psi}(Y)=1\right) + \sup_{\|p-q\|_t\geq\rho} \mathbb{P}_q\left(\psi(Y)=0\right) \\
&\leq \frac{\eta}{2} + \mathbb{P}_{Y_1^{\lfloor \bar{c}n\rfloor}\sim Poi(p)\bigotimes^{\lfloor \bar{c}n\rfloor}}\left(\psi_{Poi}(Y_1,\cdots,Y_{\lfloor \bar{c}n\rfloor})=1\right) \\
&\quad + \sup_{\|p-q\|_t\geq\rho} \mathbb{P}_{Y_1^{\lfloor \bar{c}n\rfloor}\sim Poi(q)\bigotimes^{\lfloor \bar{c}n\rfloor}}\left(\psi_{Poi}(Y_1,\cdots,Y_{\lfloor \bar{c}n\rfloor})=0\right) \\
&= \frac{\eta}{2} + R(\psi_{Poi}) \\
&= \frac{\eta}{2} + \frac{\eta}{2} = \eta
\end{aligned}$$

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

□

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

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

where we recall that for any vector $v = (v_1, \dots, v_\ell)$ with nonnegative entries, we denote by $\mathrm{Poi}(v)$ the distribution $\bigotimes_{j=1}^{\ell} \mathrm{Poi}(v_j)$. On $\mathcal{A}_2$, defined in (2.26), we have

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

so we can just apply the exact same steps to prove that, if $q = p$ then the test $\bar{\psi}$ from (2.27) has type-I error less than $\frac{\eta}{2}$ and if $\|q - p\|_{\mathcal{M},t} \geq \rho_{Poi}(\bar{c}n, p, \frac{\eta}{2})$, its type-II error is less than $\frac{\eta}{2}$.

We now prove the converse bound: $\rho^*_{Poi}(n, p^{-\max}, \eta) \lesssim_\eta \rho^*_{Mult}(n, p, \eta)$. 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

$$H_0 : q = p^{-\max} \quad \text{versus} \quad H_1 : \|q - p^{-\max}\|_t \geq \rho. \tag{2.28}$$

We exhibit a test ψ and a constant $C > 0$ such that if $\rho \geq C\rho^*_{Mult}(n, p, \eta)$, then its risk for problem (2.29) 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:

$$H_0 : q' = p \quad \text{versus} \quad H_1 : \|q' - p\|_{\mathcal{M},t} \geq \rho^*_{Mult}(p, m, \eta). \tag{2.29}$$

Now, draw $X_1 \sim \mathrm{Poi}(np_1)$ independently on $(X_2, \dots, X_N)$, so that $(X_1, \dots, X_N) \sim \mathrm{Poi}(n\bar{q})$ where $\bar{q} = (p_1, q_2, \dots, q_d)$. 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 $m := \sum_{j=1}^{N} X_j$:

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

First, if $\left| \|\bar{q}\| - 1 \right| > \frac{C'}{\sqrt{n}}$, then with probability at least $1 - \eta/100$, ψ_0 will detect it. From now on, assume that $\left| \|\bar{q}\| - 1 \right| \leq \frac{C'}{\sqrt{n}}$. We now prove that for some large enough constant C, C', if

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

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

Now, since $\|p\|_1 \leq 1$ and $r = \frac{2t}{4-t} \leq t$, we have $\|\cdot\|_r \geq \|\cdot\|_t$ so that

- $\frac{C'}{\sqrt{n}}\|p_{\leq A}\|_{\mathcal{M},t} \leq \frac{C}{\sqrt{n}}\sqrt{\|p^{-\max}\|_r} \leq c\rho^*_{Mult}(p,n,\eta)$ for some small enough $c>0$, provided that n is greater than a suitable constant depending on η.
- By Hölder's inequality, we get $\frac{C'}{\sqrt{n}}\|p_{>A}\|^t_{\mathcal{M},t} \leq \frac{C'}{\sqrt{n}}\|p_{>A}\|_1^{2-t}\|p_{>A}\|_2^{(t-1)} \leq \frac{C'}{\sqrt{n}}\|p_{>A}\|_1^{2-t}\cdot\left(\frac{1}{n^2}\right)^{(t-1)} \leq c\rho^*_{Mult}(p,n,\eta)$.

Therefore, we get:

$$\left\|\frac{\bar{q}}{\|\bar{q}\|_1}-p\right\|_{\mathcal{M},t} \geq C\rho^*_{Mult}(p,n,\eta). \tag{2.30}$$

Now, choose n larger than a suitable constant depending only on η such that $\mathbb{P}\left(\mathrm{Poi}(n) \geq \frac{n}{2}\right) \geq 1-\eta/100$. Conditional on $m = \sum_{j=1}^N X_j$, the observations $(X_1,\dots,X_N)$ follow a multinomial distribution $\mathcal{M}(m, \frac{\bar{q}}{\|\bar{q}\|_1})$. Hence, with probability at least $1-\eta/2$, the test ψ_m will conclude in favor of H_1 in view of (2.30) whenever $m \geq \frac{n}{2}$, since $\rho^*_{Mult}(p,n,\eta) \geq C\rho^*_{Mult}(p,\frac{n}{2},\frac{\eta}{2})$. We now prove that the risk of ψ for Problem (2.29) is at most η.

On the other hand, if $\bar{q}=\bar{p}$, then with probability $\geq 1-\eta/100$: $\psi_0(X_1,\dots,X_N)=0$ and whenever $m \geq \frac{n}{2}$, we have $\psi_m(X_1,\dots,X_N)=0$ with probability at least $1-\frac{\eta}{4}$ by definition of ψ_m, since $\rho^*_{Mult}(p,n,\eta) \geq C\rho^*_{Mult}(p,\frac{n}{2},\frac{\eta}{4})$.

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

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

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

□

2.D Tightness of [104] in the multinomial case

For fixed n and for two absolute constants $C, c > 0$, define ϵ_+ as the largest quantity satisfying $\epsilon_+ \leq C\sqrt{\frac{\|p^{-\max}_{-\epsilon_+/16}\|_{2/3}}{n}} + \frac{C}{n}$ and ϵ_- as the smallest quantity satisfying $\epsilon_- \geq c\sqrt{\frac{\|p^{-\max}_{-\epsilon_-}\|_{2/3}}{n}} + \frac{c}{n}$. By [104], the critical radius ρ^* satisfies $\epsilon_- \lesssim \rho^* \lesssim \epsilon_+$.

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

2. Second case: otherwise, $\epsilon_+ \leq C\sqrt{\frac{\|p^{-\max}_{-\epsilon_+/16}\|_{2/3}}{n}} + \frac{C}{n} \leq C\sqrt{\frac{\|p^{-\max}_{-\epsilon_-}\|_{2/3}}{n}} + \frac{C}{n} \leq \frac{C}{c}\epsilon_-$ so that the bounds also match in this case.

Chapter 3

Goodness-of-Fit Testing for Hölder-Continuous Densities: Sharp Local Minimax Rates

This Chapter is based on the paper "Goodness-of-Fit Testing for Hölder-Continuous Densities: Sharp Local Minimax Rates" [158] by Julien Chhor and Alexandra Carpentier (arXiv:2109.04346).

Abstract

We consider the goodness-of fit testing problem for Hölder smooth densities over $\mathbb{R}^d$: given n iid observations with unknown density p and given a known density p_0, we investigate how large ρ should be to distinguish, with high probability, the case $p = p_0$ from the composite alternative of all Hölder-smooth densities p such that $\|p - p_0\|_t \geq \rho$ where $t \in [1, 2]$. The densities are assumed to be defined over $\mathbb{R}^d$ and to have Hölder smoothness parameter $\alpha > 0$. In the present work, we solve the case $\alpha \leq 1$ and handle the case $\alpha > 1$ using an additional technical restriction on the densities. We identify matching upper and lower bounds on the local minimax rates of testing, given explicitly in terms of p_0. We propose novel test statistics which we believe could be of independent interest. We also establish the first definition of an explicit cutoff u_B allowing us to split $\mathbb{R}^d$ into a bulk part (defined as the subset of $\mathbb{R}^d$ where p_0 takes only values greater than or equal to u_B) and a tail part (defined as the complementary of the bulk), each part involving fundamentally different contributions to the local minimax rates of testing.

3.1 Introduction

This paper studies the local Goodness-of-Fit testing problem for α-Hölder densities over $\Omega = \mathbb{R}^d$. For all $\alpha, L > 0$, $H(\alpha, L)$ denotes the class of α-Hölder densities over Ω. We place ourselves on a subclass $\mathcal{P}(\alpha, L)$ of $H(\alpha, L)$. The classes $\mathcal{P}(\alpha, L)$ and $H(\alpha, L)$ are defined in Section 3.2. We endow $\mathcal{P}(\alpha, L)$ with some distance denoted by $\mathrm{dist}(\cdot, \cdot)$, which in our setting, can be any L_t distance for $t \in [1, 2]$: $\mathrm{dist}(p, q) = \|p - q\|_t$. Given the iid observations $X_1, \ldots, X_n$ with same unknown density $p \in \mathcal{P}(\alpha, L)$, and given a known density $p_0 \in \mathcal{P}(\alpha, L)$, we consider the non-parametric testing

problem:

$$H_0 : p = p_0 \qquad \text{vs} \qquad H_1(\rho) : p \in \mathcal{P}(\alpha, L) \text{ and } \text{dist}(p, p_0) \geq \rho. \tag{3.1}$$

This problem is called the goodness-of-fit problem for continuous densities, which has been thoroughly studied in many works [19, 17, 16, 183, 23, 48, 44, 29].

Following [16, 17, 183], we will focus on establishing, up to a multiplicative constant, the smallest possible separation distance $\rho^* = \rho^*(p_0, n, \text{dist})$ in a minimax sense such that a uniformly consistent test exists for Problem (3.1) - this condition will be specified in more details in Section 3.2.

Problem (3.1) has most often been studied for the uniform density p_0 over a bounded domain, e.g. $[0,1]^d$ [17], [179]. It has been extended to the case of densities p_0 constrained to be bracketed between two constants, still on a bounded domain [16], [23]. See [179, Chapter 6.2] for a more recent overview. In the case where p_0 is the uniform density on $[0,1]^d$ and for α-Hölder densities with $L = 1$, and when the distance is defined as $d(p,q) = \|p - q\|_t$ where $\|.\|_t$ is the L_t norm with $t \in [1, \infty]$, the minimax-optimal separation radius for Problem (3.1) is

$$n^{-2\alpha/(4\alpha+d)}. \tag{3.2}$$

See e.g. [183, Theorem 4.2] for the case where $d = 1$ and in the related sequence space model over Besov balls. However, these results hinge on the assumption that p_0 is lower bounded by a positive constant. Hence, they cannot be extended to null densities on unbounded domains.

In fact, there is a fundamental gap between testing on bounded or unbounded domains. This was recently illustrated in the paper [104] which considers the case of Lipschitz densities ($\alpha = 1$) with separation in total variation distance (L_1 distance). The authors prove that there can be substantial heterogeneity when it comes to the minimax-optimal radius ρ, depending on p_0: testing some null hypotheses can be much easier than testing others. More precisely, they prove that uniformly over the class of L-Lipschitz densities, the minimax separation distance is bracketed as follows:

$$\left(\frac{L^{d/2} \left(\int_{p_0 \geq a(p_0)} p_0^{\frac{2}{3+d}} \right)^{\frac{3+d}{2}}}{n} \right)^{\frac{2}{4+d}} \lesssim \rho^*(p_0, n, \|.\|_1) \lesssim \left(\frac{L^{d/2} \left(\int_{p_0 \geq b(p_0)} p_0^{\frac{2}{3+d}} \right)^{\frac{3+d}{2}}}{n} \right)^{\frac{2}{4+d}},$$

where $a(p_0) > b(p_0) > 0$ are quantities - that are small and matching in order of magnitude for many cases, albeit not all - that depend only on n, p_0 and that are defined implicitly. See Section 3.6.3 for a thorough description of their results. The authors formally prove the interesting fact that the minimax separation distance depends on p_0 and they provide a test adapted to the shape of the density. For instance, if the density p_0 defined over $\mathbb{R}$ has essentially all its mass on e.g. $[0,1]$, then the minimax optimal ρ is $L^{1/5} n^{-2/5}$ - unsurprisingly comparable with in [183]. However, if p_0 is heavy tailed, e.g. corresponds to the Pareto distribution with parameter β, then the minimax optimal ρ is $L^{1/5} n^{-2\beta/(2+3\beta)}$ - differing considerably from the rate of [183]. This example highlights a specificity of testing heavy-tailed distributions, and by extension, distributions with unbounded

support. To encompass all cases, it is therefore important to derive *local* results where both the separation distance and the associated tests depend on p_0 in a refined way. The results in [104] follow on ideas from a stream of literature concerning property testing. For goodness-of-fit testing in the discrete (multinomial) setting, see [4, 10, 32] for global results and [81, 66, 95, 88, 71, 132] for local results - see also [96] for an excellent survey. In the related setting of goodness-of-fit testing under local differential privacy, see [124, 159]. Closest to our setting is [132], which studies the problem of goodness-of-fit for multinomials in L_t norm for $t \in [1, 2]$ - see Section 3.6 for a thorough description of their results, and comparison.

In this paper, we focus on the problem of goodness-of-fit testing for Hölder smooth densities p_0, defined on unbounded domains, extending over classical goodness-of-fit testing results following [183]. We find how the minimax separation distance ρ depends on p_0 and therefore provide local results. We consider a variety of separation distances, going beyond the $\|.\|_1$ distance from [104]: namely we consider all the L_t distances for $t \in [1, 2]$, as in [132] for the multinomial case. We cover all the scale of Hölder classes $H(\alpha, L)$ for all $\alpha > 0$, under technical assumptions for $\alpha > 1$, extending from the case of testing Lipschitz densities ($\alpha = 1$) studied in [104]. We identify the matching upper and lower bounds on $\rho(p_0, n, \text{dist})$ and provide the corresponding optimal tests in all the cases described above. In our results, the radius $\rho(n, p_0, \text{dist}_{L_t})$ is given explicitly as a function of p_0.

We now give a brief overview of the related literature on density testing, and explain more in details our contributions.

1. **Testing for α-Hölder distributions:** In the setting where p_0 is the uniform distribution over $[0, 1]^d$ the global minimax rates are well understood, see Equation (3.2) and [183] for an adaptation of these results to the setting where p_0 is uniform over $[0, 1]^d$. In the local setting, however, little is known. In the breakthrough results of [104], the case of Lipschitz densities ($\alpha = 1$) is almost completely solved. However, the general case $\alpha > 0$ is not considered and from the construction of the tests in [104], it is clear that the case $\alpha > 1$ is far from being a trivial extension. Indeed, the test statistics in [104] are built using heterogeneous histograms, which are not smooth enough when $\alpha > 1$. In the present paper, we solve the case of α-Hölder densities for any α - but we need to introduce a technical condition for $\alpha > 1$. This assumption is akin to assuming that the density p_0 has all its derivative which take value 0 in its inflexion points whose value is close to 0. We introduce novel test statistics, based on kernel estimators with heterogeneous bandwidth, simpler than the test statistic defined on an heterogeneous partition of the space from [104]. We believe that this test statistic can also be of independent interest. See Section 3.6 for a comparison with [104].

2. **Extension to L_t distance:** The choice of distance influences the geometry of the alternative and consequently the nature of optimal tests as well as the expression for the minimax separation radius. In Section 3.6, we highlight that changing the norm can actually change the null densities p_0 which are the easiest or most difficult ones to test. The distances considered in the density testing literature are often either the L_1 distance - as is the case of local results in [104] where the separation is only considered in L_1 distance - or the L_∞ distance [179, Chapter 6.2]. In the discrete setting, the L_1 norm is often considered [95], as well as the L_2 norm

[124]. The paper [135] considers the two sample testing problem in inhomogeneous random graphs, in order to study the effect of various distances (total variation, Frobenius distance, operator norm, Kullback-Leibler divergence). The paper [134] considers the goodness-of-fit testing problem in inhomogeneous random graphs for the Frobenius and operator norm distances. However, as will appear in the present paper, the phenomena occurring for testing in L_t distances are similar for all $t \in [1, 2]$. This property has already been identified in [132] in the discrete setting (multivariate Poisson families, inhomogeneous random graphs, multinomials). The present paper extends the results from [132] to the continuous case, highlighting a deep connection between the two settings. However, as discussed in Section 3.6, the results from [132] cannot be directly transferred to the density setting. In our paper, we extend the results of [104] to the case of more general norms. This impacts the choice of test statistics and new regimes appear, see Section 3.6 for a comparison with [104] .

3. **Matching upper and lower bounds:** The local rates established by [104] provide the first upper and lower bounds for density testing in the local case. Although matching in most usual cases, the authors discuss quite pathological cases for which their upper and lower bounds do not match. Indeed, the method proposed in [104] builds on the well-known multinomial identity testing analysis from [95], which identifies upper and lower bounds on the minimax separation radius for testing in total variation distance. However, even in the discrete setting, some specific cases can be found for which these upper and lower bounds do not match, explaining the untightness of [104] in some cases. In the present paper we bridge the gap, by proposing a new way to define a cut-off between bulk (set of large values of p_0) and tail (set of small values of p_0). This approach leads us to provably matching upper and lower bounds on the minimax separation radius. As opposed to [104], our result can moreover be expressed as an explicit function of p_0.

The paper is organized as follows. In Section 3.2, we define the testing problem. In Section 3.3, we state our main theorem identifying the sharp minimax rate for the testing problem. We then analyse separately two different regimes, namely the bulk regime (in Section 3.4) and the tail regime (in Section 3.5). We finally discuss our results in Section 3.6.

3.2 Problem Statement

3.2.1 Definition of the class of densities $\mathcal{P}(\alpha, L)$

To ensure the existence of consistent tests for Problem (3.1), structural assumptions need to be made on the class of densities we consider. Indeed, as shown in [9] and [19], no consistent test can distinguish between an arbitrary p_0 and alternatives separated in l_t norm if no further assumption is imposed on the set of alternatives. Throughout the paper, we place ourselves on a restricted subclass of the Hölder class of functions. Our class corresponds to the densities on $\Omega = \mathbb{R}^d$ ($d \in \mathbb{N}^*$), with Hölder regularity and satisfying Assumption ($\star$) defined below.

Let $\alpha, L > 0$ and denote by $\|\cdot\|$ the Euclidean norm over $\mathbb{R}^d$. We recall the definition of the Hölder class over Ω. Set* $m = \lceil \alpha \rceil - 1$ and consider a function $p : \Omega \longrightarrow \mathbb{R}$ that is m times differentiable. Write $z \mapsto P_p(x, z)$ for the Taylor polynomial of degree m of p at x. The Hölder class is defined as:

$$H(\alpha, L) = \Big\{ p \mid \Omega \longrightarrow \mathbb{R} : p \text{ is } m \text{ times differentiable and } \forall x, y \in \Omega : |p(x) - P_p(x, y-x)| \leq L\|x-y\|^\alpha \Big\}.$$

Our class of densities is obtained by intersecting $H(\alpha, L)$ with the set of densities $p : \Omega \longrightarrow \mathbb{R}_+$ satisfying:

$$\forall (x, y) \in \Omega : \quad |p(x) - p(y)| \leq \ c_\star p(x) \ + \ L\|x-y\|^\alpha, \tag{$\star$}$$

for some fixed constant $c_\star \in (0, \frac{1}{2})$. Note that Assumption ($\star$) is automatically satisfied for $\alpha \leq 1$. We discuss Assumption ($\star$) in Section 3.6. The class of densities therefore considered throughout the paper is defined as:

$$\mathcal{P}_\Omega(\alpha, L, c_\star) = \Big\{ p \in H(\alpha, L) \ \Big| \ \int_\Omega p = 1, \ p \geq 0 \text{ and } p \text{ satisfies } (\star) \ \Big\}. \tag{3.3}$$

When no ambiguity arises, we will drop the lower index Ω since it is assumed to be equal to $\mathbb{R}^d$.

3.2.2 Minimax testing framework

Throughout the paper, we fix $t \in [1, 2]$. For $f \in L_t(\Omega)$, we denote by $\|f\|_t$ the L_t norm of f with respect to the Lebesgue measure:

$$\|f\|_t = \Big(\int_\Omega |f|^t \, dx \Big)^{1/t}.$$

Assume *wlog* that the number of observations n is even: $n = 2k$ ($k \in \mathbb{N}^*$). We fix two constants $\alpha, L > 0$. Assume moreover that we observe n *iid* random variables $X_1, \cdots, X_n$ with the same *unknown* density $p \in \mathcal{P}(\alpha, L, c_\star)$. Let p_0 be one particular *known* density in $\mathcal{P}(\alpha, L, c_\star)$ and fix $\delta > 0$. For some $\rho > 0$, the *goodness-of-fit* testing problem is defined as:

$$\begin{aligned} H_0 \quad &: \ p = p_0 \qquad \text{versus} \\ H_1(\rho, t) &: \ p \in \mathcal{P}(\alpha, L', c'_\star) \text{ s.t. } \|p - p_0\|_t \geq \rho, \end{aligned} \tag{3.4}$$

where $L' = (1+\delta)L$ and $c'_\star = (1+\delta)c_\star$. The parameter $\delta > 0$ can be chosen arbitrarily small. This point is specific for obtaining local minimax lower bounds (see also [104]). Without this assumption, arbitrarily small perturbations of p_0 when p_0 is on the boundary of the class are out of the class. Thus, putting $\delta = 0$ is problematic as the least favorable functions in the lower bounds are small perturbations of p_0.

*Here $\lceil x \rceil$ is the smallest integer greater than or equal to a given real number x.

Our goal is to establish how large ρ should be for (3.4) to be feasible in a sense we now formally specify.

A *test function* $\psi : \Omega^n \longrightarrow \{0, 1\}$ is defined as a measurable function of the observations $(X_1, \dots, X_n)$ taking only values 0 or 1. The quality of any given test ψ is measured by its *risk*, defined as the sum of its type-I and type-II errors:

$$R(\psi, \rho) \stackrel{def}{=} \mathbb{P}_{p_0}(\psi = 1) + \sup_{p \in \mathcal{P}_1(\rho,t)} \mathbb{P}_p(\psi = 0) \tag{3.5}$$

where $\mathcal{P}_1(\rho, t) = \{p \in \mathcal{P}(\alpha, L', c'_\star) : \|p - p_0\|_t \geq \rho\}$ is the set of all p satisfying $H_1(\rho, t)$. We are looking for a test with smallest possible risk, if it exists. We therefore introduce the *minimax risk* as:

$$R^*(\rho) = R^*(n, p_0, t, \rho) = \inf_\psi R(\psi) = \inf_\psi \left\{ \mathbb{P}_{p_0}(\psi = 1) + \sup_{p \in \mathcal{P}_1(\rho,t)} \mathbb{P}_p(\psi = 0) \right\}, \tag{3.6}$$

which corresponds to the risk of the best possible test. Here, $\inf_\psi$ denotes the infimum over all tests. Note that if $R^*(\rho) = 1$, then random guessing is optimal. Hence, to have a non-trivial testing problem, it is necessary to guarantee $R^*(\rho) \leq \eta$ for some fixed constant $\eta \in (0, 1)$. Noting that this bound on R^* can only hold for ρ large enough, we introduce the *minimax separation radius*, also called *minimax (testing) rate* or *critical radius*, defined as the smallest $\rho > 0$ ensuring $R^*(\rho) \leq \eta$.

Definition 3.1 (Minimax separation radius)**.** *We define the minimax separation radius, or minimax (testing) rate, as:*

$$\rho^* := \inf \left\{ \rho > 0 : R^*(n, p_0, t, \rho) \leq \eta \right\}. \tag{3.7}$$

In the following, we fix $\eta \in (0, 1)$. The aim of the paper is two-fold.

1. Find the minimax rate $\rho^* = \rho^*(n, p_0, \alpha, L, t)$ defined in (3.7) and associated to problem (3.4), up to multiplicative constants which are allowed to depend on t, η, α and d.

2. Find a test ψ^* and a constant $C > 0$ such that $R(\psi^*, C\rho^*) \leq \eta$. This ensures that if the hypotheses are separated by $C\rho^*$, then Problem (3.4) is guaranteed to have a decision procedure with risk at most η, namely ψ^*.

3.2.3 Notation

In what follows, we will define $x \vee y = \max(x, y)$, $x \wedge y = \min(x, y)$ and $x_+ = x \vee 0$. We will use $\|\cdot\|$ to denote the Euclidean norm over $\mathbb{R}^d$. The **support** of a function $p : \Omega \longrightarrow \mathbb{R}$ is defined as $\{x \in \Omega : p(x) \neq 0\}$. We write $\lceil x \rceil$ the smallest integer greater than or equal to a given real number x. For any set $A \subset \Omega$ and any function $f \in L_t$, we also define $\|f_A\|_t = \left(\int_A |f|^t \right)^{1/t}$. **Throughout the paper, we will call "constant" any strictly positive constant depending only on η, α, t and d.** For any two nonnegative functions f, g, we will write $f \lesssim g$ if there exists a constant $C > 0$ such that $f \leq Cg$, where $C = C(\eta, d, t, \alpha)$. We will also write $f \gtrsim g$ if $g \lesssim f$ and $f \asymp g$ if $f \lesssim g$ and $f \gtrsim g$. For any two real numbers a, b we will write $[a \pm b] = [a - b, a + b]$. For any

set $A \subset \Omega$, we will denote by A^c the complement of A in Ω: $A^c = \Omega \setminus A$. Denote by $\mathfrak{B}(\Omega)$ the Borel σ-algebra over Ω. For any two probability distributions P, Q over $(\Omega, \mathfrak{B}(\Omega))$, we will denote by $d_{TV}(P,Q) = \sup_{A \in \mathfrak{B}(\Omega)} |P(A) - Q(A)|$ the total variation distance between P and Q. If P and Q are absolutely continuous with respect to some measure μ over $(\Omega, \mathfrak{B}(\Omega))$ with densities p and q respectively, we will also write

$$d_{TV}(p,q) = \frac{1}{2}\int_\Omega |p-q|\, d\mu = d_{TV}(P,Q).$$

We also denote by $\mathrm{Unif}(A)$ the uniform distribution over any bounded Borel set $A \subset \mathbb{R}^d$.

3.3 Results

We fix $p_0 \in \mathcal{P}(\alpha, L)$ defined in (3.3), along with some constant $\eta \in (0,1)$. We first give an overview of our results. The domain Ω will be split into two parts, namely the *bulk* part, where p_0 takes only large values, and the *tail* part, where p_0 takes only small values. The explicit definitions of $\mathcal{B}(u)$ and $\mathcal{T}(u)$ are given below.

We will analyze separately the bulk and the tail regimes. In each case, we will restrict p_0 to each particular set, and separately establish the minimax separation radii ρ^*_{bulk} and ρ^*_{tail}. Likewise, we will identify the optimal tests ψ^*_{bulk} and ψ^*_{tail} independently on each set. The overall minimax separation radius (3.7) will be given - up to multiplicative constants - by the sum of the two terms: $\rho^* \asymp \rho^*_{bulk} + \rho^*_{tail}$, and the overall optimal test by the combination of the two tests: $\psi^* = \psi^*_{bulk} \vee \psi^*_{tail}$.

3.3.1 Partitioning the domain Ω

Splitting the domain into bulk and tail

It has been well known since [95] that, in the multinomial setting, the local goodness-of-fit problem involves splitting the null distribution into bulk and tail. In our analysis, we also divide Ω into a bulk $\mathcal{B} = \{x \in \Omega \mid p_0(x) \geq u_B\}$ and a tail $\mathcal{T} = \mathcal{B}^c = \{x \in \Omega \mid p_0(x) < u_B\}$ for some value u_B specified later. On the other hand, like in [104], a further key idea is to divide Ω into smaller cubes with possibly varying edge lengths. Each cube will be considered as a single coordinate of a multinomial distribution, allowing us to (approximately) represent our continuous density as a discrete multinomial distribution.

The fundamental idea of our tail definition is to ensure the following condition. Assume the tail has been split into cubes with suitable edge length h_{tail} (specified below). If H_0 holds, then with high probability, none of the tail cubes will contain 2 observations or more. The cut-off u_B is designed to ensure this condition. Before giving its expression, we first introduce:

$$\forall u \geq 0: \quad \mathcal{B}(u) := \{x \in \Omega : p_0(x) \geq u\} \quad \text{and} \quad \mathcal{T}(u) := \{x \in \Omega : p_0(x) < u\} = \mathcal{B}(u)^c. \tag{3.8}$$

For any Borel set $A \subset \Omega$ and any measurable nonnegative function $f : \Omega \longrightarrow \mathbb{R}_+$, we write $f[A] = f(A) = \int_A f$. We now introduce an auxiliary value u_{aux}, used to define the cut-off u_B:

$$u_{aux} := \sup \left\{ u \geq 0 : \frac{p_0^2\left[\mathcal{T}(u)\right]}{\left(p_0\left[\mathcal{T}(u)\right]\right)^{d/(\alpha+d)}} \leq c_I \widetilde{L}^{1/(\alpha+d)} \right\}, \quad \text{where } \widetilde{L} = \frac{L^d}{n^{2\alpha}} \tag{3.9}$$

and c_I is a small enough constant. We will also refer to the following notation throughout the paper:

$$\mathcal{I} := \int_{\mathcal{B}(u_{aux})} p_0^r =: p_0^r\Big[\mathcal{B}(u_{aux})\Big], \quad \text{where } r = \frac{2\alpha t}{(4-t)\alpha + d}. \tag{3.10}$$

We now introduce the value u_B defining our cut-off as

$$u_B = u_{aux} \vee \left[c_A \frac{L^{\frac{d}{4\alpha+d}}}{(n^2\mathcal{I})^{\frac{\alpha}{4\alpha+d}}} \right]^{\frac{(4-t)\alpha+d}{(2-t)\alpha+d}}, \tag{3.11}$$

where c_A is a small enough constant. The constants c_I and c_A can be chosen arbitrarily small, as long as they only depend on η, α, t and d. The value u_B can be understood as the smallest value $u \geq u_{aux}$ such that when $p_0(x) = u$ then $u \geq c_A\, L\, h_b(x)^\alpha$. Therefore, on the bulk, it is easily checked that $\forall x \in \mathcal{B}(u_B) : p_0(x) \geq c_A\, L\, h_b(x)^\alpha$. In the sequel we will write

$$\mathcal{B} = \mathcal{B}(u_B) \quad \text{and} \quad \mathcal{T} = \mathcal{T}(u_B). \tag{3.12}$$

We now state our main theorem:

Theorem 3.1. *Set $\widetilde{L} = L^d/n^{2\alpha}$ and $r = \frac{2\alpha t}{(4-t)\alpha+d}$. There exists a constant $\overline{n} = \overline{n}(d, \eta, t, \alpha)$ independent of p_0 such that, for all $n \geq \overline{n}$, the minimax separation radius associated to problem (3.4) is given by*

$$\rho^* \asymp \rho^*_{bulk} + \rho^*_{tail} + \rho^*_r, \tag{3.13}$$

where

$$\rho^*_{bulk} = \left[\widetilde{L} \left\| p_{0,\mathcal{B}(u_{aux})} \right\|_r^{2\alpha} \right]^{\frac{1}{4\alpha+d}}, \qquad \rho^*_{tail} = \left[\widetilde{L}^{t-1}\, p_0[\mathcal{T}]^{(2-t)\alpha+d} \right]^{\frac{1}{t(\alpha+d)}} \tag{3.14}$$

$$\text{and} \quad \rho^*_r = \left[\frac{L^{d(t-1)}}{n^{\alpha t+d}} \right]^{\frac{1}{t(\alpha+d)}}.$$

In the above Theorem, $p_{0,\mathcal{B}(u_{aux})} = p_0\, \mathbb{1}\{\mathcal{B}(u_{aux})\}$. Note that ρ^*_{bulk} depends on n as $n^{-2\alpha/(4\alpha+d)}$. Note moreover that $\rho^*_{tail} + \rho^*_r \asymp \left[\widetilde{L}^{\frac{t-1}{\alpha+d}} \left(\frac{1}{n} + \int_{\mathcal{T}} p_0 \right)^{\frac{(2-t)\alpha+d}{\alpha+d}} \right]^{1/t}$. The quantity ρ^*_r is a remainder term which is analogous to $\frac{1}{n}$ in discrete testing (see e.g. [132]). The optimal test achieving this rate is given by $\psi^* = \psi^*_{bulk} \vee \psi_1 \vee \psi_2$ where ψ^*_{bulk} is defined in (3.22), ψ_1 in (3.26) and ψ_2 in (3.27). We

now successively study the bulk and the tail regimes individually.

3.4 Bulk regime

In this subsection, we place ourselves on the bulk and analyze separately the upper bound and the lower bound on the minimax separation radius. For the upper bound, we identify a constant $C'_b = C'_b(\eta, \alpha, t, d)$ and a test ψ^*_{bulk} with risk $R(\psi^*_{bulk}, C'_b\rho^*_{bulk}) \leq \eta$. For the lower bound, we build a prior distribution $p_\epsilon^{(n)} \in \mathcal{P}(\alpha, L', c'_\star)$ and identify a constant C^{LB}_{bulk} such that almost surely $\|p_0 - p_\epsilon^{(n)}\|_t \geq C^{LB}_{bulk}\,\rho^*_{bulk}$ and $d_{TV}(p_0^{\otimes n}, \mathbb{E}_\epsilon(p_\epsilon^{(n)})^{\otimes n}) < 1 - \eta$ where $\mathbb{E}_\epsilon$ is the expectation with respect to the prior distribution.

3.4.1 Bulk upper bound

We will construct a test statistic possibly over the enlarged set $\mathcal{B}(\frac{u_B}{2})$ instead of the bulk $\mathcal{B}(u_B)$. For each $x \in \mathcal{B}(\frac{u_B}{2})$, introduce the following bandwidth value depending on $p_0(x)$:

$$h_b(x) = \frac{p_0(x)^{\frac{2}{(4-t)\alpha+d}}}{\left(n^2L^4\mathcal{I}\right)^{\frac{1}{4\alpha+d}}}. \tag{3.15}$$

Throughout the paper, we will say that *the bulk dominates (over the tail)* whenever

$$C_{BT}\rho^*_{bulk} \geq \rho^*_{tail}, \tag{3.16}$$

for some sufficiently large constant C_{BT}. In the converse case, (when $C_{BT}\rho^*_{bulk} < \rho^*_{tail}$), we will say that the *tail dominates*. We recall that "*constant*" denotes any positive real number allowed to depend only on η, t, d and α. To define our bulk test, we distinguish between two cases. We set

$$\widetilde{u_B} = \begin{cases} \frac{u_B}{2} & \text{if the bulk dominates} \\ u_B & \text{if the tail dominates,} \end{cases} \qquad \text{and} \qquad \widetilde{\mathcal{B}} = \mathcal{B}(\widetilde{u_B}). \tag{3.17}$$

Note that over $\widetilde{\mathcal{B}}$, it always holds $p_0 \geq \widetilde{c_A}Lh_b^\alpha$ where $\widetilde{c_A} = 2^{\frac{2\alpha}{(4-t)\alpha+d}-1}c_A$. The bulk upper will be analyzed over $\widetilde{\mathcal{B}}$ rather than over $\mathcal{B}$.

Define a kernel K over $\mathbb{R}^d$ and introduce the usual notation $K_h(x) = \frac{1}{h^d}K(\frac{x}{h})$ for all $h > 0$ and $x \in \mathbb{R}^d$. We choose K such that

- K is of order α, i.e. for any $f \in \mathcal{H}(\alpha, L)$ and $h > 0$: $\|f - K_h(x - \cdot) * f\|_\infty \leq C_K L h^\alpha$.
- K is bounded in absolute value by a constant depending on α and d.
- K is 0 over $\{x \in \mathbb{R}^d : \|x\|_2 > \frac{1}{2}\}$.

In the above definition, we set, for $m = \lceil \alpha \rceil - 1$:

$$C_K = \frac{1}{m!} \int_{\mathbb{R}^d} \|u\|_2^m \, |K(u)| \, du. \tag{3.18}$$

We first split the data $(X_1, \dots, X_{2k})$ in two equal-sized parts $(X_1, \dots, X_k)$ and $(X_{k+1}, \dots, X_n)$. We set $h(x) = c_h h_b(x)$ where $c_h = (\widetilde{c_A}/4)^{\frac{1}{\alpha}}$ and build for each batch an estimator of the true underlying distribution p over $\widetilde{\mathcal{B}}$:

$$\hat{p}(x) = \frac{1}{k} \sum_{i=1}^{k} K_{h(x)}(x - X_i), \qquad \hat{p}'(x) = \frac{1}{k} \sum_{i=k+1}^{2k} K_{h(x)}(x - X_i). \tag{3.19}$$

For all $x \in \widetilde{\mathcal{B}}$, $\hat{p}(x)$ and $\hat{p}'(x)$ are independent random variables. Note moreover the variable bandwidth $h(x)$ depends on x. We propose the following test statistic:

$$T_{bulk} = \int_{\widetilde{\mathcal{B}}} \omega(x) \left[\hat{p}(x) - p_0(x)\right] \left[\hat{p}'(x) - p_0(x)\right] dx, \tag{3.20}$$

$$\text{where} \quad \omega(x) = p_0(x)^{\frac{2\alpha t - 4\alpha}{(4-t)\alpha + d}}. \tag{3.21}$$

We can now define the optimal test on the bulk:

$$\psi^*_{bulk} = \mathbb{1}\Big\{T_{bulk} > C_{\psi_b} t_n\Big\} \quad \text{where } t_n = C_{t_n} \frac{L^{\frac{2d}{4\alpha+d}} \mathcal{I}^{\frac{2\alpha+d}{4\alpha+d}}}{n^{\frac{4\alpha}{4\alpha+d}}}, \tag{3.22}$$

where C_{ψ_b} and $C_{t_n} > 1$ are sufficiently large constants.

The re-weighting $\omega(x)$ is a re-normalizing factor whose role is to balance the expectation and variance of T_{bulk} - in a way that is adapted to, and depends on, the index t of the norm. Note that $\omega(x)$ increases as $p_0(x)$ decreases. Therefore, a large dispersion observed at some $x \in \widetilde{\mathcal{B}}$ for which $p_0(x)$ is small will be amplified by $\omega(x)$ and contribute to a larger increase in T_{bulk}. The threshold t_n corresponds, up to a constant, to the standard deviation of T_{bulk} under H_0.

The following proposition yields an upper bound on the minimax separation radius ρ^*_{bulk} in the bulk regime. Recall that $L' = (1+\delta)L$ and $c'_\star = (1+\delta)c_\star$ where $\delta \in (0,1)$ is a constant.

Proposition 3.1. *For all $\rho > 0$, define $\mathcal{P}_{Bulk}(\rho) = \Big\{p \in \mathcal{P}(\alpha, L', c'_\star) \mid \int_{\widetilde{\mathcal{B}}} |p - p_0|^t \geq \rho^t\Big\}$. There exists a constant $C'_b = C'_b(\eta, \alpha, d, t) > 0$, such that:*

$$\mathbb{P}_{p_0}(\psi^*_{bulk} = 1) + \sup_{p \in \mathcal{P}_{Bulk}(C'_b \rho^*_{bulk})} \mathbb{P}_p(\psi^*_{bulk} = 0) \leq \frac{\eta}{2}.$$

Proposition 3.1 ensures that $C'_b \rho^*_{bulk}$ is an upper bound on the contribution of the bulk to the critical radius. Moreover, it states that in this regime, ψ^*_{bulk} is an optimal test.

3.4.2 Bulk lower bound

Throughout Subsection 3.4.2, we assume that the bulk dominates. In the following sections, we will justify why, *here*, we can make this assumption without loss of generality.

To analyze the bulk lower bound, we split the bulk into small cubes using Algorithm 3. On each cube, we define an undetectable perturbation in a sense specified below. We first define a *bounded* cubic domain $\widetilde{\Omega} \subset \Omega$ containing $\mathcal{B}$. We apply Algorithm 3 from Appendix 3.B with inputs given in Appendix 3.D, which yields a covering of the bulk denoted throughout as $(B_1, \dots, B_N)$, for some $N \in \mathbb{N}$. The cells B_j do not intersect and by construction $\mathcal{B} \subset \cup_j B_j$, while the reverse inclusion may not necessarily hold. Note that our prior is supported on $\cup_j B_j$ rather than on $\mathcal{B}$, which will nonetheless yield the desired rate provided that the bulk dominates.

Following Le Cam's two-point method, the lower bound is obtained by designing a mixture of densities in $\mathcal{P}(\alpha, L', c'_\star)$, indistinguishable from p_0 with risk at most η. By "indistinguishable", we mean that the total variation between n data from p_0 and n data from the mixture is constrained to be $\leq 1-\eta$. The bulk prior is defined in Appendix 3.D. We here give a general idea of its construction. On each cell B_j, we add to p_0 a random perturbation $\epsilon_j \phi_j$ where $(\epsilon_j)_j$ are iid centered Rademacher random variables and where $\phi_j \in H(\alpha, \delta L)$ is deterministic and supported on B_j. Each perturbation ϕ_j is chosen so that $\int_{B_j} \phi_j = 0$ and $p_0 \pm \phi \geq 0$ over B_j. Hence, for all ϵ, $p_\epsilon^{(n)} := p_0 + \sum_{j=1}^N \epsilon_j \phi_j$ is by construction a density and can be shown to belong to $\mathcal{P}(\alpha, L', c'_\star)$. The magnitude of $(\phi_j)_j$ and the edge lengths of $(B_j)_j$ are optimized so as to maximize the L_t discrepancy $\|p_\epsilon^{(n)} - p_0\|_t = \|\sum_j \phi_j\|_t$, subject to $d_{TV}\left(p_\epsilon^{(n)}, p_0^{\otimes n}\right) \leq 1-\eta$. Moreover, the cut-off u_B can be understood as the smallest value (up to multiplicative constant) ensuring the condition $p_0 \pm \phi_j \geq 0$ over $\mathcal{B}(\frac{u_B}{2})$ once the (ϕ_j) have been optimally chosen.

Proposition 3.2. *In the case where the bulk dominates, i.e. when (3.16) holds, there exists a constant C^{LB}_{bulk} such that $\rho^* \geq C^{LB}_{bulk}\, \rho^*_{bulk}$.*

Proposition 3.2 is proved in Appendix 3.D, by showing that $\|\sum_j \phi_j\|_t = C^{LB}_{bulk}\, \rho^*_{bulk}$ and bounding from above by $1-\eta$ the total variation distance between the null distribution and the bulk prior.

3.5 Tail regime

In this section, we place ourselves on the tail and analyse separately the upper bound and the lower bound. For the upper bound, we identify a constant C'' and a test ψ^*_{tail} with risk $R(\psi^*_{tail}, C''\rho^*_{tail}) \leq \eta$. For the lower bound, we build a prior distribution $p_b^{(n)}$ such that $\|p_0 - p_b^{(n)}\|_t \geq C^{LB}_{tail}\, \rho^*_{tail}$ with high probability and $d_{TV}(p_0^{\otimes n}, \mathbb{E}_b[p_b^{(n)}]) < 1-\eta$ where $\mathbb{E}_b$ is the expectation with respect to the prior distribution. We recall that the tail is defined such that, with high probability under H_0,

when split into cubes with suitable edge length h_{tail}, no cube contains more than one observation. Recalling that $\mathcal{T} = \mathcal{T}(u_B)$, define:

$$h_{tail} := \left(n^2 L\, p_0[\mathcal{T}]\right)^{-\frac{1}{\alpha+d}}. \tag{3.23}$$

For both the upper and the lower bound, we define a binning of the tail domain. This is done using the following algorithm. As inputs, the algorithm takes a value $u > 0$ and a length $\tilde{h} > 0$. It (implicitly) defines a grid of cubes $C_{j_1,\dots,j_d} := [j_1\tilde{h}, (j_1+1)\tilde{h}] \times \cdots \times [j_d\tilde{h}, (j_d+1)\tilde{h}]$ for all $(j_1, \dots, j_d) \in \mathbb{Z}^d$. It returns the indices of all such cubes whose intersection with $\mathcal{T}(u)$ is empty (hence indices of cubes to be removed from the tail covering).

Algorithm 1: Tail splitting

1. **Input:** $u, \tilde{h}$.
2. Set $\lambda \in \mathbb{N}$ such that $\mathcal{B}(u) \subset [-\lambda\tilde{h}, \lambda\tilde{h}]^d$. Set $P = \emptyset$.
3. For $(j_1, \dots, j_d) \in ([-\lambda, \lambda] \cap \mathbb{Z})^d$:
   if $\mathcal{T}(u) \cap \left[j_1\tilde{h}, (j_1+1)\tilde{h}\right] \times \cdots \times \left[j_d\tilde{h}, (j_d+1)\tilde{h}\right] = \emptyset$: **then** $P \longleftarrow P \cup \{(j_1, \dots, j_d)\}$.
4. **Return** P.

The tail splitting is defined as follows. We denote by P the output of Algorithm 1 and set $I = \mathbb{Z}^d \setminus P$. Since the bulk is a bounded subset of Ω, P is finite so that I is infinite. Moreover, since the sum $\sum_{j\in I} \int_{\widetilde{C}_j} p_0$ is finite (≤ 1), it is possible to sort the cubes $(C_j)_{j\in I}$ as $(\widetilde{C}_l)_{l\in\mathbb{N}^*}$, while ensuring that $\left(\int_{\widetilde{C}_l} p_0\right)_{l\in\mathbb{N}^*}$ is sorted in decreasing order. Note that $\mathcal{T}(u) \subset \bigcup_{j\in\mathbb{N}^*} \widetilde{C}_j$, but that the reverse inclusion does not necessarily hold. Moreover, for $j \neq l$, $\widetilde{C}_j \cap \widetilde{C}_l$ has Lebesgue measure 0. Therefore almost surely, any observation X_i belongs to at most one of the cubes $(\widetilde{C}_j)_j$.

3.5.1 Tail upper bound

To define our tail test, we distinguish between two cases. In the sequel, C_{BT} denotes a large constant.

- If the tail dominates, i.e. if $C_{BT}\rho^*_{bulk} \leq \rho^*_{tail}$, then we set $(\widetilde{C}_j)_{j\in\mathbb{N}^*}$ to be the covering defined by Algorithm 1 with inputs $\widetilde{u_B} = u_B$ and $\tilde{h} = h_{tail}$.
- If the bulk dominates, i.e. if $C_{BT}\rho^*_{bulk} \geq \rho^*_{tail}$, then we set $(\widetilde{C}_j)_{j\in\mathbb{N}^*}$ to be the covering defined by Algorithm 1 with inputs $\widetilde{u_B} = \frac{u_B}{2}$ and $\tilde{h} = h_m$ where

$$h_m = \frac{c_m}{(n^2L^4\mathcal{I})^{\frac{1}{4\alpha+d}}} \left[c_A \frac{L^{\frac{d}{4\alpha+d}}}{(n^2\mathcal{I})^{\frac{\alpha}{4\alpha+d}}}\right]^{\frac{2}{(2-t)\alpha+d}}, \tag{3.24}$$

and $c_m = \left((\frac{1}{2} - \frac{c_\star}{2}) \frac{c_A}{\sqrt{d}^\alpha} \right)^{\frac{1}{\alpha}}$. To understand why h_m is a natural bandwidth to introduce, set $u_m = \left[c_A \frac{L^d}{(n^2\mathcal{I})^\alpha} \right]^{\frac{1}{4\alpha+d} \frac{(4-t)\alpha+d}{(2-t)\alpha+d}}$ and note that $u_B = u_{aux} \vee u_m$. Observe moreover that u_m is the unique value ensuring that, if for $x \in \Omega$, $p_0(x) = u_m$, then $c_m h_b(x) = h_m$.

The tail test ψ^*_{tail} is defined as a combination of two tests:

- The first test ψ_1 counts the *total number of observations* on the tail, i.e. on the union of the sets $\widetilde{C}_j$, and rejects H_0 when this total mass is substantially different from its expectation under H_0.
- The second test ψ_2 rejects H_0 whenever there exists one cell $\widetilde{C}_j$ containing two observations or more.

For each cube $\widetilde{C}_j$, N_j is defined as the total number of observations on $\widetilde{C}_j$:

$$N_j = \sum_{i=1}^{n} \mathbb{1}\{X_i \in \widetilde{C}_j\}. \tag{3.25}$$

We call $(N_j)_j$ the *histogram* of $(X_i)_i$ on the tail. Note that the family $(\widetilde{C}_j)_j$ is infinite but that there is only a finite number of sets $\widetilde{C}_j$ that contain observations. Thus, the number of values N_j that are nonzero is finite. Recalling that $\mathcal{T} = \mathcal{T}(u_B)$, our tail test ψ^*_{tail} is defined as $\psi^*_{tail} = \psi_1 \vee \psi_2$ where:

$$\psi_1 = \mathbb{1}\left\{ \left| \frac{1}{n} \sum_{j \in \mathbb{N}^*} N_j - p_0[\mathcal{T}] \right| > C_{\psi_1} \sqrt{\frac{p_0[\mathcal{T}]}{n}} \right\}, \tag{3.26}$$

$$\psi_2 = \begin{cases} 1 \text{ if } N_j \geq 2 \text{ for some } j, \\ 0 \text{ otherwise.} \end{cases} \tag{3.27}$$

Here, C_{ψ_1} is a sufficiently large constant. The following proposition yields an upper bound on the minimax separation radius ρ^*_{tail} on the tail.

Proposition 3.3. *For all $\rho > 0$, define $\mathcal{P}_{Tail}(\rho) = \left\{ p \in \mathcal{P}(\alpha, L', c'_\star) \;\middle|\; \int_{\mathcal{T}(\widetilde{u_B})} |p - p_0|^t \geq \rho^t \right\}$. There exists a constant $C'' = C''(\eta, \alpha, d, t) > 0$, such that*

$$\mathbb{P}_{p_0}(\psi^*_{tail} = 1) \;+ \sup_{p \in \mathcal{P}_{Tail}(C''\rho^*)} \mathbb{P}_p(\psi^*_{tail} = 0) \;\leq\; \frac{\eta}{2}.$$

This proposition ensures that $C''\rho^*$ is an upper bound on the minimax separation radius when one restricts to the tail coefficients. Moreover, it states that ψ^*_{tail} is a test reaching this bound. Note that in Proposition 3.3, the separation radius on the tail is $\rho^* \asymp \rho^*_{bulk} + \rho^*_r$ if the bulk dominates, or $\rho^* \asymp \rho^*_{tail} + \rho^*_r$ if the tail dominates.

3.5.2 Tail lower bound

To begin with, we state Proposition 3.4 which handles the case where $\int_{\mathcal{T}} p_0 < \frac{c_{tail}}{n}$ for a large constant c_{tail}.

Proposition 3.4. *There exists a constant n_0 such that whenever $n \geq n_0$ and $\int_{\mathcal{T}} p_0 < \frac{c_{tail}}{n}$, it holds $\rho^* \gtrsim \rho_r^* := L^{\frac{d(t-1)}{t(\alpha+d)}} n^{-\frac{\alpha t+d}{t(\alpha+d)}}$.*

To analyze the tail lower bound, Proposition 3.4 allows us to make the following two assumptions *wlog*:

(a) $C_{BT}\rho_{bulk}^* \leq \rho_{tail}^*$ i.e. the tail dominates,

(b) $\int_{\mathcal{T}} p_0 \geq c_{tail}/n$.

Indeed, for (a), Propositions 3.1 and 3.3 already establish that $\rho_{bulk}^* + \rho_r^*$ is an upper bound over ρ^*. If the bulk dominates, then Propositions 3.2 and 3.4 yield that $\rho_{bulk}^* + \rho_r^*$ is also a lower bound over ρ^*. Therefore, (a) can from now be assumed *wlog*.

As for (b), Propositions 3.1 and 3.3 already establish that $\rho_{tail}^* + \rho_r^*$ is an upper bound over ρ^* when (a) holds. If $\int_{\mathcal{T}} p_0 < \frac{c_{tail}}{n}$, this upper bound further simplifies as $\rho_{tail}^* + \rho_r^* \asymp \rho_r^*$ and Proposition 3.4 yields the matching lower bound $\rho^* \gtrsim \rho_r^*$.

We now define $(\widetilde{C}_j)_{j\in\mathbb{N}^*}$ the covering of $\mathcal{T}(u_B)$ given by Algorithm 1 with inputs $u = u_B$ and $h = c_h h_{tail}(u_B)$ for a small constant c_h. For all $j \in \mathbb{N}^*$, define $p_j = \int_{\widetilde{C}_j} p_0$. We recall that the cells $(\widetilde{C}_j)_j$ are ordered such that the nonnegative real numbers $(p_j)_{j\in\mathbb{N}^*}$ are sorted in decreasing order. Set:

$$U = \min\left\{ j \in \mathbb{N}^* \;\middle|\; n^2 p_j \sum_{l\geq j} p_l \leq c_u \right\}. \tag{3.28}$$

Lemma 46 proves that, when (a) and (b) hold, the union $D(U) := \bigcup_{j\geq U} \widetilde{C}_j$ is not empty and that $S_U = \sum_{j\geq U} p_j > 0$. For $j \geq U$ and for a sufficiently small constant $c_u > 0$, we now set

$$\pi_j = \frac{p_j}{\bar{\pi}} \quad \text{and} \quad \bar{\pi} = \frac{2c_u}{n^2 \sum_{j\geq U} p_j}. \tag{3.29}$$

Index U has no further meaning than to guarantee that $\pi_j \in [0, \frac{1}{2}]$ for all $j \geq U$. In particular, π_j is a Bernoulli parameter.

We now give high-level explanations regarding the construction of the tail prior. To start with, this prior will be supported over $D(U)$ rather than $\mathcal{T}(u_B)$. First, one sparse subset of indices $J_S \subset \{U, \dots, M\}$ is drawn by setting $J_S = \{j : b_j = 1\}$ where for each $j \geq U$, $b_j \sim Ber(\pi_j)$ are independent Bernoulli random variables with parameter π_j. The random elements of J_S represent the indices of the cubes $\widetilde{C}_j$ denoted here as the *selected cubes*. On each selected cube $\left(\widetilde{C}_j\right)_{j\in J_S}$

one large (deterministic) perturbation $\gamma_j^{(\uparrow)} \in H(\alpha, \delta' L)$ is added to p_0. Conversely, on each non-selected cube $\left(\widetilde{C}_j\right)_{j \notin J_S}$, one small perturbation $\gamma_j^{(\downarrow)} \in H(\alpha, \delta' L)$ is removed from p_0. We consider the random function defined by

$$q_b := p_0 + \sum_{j \geq U} \left[b_j \gamma_j^{(\uparrow)} - (1 - b_j) \gamma_j^{(\downarrow)} \right]. \tag{3.30}$$

Since q_b may not necessarily be a probability density, we rescale q_b to define the prior as follows:

$$p_b^{(n)} = \frac{q_b}{\left\| q_b \right\|_1}. \tag{3.31}$$

The definitions of $\gamma_j^{(\downarrow)}$ and $\gamma_j^{(\uparrow)}$ are given in Equations (3.120) and (3.121) in Appendix 3.F. We show in Proposition 3.13 that with high probability, our prior satisfies $\|p_0 - p_b\|_t \geq C_{tail}^{LB} \rho_{tail}^*$ for a constant C_{tail}^{LB}. The following proposition yields a lower bound in the tail regime.

Proposition 3.5. *If the tail dominates, i.e.* $\rho_{tail}^* \geq C_{BT} \rho_{bulk}^*$, *and if* $\int_{\mathcal{T}(u_B)} p_0 \geq \frac{c_{tail}}{n}$, *there exists a constant* C_{tail}^{LB} *such that* $\rho^* \geq C_{tail}^{LB} \rho_{tail}^*$.

Proposition 3.5 is a corollary of Proposition 3.14 proved in Appendix 3.F.

3.6 Discussion

3.6.1 Discussion of the results

Rates

For n larger than a constant n_0, we prove matching upper and lower bounds leading to the following expression for the critical radius

$$\rho^*(p_0, \alpha, L, n) \asymp \widetilde{L}^{\frac{1}{4\alpha+d}} \left(\int_{\mathcal{B}(u_{aux})} p_0^r \right)^{\frac{(4-t)\alpha+d}{t(4\alpha+d)}} + \widetilde{L}^{\frac{t-1}{\alpha+d}} \left(\frac{1}{n} + p_0\Big[\mathcal{T}(u_B)\Big] \right)^{\frac{(2-t)\alpha+d}{t(\alpha+d)}},$$

where $r = \frac{2\alpha t}{4\alpha+d}$ and $\widetilde{L} = L^d / n^{2\alpha}$. The bulk term involves the quantity $n^{-2\alpha/(4\alpha+d)}$ which is the classical non-parametric rate for testing the null hypothesis of the uniform distribution on $[0,1]^d$ against the alternative composed of (α, L)-Hölder densities separated from p_0 in $\|\cdot\|_t$ norm (see e.g. [183]). Recall that this rate is faster than the non-parametric rate of estimation $n^{-\frac{\alpha}{2\alpha+d}}$. We observe that both u_B and u_{aux} decrease as n increases. If a fixed p_0 is supported on a fixed bounded domain, then the bulk eventually dominates for n larger than a critical value (depending on p_0). The asymptotic rate therefore simplifies as $\rho^*(p_0, n, \alpha, L) \asymp \widetilde{L}^{\frac{1}{4\alpha+d}} \left(\int_\Omega p_0^r \right)^{\frac{(4-t)+d}{t(4\alpha+d)}}$ which decays with n at the non-parametric rate of testing $n^{-\frac{2\alpha}{4\alpha+d}}$. However, when Ω is not bounded, we give in Subsection 3.6.2 examples of fixed null densities p_0 for which the tail always dominates, leading to

critical radii $\rho^*(p_0, n, \alpha, L)$ decaying with n at slower rates than $n^{-\frac{2\alpha}{4\alpha+d}}$.

In the tail test statistic, we combine the tests ψ_1 and ψ_2 from (3.26) and (3.27). Test ψ_1 compares the first order moment of p with that of p_0. Test ψ_2 implicitly checks that the second moment of p is no larger than that of p_0. Indeed, on the tail, the second moment of p_0 is so small that, *whp*, any cell $\widetilde{C}_j$ contains at most one observation under H_0. Conversely, if the second moment of p is substantially larger than that of p_0, then *whp* one of the cells will contain at least two observations.

Cut-offs

Here is some intuition on why in some cases we consider $\mathcal{B}(\frac{u_B}{2})$ or $\mathcal{T}(\frac{u_B}{2})$ instead of $\mathcal{B}(u_B)$ and $\mathcal{T}(u_B)$. Indeed, it is not always possible to set the optimal bulk prior (3.97) over $\mathcal{B}(u_{aux})$. This leads us to introduce the smallest cut-off $u_B \geq u_{aux}$ ensuring that (3.97) can be supported on $\mathcal{B}(u_B)$. A nice property of u_B (see Lemma 24) is that if the tail dominates, then after splitting $\mathcal{T}(u_B)$ into cubes with edge length h_{tail}, any tail cube should contain either zero or one observation *whp* under H_0. Unfortunately, this condition no longer holds if the bulk dominates, which is the reason why we split $\mathcal{T}(u_B/2)$ into cubes of edge length $\asymp h_m \leq \min_{\mathcal{B}(u_B)} h_b$ in this case.

Discussion on the regularity conditions - Assumption ($\star$)

Our results constitute an attempt to address the case of arbitrary α-Hölder densities over $\mathbb{R}^d$. Our analysis relies on Assumption ($\star$). For $\alpha \leq 1$, Assumption ($\star$) is automatically satisfied and does not affect our result's generality. For $\alpha > 1$, Assumption ($\star$) essentially implies two limitations.

- **Limitation for two points which are close:** First, any p satisfying ($\star$) should be "approximately constant" over the balls $B(x, h(x))$ - namely the Euclidean balls centered at x with radius $h(x) \asymp (p(x)/L)^{1/\alpha}$. Formally, for any $y \in B(x, h(x))$ and for all $c \in [\frac{1}{2}, 1)$, it imposes $\frac{p(y)}{p(x)} \in \left[1 \pm c\right]$ whenever $y \in B(x, h(x))$ where $h(x) = \left(\frac{c-c_\star}{L} p(x)\right)^{1/\alpha}$. Noting that the bulk precisely consists of all x such that $Ch_b(x) \leq h(x)$ for some $C > 0$, this condition allows us to exclude fast variations of p and p_0 over the bulk.
- **Limitation for two points which are far:** Second, when $y \notin B(x, h(x))$, Assumption ($\star$) bounds the maximum deviations of p as $|p(x) - p(y)| \lesssim L\|x - y\|^\alpha$. This condition naturally arises for x corresponding to small values of $p(x)$. In particular, it allows us to exclude fast variations of p and p_0 over the tail.

This assumption is therefore implied by - and in fact, up to multiplicative constants, equivalent to - assuming that for any m, M such that $0 \leq 2(1 + c_\star)m < M$ and such that the level sets $\{p \leq m\}$ and $\{p \geq M\}$ are not empty, the smallest distance between any two points in these level sets should be at least $(M/L)^{1/\alpha}$. This is implied by assuming that whenever p has a local minimum at a point x where $p(x)$ is close to 0, then all its derivatives up to order $\lfloor \alpha \rfloor$ are null in this point. Therefore Assumption ($\star$) is not very restrictive - e.g. it is always satisfied for unimodal densities, or any densities that are monotone outside of a fixed compact, such that the ratio of the upper and

lower bound of p on this compact is bounded by a constant. Whether or not Assumption ($\star$) can be removed remains an open question.

Influence of the norm

We cover the scale of all L_t distances for $t \in [1, 2]$. Among these distances, only the L_1 distance is an f-divergence. We also identify a duality between the norms: When testing in L_t norm, the bulk radius is expressed in terms of the L_r norm, where r and t are linked through the relation $r = \frac{2\alpha t}{(4-t)\alpha+d}$. Depending on the value of r, the hardest and easiest null p_0 to test are different. If $r < 1$, then $\|p_0\|_r$ can be made arbitrarily large if the density p_0 is sufficiently small everywhere on $\mathbb{R}^d$. Conversely, $\|p_0\|_r$ is minimal for spiked p_0 and so is $\rho^*_{bulk}(p_0)$ (see the example of the spiky null in Subsection 3.6.2). This hierarchy is reversed when $r \geq 1$.

3.6.2 Examples

To illustrate our results, we give examples of null densities p_0 and of the associated radii $\rho^*(p_0, n, \alpha, L, t)$.

Example 1 (Uniform null distribution over $\Lambda = [0, \lambda]^d$).

We consider $p_0 = \lambda^{-d}$ over $[0, \lambda]^d$ and set $|\Lambda| = \lambda^d$. Note that this example cannot be handled by our present results since p_0 is not defined over $\mathbb{R}^d$. However, this case has already been analyzed in [183], which we give here for a comparison with our results. The asymptotic minimax rate (as $n \to \infty$) writes:

$$\rho^*(\alpha, L, n, p_0) \asymp \widetilde{L}^{\frac{1}{4\alpha+d}} |\Lambda|^{\frac{(4-3t)\alpha+d}{(4\alpha+d)t}}, \quad \text{where } \widetilde{L} = L^d/n^{2\alpha}. \tag{3.32}$$

This is the most commonly studied setting in the literature [183], [23], [17], [16], [179]. Indeed, for fixed constants $C > c > 0$, and any smooth density p_0 satisfying $c \leq p_0 \leq C$, its critical radius is given by (3.32). However, when we relieve this last assumption and let L be arbitrary, $\rho^*(\alpha, L, n, p_0)$ can substantially deviate from (3.32).

Example 2 (Gaussian null)

Suppose p_0 is the density of $\mathcal{N}(0, \sigma^2 I_d)$ over $\mathbb{R}^d$, where $\sigma > 0$. Fix α, L and σ, and consider the asymptotics as $n \to +\infty$. The *asymptotic* minimax rate associated to p_0 is

$$\rho^*(\alpha, L, n, p_0) \asymp \widetilde{L}^{\frac{1}{4\alpha+d}} (\sigma^d)^{\frac{(4-3t)\alpha+d}{t(4\alpha+d)}}. \tag{3.33}$$

This asymptotic rate exclusively corresponds to the bulk rate ρ^*_{bulk}. It decays with n at the classical non-parametric rate of testing $n^{-\frac{2\alpha}{4\alpha+d}}$. Note the similarity between (3.32) and (3.33) when σ^d plays the role of $|\Lambda|$. Regardless of the fixed constant σ, testing equality to $\mathcal{N}(0, \sigma^2 I_d)$ or to $\mathrm{Unif}\left([-\sigma, \sigma]^d\right)$ are asymptotically equally difficult.

Example 3 (Arbitrary p_0 with support over $\Omega' = [-1, 1]^d$ and $L = 1$.)

We recall that the support of $p : \Omega \longrightarrow \mathbb{R}$ is $\{x \in \Omega : p(x) \neq 0\}$. In this example, we therefore consider an arbitrary null density p_0 defined over $\Omega = \mathbb{R}^d$ which is zero outside $\Omega' = [-1, 1]^d$. For any such $p_0 \in \mathcal{P}(\alpha, 1, c_\star)$, the rate simplifies as:

$$\rho^*(\alpha, 1, n, p_0) \asymp n^{-\frac{2\alpha}{4\alpha+d}}. \tag{3.34}$$

Noticeably, this rate is independent of p_0 and coincides with (3.32). Indeed, fixing the support $\Omega' = [-1, 1]^d$ and $L = 1$ constrains p_0 to have only limited variations. In this case, all its mass cannot be concentrated on a small part of the domain. This example illustrates that, for fixed *bounded* support and for $L = 1$, all of the testing problems (3.4) are equally difficult regardless of p_0. Conversely, when L or Ω' are allowed to depend on n or when the support Ω' is unbounded, the local rate $\rho^*(\alpha, L, n, p_0)$ can significantly deviate from (3.32). This is illustrated in the following examples.

Example 4 (Spiky null).

Let $\Omega = \mathbb{R}^d$. Define $f \geq 0$ such that $f \in H(\alpha, 1) \cap C^\infty$ over $\mathbb{R}^d$ and f is nonzero only over $\left\{x \in \mathbb{R}^d : \|x\| < 1/2\right\}$. We here moreover assume that f satisfies

$$\forall x, y \in \mathbb{R}^d : |f(x) - f(y)| \leq c_\star f(x) + \|x - y\|^\alpha.$$

The spiky null density is defined as follows:

$$p_0(x) = La^\alpha f\left(\frac{x}{a}\right) \tag{3.35}$$

where $a = (\|f\|_1 L)^{-\frac{1}{\alpha+d}}$. Informally, this corresponds to an approximation of the Dirac distribution δ_0 by a density in $\mathcal{P}(\alpha, L, c_\star)$. In this case we have:

$$\rho^*(p_0, \alpha, L, n) \asymp L^{\frac{d(t-1)}{t(\alpha+d)}} n^{-\frac{2\alpha}{4\alpha+d}}. \tag{3.36}$$

Note that the density (3.35) is supported on $[-a, a]^d$ and the corresponding minimax rate is the same as for the uniform density over $[-a, a]^d$. Note that there is only one regime: Indeed, by the choice of a, the value $|\Omega| \asymp L^{-d(\alpha+d)}$ is always smaller than $\widetilde{L}^{-\frac{1}{\alpha+d}}$ up to constants. Assume now that $L \to \infty$ as $n \to \infty$, so that p_0 is supported over $\widetilde{\Omega} = [-1, 1]^d$ for n large enough. Suppose moreover that $\widetilde{L} \to 0$ so that, over $\widetilde{\Omega}$, the rate (3.32) simplifies as $\widetilde{L}^{\frac{1}{4\alpha+d}}$. We then note that (3.36) is faster than $\widetilde{L}^{\frac{1}{4\alpha+d}}$ if, and only if, $r \leq 1$. It is possible to show that over a bounded domain, for all L_t norms such that $r \leq 1$, the uniform null distribution has maximum separation radius whereas the spiky null has the smallest one. Conversely for t such that $r > 1$, the uniform distribution has minimum separation radius whereas the spiky null has the largest one.

Note that by rescaling (see Proposition 3.15), letting $L \to +\infty$ for fixed n and fixed support is equivalent to letting the support size go to infinity while L, n are fixed. This example illustrates that over growing domains, some compactly supported densities can substantially deviate from the

uniform distribution over their support.

Example 5 (Pareto null)

We place ourselves over $\mathbb{R}$ (hence $d = 1$) and consider for $x_0 > 0$ and $\beta \in (0, 1)$ the null density $p_0(x) = \frac{\beta x_0^\beta}{x^{\beta+1}}$ over $[x_1, +\infty)$. Here $x_1 > x_0$ is chosen so that p_0 can be extended over $(-\infty, x_1]$ to get a density in $\mathcal{P}(\alpha, L, c_\star)$ which is 0 over $(-\infty, x_{-1}]$ for $x_{-1} < x_0$. For simplicity we only give the rate for $\alpha \leq 1$ and $t = 1$ i.e. for the total variation distance, although the general rate can be established for all $\alpha > 0$ and $t \in [1, 2]$. The minimax separation radius simplifies as

$$\rho^*(\alpha, L, n, p_0) \asymp \widetilde{L}^{\frac{\beta}{3\beta+1}} = \left(\frac{L^d}{n^{2\alpha}}\right)^{\frac{\beta}{3\beta+\alpha+1}}. \tag{3.37}$$

Interestingly, *for all* L, n, this rate exclusively corresponds to the dominating term ρ^*_{tail}. This example illustrates that for some heavy-tailed densities defined on unbounded domains, the separation distance substantially deteriorates compared to the spiky null with $L = 1$ - and justifies the importance of establishing a tight rate in the tail regime. Noticeably, the whole scale of rates from 1 to $n^{-\frac{2\alpha}{4+\alpha}}$ can be obtained for $\alpha \leq 1$ and are all slower than the classical non-parametric rate of testing $n^{-\frac{2\alpha}{4\alpha+1}}$. For $t > 1$, a slower rate (depending on α, β, t) can similarly be observed as compared to the case of a spiky null density.

3.6.3 Comparison with prior work

Special case of the $\|\cdot\|_1$ norm (total variation)

The case of the L_1 norm has been studied in [104]. Here we state their main result for density testing. Suppose that we observe $X_1, \ldots, X_n$ with density p over Ω and fix a particular density p_0 over Ω. Assume that p and p_0 are L-Lipschitz and consider the identity testing problem

$$H_0 : p = p_0 \quad \text{vs} \quad H_1 : \|p - p_0\|_1 \geq \rho \text{ and } p \text{ is } L\text{-Lipschitz}. \tag{3.38}$$

Problem (3.38) is a special case of our setting where $\alpha = 1$ and $t = 1$. For all $\sigma > 0$, let $\mathcal{B}_\sigma = \{B : \mathbb{P}_{p_0}(B) \geq 1 - \sigma\}$ and define the functional

$$T_\sigma(p_0) = \inf_{B \in \mathcal{B}_\sigma} \left(\int_B p_0^\gamma\right)^{1/\gamma}, \tag{3.39}$$

where $\gamma = \frac{2}{3+d}$. For two explicit constants $c, C > 0$, define the upper and lower critical radii as the solutions of the fixed-point equations:

$$v_n(p_0) = \left(\frac{L^{d/2} T_{C v_n(p_0)}(p_0)}{n}\right)^{\frac{2}{4+d}} \quad \text{and} \quad w_n(p_0) = \left(\frac{L^{d/2} T_{c w_n(p_0)}(p_0)}{n}\right)^{\frac{2}{4+d}}. \tag{3.40}$$

Theorem 3.2. *[Balakrishnan, Wasserman (2017)] There exist two constants $c, C > 0$ such that the critical radius for problem (3.38) satisfies $\rho^*(p_0, 1, L, n) \leq C w_n(p_0)$. Moreover, if p_0 is $c'L$-Lipschitz where $c' \in (0,1)$, then it holds $\rho^*(p_0, 1, L, n) \geq c v_n(p_0)$.*

We now state our Theorem 3.1 in the special case $\alpha = 1$ and $t = 1$. For all $L > 0$, for all cubic domain Ω, for all p_0 L-Lipschitz over Ω, we have:

$$\rho^*(p_0, L, n) \asymp \left(\frac{L^{d/2}}{n} \left(\int_{\mathcal{B}(u_{aux})} p_0^r \right)^{1/r} \right)^{\frac{2}{4+d}} + p_0\Big[\mathcal{T}(u_B)\Big] + \frac{1}{n}, \quad \text{where } r = \frac{2}{3+d} = \gamma. \tag{3.41}$$

We first note that our bulk term $\left(\frac{L^{\frac{d}{2}}}{n} \left(\int_{\mathcal{B}(u_{aux})} p_0^r \right)^{1/r} \right)^{\frac{2}{4+d}}$ is the analog of $v_n(p_0)$ and $w_n(p_0)$ in Theorem 3.2. However, it is defined explicitly in terms of p_0 and does not involve solving a fixed-point equation. In Theorem 3.2 the critical radius is bracketed between $v_n(p_0)$ and $w_n(p_0)$. Although these two quantities are of the same order in most usual cases, the authors in [104] discuss pathological cases for which $w_n(p_0) \ll v_n(p_0)$. This non-tightness can be attributed to possibly large discrepancies between $T_{\sigma_1}(p_0)$ and $T_{\sigma_2}(p_0)$ (with $\sigma_1 \neq \sigma_2$) for some carefully chosen p_0. In the present paper, we bridge this gap by identifying matching upper and lower bounds in the considered class and also in more general classes corresponding to any $\alpha > 0$.

In the case of separation in $\|\cdot\|_1$, we identify a tail contribution given by $\rho^*_{tail} \asymp \int_{\mathcal{T}(u_B)} p_0$. As $\Omega = \mathbb{R}^d$, this allows us to pick p_0 depending on L and n so that $\rho^*_{tail} \asymp 1$. Indeed, for fixed n, L consider a suitably smooth density p_0 such that $\max_\Omega p_0 \leq c_I \widetilde{L}^{\frac{1}{\alpha+d}}$. Then by the definition of u_{aux} from equation (3.9), $\int_{\mathcal{T}(u_B)} p_0 \geq \int_{\mathcal{T}(u_{aux})} p_0 = 1$. This illustrates that even in the most favorable regime $L \to 0$ and $n \to \infty$, there exist smooth null densities p_0 over Ω associated to the trivial maximal separation radius $\rho^*(p_0, n, \alpha, L) \asymp 1$. These correspond to the worst case densities over the class. On unbounded domains, it is therefore crucial to identify local results since the global problem (i.e. the worst case over the class) has a trivial rate. The fact that *estimation* of α-Hölder densities in total variation over unbounded domains has a trivial rate was already highlighted in [15], [39], [69]. Recalling that estimation is more difficult than testing, we here recover this result.

Comparison with the discrete setting [132]

The paper [132] considers a discrete analog of the present problem. Suppose we observe iid $X_1, \ldots, X_n$ distributed as $\mathcal{M}(p)$ where $\mathcal{M}(p)$ denotes the multinomial distribution over $\{1, \ldots, d\}$. When $X \sim \mathcal{M}(p)$, we have $\forall j \in \{1, \ldots, d\} : \mathbb{P}(X = j) = p(j)$. Suppose we are given a known discrete distribution p_0 over $\{1, \ldots, d\}$ and assume *wlog* that the entries of p_0 are sorted in decreasin order. We consider the following testing problem:

$$H_0 : p = p_0 \quad \text{vs} \quad H_1 : \|p - p_0\|_t \geq \rho, \tag{3.42}$$

where $\|p-p_0\|_t = \Big(\sum_{j=2}^{d} \big|p(j)-p_0(j)\big|^t\Big)^{1/t}$ and $t \in [1,2]$. Introduce the index I as

$$I = \min\Big\{ j \in \{1,\dots,d\} \;\Big|\; \sum_{j>I} n^2 p^2(j) \le c'_I \Big\}, \tag{3.43}$$

for some constant $c'_I = c'_I(\eta, t)$. Moreover, define the index A as:

$$A = \min\Big\{ j \le I \;\Big|\; p^{b/2}(j) \ge \frac{c'_A}{\sqrt{n}\Big(\sum\limits_{j\le I} p_0^{r'}(j)\Big)^{1/4}} \Big\} \tag{3.44}$$

where $b = \frac{4-2t}{4-t}$ and $r' = \frac{2t}{4-t}$. The following theorem is valid:

Theorem 3.3. *[Chhor, Carpentier (2020)] It holds:*

$$\rho^*(p_0, n, d, t) \;\asymp\; \rho^{*\mathcal{M}}_{bulk} \;+\; \rho^{*\mathcal{M}}_{tail} \;+\; \rho^{*\mathcal{M}}_{remain},$$

where $\rho^{*\mathcal{M}}_{bulk} = \sqrt{\frac{1}{n}\Big\|(p_0)^{-\max}_{\le I}\Big\|_{r'}}$, $\rho^{*\mathcal{M}}_{tail} = n^{\frac{2}{t}-2}\|(p_0)_{>A}\|_1^{1-2/t}$, $\rho^{*\mathcal{M}}_{remain} = 1/n$,

$\Big\|(p_0)^{-\max}_{\le I}\Big\|_{r'} = \Big(\sum\limits_{j=2}^{I} p_0^{r'}(j)\Big)^{1/r'}$ *and* $\|(p_0)_{>A}\|_1 = \sum\limits_{j>A} p_0(j)$.

Upper and lower bounds for multinomial identity testing were previously known (see [95] for the L_1 norm), but did not match in some specific cases. The above theorem provides a new way to define the tail, leading to the matching upper and lower bounds for all L_t norms for $t \in [1,2]$, which were missing in [95].

This discrete setting and our present continuous setting involve many similar phenomena. The discrete tail, defined as $\{p_0(j) \mid j > A\}$, is designed so that *whp* under H_0, no coordinate $j > A$ is observed twice among the n data $X_1, \dots, X_n$. Our approach in the present paper aims at transferring this tail definition to the continuous setting.

Theorem 3.3 identifies a three-fold contribution to ρ^* (bulk, tail and remainder term) which is similar to ours. However, there are some substantial challenges in our continuous setting compared to the discrete testing problem.

- First, the discretization that we adopt (for both the lower bounds and for the tail statistic), as well as the bandwidth of the kernel (for the bulk statistic) must depend on p_0. Finding this optimal discretization/bandwith is challenging in itself, and raises some fundamental information theoretic questions, as well as some difficult technical issues. For instance, the bulk test statistic is based on the integral of a functions of a in-homogeneous kernel approximation of p, which is very different from what is done in the discrete setting.
- Second, even when this discretization/kernelisation has been done, the test statistics are not direct analogs of the discrete test statistics from [132]. In both cases - discrete and continuous -

the bulk test statistics is a reweighted χ^2 test statistic with inhomogeneous weights, depending on each coordinate of p_0. However in the continuous case, the reweighting factor $\omega(x)$ cannot be directly deduced from the discrete setting. Indeed, there is a distortion in the integral coming from the non-homogeneity of the Kernel bandwidth, whose effect has to be taken into account on top of the non-homegeneity coming from $\omega(x)$.

Appendix

3.A Relations between the cut-offs

We will also use the following notation:

$$\widetilde{\mathcal{T}}(u_{aux}) = \{x \in \Omega : p_0(x) \leq u_{aux}\}. \tag{3.45}$$

Note that in the above definition, the inequality $p_0(x) \leq u_{aux}$ is not strict, whereas the inequalities in the definitions of $\mathcal{T}(u_{aux})$ and $\mathcal{T}(u_B)$ are strict. Furthermore, we define three different lengths $h_{tail}(u_{aux}), h_{tail}(u_B)$ and $\bar{h}_{tail}$ as follows:

$$h_{tail}(u_{aux}) = \left(n^2 L \int_{\mathcal{T}(u_{aux})} p_0\right)^{-\frac{1}{\alpha+d}}. \tag{3.46}$$

$$h_{tail}(u_B) = \left(n^2 L \int_{\mathcal{T}(u_B)} p_0\right)^{-\frac{1}{\alpha+d}}. \tag{3.47}$$

$$\bar{h}_{tail}(u_{aux}) = \left(n^2 L \int_{\widetilde{\mathcal{T}}(u_{aux})} p_0\right)^{-\frac{1}{\alpha+d}}. \tag{3.48}$$

and prove in Subsection 3.E.4 that they all differ at most by a multiplicative constant.

Lemma 1. *Let $\widetilde{\mathcal{T}}(u_{aux})$ be defined as in (3.45).*

- *We have*
$$\int_{\mathcal{T}(u_{aux})} p_0^2 \leq \frac{c_I}{n^2 h_{tail}^d(u_{aux})}.$$
- *Moreover, if $\max_{\Omega} p_0 \geq u_{aux}$, then it holds:*
$$\frac{\int_{\widetilde{\mathcal{T}}(u_{aux})} p_0^2}{\left(\int_{\widetilde{\mathcal{T}}(u_{aux})} p_0\right)^{d/(\alpha+d)}} \geq c_I \frac{L^{d/(\alpha+d)}}{n^{2\alpha/(\alpha+d)}},$$
implying:
$$\int_{\widetilde{\mathcal{T}}(u_{aux})} p_0^2 \geq \frac{c_I}{n^2 \bar{h}_{tail}^d}.$$

Proof of Lemma 1. By definition of u_{aux}, there exists a sequence $(u_j)_{j\in\mathbb{N}}$ such that $u_j \uparrow u_{aux}$ and $\frac{\int_{\mathcal{T}(u_j)} p_0^2}{\left(\int_{\mathcal{T}(u_j)} p_0\right)^{d/(\alpha+d)}} \leq c_I \widetilde{L}^{\frac{1}{\alpha+d}}$. The dominated convergence theorem yields the result.

For the second part, when $u \downarrow u_{aux}$, we have by the dominated convergence theorem that $\int_{\mathcal{T}(u)} p_0^2 \to \int_{\widetilde{\mathcal{T}}(u_{aux})} p_0^2$ and $\int_{\mathcal{T}(u)} p_0 \to \int_{\widetilde{\mathcal{T}}(u_{aux})} p_0$ so that $\dfrac{\int_{\mathcal{T}(u)} p_0^2}{\left(\int_{\mathcal{T}(u)} p_0\right)^{d/(\alpha+d)}} \longrightarrow \dfrac{\int_{\widetilde{\mathcal{T}}(u_{aux})} p_0^2}{\left(\int_{\widetilde{\mathcal{T}}(u_{aux})} p_0\right)^{d/(\alpha+d)}}$. Moreover, by definition of u_{aux}, for all $u > u_{aux}$ we have $\dfrac{\int_{\mathcal{T}(u)} p_0^2}{\left(\int_{\mathcal{T}(u)} p_0\right)^{d/(\alpha+d)}} > c_I \dfrac{L^{d/(\alpha+d)}}{n^{2\alpha/(\alpha+d)}}$ (since $u_{aux} \leq \max_{\Omega} p_0$), which yields the result. □

We now define

$$\overline{\rho^*_{tail}} = \left[\widetilde{L}^{\frac{t-1}{\alpha+d}} \frac{\left(\int_{\mathcal{T}(u_B)} p_0\right)^{\frac{(2-t)\alpha+d}{\alpha}}}{\left(\int_{\widetilde{\mathcal{T}}(u_{aux})} p_0\right)^{\frac{(2-t)\alpha+d}{\alpha}\frac{d}{\alpha+d}}} \right]^{1/t}. \tag{3.49}$$

Lemma 2. *If $u_B > u_{aux}$ and $\max_{\Omega} p_0 \geq u_{aux}$ then $\overline{\rho^*_{tail}} \geq C_2\, \rho^*_{bulk}$ where $C_2 = c_I^{\frac{(2-t)\alpha+d}{\alpha t}}\, c_A^{-\frac{(4-t)\alpha+d}{\alpha t}}$.*

Proof of Lemma 2. If $u_B > u_{aux}$ then by definition of u_B given in (3.11), we have

$$u_B = \left[c_A \frac{L^{\frac{d}{4\alpha+d}}}{(n^2\mathcal{I})^{\frac{\alpha}{4\alpha+d}}} \right]^{\frac{(4-t)\alpha+d}{(2-t)\alpha+d}}.$$

By Lemma 1, we have:

$$u_B \int_{\widetilde{\mathcal{T}}(u_{aux})} p_0 \geq \int_{\widetilde{\mathcal{T}}(u_{aux})} p_0^2 \geq \frac{c_I}{n^2 \bar{h}_{tail}^d} = c_I \left[\widetilde{L} \left(\int_{\widetilde{\mathcal{T}}(u_{aux})} p_0 \right)^d \right]^{\frac{1}{\alpha+d}}. \tag{3.50}$$

Therefore:

$$u_B \int_{\mathcal{T}(u_B)} p_0 \geq c_I \left[\widetilde{L} \left(\int_{\widetilde{\mathcal{T}}(u_{aux})} p_0 \right)^d \right]^{\frac{1}{\alpha+d}}.$$

Raising this relation to the power $\frac{(2-t)\alpha+d}{\alpha t}$ and recalling the expressions of ρ^*_{bulk} and $\overline{\rho^*_{tail}}$, we get $\overline{\rho^*_{tail}} \geq C_2\, \rho^*_{bulk}$. □

Lemma 3. *Whenever $u_B > u_{aux}$ and $\max_{\Omega} p_0 \geq u_{aux}$, we have*

$$\mathcal{I} u_B^{2-r} \leq \frac{C_3}{n^2} \frac{\bar{h}_{tail}^{d^2/\alpha}}{h_{tail}^{d(\alpha+d)/\alpha}},$$

where the constant $C_3 = c_A^{2\frac{(4-2t)\alpha+d}{(2-t)\alpha+d}}\, C_2^{-\frac{td}{(2-t)\alpha+d}}$ can be made arbitrarily small by taking c_A small enough.

Proof of Lemma 3. We have $2-r = 2\frac{(4-2t)\alpha+d}{(4-t)\alpha+d}$, so that

$$\mathcal{I}u_B^{2-r} = c_A^{2\frac{(4-2t)\alpha+d}{(2-t)\alpha+d}} \widetilde{L}^{\frac{2}{4\alpha+d}\frac{(4-2t)\alpha+d}{(2-t)\alpha+d}} \mathcal{I}^{\frac{d}{(2-t)\alpha+d}\frac{(4-t)\alpha+d}{4\alpha+d}}. \tag{3.51}$$

On the other hand, by Lemma 1:

$$\int_{\widetilde{\mathcal{T}}(u_{aux})} p_0^2 \geq \frac{c_I}{n^2\bar{h}_{tail}^d} = c_I\left[\widetilde{L}\Big(\int_{\widetilde{\mathcal{T}}(u_{aux})} p_0\Big)^d\right]^{\frac{1}{\alpha+d}}. \tag{3.52}$$

Moreover, when $u_B > u_{aux}$, we have by Lemma 2, $C_2\rho^*_{bulk} \leq \overline{\rho^*_{tail}}(u_{aux})$. We now raise this relation to the power $\frac{td}{(2-t)\alpha+d}$:

$$\widetilde{L}^{\frac{2}{4\alpha+d}\frac{(4-2t)\alpha+d}{(2-t)\alpha+d}} \mathcal{I}^{\frac{d}{(2-t)\alpha+d}\frac{(4-t)\alpha+d}{4\alpha+d}} \leq \frac{\widetilde{L}^{\frac{1}{\alpha+d}}}{C_2^{\frac{td}{(2-t)\alpha+d}}} \frac{\left(\int_{\mathcal{T}(u_B)} p_0\right)^{\frac{d}{\alpha}}}{\left(\int_{\widetilde{\mathcal{T}}(u_{aux})} p_0\right)^{\frac{d}{\alpha}\frac{d}{\alpha+d}}} = \frac{1}{C_2^{\frac{td}{(2-t)\alpha+d}}} \frac{1}{n^2} \frac{\bar{h}_{tail}^{d^2/\alpha}}{h_{tail}^{d(\alpha+d)/\alpha}}. \tag{3.53}$$

Equations (3.51) and (3.53) yield the result. □

Lemma 4. *Set $\bar{C} = c_I + C_3$. The constant $\bar{C}$ can be made arbitrarily small by choosing successively c_I and c_A small enough. This can be done by taking $c_I^{\alpha+d} = c_A^{4\alpha+d}$. If $u_B > u_{aux}$ and $\max_\Omega p_0 \geq u_{aux}$, then it holds:*

$$\int_{\mathcal{T}(u_B)} p_0^2 \leq \frac{\bar{C}}{n^2\overline{h_{tail}}^d}.$$

Proof of lemma 4. Suppose that $u_B > u_{aux}$.

$$\begin{aligned}
\int_{\mathcal{T}(u_B)} p_0^2 &= \int_{\mathcal{T}(u_{aux})} p_0^2 + \int_{\mathcal{B}(u_{aux})\cap\mathcal{T}(u_B)} p_0^2 \leq \frac{c_I}{n^2 h_{tail}^2(u_{aux})} + \mathcal{I}u_B^{2-r} \\
&\leq \frac{c_I}{n^2 h_{tail}^2(u_{aux})} + \frac{C_3}{n^2}\frac{\bar{h}_{tail}^{d^2/\alpha}}{h_{tail}^{d(\alpha+d)/\alpha}(u_{aux})} \text{ by Lemma 3} \qquad (3.54)\\
&\leq \bar{C}\frac{1}{n^2\bar{h}_{tail}^d}. \qquad (3.55)
\end{aligned}$$

□

Lemma 5. *Regardless of whether $u_{aux} > \max_\Omega p_0$ or not, and regardless of whether $u_{aux} = u_B$ or not, it always holds :*

$$\int_{\mathcal{T}(u_B)} p_0^2 \leq \frac{\bar{C}}{n^2 h_{tail}^d(u_B)}.$$

Proof of Lemma 5. If $u_{aux} = u_B$ then $h_{tail}(u_{aux}) = h_{tail}(u_B)$. Moreover, by definition of u_{aux}, we have: $\int_{\mathcal{B}(u_{aux})} p_0^2 \leq \frac{c_I}{n^2 h^d}$ so the result holds by recalling $\bar{C} = c_I + C_3$. Now If $u_B > u_{aux} > \max_\Omega p_0$,

then $h_{tail}(u_B) \leq \bar{h}_{tail}(u_{aux})$, so the result holds as well. Finally, if $u_{aux} > \max_{\Omega} p_0$, then $\mathcal{T}(u_{aux}) = \mathcal{T}(u_B)$ so the result holds as well by Lemma 1 item 1. □

We now show that $h_{tail}(u_B) \asymp \bar{h}_{tail}(u_{aux})$ when $u_B > u_{aux}$.

Lemma 6. *If $u_B > u_{aux}$ then $C_6 h_{tail}(u_B) \geq \bar{h}_{tail}(u_{aux})$ where $h_{tail}(u_B)$, $\bar{h}_{tail}(u_{aux})$ are defined in (3.47), (3.48) and C_6 is a constant. Hence it always holds $h_{tail}(u_B) \asymp \bar{h}_{tail}(u_{aux})$.*

Proof of Lemma 6. Suppose that $u_B > u_{aux}$. Then by Lemma 4, we have $\int_{\mathcal{T}(u_B)} p_0^2 \leq \bar{C} \frac{1}{n^2 \bar{h}_{tail}^d}$. Moreover, by definition of u_{aux}, since $u_B > u_{aux}$ we can write: $\int_{\mathcal{T}(u_B)} p_0^2 \geq c_I \frac{1}{n^2 h_{tail}(u_B)^d}$, hence: $(1 + \frac{C_3}{c_I}) h_{tail}(u_B) \geq \bar{h}_{tail}(u_{aux})$. Moreover, it directly follows from the definition of $h_{tail}(u_B)$ and $\bar{h}_{tail}(u_{aux})$ that $h_{tail}(u_B) \leq \bar{h}_{tail}(u_{aux})$. Hence $h_{tail}(u_B) \asymp \bar{h}_{tail}(u_{aux})$. □

Lemma 7. *Assume $C_{BT} \rho^*_{bulk} \geq \rho^*_{tail}$. There exists a constant $C_7 > 1$ depending only on C_{BT}, c_I and c_A, such that: $\int_{\mathcal{B}(\frac{u_B}{2})} p_0^r \leq C_7 \mathcal{I}$.*

Proof of Lemma 7.

$$\begin{aligned}\int_{\mathcal{B}(\frac{u_B}{2}) \setminus \mathcal{B}(u_B)} p_0^r &\leq u_B^r \Big| \mathcal{B}\Big(\frac{u_B}{2}\Big) \setminus \mathcal{B}(u_B) \Big| \leq u_B^{r-2} \int_{\mathcal{B}(\frac{u_B}{2}) \setminus \mathcal{B}(u_B)} 4 p_0^2 \\ &\leq 4 u_B^{r-2} \frac{\bar{C}}{n^2 h_{tail}(u_B)^d} \qquad \text{by Lemma 5.} \end{aligned} \tag{3.56}$$

Moreover, the condition $C_{BT} \rho^*_{bulk} \geq \rho^*_{tail}$ exactly rewrites

$$\frac{u_B^{r-2}}{n^2 h_{tail}(u_B)^d} \leq C_{BT}^{\frac{td}{(2-t)\alpha+d}} c_A^{-2\frac{(4-2t)\alpha+d}{(4-t)\alpha+d}} \cdot \mathcal{I},$$

so that (3.56) gives:

$$\int_{\mathcal{B}(\frac{u_B}{2}) \setminus \mathcal{B}(u_B)} p_0^r \leq 4 \bar{C} C_{BT}^{\frac{td}{(2-t)\alpha+d}} c_A^{-2\frac{(4-2t)\alpha+d}{(4-t)\alpha+d}} \cdot \mathcal{I}$$

so that

$$\int_{\mathcal{B}(\frac{u_B}{2})} p_0^r \leq \left(1 + 4 \bar{C} C_{BT}^{\frac{td}{(2-t)\alpha+d}} c_A^{-2\frac{(4-2t)\alpha+d}{(4-t)\alpha+d}}\right) \cdot \mathcal{I} =: C_7 \mathcal{I}.$$

□

Lemma 8. *If the tail dominates, i.e. if $\rho^*_{tail} \geq C_{BT} \rho^*_{bulk}$ where C_{BT} is a large constant, then there exists a small constant C_8 depending only on C_{BT} and decreasing with respect to C_{BT}, such that*

$$\int_{\mathcal{T}(2u_B)} p_0 \leq (1 + C_8) \int_{\mathcal{T}(u_B)} p_0.$$

Proof of Lemma 8. We show that $\int_{\mathcal{T}(2u_B)} p_0 - \int_{\mathcal{T}(u_B)} p_0 \leq C_8 \int_{\mathcal{T}(u_B)} p_0$. Note that r can be greater or smaller than 1, so that on $\mathcal{T}(2u_B) \setminus \mathcal{T}(u_B)$ we have: $p_0 \leq (2^{1-r} \vee 1) u_B^{1-r} p_0^r$. We therefore have

$$\begin{aligned}(2^{r-1} \wedge 1) \int_{\mathcal{T}(2u_B)\setminus\mathcal{T}(u_B)} p_0 &\leq u_B^{1-r} \int_{\mathcal{T}(2u_B)\setminus\mathcal{T}(u_B)} p_0^r \leq u_B^{1-r} \mathcal{I} \\ = \mathcal{I}^{\frac{(\alpha+d)((4-t)\alpha+d)}{(4\alpha+d)((2-t)\alpha+d)}} \frac{L^{\frac{d}{4\alpha+d}\frac{(4-3t)\alpha+d}{(2-t)\alpha+d}}}{n^{\frac{2\alpha}{4\alpha+d}\frac{(4-3t)\alpha+d}{(2-t)\alpha+d}}} &\leq C_{BT}^{-\frac{\alpha+d}{(2-t)\alpha+d}} \int_{\mathcal{T}(u_B)} p_0,\end{aligned}$$

where the last inequality is obtained by using the assumption $\rho^*_{tail} \geq C_{BT}\rho^*_{bulk}$ and the expressions of ρ^*_{bulk} and ρ^*_{tail}. □

Lemma 9. *If the tail dominates, i.e. if $\rho^*_{tail} \geq C_{BT}\rho^*_{bulk}$ where C_{BT} is a large enough constant, then there exists a small constant C_9 depending only on C_{BT} and decreasing with respect to C_{BT}, such that*

$$\int_{\mathcal{T}(2u_B)} p_0^2 \leq \frac{C_9}{n^2 h_{tail}^d}.$$

Proof of Lemma 9. We have, recalling equations (3.51), (3.53), (3.55):

$$2^{r-2} \int_{T(2u_B)\setminus\mathcal{T}(u_B)} p_0^2 \leq u_B^{2-r} \int_{\mathcal{B}(u_B)} p_0^r \leq \frac{\bar{C}}{n^2 \bar{h}_{tail}^d}.$$

Moreover, $\frac{\bar{C}}{n^2 \bar{h}_{tail}^d} \leq \frac{\bar{C}}{n^2 h_{tail}^d(2u_B)}$ and $h_{tail}(2u_B) \geq (1 + C_8)^{-\frac{1}{\alpha+d}} h_{tail}(u_B)$ by Lemma 8. Now:

$$\int_{\mathcal{T}(2u_B)} p_0^2 \leq \frac{\bar{C}}{n^2 h_{tail}^d} + \frac{\bar{C} 2^{2-r} (1 + C_8)^{\frac{d}{\alpha+d}}}{n^2 h_{tail}^d}.$$

which yields the result. □

Lemma 10. *In the case $\int_{\mathcal{T}} p_0 \geq \frac{c_{tail}}{n}$, there exists constants $C_{BT}, C_{BT}^{(2)}$ such that we have*

$$\rho^*_{tail} \geq C_{BT}\rho^*_{bulk} \iff h_m \geq C_{BT}^{(2)} h_{tail}(u_B),$$

where h_m is defined in Equation (3.24). In particular we have

$$\rho^*_{tail} \geq C_{BT}\rho^*_{bulk} \implies \inf_{x\in\mathcal{B}} h_b(x) \geq C_{BT}^{(2)} h_{tail}(u_B),$$

where $C_{BT}^{(2)}$ can be made arbitrarily large by choosing C_{BT} large enough.

Proof of Lemma 10. The result can be proved by direct calculation, recalling the expression of $h_{tail}(u_B)$ from (3.47), the expressions of ρ^*_{bulk} and ρ^*_{tail} from (3.14) and that $\inf_{x\in\mathcal{B}} h_b(x) \geq h_m$. □

3.B Partitioning algorithm

We now introduce the recursive partitioning scheme, inspired from [104]. For any cube $A \subset \Omega$, denote by $e(A)$ its edge length. For any function $h : \Omega \to \mathbb{R}_+$, denoting by x_A the center of A, define $h(A) = h(x_A)$. The partitionning algorithm takes as input a cubic domain $\widetilde{\Omega} \subset \Omega$, a parameter $\beta \geq \alpha$, a value $u > 0$ and a constant $c_\beta > 0$. Defining the bandwidth function $h : \widetilde{\Omega} \to \mathbb{R}_+$ such that $p_0(x) = c_\beta\, h^\beta(x)$ over $\widetilde{\Omega}$, as well as the set $\mathcal{D}(u) = \{x \in \widetilde{\Omega} : p_0(x) \geq u\}$ to be split into cubes, the algorithm returns a family $P = \{A_1, \ldots, A_N\}$ of *disjoints* cubes of $\widetilde{\Omega}$ *covering* $\mathcal{D}(u)$ (i.e. such that $\mathcal{D}(u) \subset \cup_{j=1}^N A_j$), and such that, for all $j = 1, \ldots, N$, $h(A_j) \geq e(A_j) \geq \frac{1}{2^{\beta+1}} h(A_j)$. Note that the center of A_j need not belong to $\mathcal{D}(u)$. Algorithm 2 corresponds to an auxiliary algorithm called by the actual partitioning algorithm defined in Algorithm 3.

Algorithm 2: Recursive auxiliary algorithm

1. **Input:** A, h, D, P.
2.
   - **If** $A \cap D = \emptyset$: **return** P.
   - **If** $e(A) \leq h(A)$: **return** $P \cup \{A\}$.
   - **Else:**
     (a) Split A into 2^d cubes $A_1, \ldots, A_{2^d}$ obtained by halving A along each of its axes.
     (b) **return** $\bigcup_{i=1}^{2^d}$ Algorithm 2(A_i, h, D, P).

Algorithm 3: Adaptive partition

1. **Input:** $\widetilde{\Omega}, \beta, u, c_\beta$.
2. **Initialization:** $P = \emptyset$, $\mathcal{D}(u) = \{x \in \widetilde{\Omega} : p_0(x) \geq u\}$, $h = \left(\frac{p_0}{c_\beta}\right)^{\frac{1}{\beta}}$.
3. **Return** Algorithm 2$\Big(\widetilde{\Omega}, h, \mathcal{D}(u), P\Big)$.

We have the following guarantees for Algorithm 3.

Proposition 3.6. *Algorithm 3 terminates. Assume moreover that Algorithm 3 splits the domain at least once and that there exists a constant $c_\alpha > 0$ such that $c_\star + \frac{\sqrt{d}^\alpha}{c_\alpha}(2^{1-\alpha} \vee 1) \leq 1/2$ and such that $\forall x \in \mathcal{D}(u) : p_0(x) \geq c_\alpha L h^\alpha(x)$. Then:*

1. *Denoting by P the output of Algorithm 3 with inputs $\widetilde{\Omega}, \beta, u, c_\beta$, it holds: $\mathcal{D}(u) \subset \cup_{A \in P} A$;*
2. *For all cube $A \in P$, it holds: $h(A) \geq e(A) \geq \frac{1}{2^{\beta+1}} h(A)$.*
3. *For all cell $A \in P$ we have $\min_A p_0 \geq \frac{1}{2} \max_A p_0$. Consequently, it holds $\cup_{A \in P} A \subset \mathcal{D}(\frac{u}{2})$.*

Proof of Proposition 3.6. Fix a cube $\widetilde{\Omega} \subset \Omega$, $\beta > 0$, $u > 0$ $c_\beta > 0$.

Termination: Suppose that Algorithm 3 does not terminate. Then, among the cubes defined at some step by the algorithm, there would exist an infinite sequence $(A_l)_{l\in\mathbb{N}}$ of nested cubes of $\widetilde{\Omega}$ satisfying:

(i) $A_0 = \widetilde{\Omega}$,
(ii) $\forall l \in \mathbb{N}$: $A_{l+1} \subset A_l$,
(iii) $\forall l \in \mathbb{N} : e(A_{l+1}) = \frac{1}{2}e(A_l)$,
(iv) $\forall l \in \mathbb{N} : A_l \cap \mathcal{D}(u) \neq \emptyset$.

Denote by x_l the center of A_l for all $l \in \mathbb{N}$. Then by (ii) and (iii), $(x_l)_l$ is a Cauchy sequence of $[0,1]^d$ and thus converges to some $x_\infty \in \Omega$. Moreover, denoting by $d_{\|\cdot\|_2}(x, \mathcal{D}(u))$ the Euclidean distance of x to $\mathcal{D}(u)$, we have: $d_{\|\cdot\|_2}(x_l, \mathcal{D}(u)) \leq e(A_l)\frac{\sqrt{d}}{2} = 2^{-l}e(\widetilde{\Omega})\frac{\sqrt{d}}{2} \to 0$. Since $\mathcal{D}(u)$ is closed by continuity of p_0, it holds $p_0(x_\infty) \geq u > 0$. However, at each step, A_l is split, imposing $e(A_l) \geq h(x_l) \to 0$, hence $h(x_\infty) = 0$, yielding $p_0(x_\infty) = 0$ since $p_0 = c_\beta h^\beta$ over $\widetilde{\Omega}$. This leads to a contradiction.

1. It is straightforward to check that when the algorithm terminates, $\mathcal{D}(u) \subset \cup_{A\in P} A$.

2. Let $A \in P$. Denote by A' the parent of A in the hierarchical splitting performed by Algorithm 2. Since by assumption the domain is split at least once, A' exists. Since A' was split and A was kept we necessarily have: $2h(A) \geq 2e(A) = e(A') > h(A')$. Denote by x_A and $x_{A'}$ the respective centers of A and A'. Since by definition of A, $x_{A'}$ is a vertex of A, we have $\|x_A - x_{A'}\| = e(A)\frac{\sqrt{d}}{2} \leq h(A)\frac{\sqrt{d}}{2}$. By Assumption ($\star$):

$$|p_0(x_A) - p_0(x_{A'})| \leq c_\star p_0(x_A) + L\|x_A - x_{A'}\|^\alpha = c_\star p_0(x_A) + Lh(A)^\alpha \left(\frac{\sqrt{d}}{2}\right)^\alpha. \tag{3.57}$$

There are two cases:

- If $x_A \in \mathcal{D}(u)$ then $|p_0(x_A) - p_0(x_{A'})| \leq c_\star p_0(x_A) + \frac{1}{c_\alpha}p_0(x_A)\left(\frac{\sqrt{d}}{2}\right)^\alpha \leq p_0(x_A)/2$, hence $p_0(x'_A) \geq p_0(x_A)/2$.
- Otherwise, $x_A \notin \mathcal{D}(u)$. Let $x_u \in A \cap \mathcal{D}(u)$. Since A was kept by Algorithm 3, x_u exists. By the definition of $D(u)$, it holds $p_0(x_A) < p_0(x_u)$, hence $h(x_A) < h(x_u)$. We have $\|x_A - x_u\| \leq e(A)\frac{\sqrt{d}}{2} \leq h(A)\frac{\sqrt{d}}{2} \leq h(x_u)\frac{\sqrt{d}}{2}$. By Assumption ($\star$): $|p_0(x_A) - p_0(x_u)| \leq c_\star p_0(x_u) + Lh(x_u)^\alpha\left(\frac{\sqrt{d}}{2}\right)^\alpha \leq \frac{p_0(x_u)}{2}$, hence $p_0(x_A) \geq \frac{p_0(x_u)}{2}$. Therefore, it holds:

$$Lh^\alpha(x_A) \leq Lh^\alpha(x_u) \leq \frac{1}{c_\alpha}p_0(x_u) \leq \frac{2}{c_\alpha}p_0(x_A).$$

Injecting this relation into (3.57), we get: $|p_0(x_A) - p_0(x_{A'})| \leq \frac{1}{2}p_0(x_A)$ hence $p_0(x_{A'}) \geq p_0(x_A)/2$.

In both cases, it holds $p_0(x_{A'}) \geq p_0(x_A)/2$, hence $h(x_{A'}) \geq h(x_A)/2^{1/\beta}$. We therefore get $h(A) \geq e(A) \geq \frac{h(A')}{2} \geq h(A)/2^{1+1/\beta}$.

3. Let $A \in P$ and $x = \arg\max_A p_0$, $y = \arg\min_A p_0$. We have by Assumption ($\star$):

$$|p_0(x) - p_0(y)| \leq c_\star p_0(x) + L(e(A)\sqrt{d})^\alpha \leq \left(c_\star + \frac{\sqrt{d}^\alpha}{c_\alpha}\right) p_0(x) \leq \frac{p_0(x)}{2}.$$

□

3.C Upper bound in the bulk regime

3.C.1 Technical lemmas in the bulk regime

Throughout the Appendix, we will denote by $B(x,h)$ the Euclidean ball of $\mathbb{R}^d$ centered at x and of radius h. When no ambiguity arises, $\|\cdot\|$ will denote the Euclidean norm over $\mathbb{R}^d$. **In Appendix 3.C only,** we will denote by $h(x)$ the quantity $c_h h_b(x)$ where the constant c_h can be chosen arbitrarily small. We first prove the following result, stating that for any p satisfying Assumption ($\star$), p can be considered as approximately constant over the balls $B(x,h(x))$ for all $x \in \widetilde{\mathcal{B}}$ where $\widetilde{\mathcal{B}} = \mathcal{B}(\frac{u_B}{2})$ if the bulk dominates and $\widetilde{\mathcal{B}} = \mathcal{B}$ if the tail dominates.

Lemma 11. *Recall that $h(x) = c_h h_b(x) = (\widetilde{c_A}/4)^{1/\alpha} h_b(x)$. Let $x \in \widetilde{\mathcal{B}}$ and $y \in B(x,h(x))$. For all p satisfying Assumption ($\star$), we have:*

$$\frac{p(y)}{p(x)} \in [c^{(p)}, C^{(p)}] \qquad \text{where} \qquad C^{(p)}, c^{(p)} = 1 \pm \left(c_\star + \frac{c_h^\alpha(1+\delta)}{\widetilde{c_A}}\right).$$

It follows that:

$$\frac{\omega(y)}{\omega(x)} \in [c^{(\omega)}, C^{(\omega)}] \qquad \text{and} \qquad \frac{h_b(y)}{h_b(x)} \in [c^{(h)}, C^{(h)}],$$

where $C^{(h)} = C^{(p)\,\frac{2}{(4-t)\alpha+d}}$, $c^{(h)} = c^{(p)\,\frac{2}{(4-t)\alpha+d}}$, $C^{(\omega)} = C^{(p)\,\frac{2\alpha t-4\alpha}{(4-t)\alpha+d}}$ and $c^{(\omega)} = C^{(p)\,\frac{2\alpha t-4\alpha}{(4-t)\alpha+d}}$.

Proof. Let $x \in \widetilde{\mathcal{B}}$ and $y \in B(x,h(x))$ and assume that p satisfies Assumption ($\star$) with the constants $c'_\star$ and L'. We have

$$\begin{aligned}|p(x) - p(y)| &\leq c'_\star p(x) + L'\|x-y\|^\alpha \leq c'_\star p(x) + L'\, c_h^\alpha\, h_b^\alpha(x) \\ &\leq c'_\star p(x) + \frac{(1+\delta)c_h^\alpha}{\widetilde{c_A}} p(x) \qquad \text{since on } \widetilde{\mathcal{B}} : \widetilde{c_A} L h_b(x)^\alpha \leq p(x).\end{aligned}$$

We then take c_h small enough to guarantee that $\frac{(1+\delta)c_h^\alpha}{\widetilde{c_A}} \leq \frac{c_\star}{2}$. Therefore, $C^{(p)}\, p_0(x) \geq p_0(y) \geq c^{(p)}\, p_0(x)$. The analogous relations for h_b and ω directly follow from their definitions (3.15), (3.21). □

In the sequel, we define the following notation:

$$\widetilde{L} = \frac{L^d}{n^{2\alpha}}, \tag{3.58}$$

and for all $u \geq 0$:

$$\mathcal{I}(u) := \int_{\mathcal{B}(u)} p_0^{\frac{2\alpha t}{(4-t)\alpha+d}}. \tag{3.59}$$

so that $\mathcal{I} = \mathcal{I}(u_{aux})$.

Lemma 12. *It holds*

$$\widetilde{u_B} \geq \left[\widetilde{c_A}\frac{\widetilde{L}}{\mathcal{I}^\alpha}\right]^{\frac{1}{4\alpha+d}\frac{(4-t)\alpha+d}{(2-t)\alpha+d}}. \tag{3.60}$$

The proof follows directly from the definition of u_B in (3.11) and (3.17).

Lemma 13. *We have:* $\frac{1}{n}\int_{\widetilde{\mathcal{B}}}\omega(x)p_0(x)^2dx \leq t_n$.

Proof of Lemma 13. We have by the Cauchy-Schwarz inequality:

$$\begin{aligned}\frac{1}{n}\int_{\widetilde{\mathcal{B}}}\omega p_0^2 &= \frac{1}{n}\int_{\widetilde{\mathcal{B}}} p_0\, h_b^{d/2}\, p_0^{r/2} \times \left(n^2L^4\mathcal{I}\right)^{\frac{d/2}{4\alpha+d}} \leq \frac{L^{\frac{2d}{4\alpha+d}}\mathcal{I}^{\frac{d/2}{4\alpha+d}}}{n^{\frac{4\alpha}{4\alpha+d}}}\left(\int_{\widetilde{\mathcal{B}}} p_0^2h_b^d\int_{\widetilde{\mathcal{B}}} p_0^r\right)^{1/2}\\ &\leq \frac{\sqrt{C_7}}{C_{t_n}}t_n\left(\int_{\widetilde{\mathcal{B}}} p_0^2h_b^d\right)^{1/2}.\end{aligned}$$

Moreover by Lemma 11,

$$p_0(x)h_b^d(x) = p_0(x)\frac{h(x)^d}{c_h^d} \leq \frac{1}{c_h^d c^{(p)}}\int_{B(x,h(x))} p_0 \leq \frac{1}{c^{(p)}c_h},$$

so that

$$\int_{\widetilde{\mathcal{B}}} p_0^2h_b^d \leq \frac{1}{c^{(p)}c_h}\int_{\widetilde{\mathcal{B}}} p_0 \leq \frac{1}{c^{(p)}c_h}.$$

Taking $C_{t_n} \geq \sqrt{\frac{C_7}{c^{(p)}c_h}}$ yields the result. □

Lemma 14. *It holds:* $\int_{\widetilde{\mathcal{B}}} L^2\omega(x)h_b^{2\alpha}(x)dx \leq t_n$.

Proof of Lemma 14. We have:

$$\int_{\widetilde{\mathcal{B}}} L^2\omega(x)h_b^{2\alpha}(x)dx = \frac{L^{\frac{2d}{4\alpha+d}}\mathcal{I}(\widetilde{u_B})}{n^{\frac{4\alpha}{4\alpha+d}}\mathcal{I}(u_{aux})^{\frac{2\alpha}{4\alpha+d}}} \leq \frac{C_7}{C_{t_n}}t_n,$$

recalling the expressions of $\omega(x)$ from (3.21), $h_b(x)$ from (3.15), t_n from (3.22), and using at the last step (if $\widetilde{u_B} = u_B$) $\mathcal{I}(u_B) \leq \mathcal{I}(u_{aux})$ since $u_B \geq u_{aux}$. Taking $C_{t_n} \geq C_7$ yields the result. □

In the remaining of the analysis of the upper bound in the bulk regime, we fix $X_1, \ldots, X_n$ a family of iid random variables with density either $p_0 \in \mathcal{P}(\alpha, L, c_\star)$ or $p \in \mathcal{P}(\alpha, L', c'_\star)$. In the whole analysis

of the upper bound, we will use the following notation:

$$\Delta(x) := p(x) - p_0(x) \qquad \hat{\Delta}(x) := \hat{p}(x) - p_0(x) \qquad \text{and} \qquad \hat{\Delta}'(x) := \hat{p}'(x) - p_0(x). \tag{3.61}$$

We also define

$$J := \int_{\widetilde{\mathcal{B}}} \omega(x)\Delta(x)^2 dx. \tag{3.62}$$

We will denote by C'_K the constant $(1+\delta)C_K$ and by $C_K^{(2)}$ the constant $\int_{\mathbb{R}^d} K^2$.

Lemma 15. *For $p \in \mathcal{P}(\alpha, L', c'_\star)$, it holds $\frac{1}{n}\int_{\widetilde{\mathcal{B}}} \omega p^2 \leq A_{15} t_n + \frac{B_{15}}{n} J$, where A_{15} and B_{15} are two constants.*

Proof of Lemma 15. Using $(a+b)^2 \leq 2a^2 + 2b^2$ and the triangle inequality, we get:

$$\frac{1}{n}\int_{\widetilde{\mathcal{B}}} \omega p^2 \leq \frac{1}{n}\int_{\widetilde{\mathcal{B}}} \omega[2p_0^2 + 2\Delta^2] \leq 2t_n + \frac{2}{n}J =: A_{15}t_n + \frac{B_{15}}{n}J \quad \text{by Lemma 13.}$$

□

Lemma 16. *If $J \geq t_n$ then we have: $\mathbb{E}T_{bulk} \geq \left(\sqrt{J} - \sqrt{t_n}\right)^2$.*

Proof of Lemma 16. By the Minkowski inequality:

$$\begin{aligned} J &= \int_{\widetilde{\mathcal{B}}} \omega(x)\Delta(x)^2 dx = \int_{\widetilde{\mathcal{B}}} \omega(x)\Big[\Delta(x) + \mathbb{E}[\hat{p}(x) - p(x)] - \mathbb{E}[\hat{p}(x) - p(x)]\Big]^2 dx \\ &\leq \left[\left(\int_{\widetilde{\mathcal{B}}} \omega(x)\mathbb{E}^2[\hat{\Delta}(x)]dx\right)^{1/2} + \left(\int_{\widetilde{\mathcal{B}}} \omega(x)\mathbb{E}^2[\hat{p}(x) - p(x)]dx\right)^{1/2}\right]^2 \\ &\leq \left[\sqrt{\mathbb{E}T_{bulk}} + C'_K\sqrt{\frac{t_n}{C_{t_n}}}\right]^2 \leq \left[\sqrt{\mathbb{E}T_{bulk}} + \sqrt{t_n}\right]^2 \text{ by choosing } C_{t_n} \geq {C'_K}^2. \end{aligned}$$

At the last step we used $|\mathbb{E}(\hat{p}(x) - p(x))| \leq C_K L' h^\alpha(x) = C'_K L h^\alpha(x)$, by [188] Proposition 1.2. Moreover, we used Lemma 14. This yields the result, since $J \geq t_n$. □

Lemma 17. *We have: $\mathbb{V}(T_{bulk}) \leq \left[\left(\sqrt{J} + \sqrt{t_n}\right)^2 + J_2^{1/2}\right]^2 - \mathbb{E}^2 T_{bulk}$ where*

$$J_2 = \iint_{\widetilde{\mathcal{B}}^2} \omega(x)\omega(y)\frac{1}{k^2} cov^2\left(K_{h(x)}(x - X), K_{h(y)}(y - X)\right) dxdy. \tag{3.63}$$

Proof of Lemma 17.

$$\begin{aligned} \mathbb{V}(T_{bulk}) &= \mathbb{E}(T_{bulk}^2) - \mathbb{E}^2 T_{bulk} = \mathbb{E}\iint_{\widetilde{\mathcal{B}}^2} \omega(x)\omega(y)\ \hat{\Delta}(x)\hat{\Delta}(y)\ \hat{\Delta}'(x)\hat{\Delta}'(y)\ dxdy - \mathbb{E}^2 T_{bulk} \\ &= \iint_{\widetilde{\mathcal{B}}^2} \omega(x)\omega(y)\mathbb{E}\left[\hat{\Delta}(x)\hat{\Delta}(y)\right]^2 dxdy - \mathbb{E}^2 T_{bulk}. \end{aligned} \tag{3.64}$$

Recall that throughout Appendix 3.C, $h(x) = c_h h_b(x)$ where $c_h = (\widetilde{c_A}/4)^{\frac{1}{\alpha}}$. We now compute the term $\mathbb{E}\left[\hat{\Delta}(x)\hat{\Delta}(y)\right]^2$. We have:

$$\mathbb{E}\left[\hat{\Delta}(x)\hat{\Delta}(y)\right] = \mathbb{E}\left\{\left[\frac{1}{k}\sum_{i=1}^{k}\Big(K_{h(x)}(x-X_i)-p_0(x)\Big)\right]\left[\frac{1}{k}\sum_{i=1}^{k}\Big(K_{h(y)}(y-X_i)-p_0(y)\Big)\right]\right\}$$

$$= \frac{1}{k^2}\sum_{i=1}^{k}\mathbb{E}\left\{\Big[K_{h(x)}(x-X_i)-p_0(x)\Big]\Big[K_{h(y)}(y-X_i)-p_0(y)\Big]\right\} + \frac{1}{k^2}\sum_{i\neq j}\mathbb{E}\Big[K_{h(x)}(x-X_i)-p_0(x)\Big]\,\mathbb{E}\Big[K_{h(y)}(y-X_j)-p_0(y)\Big]$$

$$= \frac{1}{k}\,\mathbb{E}\left\{\Big[K_{h(x)}(x-X)-p_0(x)\Big]\Big[K_{h(y)}(y-X)-p_0(y)\Big]\right\} + \frac{k-1}{k}\mathbb{E}\Big[K_{h(x)}(x-X)-p_0(x)\Big]\,\mathbb{E}\Big[K_{h(y)}(y-X)-p_0(y)\Big]$$

$$= \mathbb{E}\Big[K_{h(x)}(x-X)-p_0(x)\Big]\,\mathbb{E}\Big[K_{h(y)}(y-X)-p_0(y)\Big] + \frac{1}{k}\mathrm{cov}\Big(K_{h(x)}(x-X),K_{h(y)}(y-X)\Big), \tag{3.65}$$

so that, by the Minkowski inequality:

$$\iint_{\widetilde{\mathcal{B}}^2}\omega(x)\omega(y)\mathbb{E}\left[\hat{\Delta}(x)\hat{\Delta}(y)\right]^2 dxdy \leq (J_1^{1/2}+J_2^{1/2})^2, \tag{3.66}$$

where

$$J_1^{1/2} = \int_{\widetilde{\mathcal{B}}}\omega(x)\mathbb{E}^2\Big[K_{h(x)}(x-X)-p_0(x)\Big]dx, \tag{3.67}$$

$$J_2 = \iint_{\widetilde{\mathcal{B}}^2}\omega(x)\omega(y)\frac{1}{k^2}\mathrm{cov}^2\Big(K_{h(x)}(x-X),K_{h(y)}(y-X)\Big)dxdy. \tag{3.68}$$

Moreover, by triangular inequality and by [188], Proposition 1.1:

$$\Big|\mathbb{E}\Big[K_{h(x)}(x-X)-p_0(x)\Big]\Big| \;\leq\; |p(x)-p_0(x)| + \Big|\mathbb{E}\left(\hat{p}(x)-p(x)\right)\Big| \;\leq\; |\Delta(x)| + C'_K L h(x)^\alpha,$$

Therefore, still by the Minkowski inequality:

$$J_1^{1/2} \leq \int_{\widetilde{\mathcal{B}}}\omega(x)\Big[\,|\Delta(x)| + C'_K L h(x)^\alpha\Big]^2 dx \leq \left[\Big(\int_{\widetilde{\mathcal{B}}}\omega(x)|\Delta(x)|^2dx\Big)^{1/2} + \Big(\int_{\widetilde{\mathcal{B}}}\omega(x){C'_K}^2L^2h(x)^{2\alpha}dx\Big)^{1/2}\right]^2$$

$$\leq \left(\sqrt{J} + \sqrt{t_n}\right)^2 \quad \text{by Lemma 14, the definition of } J \text{ and using } c_h \leq 1. \tag{3.69}$$

Equations (3.64), (3.66) and (3.69) yield the result. □

Lemma 18. *Let X be a random variable with density $p \in \mathcal{P}(\alpha, L', c'_\star)$. If $\|x-y\| > \frac{1}{2}[h(x)+h(y)]$, then*

$$\left|\text{cov}\left(K_{h(x)}(x-X), K_{h(y)}(y-X)\right)\right| \leq \left(C'_K L h(x)^\alpha + p(x)\right)\left(C'_K L h(y)^\alpha + p(y)\right).$$

Proof of Lemma 18. We recall that by definition K has bounded support $B(0, \frac{1}{2})$. In this case: $\text{Supp}K_{h(x)}(x-\cdot) \cap \text{Supp}K_{h(y)}(y-\cdot) = \emptyset$. Therefore, $K_{h(x)}(x-X)K_{h(y)}(y-X) = 0$ almost surely. Then by [188], Proposition 1.1:

$$\begin{aligned}\left|\text{cov}\left(K_{h(x)}(x-X), K_{h(y)}(y-X)\right)\right| &= \left|\mathbb{E}\left[K_{h(x)}(x-X)\right]\mathbb{E}\left[K_{h(y)}(y-X)\right]\right| \\ &\leq \left(C'_K L h(x)^\alpha + p(x)\right)\left(C'_K L h(y)^\alpha + p(y)\right).\end{aligned}$$

□

Lemma 19. *Let X be a random variable with density $p \in \mathcal{P}(\alpha, L', c'_\star)$. If $\|x-y\| \leq \frac{1}{2}[h(x)+h(y)] \leq h(x) \vee h(y)$ then*

$$\left|cov\left(K_{h(x)}(x-X), K_{h(y)}(y-X)\right)\right| \leq C_{19}\frac{p(x)}{h^d(x)}. \tag{3.70}$$

where C_{19} is a constant.

Proof of Lemma 19. If $\|x-y\| \leq \frac{1}{2}[h(x)+h(y)] \leq h(x) \vee h(y)$, we suppose by symmetry $h(y) = h(x) \vee h(y)$. Then

$$\left|\text{cov}\left(K_{h(x)}(x-X), K_{h(y)}(y-X)\right)\right| \leq \sqrt{\mathbb{V}\left(K_{h(x)}(x-X)\right)\mathbb{V}\left(K_{h(y)}(y-X)\right)}.$$

Now, by [188], Proposition 1.1, the variance of the Kernel estimator $K_{h(x)}(x-X)$ is upper bounded as:

$$\mathbb{V}\left(K_{h(x)}(x-X)\right) \leq \frac{1}{h^d(x)}\left[\sup_{B(x,h(x))} p\right]\int_{\mathbb{R}^d} K^2 \leq \frac{1}{h^d(x)} C^{(p)} C_K^{(2)}\, p(x). \tag{3.71}$$

In the last inequality, we used Lemma 11. Hence, since $h(y) \geq h(x)$ and $\|x-y\| \leq h(y)$, we have $x \in B(y, h(x))$ so that $p(y) \leq \frac{1}{c^{(p)}} p(x)$. Thus:

$$\left|\text{cov}\left(K_{h(x)}(x-X), K_{h(y)}(y-X)\right)\right| \leq \frac{\sqrt{C^{(p)}C_K^{(2)}\, p(x) \cdot C^{(p)}C_K^{(2)}\, p(y)}}{\sqrt{h^d(x) \cdot h^d(x)}}$$

$$\leq \frac{C^{(p)} C_K^{(2)}}{\sqrt{c^{(p)}}} \frac{p(x)}{h^d(x)} =: C_{19} \frac{p(x)}{h^d(x)}. \tag{3.72}$$

□

3.C.2 Analysis of the upper bound in the bulk regime

In the bulk regime, we recall that ψ^*_{bulk} rejects H_0 if, and only if: $T_{bulk} > C_{\psi_b} t_n$. We prove the bulk upper bound by showing that ψ^*_{bulk} has small type-I and type-II errors. To do so, we show that *whp* under H_0, $T_{bulk} \leq C_{\psi_b} t_n$, whereas *whp* under H_1: $T_{bulk} > C_{\psi_b} t_n$. This will be proved by computing the expectation and variance of T_{bulk} under H_0 in Proposition 3.7 and under H_1 in Proposition 3.8. In both cases, we then use Chebyshev's inequality to show, in Corollary 3.1, that under H_0 T_{bulk} is concentrated below $C_{\psi_b} t_n$, while under H_1 it is concentrated above $C_{\psi_b} t_n$.

Proposition 3.7. *Under H_0 we have:*

- $\left|\mathbb{E}(T_{bulk})\right| \leq t_n$
- $\mathbb{V}(T_{bulk}) \leq C_{\mathbb{V},H_0} t_n^2$,

where $C_{\mathbb{V},H_0}$ is a constant given in the proof.

Proposition 3.8. *There exists a constant n_{bulk} depending only on C'_b and a constant $C_{\mathbb{V},H_1}$ such that, whenever $n \geq n_{bulk}$ and $\left(\int_{\widetilde{\mathcal{B}}} |p - p_0|^t\right)^{1/t} \geq C'_b \rho^*_{bulk}$, it holds:*

- $\mathbb{E}(T_{bulk}) \geq (1 - 1/\sqrt{C'_b})^2 J$
- $\mathbb{V}(T_{bulk}) \leq \frac{C_{\mathbb{V},H_1}}{C'_b} J^2$.

Corollary 3.1. *There exist three large constants C_{ψ_b}, C'_b, n_{bulk} where n_{bulk} only depends on C'_b, such that whenever $n \geq n_{bulk}$ and $\int_{\widetilde{\mathcal{B}}} |p - p_0|^t \geq C'_b {\rho^*_{bulk}}^t$, it holds:*

1. $\mathbb{P}_{p_0}\left(\psi^*_{bulk} = 1\right) = \mathbb{P}_{p_0}\left(T_{bulk} > C_{\psi_b} t_n\right) \leq \frac{\eta}{4}$,
2. $\mathbb{P}_p\left(\psi^*_{bulk} = 0\right) = \mathbb{P}_p\left(T_{bulk} \leq C_{\psi_b} t_n\right) \leq \frac{\eta}{4}$.

Proof of Proposition 3.7. We place ourselves under H_0 and bound the expectation and variance of T_{bulk}. We recall that $\hat{p}(x)$ and $\hat{p}'(x)$ are independent for all $x \in \widetilde{\mathcal{B}}$, and so are $\hat{\Delta}(x)$ and $\hat{\Delta}'(x)$.

Expectation: By the triangle inequality:

$$\Big|\mathbb{E}[T_{bulk}]\Big| = \Big|\int_{\widetilde{\mathcal{B}}}\omega(x)\,\mathbb{E}[\hat{\Delta}(x)]\,\mathbb{E}[\hat{\Delta}'(x)]dx\Big| \le \int_{\widetilde{\mathcal{B}}}\omega(x)\Big|\mathbb{E}[\hat{\Delta}(x)]\Big|\,\Big|\mathbb{E}[\hat{\Delta}'(x)]\Big|dx.$$

Now, recalling (3.61) and (3.19), we have (see e.g. [188], Prop 1.2):

$$\Big|\mathbb{E}[\hat{\Delta}(x)]\Big| \le C_K L h_b^\alpha(x) \qquad \text{and} \qquad \Big|\mathbb{E}[\hat{\Delta}'(x)]\Big| \le C_K L h_b^\alpha(x), \tag{3.73}$$

$$\text{so that: } \Big|\mathbb{E}[T_{bulk}]\Big| \;\le\; C_K^2 L^2 \int_{\widetilde{\mathcal{B}}}\omega(x)h^{2\alpha}(x)dx \le c_h^{2\alpha}C_K^2 t_n \le t_n,$$

by Lemma 14 and taking c_h small enough.

Variance: By Lemma 17, the variance under H_0 of T_{bulk} can be upper bounded as

$$\mathbb{V}(T_{bulk}) \le \Big[\Big(\sqrt{J} + C_K\sqrt{t_n}\Big)^2 + J_2^{1/2}\Big]^2 - \mathbb{E}^2 T_{bulk} \le [C_K t_n + J_2^{1/2}]^2. \tag{3.74}$$

We now analyse the covariance term in J_2. There are two cases. To analyse them, introduce the bulk diagonal:

$$\text{Diag} = \left\{(x,y) \in \widetilde{\mathcal{B}}^2 : \|x-y\| \le \frac{1}{2}[h(x)+h(y)]\right\}. \tag{3.75}$$

First case: $\|x-y\|_2 > \frac{1}{2}[h(x)+h(y)]$ i.e. $(x,y) \notin$ Diag.

By Lemma 18, and recalling that $\forall x \in \widetilde{\mathcal{B}} : \widetilde{c_A} L h^\alpha(x) \le p_0(x)$ we have

$$\begin{aligned}\frac{1}{k}\Big|\text{cov}\Big(K_{h(x)}(x-X), K_{h(y)}(y-X)\Big)\Big| &\le \frac{1}{k}\left[C_K L h(x)^\alpha + p_0(x)\right]\left[C_K L h(y)^\alpha + p_0(y)\right] \\ &\le \frac{1}{k}\Big(\frac{C_K}{\widetilde{c_A}} + 1\Big)^2 p_0(x)p_0(y).\end{aligned} \tag{3.76}$$

Second case: $\|x-y\| \le \frac{1}{2}[h(x)+h(y)] \le h(x) \vee h(y)$, i.e. $(x,y) \in$ Diag.

We suppose by symmetry $h(y) = h(x) \vee h(y)$. Then by Lemma 19:

$$\frac{1}{k}\Big|\text{cov}\Big(K_{h(x)}(x-X), K_{h(y)}(y-X)\Big)\Big| \le C_{19}\frac{p_0(x)}{k\,h^d(x)}.$$

Putting together the above equation with (3.65) and (3.76), we get:

$$\mathbb{E}\left[\hat{\Delta}(x)\hat{\Delta}(y)\right] \leq C_K^2 L^2 h(x)^\alpha h(y)^\alpha + \mathbb{1}_{\mathrm{Diag}^c}\frac{1}{k}\left(\frac{C_K}{c_A}+1\right)^2 p_0(x)p_0(y) + \mathbb{1}_{\mathrm{Diag}} C_{19}\frac{p_0(x)}{k\,h(x)^d}. \tag{3.77}$$

We now combine (3.77) with (3.64) and use Minkowski's inequality:

$$\begin{aligned}\mathbb{V}(T_{bulk}) \leq &\left\{\left(\iint_{\widetilde{\mathcal{B}}^2}\omega(x)\omega(y)\left[C_K^2 L^2 h(x)^\alpha h(y)^\alpha\right]^2 dxdy\right)^{1/2} + \right.\\ &\left.\left(\iint_{\widetilde{\mathcal{B}}^2}\omega(x)\omega(y)\left[\mathbb{1}_{\mathrm{Diag}^c}\frac{1}{k}\left(\frac{C_K}{\widetilde{c_A}}+1\right)^2 p_0(x)p_0(y) + \mathbb{1}_{\mathrm{Diag}} C_{19}\frac{p_0(x)}{k\,h(x)^d}\right]^2 dxdy\right)^{1/2}\right\}^2\\ \leq &\left\{C_K t_n + 2\left(\frac{C_K}{\widetilde{c_A}}+1\right)^2 t_n + \left(\iint_{\widetilde{\mathcal{B}}^2}\omega(x)\omega(y)\left[\mathbb{1}_{\mathrm{Diag}} C_{19}\frac{p_0(x)}{k\,h(x)^d}\right]^2 dxdy\right)^{1/2}\right\}^2. \end{aligned} \tag{3.78}$$

The last step is obtained by using Lemmas 13 and 14. We now analyse the term $C_{19}^2\iint_D \omega(x)\omega(y)\left[\frac{p_0(x)}{k\,h(x)^d}\right]^2 dxdy$ By Lemma 20 (at the end of Appendix 3.C):

$$\iint_{\mathrm{Diag}}\omega(x)\omega(y)\left[\frac{p_0(x)}{k\,h(x)^d}\right]^2 dxdy \leq 2C^{(\omega)}\int_{\widetilde{\mathcal{B}}}\omega(x)^2 h(x)^d\left[\frac{p_0(x)}{k\,h(x)^d}\right]^2 dx = 8\frac{C^{(\omega)}}{c^{(h)^d}}t_n^2,$$

by immediate calculation, recalling that $n = 2k$. Therefore, by equation (3.78), we have

$$\mathbb{V}T_{bulk} \leq \left[C_K t_n + 2\left(\frac{C_K}{c_A}+1\right)^2 t_n + C_{19}\sqrt{8\frac{C^{(\omega)}}{c^{(h)^d}}}t_n\right]^2 =: C_{\mathbb{V},H_0}t_n^2. \tag{3.79}$$

□

Analysis of the test statistic under H_1.

Proof of Proposition 3.8.

Suppose the data $(X_1,\ldots,X_n)$ is drawn from a probability density p satisfying:

$$C_b'^{\,t}\,{\rho^*_{bulk}}^{\,t} \leq \int_{\widetilde{\mathcal{B}}}|p-p_0|^t. \tag{3.80}$$

Expectation: We first prove $J \geq t_n$ in order to apply Lemma 16. We recall that $L' = (1+\delta)L$. From Equation (3.80) we get:

$$C_b'^t \left[\frac{L^{\frac{d}{4\alpha+d}}\mathcal{I}^{\frac{1}{t}-\frac{\alpha}{4\alpha+d}}}{n^{\frac{2\alpha}{4\alpha+d}}}\right]^t = C_b'^t \rho_{bulk}^{*\,t} \leq \int_{\widetilde{\mathcal{B}}} |\Delta|^t = \int_{\widetilde{\mathcal{B}}} \left[\omega(x)\Delta(x)^2\right]^{\frac{t}{2}} \omega(x)^{-\frac{t}{2}} dx$$
$$\underset{\text{Hölder}}{\leq} \left[\int_{\widetilde{\mathcal{B}}} \omega(x)\Delta^2(x)dx\right]^{\frac{t}{2}} \left[\int_{\widetilde{\mathcal{B}}} \omega^{-\frac{t}{2-t}}\right]^{\frac{2-t}{2}}, \tag{3.81}$$

where we have applied Hölder's inequality with $u = \frac{2}{t}$ and $v = \frac{2}{2-t}$ satisfying $\frac{1}{u} + \frac{1}{v} = 1$. Hence:

$$J = \int_{\widetilde{\mathcal{B}}} \omega(x)\Delta^2(x)dx \geq C_b'^2 \left(\frac{L^{\frac{d}{4\alpha+d}}\mathcal{I}^{\frac{1}{t}-\frac{\alpha}{4\alpha+d}}}{n^{\frac{2\alpha}{4\alpha+d}}}\right)^2 \times \left(\int_{\widetilde{\mathcal{B}}} \frac{1}{\omega^{\frac{t}{2-t}}}\right)^{-\frac{2-t}{t}} \geq C_b'^2 C_7^{\frac{t-2}{t}} \frac{t_n}{C_{t_n}} \geq C_b' t_n. \tag{3.82}$$

Taking C_b' large enough yields $J \geq t_n$, hence we can apply Lemma 16 which yields

$$\mathbb{E}T_{bulk} \geq \left(\sqrt{J} - \sqrt{t_n}\right)^2 \geq \left(\sqrt{J} - \sqrt{\frac{J}{C_b'}}\right)^2 = (1 - 1/\sqrt{C_b'})^2 J, \tag{3.83}$$

where we recall that C_b' can be taken arbitrarily large.

Variance:

We still have by Lemma 17 and by (3.83):

$$\mathbb{V}(T_{bulk}) \leq \left[\left(\sqrt{J} + \sqrt{t_n}\right)^2 + J_2^{1/2}\right]^2 - \mathbb{E}^2 T_{bulk} \tag{3.84}$$
$$\leq \left[\left(\sqrt{J} + \sqrt{t_n}\right)^2 + J_2^{1/2}\right]^2 - \left(\sqrt{J} - \sqrt{t_n}\right)^4, \tag{3.85}$$

where

$$J_2 = \iint_{\widetilde{\mathcal{B}}^2} \omega(x)\omega(y)\frac{1}{k^2}\mathrm{cov}^2\left(K_{h(x)}(x-X), K_{h(y)}(y-X)\right) dxdy.$$

We now compute J_2. We have

$$J_2 = J_{\mathrm{Diag}} + J_{\mathrm{Diag}^c} \tag{3.86}$$

where

$$J_{\mathrm{Diag}} = \iint_{\mathrm{Diag}} \omega(x)\omega(y)\frac{1}{k^2}\mathrm{cov}^2\left(K_{h(x)}(x-X), K_{h(y)}(y-X)\right) dxdy, \tag{3.87}$$
$$J_{\mathrm{Diag}^c} = \iint_{\mathrm{Diag}^c} \omega(x)\omega(y)\frac{1}{k^2}\mathrm{cov}^2\left(K_{h(x)}(x-X), K_{h(y)}(y-X)\right) dxdy, \tag{3.88}$$

We examine J_{Diag} and J_{Diag^c} separately.

Term J_{Diag^c}. We have outside the diagonal Diag: $\|x-y\| > \frac{1}{2}[h(x)+h(y)]$.

By Lemma 18, and using $\frac{\widetilde{c_A}}{c_h^\alpha} L h^\alpha \le p_0$ on $\widetilde{\mathcal{B}}$:

$$\begin{aligned}\frac{1}{k}\left|\operatorname{cov}\left(K_{h(x)}(x-X), K_{h(y)}(y-X)\right)\right| &\le \frac{1}{k}\left(C'_K L h(x)^\alpha + p(x)\right)\left(C'_K L h(y)^\alpha + p(y)\right)\\ &\le \frac{1}{k}\left[\left(\frac{C'_K c_h^\alpha}{\widetilde{c_A}}+1\right)p_0(x)+|\Delta(x)|\right]\left[\left(\frac{C'_K c_h^\alpha}{\widetilde{c_A}}+1\right)p_0(y)+|\Delta(y)|\right]\\ &=: \frac{1}{k}\left[C^{(\mathrm{c})}p_0(x)+|\Delta(x)|\right]\left[C^{(\mathrm{c})}p_0(y)+|\Delta(y)|\right],\end{aligned}$$

where $C^{(\mathrm{c})} = \frac{C'_K c_h^\alpha}{\widetilde{c_A}} + 1$. Therefore, outside the diagonal (3.75), we have:

$$\begin{aligned}J_{\text{Diag}^c} &= \frac{1}{k^2}\iint_{\text{Diag}^c}\omega(x)\omega(y)\operatorname{cov}^2\left(K_{h(x)}(x-X), K_{h(y)}(y-X)\right)dxdy\\ &\le \frac{1}{k^2}\iint_{\text{Diag}^c}\omega(x)\omega(y)\left[C^{(\mathrm{c})}p_0(x)+|\Delta(x)|\right]^2\left[C^{(\mathrm{c})}p_0(y)+|\Delta(y)|\right]^2 dxdy\\ &= \left[\frac{1}{k}\int_{\widetilde{\mathcal{B}}}\omega(x)\left(C^{(\mathrm{c})}p_0(x)+|\Delta(x)|\right)^2 dx\right]^2 \le \left[\frac{1}{k}\int_{\widetilde{\mathcal{B}}}\omega(x)\left({C^{(\mathrm{c})}}^2 p_0^2(x)+|\Delta(x)|^2\right)dx\right]^2\\ &\le \frac{4}{k^2}\left(C^{(\mathrm{c})}\,t_n + J\right)^2 \quad \text{by Lemma 13}\\ &\le \frac{16}{n^2}\left(\frac{C^{(\mathrm{c})}}{C'_b}+1\right)^2 J^2 =: \frac{C_{\text{Diag}^c}}{n^2}J^2. \end{aligned} \tag{3.89}$$

For J_{Diag}: If $\|x-y\| \le \frac{1}{2}[h(x)+h(y)] \le h(x)\vee h(y)$, we suppose by symmetry $h(y) = h(x)\vee h(y)$. Then by Lemma 19, we have:

$$\frac{1}{k}\left|\operatorname{cov}\left(K_{h(x)}(x-X), K_{h(y)}(y-X)\right)\right| \le C_{19}\frac{p(x)}{k\,h^d(x)}.$$

Therefore:

$$J_{\text{Diag}} = \frac{1}{k^2}\iint_{\widetilde{\mathcal{B}}^2}\omega(x)\omega(y)\operatorname{cov}^2\left(K_{h(x)}(x-X), K_{h(y)}(y-X)\right)dxdy \le \iint_{\text{Diag}^2}\omega(x)\omega(y)\left[C_{19}\frac{p(x)}{k\,h^d(x)}\right]^2 dxdy.$$

By Lemma 20, we can upper bound the term as:

$$\frac{J_{\text{Diag}}}{C_{19}^2} \le \iint_{\text{Diag}}\omega(x)\omega(y)\left[\frac{p(x)}{k\,h(x)^d}\right]^2 dxdy \quad \le \quad 2C^{(\omega)}\int_{\widetilde{\mathcal{B}}}\omega(x)^2 h(x)^d\left[\frac{p(x)}{k\;h(x)^d}\right]^2 dx$$

$$\leq \quad 2\frac{C^{(\omega)}}{c_h^d}\int_{\widetilde{\mathcal{B}}}\frac{\omega(x)^2}{k^2h(x)^d}[2p_0(x)^2+2\Delta^2(x)]dx \quad \leq \quad 16\,\frac{C^{(\omega)}}{c_h^d}t_n^2+16\frac{C^{(\omega)}}{c_h^d}\underbrace{\int_{\widetilde{\mathcal{B}}}\frac{\omega(x)^2\Delta^2(x)}{n^2h_b^d(x)}dx}_{\text{Term I}}. \tag{3.90}$$

Term I: We recall that by definition we have $p_0 \geq \widetilde{u_B}$ on $\widetilde{\mathcal{B}}$ and we have $2\alpha t - 2d - 4\alpha < 0$ since $t \in [1,2]$. Therefore:

$$\begin{aligned}
\frac{1}{k^2}\int_{\widetilde{\mathcal{B}}}\omega^2(x)\frac{\Delta(x)^2}{h_b^d(x)}dx &= \frac{1}{n^2}\int_{\widetilde{\mathcal{B}}}(n^2L^4\mathcal{I})^{\frac{d}{4\alpha+d}}\,p_0^{\frac{2\alpha t-2d-4\alpha}{(4-t)\alpha+d}}\,\omega(x)\Delta(x)^2dx\\
&\leq \frac{(n^2L^4\mathcal{I})^{\frac{d}{4\alpha+d}}}{n^2}\int_{\widetilde{\mathcal{B}}}\omega(x)\Delta(x)^2\,\widetilde{u_B}^{\frac{2\alpha t-2d-4\alpha}{(4-t)\alpha+d}}\,dx\\
&\leq \frac{(n^2L^4\mathcal{I})^{\frac{d}{4\alpha+d}}}{n^2}\int_{\widetilde{\mathcal{B}}}\omega(x)\Delta(x)^2\left[\widetilde{c_A}\frac{L^d}{n^{2\alpha}\mathcal{I}^\alpha}\right]^{\frac{1}{4\alpha+d}\frac{(4-t)\alpha+d}{(2-t)\alpha+d}\frac{2\alpha t-2d-4\alpha}{(4-t)\alpha+d}}\\
&= \widetilde{c}\,t_n\int_{\widetilde{\mathcal{B}}}\omega(x)\Delta(x)^2 = \widetilde{c}\,t_nJ.
\end{aligned}$$

where $\widetilde{c} = \widetilde{c_A}^{-\frac{1}{4\alpha+d}\frac{-2\alpha t+2d+4\alpha}{(2-t)\alpha+d}}$. Therefore, by equation (3.90):

$$J_{\text{Diag}} \leq C_{19}\left[16\frac{C^{(\omega)}}{c_h^d}t_n^2+16\frac{C^{(\omega)}}{c_h^d}\widetilde{c}t_nJ\right] =: A_{\text{Diag}}t_n^2+B_{\text{Diag}}t_nJ, \tag{3.91}$$

for two constants A_{Diag} and B_{Diag}. By equations (3.89) and (3.91), it holds:

$$J_2 \leq A_{J_2}\,t_n^2+B_{J_2}\,t_nJ+C_{J_2}\frac{J^2}{n^2}, \tag{3.92}$$

for three constants $A_{J_2}, B_{J_2}, C_{J_2} > 0$. Recalling Equation (3.82), we can further upper bound J_2 as $J_2 \leq A_{J_2}J^2/C_b'^2 + B_{J_2}\,J^2/C_b' + C_{J_2}\frac{J^2}{n^2}$, hence taking $n_{bulk} := \left\lceil\sqrt{C_b'}\right\rceil$, we get:

$$J_2 \leq \frac{A_{J_2}+B_{J_2}+C_{J_2}}{C_b'}J^2 =: \frac{D_{J_2}^2}{C_b'}J^2. \tag{3.93}$$

□

It then follows, from Equation (3.85):

$$\begin{aligned}
\mathbb{V}T_b &\leq \left[\left(\sqrt{J}+\sqrt{t_n}\right)^2+J_2^{1/2}\right]^2-\left(\sqrt{J}-\sqrt{t_n}\right)^4\\
&\leq \left[\left(\sqrt{J}+\sqrt{\frac{J}{C_b'}}\right)^2+\frac{D_{J_2}}{\sqrt{C_b'}}J\right]^2-\left(\sqrt{J}-\sqrt{J/C_b'}\right)^4 \quad \text{by Equation (3.82) and (3.93)}
\end{aligned}$$

$$
\begin{aligned}
&= J^2 \left\{ \left[1 + \frac{2 + D_{J_2}}{C'_b} + \frac{1^2}{C'_b} \right]^2 - \left[1 - 4\frac{1}{\sqrt{C'_b}} + O\left(\frac{1}{C'_b}\right) \right] \right\} \\
&\leq J^2 \left[\frac{8 + 2D_{J_2} + 1}{C'_b} \right] \quad \text{for } C'_b \text{ large enough} \\
&=: \frac{C_{\mathbb{V},H_1}}{C'_b} J^2.
\end{aligned}
$$

3.C.3 Proof of Corollary 3.1

Proof of Corollary 3.1.

1. Set $C_{\psi_b} > 1$. It holds:

$$
\begin{aligned}
\mathbb{P}_{p_0}\left(T_{bulk} > C_{\psi_b} t_n\right) &\leq \mathbb{P}_{p_0}\left(|T_{bulk} - \mathbb{E}T_{bulk}| > (C_{\psi_b} - 1)t_n\right) \quad \text{by Proposition 3.7} \\
&\leq \frac{C_{\mathbb{V},H_0} t_n^2}{(C_{\psi_b} - 1)^2 t_n^2} \quad \text{by Proposition 3.7 and Chebyshev's inequality} \\
&\leq \frac{\eta}{4} \quad \text{for } C_{\psi_b} \text{ larger than a suitable constant.}
\end{aligned}
$$

2. Assume C'_b is large enough to ensure

$$
(1 - 1/\sqrt{C'_b})^2 > \frac{C_{\psi_b}}{C'_b}. \tag{3.94}
$$

The value of the constant C'_b being given, assume moreover that $n \geq n_{bulk}$. We then have:

$$
\begin{aligned}
\mathbb{P}_p\left(T_{bulk} \leq C_{\psi_b} t_n\right) &\leq \mathbb{P}_p\left(T_{bulk} - \mathbb{E}T_{bulk} \leq C_{\psi_b} t_n - (1 - 1/\sqrt{C'_b})^2 J\right) \quad \text{by Proposition 3.8} \\
&\leq \mathbb{P}_p\left(T_{bulk} - \mathbb{E}T_{bulk} \leq \frac{C_{\psi_b}}{C'_b} J - (1 - 1/\sqrt{C'_b})^2 J\right) \quad \text{by Equation (3.82)} \\
&\leq \mathbb{P}_p\left(|T_{bulk} - \mathbb{E}T_{bulk}| \leq (1 - 1/\sqrt{C'_b})^2 J - \frac{C_{\psi_b}}{C'_b} J\right) \quad \text{by Equation (3.94)} \\
&\leq \frac{C_{\mathbb{V},H_1} J^2 / C'_b}{\left((1 - 1/\sqrt{C'_b})^2 - \frac{C_{\psi_b}}{C'_b}\right)^2 J^2} \quad \text{by Chebyshev's inequality} \\
&\leq \frac{\eta}{4} \quad \text{for } C'_b \text{ large enough.}
\end{aligned}
$$

□

Lemma 20. *For any $p \in \mathcal{P}(\alpha, L)$ it holds:* $\iint_{Diag} \omega(x)\omega(y) \left[\frac{p(x)}{k\, h(x)^d}\right]^2 dxdy \leq 2C^{(\omega)} \int_{\widetilde{\mathcal{B}}} \omega(x)^2 h(x)^d \left[\frac{p(x)}{k\, h(x)^d}\right]^2 dx.$

Proof of Lemma 20. We set:

$$\mathrm{Diag}_+ = \{(x,y) \in \mathrm{Diag} : p_0(x) \geq p_0(y)\}\,. \tag{3.95}$$

On Diag_+, we have $\|x-y\| \leq h(x) \vee h(y) = h(x)$ so in particular: $y \in B(x, h(x))$. We therefore have:

$$\iint_{\mathrm{Diag}} \frac{\omega(x)\omega(y)p^2(x)}{k^2\, h(x)^{2d}} dxdy = 2 \iint_{\mathrm{Diag}_+} \frac{\omega(x)\omega(y)p^2(x)}{k^2\, h(x)^{2d}} dxdy \leq 2 \int_{\widetilde{\mathcal{B}}} \omega(x) \left[\frac{p(x)}{k\, h(x)^d}\right]^2 \left[\int_{B(x,h(x))} \omega(y)dy\right] dx$$
$$\leq 2 \int_{x\in\widetilde{\mathcal{B}}} \omega(x) \left[\frac{p(x)}{k\, h(x)^d}\right]^2 \left\{h(x)^d C^{(\omega)}\omega(x)\right\} dx = 2C^{(\omega)} \int_{\widetilde{\mathcal{B}}} \omega(x)^2 h(x)^d \left[\frac{p(x)}{k\, h(x)^d}\right]^2 dx.$$

□

3.D Lower bound in the bulk regime: Proof of Proposition 3.2

We here define the bulk prior. Fix c a constant, allowed to be arbitrarily small. We apply Algorithm 3, with $\widetilde{\Omega}$, $\beta = \frac{2}{(4-t)\alpha+d}$, $u = u_B$, $c_\beta = c^{-\beta}\left(n^2L^4\mathcal{I}\right)^{\frac{\beta}{4\alpha+d}}$ and set $c_\alpha = c_A c^{-\alpha}$. Following the notation from Algorithm 3, let $h = (p_0/c_\beta)^{1/\beta}$. The choice of the constants ensures $h \leq ch_b$ and $p_0 \geq c_\alpha L h^\alpha$ over $\mathcal{B}(u_B)$. Moreover, c is chosen small enough to ensure $c_\alpha \geq \frac{\sqrt{d}^\alpha(2^{1-\alpha}\vee 1)}{1/2-c_\star}$. Since Ω is unbounded, we can moreover choose a subset $\widetilde{\Omega}$ large enough to ensure that it is split at least once by Algorithm 3. Therefore, the guarantees of Proposition 3.6 are ensured. Let $j \in \{1,\dots,N\}$ and consider the cell B_j. Its center is denoted by x_j and we also set $h_j = ch_b(x_j)/4$ where c is the constant used to define the constants c_α, c_β taken as inputs for Algorithms 3. We also set $\vec{\mathbb{1}} = (1,\dots,1)$. Define the perturbation function $f \geq 0$ over $\mathbb{R}^d$, such that $f \in H(\alpha,1)\cap C^\infty$, f is supported over $\left\{x \in \mathbb{R}^d : \|x\| < 1/2\right\}$. We define the perturbations $(\phi_j)_{j=1}^N$ as follows:

$$\phi_j(x) = C^{(\phi)} L h_j^\alpha\, f\left(\frac{x - x_j - \frac{h_j}{\sqrt{d}}\vec{\mathbb{1}}}{h_j}\right) - C^{(\phi)} L h_j^\alpha\, f\left(\frac{x - x_j + \frac{h_j}{\sqrt{d}}\vec{\mathbb{1}}}{h_j}\right), \tag{3.96}$$

where $C^{(\phi)}$ is a small enough constant. For $\epsilon = (\epsilon_1,\dots,\epsilon_N)$ where $\epsilon_j \overset{iid}{\sim} Rad(\frac{1}{2})$, the prior is defined as follows:

$$p_\epsilon^{(n)} = p_0 + \sum_{j=1}^N \epsilon_j \phi_j. \tag{3.97}$$

For clarity, we give the probability density over Ω^n corresponding to data drawn from this prior distribution. Assume that $(X_1',\dots,X_n')$ are drawn from (3.97). The data is therefore *iid* with the *same* density q, itself uniformly drawn in the set $\{p_\epsilon \mid \epsilon \in \{\pm 1\}^n\}$. In other words, the density of

$(X'_1, \dots, X'_n)$ corresponds to the mixture

$$\widetilde{p} = \frac{1}{2^N} \sum_{\epsilon \in \{\pm 1\}^N} \left(p_0 + \sum_{j=1}^{N} \epsilon_j \phi_j \right)^{\otimes n},$$

where, for any $q \in \mathcal{P}(\alpha, L)$, $q^{\otimes n}$ is defined by $q^{\otimes n}(x_1, \dots, x_n) = q(x_1) \dots q(x_n)$ and represents the density of $(Y_1, \dots, Y_n)$ when $Y_i \overset{iid}{\sim} q$.

The lower bound will be proved by showing that there exist no test with risk smaller than η for the testing problem $H_0 : (X'_1, \dots, X'_n) \sim p_0^{\otimes n}$ vs $H'_1 : (X'_1, \dots, X'_n) \sim \widetilde{p}$. Whenever no ambiguity arises, we will just write p_ϵ instead of $p_\epsilon^{(n)}$. Recalling that $L' = (1+\delta)L$ and $c'_\star = (1+\delta)c_\star$, the following proposition states that this prior is admissible, *i.e.* that each one of these densities belongs to $\mathcal{P}(\alpha, L', c'_\star)$.

Proposition 3.9. *For all $\epsilon = (\epsilon_1, \dots, \epsilon_N) \in \{\pm 1\}$: $p_\epsilon \in \mathcal{P}(\alpha, L', c'_\star)$.*

We now prove that this prior distribution gives a lower bound on ρ^*_{bulk}. This lower bound will be denoted by ρ^{LB}_{bulk}, defined as the L_t norm of the perturbation:

$$\rho^{LB}_{bulk} = \Big\| \sum_{j=1}^{N} \phi_j \Big\|_t. \tag{3.98}$$

Then by definition, $\forall \epsilon \in \{\pm 1\}^N : p_\epsilon \in H_1(\rho^{LB}_{bulk})$. Moreover, the following Proposition states that the prior we consider yields a lower bound of order ρ^*_{bulk}:

Proposition 3.10. *There exists a constant C^{LB}_{bulk} given in the Appendix, such that*

$$\rho^{LB}_{bulk} = C^{LB}_{bulk}\, \rho^*_{bulk}.$$

We now introduce the *Bayes risk* associated with the prior distribution (3.97):

Definition 3.2. *Define*

$$R_B^{bulk} = \inf_{\psi \ test} \left\{ \mathbb{P}_{p_0}(\psi = 1) + \mathbb{E}_\epsilon \Big[\mathbb{P}_{p_\epsilon}(\psi = 0) \Big] \right\},$$

where the expectation is taken with respect to the realizations of ϵ and $\mathbb{P}_{p_e}$ denotes the probability distribution when the data is drawn with density (3.97).

As classical in the minimax framework, we have $R^*(\rho^{LB}_{bulk}) \geq R_B^{bulk}$ (indeed, the supremum in (3.6) can be lower bounded by the expectation over ϵ). The following proposition states that ρ^{LB}_{bulk} is indeed a lower bound on ρ^*_{bulk}:

Proposition 3.11. *It holds $R_B^{bulk} > \eta$.*

Indeed, Proposition 3.11 proves that $R^*(\rho^{LB}_{bulk}) > \eta$. Since $R^*(\rho)$ is a decreasing function of ρ, we therefore have $\rho^* > \rho^{LB}_{bulk}$ by the definition of ρ^* in equation (3.7). This ends the proof of Proposition 3.2

3.D.1 Proof of Proposition 3.9

Proof of Proposition 3.9. First, for each $j = 1, \dots, N$ the functions $C^{(\phi)}L\left(h_j - \left\|x - x_j \pm \frac{h_j}{\sqrt{d}}\vec{\mathbb{1}}\right\|\right)_+^\alpha$ are in $H(\alpha, C^{(\phi)}L)$ and have disjoint support so that their sum also belongs to $H(\alpha, C^{(\phi)}L)$. Hence for all $\epsilon \in \{\pm 1\}^N$, $\sum_{j=1}^N \epsilon_j \phi_j \in H(\alpha, C^{(\phi)}L)$, proving that $p_\epsilon \in H(\alpha, (1 + C^{(\phi)})L)$.

Now, note that we have:

$$\forall x \in \mathcal{B}\left(\frac{u_B}{2}\right) : \quad Lh_b(x)^\alpha \leq \frac{2^{\frac{(2-t)\alpha+d}{(4-t)\alpha+d}}}{c_A} p_0(x). \tag{3.99}$$

Let $\epsilon \in \{\pm 1\}^N$ and $x \in B_j$, for some $j \in \{1, \dots, N\}$. Recalling that $h_j = c\, h_b(x_j)$ we have:

$$\begin{aligned}
|\phi_j(x)| &\leq C^{(\phi)}\|f\|_1 L h_j^\alpha = C^{(\phi)}\|f\|_1 c^\alpha L h_b(x_j)^\alpha \\
&\leq C^{(\phi)}\|f\|_1 c^\alpha C_{21}^{(p_0)} L h_b(x)^\alpha && \text{by Lemma 21} \\
&\leq \frac{2^{\frac{(2-t)\alpha+d}{(4-t)\alpha+d}}}{c_A} C^{(\phi)}\|f\|_1 c^\alpha C_{21}^{(p_0)} p_0(x) && \text{by equation (3.99)} \\
&=: \lambda\, p_0(x).
\end{aligned}$$

Therefore, by Lemma 22, we have $p + \sum \epsilon_j \phi_j \in \mathcal{P}(\alpha, (1+\lambda)L, \frac{c_\star + 2\lambda + \lambda c_\star}{1-\lambda})$. Choosing the constant λ small enough (by adjusting $C^{(\phi)}$), we can ensure $p_\epsilon \in \mathcal{P}(\alpha, L', c'_\star)$ where $L' = (1+\delta)L$ and $c'_\star = (1+\delta)c_\star$. □

3.D.2 Proof of Proposition 3.10

Proof of Proposition 3.10. Since the ϕ_j have disjoint support: $\left\|\sum_{j=1}^N \phi_j\right\|_t = \sum_{j=1}^N \left\|\phi_j\right\|_t$. Let $j \in \{1, \dots, N\}$. We have:

$$\begin{aligned}
\|\phi_j\|_t^t &= 2\int_{B_j} \left\{C^{(\phi)} L h_j^\alpha\, f\left(\frac{x - x_j - \frac{h_j}{\sqrt{d}}\vec{\mathbb{1}}}{h_j}\right)\right\}^t dx \\
&= 2\left(C^{(\phi)} L \|f\|_t\right)^t h_j^{\alpha t + d} \\
&\geq 2\left(C^{(\phi)} L \|f\|_t\right)^t \frac{c^{\alpha t}}{C_{21}^{(h)\alpha t}} \int_{B_j} h_b^{\alpha t} \quad \text{by Lemma 21} \\
&=: C_{bulk}^{LB} L^t \int_{B_j} h_b^{\alpha t}
\end{aligned}$$

where $C_{bulk}^{LB} = \frac{2\, c^{\alpha t}\, C^{(\phi)^t}}{C_{21}^{(h)^{\alpha t}}} \|f\|_t^t$, so that

$$\rho_{Bulk}^{LB}{}^t = C_{bulk}^{LB} L^t \int_{\cup_{j=1}^N B_j} h_b^{\alpha t} \geq C_{bulk}^{LB} L^t \int_{\mathcal{B}(u_B)} h_b^{\alpha t} = C_{bulk}^{LB} \overline{\rho_{bulk}^*}^t,$$

where

$$\overline{\rho_{bulk}^*} = \left(\frac{L^d}{n^{2\alpha} \mathcal{I}^\alpha}\right)^{\frac{t}{4\alpha+d}} \int_{\mathcal{B}(u_B)} p_0^r. \tag{3.100}$$

Now, if $u_B = u_{aux}$ then $\overline{\rho_{bulk}^*} = \rho_{bulk}^*$. Otherwise, if $u_B > u_{aux}$,we have by Lemma 2 and Lemma 6: $\rho_{tail}^* \asymp \overline{\rho_{tail}^*} \geq C_2 \rho_{bulk}^* \geq \overline{\rho_{bulk}^*}$. Therefore: $\overline{\rho_{bulk}^*} + \rho_{tail}^* \asymp \rho_{bulk}^* + \rho_{tail}^*$. □

3.D.3 Proof of Proposition 3.11

Proof of Proposition 3.11. As classical in the minimax literature we always have $R^*(\rho_{bulk}^{LB}) \geq R_B^{bulk} = 1 - d_{TV}(p_0^{\otimes n}, p_\epsilon^{(n)})$. Moreover, by Pinsker's inequality (see e.g. [188]) we have $d_{TV}(p_0^{\otimes n}, p_\epsilon^{(n)}) \leq \frac{1}{2}\sqrt{\chi^2\left(p_\epsilon^{(n)} || p_0^{\otimes n}\right)}$, therefore: $R_B^{bulk} \geq 1 - \frac{1}{2}\sqrt{\chi^2\left(p_\epsilon^{(n)} || p_0^{\otimes n}\right)}$. To prove that $R_B^{bulk} > \eta$, it therefore suffices to prove that $\chi^2\left(p_\epsilon^{(n)} || p_0^{\otimes n}\right) < 4(1-\eta)^2$. We recall that by Proposition 3.6 item 3, we have:

$$\mathcal{B}(u_B) \subset \bigcup_{j=1}^N B_j \subset \mathcal{B}(\frac{u_B}{2}). \tag{3.101}$$

We now compute $1 + \chi^2\left(p_\epsilon^{(n)} || p_0^{\otimes n}\right)$.

$$\begin{aligned}
1 + \chi^2\left(p_\epsilon^{(n)} || p_0^{\otimes n}\right) &= \int_{\Omega^n} \frac{\left(\frac{1}{2^N} \sum_{\epsilon \in \{\pm 1\}^N} \prod_{i=1}^n p_\epsilon(y_i)\right)^2}{\prod_{i=1}^n p_0(y_i)} dy_1 \dots dy_n \\
&= \frac{1}{4^N} \int \sum_{\epsilon, \epsilon' \in \{\pm 1\}^N} \prod_{i=1}^n \frac{p_\epsilon(y_i) p_{\epsilon'}(y_i)}{p_0(y_i)} dy_1 \dots dy_n \\
&= \frac{1}{4^N} \int_{\Omega^n} \sum_{\epsilon, \epsilon' \in \{\pm 1\}^N} \prod_{i=1}^n \frac{\left(p_0(y_i) + \sum_{j=1}^N \epsilon_j \phi_j(y_i)\right)\left(p_0(y_i) + \sum_{j=1}^N \epsilon'_j \phi_j(y_i)\right)}{p_0(y_i)} dy_1 \dots dy_n \\
&= \frac{1}{4^N} \sum_{\epsilon, \epsilon' \in \{\pm 1\}^n} \left(\int_\Omega p_0(x) + \sum_{j=1}^N (\epsilon_j + \epsilon'_j)\phi_j(x) + \sum_{j=1}^N \epsilon_j \epsilon'_j \frac{\phi_j^2(x)}{p_0(x)} dx\right)^n \\
&= \frac{1}{4^N} \sum_{\epsilon, \epsilon' \in \{\pm 1\}^n} \left(1 + \sum_{j=1}^N \epsilon_j \epsilon'_j \int_\Omega \frac{\phi_j^2(x)}{p_0(x)} dx\right)^n
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{4^N} \sum_{\epsilon,\epsilon' \in \{\pm 1\}^n} \exp\Big(n \sum_{j=1}^N \epsilon_j \epsilon'_j \int_\Omega \frac{\phi_j^2(x)}{p_0(x)} dx\Big) \\
&= \prod_{j=1}^N \left(\frac{1}{4} \sum_{\epsilon_j,\epsilon'_j \in \{\pm 1\}} \exp\Big(n\epsilon_j \epsilon'_j \int_\Omega \frac{\phi_j^2(x)}{p_0(x)} dx\Big) \right) = \prod_{j=1}^N \cosh\Big(n \int_\Omega \frac{\phi_j^2(x)}{p_0(x)} dx\Big) \\
&\leq \exp\left(\frac{1}{2} \sum_{j=1}^N n^2 \Big(\int_\Omega \frac{\phi_j^2(x)}{p_0(x)} dx \Big)^2 \right) \qquad (3.102) \\
&\leq \exp\left(\frac{1}{2} \sum_{j=1}^N n^2 C^{(\phi)^4} \widetilde{C} L^4 \int_{B_j} \frac{{h_b}^{4\alpha+d}}{p_0^2} \right) \qquad \text{by Lemma 23} \\
&\leq \exp\left(\frac{1}{2} n^2 C^{(\phi)^4} \widetilde{C} L^4 \frac{1}{n^2 L^4 \mathcal{I}} \int_{\mathcal{B}(\frac{u_B}{2})} p_0^r \right) \qquad \text{by equation (3.101)} \\
&\leq \exp\left(\frac{1}{2} n^2 C^{(\phi)^4} \widetilde{C} L^4 C_7 \frac{\mathcal{I}}{n^2 L^4 \mathcal{I}} \right) \qquad \text{by Lemma 7} \\
&= \exp\Big(C^{(\phi)^4} \widetilde{C} \Big) \ \leq \ 1 + 4(1-\eta)^2 \quad \text{for } C^{(\phi)} \leq \left(\frac{1}{\widetilde{C}} \log(1 + 4(1-\eta)^2) \right)^{\frac{1}{4}}.
\end{aligned}
$$

□

3.D.4 Technical results for the LB in the bulk regime

We recall that **in this section**, $c_\alpha = c_A c^{-\alpha}$, where c is a constant chosen small enough to ensure $c_\alpha \geq \frac{\sqrt{d}^\alpha (2^{1-\alpha} \vee 1)}{1/2 - c_\star}$.

Lemma 21. *Let $j \in \{1, \dots, M\}$ and $x \in B_j$. Denote by x_j the center of B_j. Then $\frac{p_0(x)}{p_0(x_j)} \in \Big[c_{21}^{(p_0)}, C_{21}^{(p_0)}\Big]$ where $c_{21}^{(p_0)} = \frac{1}{2}$ and $C_{21}^{(p_0)} = \frac{3}{2}$ are two constants. It follows that $\frac{h_b(x)}{h_b(x_j)} \in \Big[c_{21}^{(h)}, C_{21}^{(h)}\Big]$ where $c_{21}^{(h)} = \Big(\frac{1}{2}\Big)^{\frac{2}{(4-t)\alpha+d}}$ and $C_{21}^{(h)} = \Big(\frac{3}{2}\Big)^{\frac{2}{(4-t)\alpha+d}}$ are two constants.*

Proof of Lemma 21. The proof follows from Assumption ($\star$):

$$
\begin{aligned}
|p_0(x) - p_0(x_j)| &\leq c_\star p_0(x_j) + L(e(B_j)\sqrt{d})^\alpha \leq c_\star p_0(x_j) + L h^\alpha(x_j) \sqrt{d}^\alpha \\
&\leq \left(c_\star + \frac{\sqrt{d}^\alpha}{c_\alpha} \right) p_0(x_j) \ \leq \ \frac{p_0(x_j)}{2}.
\end{aligned}
$$

□

Lemma 22. *Let $p : \Omega \longrightarrow \mathbb{R}_+$ satisfying Assumption ($\star$). Let $\phi : \Omega \to \mathbb{R}$ in $H(\alpha, \mu L)$ for some constant $\mu > 0$ and such that $|\phi| \leq \lambda p$ over Ω for some constant $\lambda > 0$. Then*

$$
p + \phi \in \mathcal{P}\Big(\alpha,\ (1 + \lambda \vee \mu) L,\ \frac{c_\star + 2\lambda + \lambda c_\star}{1 - \lambda}\Big).
$$

Proof of Lemma 22. Clearly, $p+\phi \in H(\alpha,(1+\mu)L) \subset H(\alpha,(1+\mu\vee\lambda)L)$. Now, let $x,y\in\Omega$. By Assumption ($\star$) and the triangular inequality, we have:

$$\begin{aligned}
|p(x)+\phi(x)-p(y)-\phi(y)| &\le |p(x)-p(y)|+|\phi(x)|+|\phi(y)| \\
&\le c_\star p(x)+L\|x-y\|^\alpha+2\lambda p(x)+\lambda[p(y)-p(x)] \\
&\le (c_\star+2\lambda)p(x)+L\|x-y\|^\alpha+\lambda(c_\star p(x)+L\|x-y\|^\alpha) \\
&\le (c_\star+2\lambda+\lambda c_\star)p(x)+(1+\lambda)L\|x-y\|^\alpha \\
&\le \frac{c_\star+2\lambda+\lambda c_\star}{1-\lambda}[p(x)+\phi(x)]+(1+\lambda\vee\mu)L\|x-y\|^\alpha.
\end{aligned}$$

□

Lemma 23. *There exist two constants $c^{(\phi)}, C^{(\phi)}>0$ such that for all $j=1,\dots,N$:*

$$\Big(\int_\Omega \frac{\phi_j^2}{p_0}\Big)^2 \le \widetilde{C}\, C^{(\phi)4} L^4 \int_{B_j} \frac{{h_b}^{4\alpha+d}}{p_0^2},$$

where $\widetilde{C}$ is a constant given in the proof.

Proof of Lemma 23. Recall that ϕ_j is supported on B_j. By Lemma 21,

$$\begin{aligned}
\Big(\int_\Omega \frac{\phi_j^2}{p_0}\Big)^2 = \Big(\int_{B_j} \frac{\phi_j^2}{p_0}\Big)^2 &\le \int_{B_j}\frac{\phi_j^4}{p_0^2}h_b^d \cdot \int_{B_j}\frac{1}{h_b^d} \quad \text{by Cauchy-Schwarz' inequality} \\
&\le \frac{h_j^d}{p_0(x_j)^2}\frac{{C_{21}^{(h)}}^d}{c^d\, {c_{21}^{(p_0)}}^2}\int_{B_j}\phi_j^4 \times \frac{1}{{c_{21}^{(h)}}^d}\frac{c^d}{h_j^d}|B_j| \\
&\le \frac{h_j^d}{p_0(x_j)^2}\frac{{C_{21}^{(h)}}^d}{{c_{21}^{(p_0)}}^2 {c_{21}^{(h)}}^d}\int_{B_j}\phi_j^4. \qquad (3.103)
\end{aligned}$$

Moreover, by the change of variable $y=(x-x_j)/h_j$ we have

$$\int_{B_j}\phi_j^4 = 2\int_{\mathbb{R}^d}\Big\{C^{(\phi)}Lh_j^\alpha\ f(y)\Big\}^4 h_j^d dy = 2\Big(C^{(\phi)}L\Big)^4\|f\|_4^4\ h_j^{4\alpha+d}.$$

Injecting into (3.103) we get:

$$\begin{aligned}
\int_\Omega \frac{\phi_j^2}{p_0} &\le 2C^{(\phi)4}\frac{{C_{21}^{(h)}}^d}{{c_{21}^{(p_0)}}^2 {c_{21}^{(h)}}^d}\cdot\|f\|_4^4\ L^4\frac{h_j^{4\alpha+d}}{p_0(x_j)^2}h_j^d \\
&\le 2C^{(\phi)4}\frac{{C_{21}^{(h)}}^d}{{c_{21}^{(p_0)}}^2 {c_{21}^{(h)}}^d}\cdot\|f\|_4^4\ L^4\int_{B_j}\frac{h_b(x)^{4\alpha+d}}{p_0^2(x)}dx\ \frac{c^{4\alpha+d}{C_{21}^{(p_0)}}^2}{{c_{21}^{(h)}}^{4\alpha+d}} \\
&=: C^{(\phi)4}\widetilde{C}L^4\int_{B_j}\frac{{h_b}^{4\alpha+d}}{p_0^2}.
\end{aligned}$$

□

3.E Upper bound in the tail regime

In the tail regime, we show that the combination of the tests ψ_1 and ψ_2 has both type-I and type-II errors upper bounded by $\eta/4$ when $\left(\int_{\mathcal{T}} |p-p_0|^t\right)^{1/t} \geq C''\rho^*$ for some constant C''. We defer to Subsection 3.E.4 the technical results needed for proving this upper bound. We recall that $\mathcal{T} = \mathcal{T}(u_B)$ but that we place ourselves over a covering of $\mathcal{T}(\widetilde{u_B})$. **Until the end of the proof, whenever no ambiguity arises, we drop the indexation in $\bigcup_{j=1}^M \widetilde{C}_j$ and only write $\|p_0\|_1, \|p\|_1, \|\Delta\|_1$ to denote $\int_{\bigcup_{j=1}^M \widetilde{C}_j} p_0$, $\int_{\bigcup_{j=1}^M \widetilde{C}_j} p$, and $\int_{\bigcup_{j=1}^M \widetilde{C}_j} |\Delta|$ respectively. Moreover, in Appendix 3.E only, we will write h for $h_{tail}(u_B)$ when the tail dominates and $h = h_m$ when the bulk dominates.**

3.E.1 Under H_0

We here prove that $\psi_1 \vee \psi_2$ has a type-I error upper bounded by $\eta/2$, no matter whether the bulk or the tail dominates.

By Lemma 24, the type-I error of ψ_2 is upper bounded by

$$\mathbb{P}_{p_0}(\psi_2 = 1) \leq n^2 h^d \int_{\bigcup_{j\in\mathbb{N}^*} \widetilde{C}_j} p_0^2 \leq C_{24} \leq \frac{\eta}{8} \quad \text{taking } C_{24} \text{ small enough.} \tag{3.104}$$

As to the type-I error of ψ_1, we have under H_0:

$$\mathbb{E}\Big[\sum_{j\in\mathbb{N}^*} \frac{N_j}{n}\Big] = \|p_0\|_1 \qquad \text{and} \qquad \mathbb{V}\Big[\sum_{j\in\mathbb{N}^*} \frac{N_j}{n}\Big] \leq \frac{\|p_0\|_1}{n}.$$

Recalling that we write $\|p_0\|_1$ for $\int_{\bigcup_{j\in\mathbb{N}^*} \widetilde{C}_j} p_0$, we therefore have by Chebyshev's inequality:

$$\mathbb{P}_{p_0}\left[\left|\sum_{j\in\mathbb{N}^*} \frac{N_j}{n} - \|p_0\|_1\right| > C_{\psi_1}\sqrt{\frac{\|p_0\|_1}{n}}\right] \leq \frac{\eta}{4}. \tag{3.105}$$

for $C_{\psi_1} = 2\sqrt{2}/\sqrt{\eta}$. Combining (3.104) and (3.105), we conclude that $\psi_1 \vee \psi_2$ has type-I error upper bounded by $\eta/4$.

3.E.2 Under the alternative when the tail dominates

We now prove that when the tail dominates, $\rho^*_{tail} + \rho^*_r$ is an upper bound on the minimax separation radius. To do so, we show that when p is such that $\int_{\mathcal{T}} |p-p_0|^t \geq C''\left({\rho^*_{tail}}^t + {\rho^*_r}^t\right)$, one of the two

tests ψ_1 or ψ_2 rejects H_0, *whp*. Fix a density p satisfying:

$$\int_{\mathcal{T}} |p-p_0|^t \geq C'' {\rho^*_{tail}}^t \geq \widetilde{C''}\left(\frac{L^d}{n^{2\alpha}}\right)^{\frac{t-1}{\alpha+d}} \left(\int_{\mathcal{T}} p_0 + \frac{1}{n}\right)^{\frac{(2-t)\alpha+d}{\alpha+d}}$$
$$\geq \widetilde{C''}\left(\int_{\mathcal{T}} p_0 + \frac{1}{n}\right)^{2-t} \left[\frac{L^d}{n^{2\alpha}}\left(\int_{\mathcal{T}} p_0\right)^d\right]^{\frac{t-1}{\alpha+d}}. \tag{3.106}$$

where $\widetilde{C''} = C''/2^{\frac{(2-t)\alpha+d}{\alpha+d}}$ and C'' is a large enough constant.
Setting $u = 2-t$ and $v = t-1$ satisfying $u+v=1$ and $u+2v=t$, we have by Hölder's inequality:

$$\int_{\mathcal{T}} |p-p_0|^t = \int_{\mathcal{T}} |p-p_0|^{u+2v} \leq \left[\int_{\mathcal{T}} |p-p_0|\right]^u \left[\int_{\mathcal{T}} |p-p_0|^2\right]^v \leq \left[\int_{\bigcup_{j=1}^M \widetilde{C}_j} |p-p_0|\right]^u \left[\int_{\bigcup_{j=1}^M \widetilde{C}_j} |p-p_0|^2\right]^v.$$

Then by (3.106), one of the following two inequalities must hold:

$$\text{(i)} \quad \int_{\bigcup_{j=1}^M \widetilde{C}_j} |p-p_0| \;\geq\; C_1''\left(\int_{\mathcal{T}} p_0 + \frac{1}{n}\right)$$

$$\text{(ii)} \quad \int_{\bigcup_{j=1}^M \widetilde{C}_j} |p-p_0|^2 \geq\; C_2''\left[\frac{L^d}{n^{2\alpha}}\left(\int_{\mathcal{T}} p_0\right)^d\right]^{\frac{1}{\alpha+d}} = \frac{C_2''}{n^2 h_{tail}^d(u_B)}.$$

where C_1'' and C_2'' are two constants given in the proof, such that $C_1''C_2'' = \widetilde{C''}$.

First case: Suppose that (i) holds, i.e. $\|\Delta\|_1 \geq C_1''\left(\int_{\mathcal{T}} p_0 + \frac{1}{n}\right)$. We then have $\|\Delta\|_1 \geq C_1''\left(\frac{1}{n} + \frac{1}{1+C_8}\int_{\mathcal{T}(2u_B)} p_0\right) \geq C_{27}(\|p_0\|_1 + 1/n)$ for C_1'' large enough. Therefore, by Lemma 27, we have $\mathbb{P}_p(\psi_1 = 0) \leq \frac{\eta}{8}$.

Second case: Suppose (i) does not hold. Then (ii) holds. By Lemma 28, we can write:

$$\int_{\bigcup_{j=1}^M \widetilde{C}_j} (p-p_0)^2 \leq \frac{A_{28}}{h^d}\sum_{j=1}^M \left(\int_{\widetilde{C}_j} p\right)^2 + \frac{B_{28}}{n^2h^d} + C_{28}Lh^\alpha \int_{\bigcup_{j=1}^M \widetilde{C}_j} |p-p_0|,$$

Since (i) does not hold we can further upper bound this expression as:

$$\int_{\bigcup_{j=1}^M \widetilde{C}_j} (p-p_0)^2 \leq \frac{A_{28}}{h^d}\sum_{j=1}^M \left(\int_{\widetilde{C}_j} p\right)^2 + \frac{B_{28}}{n^2h^d} + C_{28}Lh^\alpha \cdot C_1'' \int_{\mathcal{T}} p_0$$
$$\leq \frac{A_{28}}{h^d}\sum_{j=1}^M \left(\int_{\widetilde{C}_j} p\right)^2 + \frac{B_{28}}{n^2h^d} + C_{28}\frac{C_1''}{n^2h^d}.$$

By (ii) we therefore have:

$$\frac{A_{28}}{h^d}\sum_{j=1}^{M}\left(\int_{\widetilde{C}_j} p\right)^2 + \frac{B_{28}}{n^2h^d} + \frac{C_{28}C_1''}{n^2h^d} \geq \frac{C_2''}{n^2h^d}$$

hence:

$$\frac{A_{28}}{h^d}\sum_{j=1}^{M}\left(\int_{\widetilde{C}_j} p\right)^2 \geq \frac{C_2'' - C_{28}C_1'' - B_{28}}{n^2h^d} =: \frac{A_{28}C_3''}{n^2h^d}$$

$$\text{i.e.} \quad \sum_{j=1}^{M}\left(\int_{\widetilde{C}_j} p\right)^2 \geq \frac{C_3''}{n^2} \tag{3.107}$$

Taking C_2'' large enough ensures that $C_3'' \geq C_{29}$, so that by Lemma 29, we have $\mathbb{P}_p(\psi_2 = 0) \leq \frac{\eta}{8}$.

3.E.3 Under $H_1(C''\rho^*_{bulk})$ when the bulk dominates

We now suppose that $C_{BT}\rho^*_{bulk} \geq \rho^*_{tail}$. Moreover, we suppose $\int_\Omega |p-p_0|^t \geq C''\left({\rho^*_{bulk}}^t + {\rho^*_r}^t\right)$ for C'' large enough. If $\int_{\mathcal{B}(u_B/2)} |p-p_0|^t \geq \frac{C''}{2}\left({\rho^*_{bulk}}^t + {\rho^*_r}^t\right)$, then for C'' large enough, $\mathbb{P}_p(\psi^*_{bulk} = 0) \leq \frac{\eta}{4}$ by the analysis of the upper bound. Therefore, *wlog*, suppose that $\int_{\mathcal{B}(u_B/2)} |p-p_0|^t \leq \frac{C''}{2}\left({\rho^*_{bulk}}^t + {\rho^*_r}^t\right)$, hence that $\int_{\mathcal{T}(u_B/2)} |p-p_0|^t \geq \frac{C''}{2}\left({\rho^*_{bulk}}^t + {\rho^*_r}^t\right)$.

Assume first that $\|p_0\|_1 \leq \frac{1}{n}$. Then we have $\|p\|_t \geq \left(\frac{C''}{2}\right)^{1/t}\rho^*_r - \|p_0\|_t \geq \left(\frac{C''}{2} - A_{49}(1)\right)\rho^*_r$ by Lemma 49. Taking C'' large enough imposes $\|p\|_1 \geq 2\frac{C_{27}}{n} \geq C_{27}\left(\frac{1}{n} + \|p_0\|_1\right)$ hence $\mathbb{P}_p(\psi_1 = 0) \leq \frac{\eta}{8}$ by Lemma 27.

Now, in the remaining of the proof, assume $\|p_0\|_1 > \frac{1}{n}$. Again, there are two cases.

First case: If $\|\Delta\|_1 \geq 2C_{27}\|p_0\|_1 \geq C_{27}(\|p_0\|_1 + \frac{1}{n})$, then by Lemma 27: $\mathbb{P}_p(\psi_1 = 0) \leq \frac{\eta}{8}$.

Second case: Assume now that $\|\Delta\|_1 \leq 2C_{27}\|p_0\|_1$, hence that $\|p\|_1 \leq (2C_{27}+1)\|p_0\|_1$. By Assumption $(\star)$, the definition of h_m from (3.24) and the choice of c_m, we can immediately check that $\bigcup_{j\in\mathbb{N}^*}\widetilde{C}_j \subset \mathcal{T}(u_B)$. Hence $\|p\|_1 \leq (2C_{27}+1)\int_{\mathcal{T}(u_B)} p_0 = (2C_{27}+1)\frac{1}{Lh^{\alpha+d}_{tail}}$. We can now lower bound $\sum_{j\in\mathbb{N}^*} n^2q_j^2$ using Lemma 28:

$$\int_{\bigcup_{j=1}^M \widetilde{C}_j}(p-p_0)^2 \leq \frac{A_{28}}{h_m^d}\sum_{j\in\mathbb{N}^*}\left(\int_{\widetilde{C}_j} p\right)^2 + \frac{B_{28}}{n^2h_m^d} + C_{28}(2C_{27}+1)\frac{Lh_m^\alpha}{Lh_{tail}^{\alpha+d}} \tag{3.108}$$

$$\leq \frac{A_{28}}{h_m^d}\sum_{j\in\mathbb{N}^*}\left(\int_{\widetilde{C}_j} p\right)^2 + \left(B_{28} + \frac{C_{28}(2C_{27}+1)}{{C^{(2)}_{BT}}^{\alpha+d}}\right)\frac{1}{n^2h_m^d} \quad \text{by Lemma 10} \tag{3.109}$$

$$=: \frac{A_{\psi_2}}{h_m^d} \sum_{j\in\mathbb{N}^*} \left(\int_{\widetilde{C}_j} p \right)^2 + \frac{B_{\psi_2}}{n^2 h_m^d} \tag{3.110}$$

where A_{ψ_2} and B_{ψ_2} are two constants. We recall that **in this section,** we respectively denote by $\|\Delta\|_2^2$ and $\|\Delta\|_t^t$ the quantities $\int_{\bigcup_{j=1}^M \widetilde{C}_j} (p-p_0)^2$ and $\int_{\bigcup_{j=1}^M \widetilde{C}_j} |p-p_0|^t$. We now lower bound the term $\|\Delta\|_2^2$. By Hölder's inequality:

$$\begin{aligned}
\|\Delta\|_2^2 &\geq \left(\|\Delta\|_t^t \|\Delta\|_1^{t-2}\right)^{\frac{1}{t-1}} \geq \left(\frac{C''}{2} {\rho^*_{bulk}}^t \left\{ (2C_{27}+1)\|p_0\|_1 \right\}^{t-2} \right)^{\frac{1}{t-1}} \\
&= \left(\frac{C''}{2} {\rho^*_{bulk}}^t \left\{ (2C_{27}+1) \left(\frac{{\rho^*_{tail}}^t}{\widetilde{L}^{\frac{t-1}{\alpha+d}}} \right)^{\frac{\alpha+d}{(2-t)\alpha+d}} \right\}^{t-2} \right)^{\frac{1}{t-1}} \\
&\geq \left(\frac{C''}{2} {\rho^*_{bulk}}^t \left\{ (2C_{27}+1) \left(\frac{C_{BT}^t {\rho^*_{bulk}}^t}{\widetilde{L}^{\frac{t-1}{\alpha+d}}} \right)^{\frac{\alpha+d}{(2-t)\alpha+d}} \right\}^{t-2} \right)^{\frac{1}{t-1}} \quad \text{recalling } t-2 \leq 0 \\
&=: C_\Delta {\rho^*_{bulk}}^{\frac{td}{(2-t)\alpha+d}} \widetilde{L}^{\frac{2-t}{(2-t)\alpha+d}},
\end{aligned}$$

where C_Δ is a constant can be made arbitrarily large by choosing C'' large enough. We therefore have $h_m^d \|\Delta\|_2^2 \geq c_m C_\Delta$, hence combining with equation (3.110), we get

$$n^2 \sum_{j\in\mathbb{N}^*} q_j^2 \geq \frac{1}{A_{\psi_2}} \left(c_m C_\Delta - B_{\psi_2} \right) \geq C_{29},$$

by choosing C'' large enough, which yields $\mathbb{P}_p(\psi_2 = 0) \leq \frac{\eta}{8}$.

3.E.4 Technical results

Lemma 24. *The following result holds no matter whether the bulk or the tail dominates. Under H_0, the probability that at least one of the cells $(\widetilde{C}_j)_{j=1,\dots,M}$ contains at least two observations is upper bounded as*

$$\mathbb{P}_{p_0}[\,\exists j \in \mathbb{N}^* : N_j \geq 2\,] \leq n^2 h^d \int_{\bigcup_{j\in\mathbb{N}^*} \widetilde{C}_j} p_0^2 \leq C_{24},$$

where C_{24} is a constant which can be made arbitrarily small by choosing C_{BT} large enough.

Proof of Lemma 24. We place ourselves under H_0. For all $j \in \mathbb{N}^*$, let $p_j = \int_{\widetilde{C}_j} p_0$. By the definition of $N_j = \sum_{i=1}^n \mathbb{1}\{X_i \in \widetilde{C}_j\}$, we have $N_j \sim Bin(p_j, n)$ for all $j = 1, \dots, M$. Therefore the probability that for a fixed j we have $N_j \geq 2$ is upper bounded as:

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

The probability that at least one of the N_j is at least 2 is therefore upper bounded by $\sum_{j\in\mathbb{N}^*} n^2 p_j^2$. Now, by the Cauchy-Schwarz inequality:

$$\sum_{j\in\mathbb{N}^*} n^2 p_j^2 = \sum_{j\in\mathbb{N}^*} n^2 \left(\int_{\widetilde{C}_j} p_0\right)^2 \leq \sum_{j\in\mathbb{N}^*} n^2 h^d \int_{\widetilde{C}_j} p_0^2 = n^2 h^d \int_{\bigcup_{j\in\mathbb{N}^*} \widetilde{C}_j} p_0^2.$$

If the tail dominates, then Lemma 9 proves that the last quantity is at most C_9. Otherwise, since $\bigcup_{j\in\mathbb{N}^*} \widetilde{C}_j \subset \mathcal{T}(u_B)$, the RHS can be further upper bounded by Lemma 5 as $\frac{\bar{C} n^2 h^d}{n^2 h_{tail}^d(u_B)} \leq \bar{C}(C_{BT}^{(2)})^d$ by Lemma 10. In both cases, the constant upper bounding the RHS can be made arbitrarily small by choosing $\bar{C}$ large enough. □

Lemma 25. *If the tail dominates, i.e. if $\rho_{tail}^* \geq C_{BT}\rho_{bulk}^*$ where C_{BT} is defined in Lemma 10, then it holds $\bigcup_{j\in\mathbb{N}^*} \widetilde{C}_j \subset \mathcal{T}(2u_B)$, provided that C_{BT} is larger than a constant.*

Proof of Lemma 25. Let $y \in \bigcup_{j\in\mathbb{N}^*} \widetilde{C}_j$ and $x \in \mathcal{T}(u_B)$ such that x and y belong to the same cell $\widetilde{C}_j$. By Assumption ($\star$) and Lemma 10 we have:

$$p_0(y) \leq (1+c_\star)p_0(x) + L(h\sqrt{d})^\alpha \leq (1+c_\star)p_0(x) + L\sqrt{d}^\alpha \inf_{x\in\mathcal{B}} h_b(x)^\alpha \leq 2u_B.$$

□

Lemma 26. *Recall that $\|\Delta\|_1 = \int_{\mathcal{T}(u_B)} |\Delta|$ and $\|p_0\|_1 = \int_{\mathcal{T}(u_B)} |p_0|$. If $\|\Delta\|_1 \geq 3\|p_0\|_1$, then $\left|\int_{\mathcal{T}(u_B)} \Delta\right| \geq \frac{1}{2}\|\Delta\|_1$.*

Proof of Lemma 26. Define $J_+ = \{x \in \mathcal{T} : p(x) \geq p_0(x)\}$ and $J_- = \{x \in \mathcal{T} : p(x) < p_0(x)\}$. Define also:

$$s = \frac{\int_{\mathcal{T}} \Delta}{\int_{\mathcal{T}} p_0}, \qquad s_+ = \frac{\int_{J_+} \Delta}{\int_{\mathcal{T}} p_0}, \qquad s_- = -\frac{\int_{J_-} \Delta}{\int_{\mathcal{T}} p_0}$$

Then by assumption: $s_+ - s_- = s \geq 3$. Moreover, $s_- = \frac{\int_{J_-} p_0 - p}{\int_{\mathcal{T}} p_0} \leq 1$. Thus, $s_+ \geq 3 \geq 3s_-$ so that $2(s_+ - s_-) \geq s_+ + s_-$, which yields the result. □

Lemma 27. *The following result holds no matter whether the bulk or the tail dominates. There exists a constant C_{27} such that, whenever $\|\Delta\|_1 \geq C_{27}(\|p_0\|_1 + 1/n)$, then $\mathbb{P}_p(\psi_1 = 0) \leq \frac{\eta}{8}$.*

Proof of Lemma 27. Choose $C_{27} \geq 10$ and $c_{tail} \geq 1$ so that by the triangular inequality and recalling $\int_{\mathcal{T}(u_B)} p_0 \geq \frac{c_{tail}}{n}$ we have: $\|p\|_1 + \|p_0\|_1 \geq 5\|p_0\|_1$, hence $\|\Delta\|_1 \geq \int p - \int p_0 \geq 3\|p_0\|_1$. Therefore, the assumptions of Lemma 26 are met.

$$\mathbb{P}_p(\psi_1 = 0) = \mathbb{P}_p\Big(\Big|\sum_{j\in\mathbb{N}^*} \frac{N_j}{n} - \|p_0\|_1\Big| \leq C_{\psi_1}\sqrt{\frac{\|p_0\|_1}{n}}\Big)$$

$$
\begin{aligned}
&\leq \mathbb{P}_p\Big(\Big|\int_{\bigcup_{j\in\mathbb{N}^*}\widetilde{C}_j} p - p_0\Big| - \Big|\sum_{j\in\mathbb{N}^*}\frac{N_j}{n} - \|p\|_1\Big| \leq C_{\psi_1}\sqrt{\frac{\|p_0\|_1}{n}}\Big) \text{ by the triangular inequality}\\
&\leq \mathbb{P}_p\Big(\frac{1}{2}\|\Delta\|_1 - C_{\psi_1}\sqrt{\frac{\|p_0\|_1}{n}} \leq \Big|\sum_{j\in\mathbb{N}^*}\frac{N_j}{n} - \|p\|_1\Big|\Big) \quad \text{by Lemma 26}\\
&\leq \frac{\frac{1}{n}\|p\|_1}{\Big(\frac{1}{2}\|\Delta\|_1 - C_{\psi_1}\sqrt{\frac{\|p_0\|_1}{n}}\Big)^2} \quad \text{by Chebyshev's inequality}\\
&\leq \frac{\|p\|_1/n}{\Big(\frac{1}{2}\|p\|_1 - \frac{1}{2}\|p_0\|_1 - C_{\psi_1}\sqrt{\frac{\|p_0\|_1}{n}}\Big)^2} \text{ by the triangular inequality}\\
&\leq \frac{\|p\|_1/n}{\Big(\frac{1}{2}\|p\|_1 - \frac{1}{2}\|p_0\|_1 - C_{\psi_1}(\|p_0\|_1 + 1/n)\Big)^2} \text{ using } \sqrt{xy}\leq x+y\\
&\leq \frac{\|p\|_1/n}{\Big(\frac{1}{2}\|p\|_1 - (C_{\psi_1}+1)(\|p_0\|_1 + 1/n)\Big)^2}.
\end{aligned}
$$

Choose $C_{27} \geq 4(C_{\psi_1}+1)+1$, so that the quantity $\frac{1}{2}\|p\|_1 - (C_{\psi_1}+1)(\|p_0\|_1+1/n)$ is strictly positive. This ensures that all of the above operations are valid. Now set $z = (C_{\psi_1}+1)(\|p\|_1 + 1/n)$. The function $f : x \mapsto \frac{x}{n(x/2-z)^2}$ is decreasing over $(2z,\infty)$. For $x \geq 20z/\eta$, since $nz > 1$ and $\eta \leq 1$, we have:

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

which proves that, whenever $\|p\|_1 \geq \frac{20}{\eta}(C_{\psi_1}+1)(\|p_0\|_1+1/n)$, we have $\mathbb{P}_p(\psi_1 = 0) \leq \eta/4$. This condition is guaranteed whenever $\|\Delta\|_1 \geq \Big(1+\frac{20}{\eta}(C_{\psi_1}+1)\Big)(\|p_0\|_1+1/n) = C_{27}(\|p_0\|_1+1/n)$ for $C_{27} = 1 + \frac{20}{\eta}(C_{\psi_1}+1)$. □

Lemma 28. *We have:*

$$
\int_{\bigcup_{j\in\mathbb{N}^*}\widetilde{C}_j}(p-p_0)^2 \;\leq\; \frac{A_{28}}{h^d}\sum_{j\in\mathbb{N}^*}\Big(\int_{C_j} p\Big)^2 \;+\; \frac{B_{28}}{n^2h^d} \;+\; C_{28}Lh^\alpha\int_{\bigcup_{j\in\mathbb{N}^*}\widetilde{C}_j}|p-p_0|,
$$

where A_{28}, B_{28}, C_{28} are constants given in the proof.

Proof of Lemma 28. Let $j \in \{1,\dots,M\}$. Assume that each cube $\widetilde{C}_j$ is centered at x_j. By Assumption $(\star)$ we have for all $j = 1,\dots,M$ and $x \in \widetilde{C}_j$:

$$
p(x) \leq (1+c_\star)p(x_j) + L(h\sqrt{d})^\alpha, \tag{3.111}
$$

hence by exchanging x and x_j and integrating:

$$p(x_j) \leq \frac{1+c_\star}{h^d} \int_{\widetilde{C}_j} p + L(h\sqrt{d})^\alpha, \tag{3.112}$$

and by equations (3.111) and (3.112), we have:

$$p(x) \leq \frac{(1+c_\star)^2}{h^d} \int_{\widetilde{C}_j} p + (2+c_\star)L(h\sqrt{d})^\alpha. \tag{3.113}$$

Therefore, fixing any $j \in \mathbb{N}^*$ it holds:

$$\begin{aligned}\int_{\widetilde{C}_j} p^2 &\leq \int_{\widetilde{C}_j} p(x)dx \left[\frac{(1+c_\star)^2}{h^d} \int_{\widetilde{C}_j} p + (2+c_\star)L(h\sqrt{d})^\alpha\right] \\ &= \frac{(1+c_\star)^2}{h^d} \left(\int_{\widetilde{C}_j} p\right)^2 + (2+c_\star)L(h\sqrt{d})^\alpha \int_{\widetilde{C}_j} p. \end{aligned} \tag{3.114}$$

Now, we have for all $j \in \mathbb{N}^*$:

$$\int_{\widetilde{C}_j} (p-p_0)^2 \leq 2\int_{\widetilde{C}_j} p^2 + 2\int_{\widetilde{C}_j} p_0^2 \leq 2\frac{(1+c_\star)^2}{h^d} \left(\int_{\widetilde{C}_j} p\right)^2 + 2(2+c_\star)L(h\sqrt{d})^\alpha \int_{\widetilde{C}_j} p + 2\int_{\widetilde{C}_j} p_0^2,$$

and summing for $j \in \mathbb{N}^*$:

$$\int_{\bigcup_{j\in\mathbb{N}^*} \widetilde{C}_j} (p-p_0)^2 \leq A_{28} \sum_{j\in\mathbb{N}^*} \left(\int_{\widetilde{C}_j} p\right)^2 + C_{28}Lh^\alpha \|p\|_1 + \frac{C_9}{n^2h^d}.$$

Now,

$$Lh^\alpha\|p\|_1 \leq Lh^\alpha\left(\|\Delta\|_1 + \|p_0\|_1\right) \leq Lh^\alpha\|\Delta\|_1 + \frac{C_8}{n^2h^d}.$$

Therefore, setting $B_{28} = C_8 + C_9$ yields the result. □

Lemma 29. *The following result holds no matter whether the bulk or the tail dominates. Assume that $\sum_j n^2q_j^2 \geq C_{29}$ where C_{29} is a large constant and $q_j = \int_{\widetilde{C}_j} p$ for all j. Then $\mathbb{P}_p(\psi_2 = 0) \leq \frac{\eta}{8}$.*

Proof. We draw $\widetilde{k} \sim \mathrm{Poi}(k)$ where we recall that $k = \frac{n}{2}$. We consider the setting where we observe $\widetilde{X}_1, \ldots, \widetilde{X}_{\widetilde{k}}$ iid drawn from the density p and define $\forall j \in \mathbb{N}^*, N_j' = \sum_{i=1}^{\widetilde{k}} \mathbb{1}_{\widetilde{X}_i=j}$ the histogram of the tail in this modified setting. We recall that by the classical poissonization trick, the random variables $(N_j')_j$ are independent and distributed as $\mathrm{Poi}(kq_j)$ respectively. We first notice that

$$\begin{aligned}\mathbb{P}_{p^{\otimes\widetilde{k}}}(\forall j \in \mathbb{N}^* : N_j' = 0 \text{ or } 1) &\geq \mathbb{P}_{p^{\otimes\widetilde{k}}}(\forall j \in \mathbb{N}^* : N_j' = 0 \text{ or } 1 | \widetilde{k} \leq n)\mathbb{P}(\widetilde{k} \leq n) \\ &\geq \mathbb{P}_{p^{\otimes n}}(\forall j \in \mathbb{N}^* : N_j' = 0 \text{ or } 1)\mathbb{P}(\widetilde{k} \leq n) \end{aligned} \tag{3.115}$$

Moreover,

$$\mathbb{P}_{p^{\otimes \widetilde{k}}}(\forall j \in \mathbb{N}^* : N_j = 0 \text{ or } N_j = 1) \;=\; \prod_{j\in\mathbb{N}^*} e^{-kq_j}\,(1+kq_j)\,.$$

Let $I_- = \{j \in \mathbb{N}^* : kq_j \le \frac{1}{2}\}$ and $I_+ = \{j \in \mathbb{N}^* : kq_j > \frac{1}{2}\}$. Recall that for $x \in (0, 1/2]$, $\log(1+x) \le x - x^2/3$. Then, for $j \in I_-$:

$$e^{-kq_j}\,(1+kq_j) = \exp\{-kq_j + \log(1+kq_j)\} \le \exp\left(-\frac{k^2q_j^2}{3}\right)$$

Now, for $j \in I_+$, we have: $-kq_j + \log(1+kq_j) \le -kq_j + \log(1+kq_j) \le -\frac{1}{10}kq_j$ using the inequality $-0.9x + \log(1+x) \le 0$ true for all $x \ge \frac{1}{2}$. Therefore, we have upper bounded the type-II error of ψ_2 by:

$$\begin{aligned}
\mathbb{P}_{p^{\otimes \widetilde{k}}}(\forall j \in&\mathbb{N}^* : N_j = 0 \text{ or } N_j = 1) \le \exp\left(-\frac{1}{3}\sum_{j\in I_-} k^2q_j^2 - \frac{1}{10}\sum_{j\in I_+} kq_j\right)\\
&\le \exp\left(-\frac{1}{3}\sum_{j\in I_-} k^2q_j^2 - \frac{1}{10}\Big(\sum_{j\in I_+} k^2q_j^2\Big)^{1/2}\right)\\
&= \exp\left(-\frac{1}{3}(S - S_+) - \frac{1}{10}\,(S_+)^{1/2}\right) \text{ for } S = \sum_{j\in\mathbb{N}^*} k^2q_j^2 \text{ and } S_+ = \sum_{j\in I_+} k^2q_j^2.
\end{aligned}$$

Now, $S_+ \mapsto -\frac{S}{3} + \frac{1}{3}S_+ - \frac{\sqrt{S_+}}{10}$ is convex over $[0, S]$ so its maximum is reached on the boundaries of the domain and is therefore equal to $(-\frac{\sqrt{S}}{10}) \vee -\frac{S}{3} = -\frac{\sqrt{S}}{3}$ for $S \ge 9/100$. Now, since $\|q\|_2^2 \ge 4C_{29}/k^2 \ge C_{29}/k^2$, we have $S = k^2\|q\|_2^2 \ge \log(16/\eta)^2 \vee 9/100$ which ensures $\mathbb{P}_{p^{\otimes \widetilde{k}}}(\forall j \in \mathbb{N}^* : N_j = 0 \text{ or } N_j = 1) \le \eta/16$, hence, by equation (3.115), $\mathbb{P}_p(\psi_2 = 0) \le \frac{\eta}{16}/\mathbb{P}(\widetilde{k} \le n) \le \frac{\eta}{8}$ if n is larger than a constant. □

3.F Lower bound in the tail regime

Recall that $p_b^{(n)} = \frac{q_b}{\|q_b\|_1}$ where $q_b := p_0 + \sum_{j\ge U}\left[b_j\gamma_j^{(\uparrow)} - (1-b_j)\gamma_j^{(\downarrow)}\right]$ and that $p_j = \int_{\widetilde{C}_j} p_0$.

We now give the precise definitions of $(\gamma_j^{(\uparrow)})_j$ and $(\gamma_j^{(\downarrow)})_j$. The perturbations $(\gamma_j^{(\downarrow)})_j$ are designed to guarantee the following condition: $\forall j \ge U, \int_{\widetilde{C}_j}\gamma_j^{(\downarrow)} \ge c\,p_j$ for some small constant $c > 0$. To do this, we split $\widetilde{C}_j$ into smaller cells $(E_l^{(j)})_{l=1}^{M_j}$ on which p_0 can be considered as "approximately constant", in the sense that $\max_{E_l^{(j)}} p_0 / \max_{E_l^{(j)}} p_0 \in [c', c'']$ for two constants $c', c'' > 0$. By Assumption ($\star$), this condition is satisfied if all $E_l^{(j)}$ have edge length $\asymp \left(\frac{p_0(x)}{L}\right)^{1/\alpha}$. We now remove on each $E_l^{(j)}$ a small deterministic function $\phi_l^{(j)}$ whose total mass is at least $c\int_{E_l^{(j)}} p_0$. The role of the $\gamma_j^{(\downarrow)}$ is therefore to remove a small fraction of the mass of p_0 on each cell where $b_j = 0$. To formally

define $\gamma_j^{(\downarrow)}$, we first let for all $j \geq U$:

$$u_j = \inf\Big\{u > 0 : \int_{\widetilde{C}_j} \mathbb{1}_{p_0(x)\geq u}\, p_0 \;\geq \frac{1}{2} p_j\Big\}, \tag{3.116}$$

$$D_j = \;\Big\{x \in \widetilde{C}_j : p_0(x) \geq u_j\Big\}. \tag{3.117}$$

We therefore apply Algorithm 3 with inputs $\widetilde{\Omega} = \widetilde{C}_j$, $\beta = \alpha$, $u = u_j$ and $c_\beta = c'_\beta L$ for some large constant c'_β, and we set $c_\alpha = c'_\beta$. Taking c'_β large enough ensures $c_\star + \frac{\sqrt{d}^\alpha}{c_\alpha}(2^{1-\alpha} \vee 1) \leq 1/2$, hence, the guarantees of Proposition 3.6 are satisfied. For each cube $\widetilde{C}_j$, Algorithm 3 defines the family of smaller cells $(E_1^{(j)}, \dots, E_{M_j}^{(j)})$ for some $M_j \in \mathbb{N}$. We denote the center of each cube E_l^j by $z_l^{(j)} \in \widetilde{C}_j$ and its edge length by $h_l^{(j)} \asymp \left(\frac{1}{L\, c'_\beta} p_0(z_l^{(j)})\right)^{1/\alpha}$. Moreover, each cube has non empty intersection with D_j and $D_j \subset \bigcup_{l=1}^{M_j} E_l^{(j)}$. For some constant $c^{(\downarrow)}$ small enough, define on each cell $E_l^{(j)}$:

$$\phi_l^{(j)}(x) = c^{(\downarrow)} L \Big(h_l^{(j)}\Big)^\alpha f\Big(\frac{x - z_l^{(j)}}{h_l^{(j)}}\Big), \tag{3.118}$$

where we recall that $f \geq 0$ over $\mathbb{R}^d$, $f \in H(\alpha, 1) \cap C^\infty$, and f is supported over $\Big\{x \in \mathbb{R}^d : \|x\| < 1/2\Big\}$. We here moreover assume that f satisfies

$$\forall x, y \in \mathbb{R}^d : |f(x) - f(y)| \leq c_\star f(x) + \|x - y\|^\alpha. \tag{3.119}$$

The perturbation $\gamma_j^{(\downarrow)}$ is defined as:

$$\gamma_j^{(\downarrow)} = \sum_{l=1}^{M_j} \phi_l^{(j)}. \tag{3.120}$$

We now move to the definition of $\gamma_j^{(\uparrow)}$. Assuming that each cube $\widetilde{C}_j$ is centered at z_j, $\gamma_j^{(\uparrow)}$ is defined as:

$$\gamma_j^{(\uparrow)}(x) = c_j^{(\uparrow)} L h^\alpha \; f\left(\frac{x - z_j}{h}\right), \quad j \geq U, \tag{3.121}$$

where $h := h_{tail}(u_B)$ is defined in (3.23) and $c_j^{(\uparrow)}$ is chosen so as to ensure that $\pi_j \int \gamma_j^{(\uparrow)} = (1 - \pi_j) \int \gamma_j^{(\downarrow)}$. In other words, $c_j^{(\uparrow)}$ is chosen so that the total mass of the prior is equal to $\int_{D(U)} p_0$ in expectation over the $(b_j)_{j\geq U}$. Noticeably, when setting $c^{(\uparrow)} := c_u c^{(\downarrow)}$, Proposition 3.12 shows that $\forall j \geq U, c_j^{(\uparrow)} \in \Big[c^{(\uparrow)},\; 2c^{(\uparrow)}\Big]$ i.e. $c_j^{(\uparrow)}$ is lower- and upper bounded by two strictly positive constants.

The functions $(\gamma_j^{(\downarrow)})_{j\geq U}$ and $(\gamma_j^{(\uparrow)})_{j\geq U}$ are chosen to ensure the following properties:

Proposition 3.12. *1. For all $(b_j)_{j\geq U}$: $p_b^{(n)} \in \mathcal{P}(\alpha, L', c'_\star)$ over the whole domain $[0,1]^d$.*

2. *There exists a constant* $C^{(\downarrow)} > 0$ *independent of* p_0 *such that for all* $j \geq U$: $C^{(\downarrow)} \int_{\widetilde{C}_j} \gamma_j^{(\downarrow)} \geq \int_{\widetilde{C}_j} p_0$ *where* $C^{(\downarrow)} = c^{(\downarrow)}(1 - c_\star)\|f\|_1$.

For clarity, we now give the the probability density over the space Ω^n of the data when they are generated from prior (3.31). Assume that we observe $(X_1'', \ldots, X_n'')$ generated from $p_b^{(n)}$. Then $X_1'', \ldots, X_n''$ are all *iid* with the *same* density q, which is itself (not uniformly) drawn in the set $\{p_b \mid b_j \in \{0,1\}\ \forall j \geq U\}$. In other words, the density of $(X_1'', \ldots, X_n'')$ corresponds to the mixture

$$\overline{p}^{(n)} = \sum_{\substack{b_j \in \{0,1\} \\ j \geq U}} \prod_{j \geq U} \pi_j^{b_j}(1-\pi_j)^{1-b_j} \left(\frac{q_b}{\|q_b\|_1}\right)^{\otimes n}, \tag{3.122}$$

where q_b is defined in (3.30). The lower bound will be proved by showing that there exists no test with risk $\leq \eta$ for the testing problem $H_0' : (X_1'', \ldots, X_n'') \sim p_0^{\otimes n}$ vs $H_1' : (X_1'', \ldots, X_n'') \sim \overline{p}^{(n)}$. The following Proposition states that the prior concentrates *whp* on a zone separated away from p_0 by an L_t distance of order ρ_{tail}^*.

Proposition 3.13. *There exists a constant* C_{tail}^{LB} *such that, when* $\int_{\mathcal{T}} p_0 \geq c_{tail}/n$, *we have with probability at least* $1 - \frac{\eta}{4}$ *(over the realizations of* $b = (b_U, \ldots, b_M)$*):*

$$\|p_b^{(n)} - p_0\|_t \geq C_{tail}^{LB}\, \rho_{tail}^*.$$

We now introduce the *Bayes risk* associated with the prior distribution (3.31):

Definition 3.3. *Define*

$$R_B^{tail} = \inf_{\psi\ test} \left\{ \mathbb{P}_{p_0}(\psi = 1) + \mathbb{E}_b\Big[\mathbb{P}_{p_b}(\psi = 0)\Big] \right\},$$

where the expectation is taken with respect to the realizations of $(b_j)_{j \geq U}$ *and* $\mathbb{P}_{p_b}$ *denotes the probability distribution when the data is drawn with density* (3.122).

The Proposition below states that when $\int_{\mathcal{T}(u_B)} p_0 \geq c_{tail}/n$ and when the tail dominates, the prior (3.31) is indistinguishable from $p_0^{\otimes n}$, in the sense that there exists no test with risk $\leq \eta$ for the testing problem $H_0' : (X_1'', \ldots, X_n'') \sim p_0^{\otimes n}$ vs $H_1' : (X_1'', \ldots, X_n'') \sim \overline{p}^{(n)}$.

Proposition 3.14. $R_B^{tail} > \eta$.

Remark: Our prior concentrates only with high probability on the zone $\|p_0 - p_b\|_t \geq C_{tail}^{LB}\rho_{tail}^*$. We can here justify that this is not restrictive. Indeed, we can *wlog* modify Proposition 3.14 to get $R_B^{tail} > \eta - 2\epsilon$ for any $\epsilon > 0$ small enough. We moreover show in Lemma 42 that if instead of our prior $\mathbb{E}\left(p_b^{(n)}\right)$, we considered as prior $p_{b,cond} = \mathbb{E}(p_b^{(n)} | \mathcal{A}_{\text{sep}})$ where $\mathcal{A}_{\text{sep}} = \{b \text{ is such that } \|p_0 - p_b\|_t \geq C_{tail}^{LB}\rho_{tail}^*\}$ and where the expectation is taken according to the realizations of b, then we would have $d_{TV}(p_0^{\otimes n}, p_{b,cond}) < d_{TV}(p_0^{\otimes n}, \mathbb{E}_b(p_b^{(n)})) + 2\epsilon \leq 1 - \eta - 2\epsilon + 2\epsilon = 1 - \eta$. Now, $p_{b,cond}$ satisfies almost surely $\|p_0 - p_{b,cond}\|_t \geq C_{tail}^{LB}\rho_{tail}^*$.

3.F.1 Proof of Proposition 3.4

Proof of Proposition 3.4. Assume that $\int_{\mathcal{T}(u_B)} p_0 < \frac{c_{tail}}{n}$ and that $n > c_{tail}$. Since $\int_{\mathcal{T}(u_B)} p_0 \leq \frac{c_{tail}}{n} < 1$, we necessarily have $\mathcal{T}(u_B) \subsetneq \Omega$. Set $u = \sup\{v > 0 : \int_{\mathcal{T}(v)} p_0 \leq \frac{c_{tail}}{n}\}$. We therefore necessarily have $u \leq \max_\Omega p_0$ (since $n > c_{tail}$) and $\int_{\overline{\mathcal{T}}(u)} p_0 \geq \frac{c_{tail}}{n}$. Choose $D(u) \subset p_0^{-1}(\{u\})$ a subset such that $\int_{D(u)\cup\mathcal{T}(u)} p_0 = \frac{c_{tail}}{n}$ and define $T'(u) = D(u) \cup \mathcal{T}(u)$. By the definition of u_{aux}, we have $u > u_{aux}$ so that

$$\left(\max_\Omega p_0\right) \int_{\mathcal{T}'(u)} p_0 \geq \int_{\mathcal{T}(u)} p_0^2 \geq c_I \left[\frac{L^d}{n^{2\alpha}\left(\int_{\mathcal{T}'(u)} p_0\right)^d}\right]^{\frac{1}{\alpha+d}}$$

$$\text{hence} \quad \max_\Omega p_0 \geq c_I \left[\frac{L^d}{n^{2\alpha}\left(\int_{T'(u)} p_0\right)^\alpha}\right]^{\frac{1}{\alpha+d}} = \frac{c_I}{c_{tail}^{\frac{\alpha}{\alpha+d}}} \left[\frac{L^d}{n^\alpha}\right]^{\frac{1}{\alpha+d}} =: m. \tag{3.123}$$

Define $h_r = (nL/c_{small})^{-\frac{1}{\alpha+d}}$ for some small enough constant c_{small} and $x_0 = \arg\max_\Omega p_0$. We note that $m = c_I' L h_r^\alpha$ where $c_I' = c_{small}^{\frac{\alpha}{\alpha+d}} c_I / c_{tail}^{\frac{\alpha}{\alpha+d}}$. Set B_1 and B_2 two disjoint balls included in $\Omega \cap B\left(x_0, (c_I' c_\star)^{1/\alpha} h_r\right)$ with radius $R := \frac{1}{4\sqrt{d}} (c_I' c_\star)^{1/\alpha} h_r$. B_1 and B_2 exist no matter how close x_0 is to the boundary of Ω. Denote by $x_1^{(r)}$ and $x_2^{(r)}$ the respective centers of B_1 and B_2. By Assumption ($\star$), we have $p_0 \geq m(1-2c_\star)$ over $\Omega \cap B\left(x_0, (c_I' c_\star)^{1/\alpha} h_r\right)$ so that it is possible to set the following prior:

$$p_r(x) = p_0(x) + c_r L h_r^\alpha \, f\left(\frac{x_1^{(r)} - x}{h_r}\right) - c_r L h_r^\alpha \, f\left(\frac{x_2^{(r)} - x}{h_r}\right), \tag{3.124}$$

where c_r is a small enough constant. This prior satisfies $\int_\Omega p_r = 1$, $p_r \geq 0$, $p_r \in H(\alpha, L(1+c_r))$ and satisfies Assumption ($\star$) by Lemma 22 if we choose c_r small enough. Moreover, the L_t discrepancy between p_0 and p_r is given by

$$\|p_0 - p_r\|_t^t = 2 \int_{\mathbb{R}^d} \left\{c_r L h_r^\alpha \, f\left(\frac{x_1^{(r)} - x}{h_r}\right)\right\}^{\alpha t} dx = 2 \left(c_r L\right)^{\alpha t} \|f\|_{\alpha t}^{\alpha t} \, h_r^{\alpha t + d} \asymp L^{\frac{d(t-1)}{t(\alpha+d)}} n^{-\frac{\alpha t + d}{t(\alpha+d)}}.$$

Now, the total variation between p_r and p_0 is given by:

$$d_{TV}(p_0, p_r) = L h_r^{\alpha+d} \|f\|_1 = c_r \|f\|_1 \frac{1}{n} < 1 - \eta,$$

for c_r small enough, which proves the desired lower bound. □

3.F.2 Proof of Proposition 3.12

Lemma 30. *It holds* $\mathbb{E}[\|q_b\|_1] = 1$ *and* $\mathbb{V}[\|q_b\|_1] \leq C_{30}/n^2$, *where* C_{30} *is a constant.*

Proof of Lemma 30. First, $\mathbb{E}[\|q_b\|_1] = 1$ is true by the definition of $c_j^{(\uparrow)}$ and $c^{(\downarrow)}$. As to the variance, we recall that for all $j \geq U : \Gamma_j^{(\uparrow)} = \int_{\widetilde{C}_j} \gamma_j^{(\uparrow)}$ and $\Gamma^{(\downarrow)} = \int_{\widetilde{C}_j} \gamma_j^{(\downarrow)}$. We have:

$$\begin{aligned}\mathbb{V}[\|q_b\|_1] &= \sum_{j\geq U} \mathbb{V}\left(b_j(\Gamma_j^{(\uparrow)} + \Gamma_j^{(\downarrow)})\right) \leq \sum_{j\geq U} \pi_j \left(\frac{A_{32}}{1+B_{32}}\right)^2 {\Gamma_j^{(\uparrow)}}^2 \quad \text{by Lemma 32} \\ &\leq \left(2c^{(\uparrow)}\right)^2 \left(\frac{A_{32}}{1+B_{32}}\right)^2 \left(\sum_{j\geq U} \pi_j\right) \left(Lh^{\alpha+d}\right)^2.\end{aligned}$$

Moreover:

$$\sum_{j\geq U} \pi_j = \frac{\left(n\sum_{j\geq U} p_j\right)^2}{2c_u} \leq \frac{\left(n\int_{\mathcal{T}(u_B)} p_0\right)^2}{2c_u} {C_{47}}^2 \quad \text{by Lemma 47,}$$

and $\left(Lh^{\alpha+d}\right)^2 = n^{-4}\left(\int_{\mathcal{T}(u_B)} p_0\right)^{-2}$. Hence: $\mathbb{V}[\|q_b\|_1] \leq C_{30}/n^2$, for some constant C_{30}. □

Lemma 31. *Let I be a countable set of indices and $(x_l)_{l\in I} \in \Omega$ and $(h_l)_{l\in I} > 0$ such that the balls $(B(x_l, h_l))_l$ are disjoint. Set moreover $(\epsilon_l)_{l\in I} \in \{\pm 1\}^I$ and let $C_\alpha = 1 \vee 2^{1-\alpha}$ and $\gamma(x) = \sum_{l\in I} \epsilon_l a_l\, Lh_l^\alpha\ f\left(\frac{x-x_l}{h_l}\right)$ where $(a_l)_l \geq 0$. Then $\forall x, y \in \Omega, |\gamma(x) - \gamma(y)| \leq c_\star|\gamma(x)| + \bar{a}\, C_\alpha L\|x-y\|^\alpha$ where $\bar{a} = \sup_{l\in I} a_l$.*

Proof of Lemma 31. Set for all $l \in I : A_l = B(x_l, h_l)$ and $A_0 = \Omega \setminus \left(\bigcup_{l\in I} A_l\right)$. Let $x, y \in \Omega$. The result is direct if $x, y \in A_0$. If x, y are in the same set A_l where $l \neq 0$ then by equation (3.119) we have:

$$\begin{aligned}|\gamma(x) - \gamma(y)| &= a_l L h_l^\alpha \left|f\left(\frac{x-x_l}{h_l}\right) - f\left(\frac{y-x_l}{h_l}\right)\right| \leq a_l L h_l^\alpha \left[c_\star f\left(\frac{x-x_l}{h_l}\right) + \left\|\frac{y-x}{h_l}\right\|^\alpha\right] \\ &= c_\star\gamma(x) + a_l L\|y-x\|^\alpha.\end{aligned}$$

Assume now there exist $i \neq l$ such that $x \in A_i$ and $y \in A_l$. For $x' \in A_i$ and $y' \in A_l$ such that $d_{\|\cdot\|}(x', A_0) = 0$ and $d_{\|\cdot\|}(y', A_0) = 0$ we have by equation (3.119):

$$\left|\gamma(x)\right| = \left|\gamma(x) - \gamma(x')\right| \leq c_\star\left|\gamma(x')\right| + a_i\, Lh_i^\alpha \left\|\frac{x-x'}{h_i}\right\|^\alpha = a_i L\|x-x'\|^\alpha,$$

$$\left|\gamma(y)\right| = \left|\gamma(y) - \gamma(y')\right| \leq c_\star\left|\gamma(y')\right| + a_l\, Lh_l^\alpha \left\|\frac{y-y'}{h_l}\right\|^\alpha = a_l L\|y-y'\|^\alpha.$$

Moreover, we have $\|x-y\| \geq \|x-x'\| + \|y-y'\|$ since x and y are in two different sets, and $C_\alpha = \max\{\lambda^\alpha + (1-\lambda)^\alpha : \lambda \in [0,1]\}$ so that: $C_\alpha\|x-y\|^\alpha \geq \|x-x'\|^\alpha + \|y-y'\|^\alpha$. This yields the result. □

Proof of Proposition 3.12. 1. We first show that $p_0 - \gamma_j^{(\downarrow)} \geq 0$ for all $j \in \mathbb{N}^*$. By Lemma 44, we have $p_U \leq C_{44} L h^{\alpha+d}$ where C_{44} is a constant, so that by Assumption $(\star)$:

$$\forall j \geq U,\ \forall x \in \widetilde{C}_j,\ p_0(x) \leq \left(C_{44}(1+c_\star) + \sqrt{d}^{\alpha}\right) L h^{\alpha} =: C L h^{\alpha}. \tag{3.125}$$

Recall that $e(\widetilde{C}_j) = h_{tail}$. Therefore, the condition that Algorithm 3 splits $\widetilde{C}_j$ at least once rewrites:

$$e(\widetilde{C}_j) = h_{tail} > \left(\frac{p_0(x_j)}{c'_\beta L}\right)^{1/\alpha} \Longleftarrow h_{tail} > \left(\frac{C}{c'_\beta}\right)^{1/\alpha} h_{tail}$$

by equation (3.125), which is true if we choose c'_β large enough. This ensures that for all cell $E_l^{(j)}$ and for all $x \in E_l^{(j)}$, we have by the properties of the partitioning scheme (Proposition 3.6 item 3), that $p_0(x) \geq \frac{1}{2} p_0\left(z_l^{(j)}\right) \geq L(h_l^{(j)})^\alpha$ by taking $c_\beta = c_\alpha \geq 2$. Therefore,

$$p_0 - \phi_l^{(j)} \geq p_0 - c^{(\downarrow)} L\left(h_l^{(j)}\right)^\alpha \geq \frac{1}{2} p_0\left(z_l^{(j)}\right) - c^{(\downarrow)} L\left(h_l^{(j)}\right)^\alpha \geq \frac{1 - c^{(\downarrow)}}{2} p_0\left(z_l^{(j)}\right) \geq 0.$$

Moreover, it is clear that $\widetilde{C}_j,\ p_0 - \gamma_j^{(\downarrow)} \in H(\alpha, L(1+c^{(\downarrow)})) \subset H(\alpha, L')$ and $\widetilde{C}_j,\ p_0 + \gamma_j^{(\uparrow)} \in H(\alpha, L(1+c_j^{(\uparrow)})) \subset H(\alpha, L')$ for $c^{(\downarrow)}$ small enough. To finish, by Lemma 31, we have that for all $(b_j)_{j \geq U}$, $p_b^{(n)}$ satisfies Assumption $(\star)$ with the constants $c'_\star$ and L' by choosing $c^{(\downarrow)}$ small enough. Indeed, set $\gamma = \sum_{l \geq U} b_j \gamma_j^{(\uparrow)} - (1-b_j)\gamma_j^{(\downarrow)}$ and $\bar{a} = \sup\left(\{c^{(\downarrow)}\} \cup \{c_j^{(\uparrow)} : j \geq U\}\right)$.

$$\begin{aligned} |q_b(x) - q_b(y)| &\leq |p_0(x) - p_0(y)| + |\gamma(x) - \gamma(y)| \\ &\leq c_\star p_0(x) + c_\star |\gamma(x)| + (1 + \bar{a}\, C_\alpha) L \|x - y\|^\alpha. \end{aligned}$$

Let $j \geq U$ such that $x \in \widetilde{C}_j$. If $b_j = 1$ then $\gamma(x) = \gamma_j^{(\uparrow)}(x) \geq 0$ hence $p_0(x) + |\gamma(x)| = p_0(x) + \gamma(x)$ which proves that $(\star)$ is satisfied. Otherwise, $\gamma(x) = -\gamma^{(\downarrow)}(x)$. We have already shown that $\gamma_j^{(\downarrow)} \leq p_0$ over $\widetilde{C}_j$. Taking $c^{(\downarrow)}$ small enough, we can therefore impose, for any $\lambda > 0 : \gamma_j^{(\downarrow)} \leq \lambda p_0$ over $\widetilde{C}_j$. Therefore,

$$p_0(x) + |\gamma(x)| = p_0(x) + \gamma_j^{(\downarrow)}(x) \leq \frac{1+\lambda}{1-\lambda}\left(p_0(x) - \gamma_j^{(\downarrow)}(x)\right).$$

Taking λ and $\bar{a}$ small enough (which can be done by taking $c^{(\downarrow)}$ small enough), we get in both cases that q_b satisfies Assumption $(\star)$ with the constants $c_\star(1+\delta/2)$ and $L(1+\delta/2)$ instead of $c_\star$ and L. Now, by Lemma 30 and the Chebyshev inequality, the event $\left\{\left|\, \|q_b\|_1 - 1 \,\right| \leq C_{32}\right\}$ can have arbitrarily high probability when n is larger than a suitably chosen constant. Taking n large enough ensures that with probability arbitrarily close to 1, $p_b^{(n)}$ satisfies $(\star)$ with the constants $c'_\star$ and L'.

2. By Proposition 3.6, we have $h_l^{(j)} \geq \frac{1}{2^{\beta+1}} \left(\frac{1}{L\, c'_\beta} p_0(z_l^{(j)}) \right)^{1/\alpha}$ and that for all $l \in \{1, \dots, M_j\}$: $p_0 \geq \frac{1}{2}\, p_0(z_l^{(j)})$ over $E_l^{(j)}$. Now, for all $j \geq U$ and $l \in \{1, \dots, M_j\}$ we have

$$\begin{aligned}
\int_{\widetilde{C}_j} \gamma_j^{(\downarrow)} &= \sum_{j=1}^{M_j} \int_{E_l^{(j)}} \phi_l^{(j)} = \sum_{j=1}^{M_j} c^{(\downarrow)} L \Big(h_l^{(j)}\Big)^{\alpha+d} \|f\|_1 \\
&\geq c^{(\downarrow)} \|f\|_1 \sum_{j=1}^{M_j} L \frac{1}{2^{\beta+1}} \frac{1}{L\, c'_\beta} p_0(z_l^{(j)}) \Big(h_l^{(j)}\Big)^d \\
&\geq c^{(\downarrow)} \|f\|_1 \; \frac{1}{2^{\beta+2} c'_\beta} \int_{D_j} p_0 \;\geq\; \frac{1}{C^{(\downarrow)}} \int_{D_j} p_0,
\end{aligned}$$

where $C^{(\downarrow)}$ is a constant, which ends the proof.

□

3.F.3 Proof of Proposition 3.13

In what follows we set for all $j \geq U$:

$$\Gamma_j^{(\uparrow)} = \int_{\widetilde{C}_j} \gamma_j^{(\uparrow)} \quad \text{and} \quad \Gamma^{(\downarrow)} = \int_{\widetilde{C}_j} \gamma_j^{(\downarrow)}. \tag{3.126}$$

Lemma 32. *There exist three constants A_{32}, B_{32} and C_{32} such that for all $j \geq U$, it holds:*

1. $\Gamma_j^{(\downarrow)} \leq A_{32}\, p_j$,
2. $\Gamma_j^{(\uparrow)} \geq B_{32}\, p_j$
3. $\|\gamma_j^{(\uparrow)}\|_t^t \geq C_{32} \int_{\widetilde{C}_j} p_0^t$ *where* $C_{32} < 1$.

Proof of Lemma 32. Fix $j \geq U$.

1. We have:

$$\begin{aligned}
\int_{\widetilde{C}_j} \gamma_j^{(\downarrow)} &= \sum_{l=1}^{M_j} c^{(\downarrow)} L {h_l^{(j)}}^{\alpha+d} \|f\|_1 \leq \|f\|_1 \sum_{l=1}^{M_j} \frac{c^{(\downarrow)}}{c'_\beta} p_0\Big(z_l^{(j)}\Big) {h_l^{(j)}}^d \\
&\leq 2\|f\|_1 \frac{c^{(\downarrow)}}{c'_\beta} \sum_{l=1}^{M_j} \int_{E_l^{(j)}} p_0 \leq 2\|f\|_1 \frac{c^{(\downarrow)}}{c'_\beta} p_j =: A_{32}\, p_j.
\end{aligned}$$

2. By definition of U:

$$\begin{aligned}
p_j &\leq \frac{c_u}{n^2 \sum_{j \geq U} p_j} \leq \frac{c_u}{n^2 D \int_{\mathcal{T}(u_B)} p_0} = \frac{c_u}{D} L h^{\alpha+d} \\
&\leq \frac{1}{B_{32}} \Gamma_j^{(\uparrow)}, \qquad \text{for some constant } B_{32}.
\end{aligned} \tag{3.127}$$

3. Let $x \in \widetilde{C}_j$ and $y \in \widetilde{C}_j$ such that $p_0(y) = \frac{p_j}{h^d}$, which exists by the intermediate value theorem. We have by Assumption ($\star$):

$$\begin{aligned} p_0(x) &\leq (1+c_\star)p_0(y) + L(h\sqrt{d})^\alpha \\ &\leq \left[(1+c_\star)\frac{c_u}{D} + \sqrt{d}^\alpha\right] Lh^\alpha \qquad \text{by Equation (3.127),} \end{aligned}$$

so that:

$$\int_{\widetilde{C}_j} p_0^t \leq \left[(1+c_\star)\frac{c_u}{D} + \sqrt{d}^\alpha\right]^t L^t h^{\alpha t + d} \leq \frac{1}{C_{32}} \|\gamma_j^{(\uparrow)}\|_t^t \,, \quad \text{for some constant } C_{32}.$$

□

Proof of Proposition 3.13. Assume throughout the proof that $\int_{\mathcal{B}^c} p_0 > \frac{c_{tail}}{n}$. We show that our prior concentrates with high probability on a zone separated away from p_0 by an L_t distance of order ρ^*_{tail}, up to a constant. To lower bound the L_t separation between our prior and the null distribution, we will only consider the discrepancy accounted for by the perturbations $(\gamma_j^{(\uparrow)})_j$. We recall that $\widetilde{C}_0 = D(U)^c$. For all $j \geq U$, fix $b_j \in \{0,1\}$ as well as n large enough, such that

$$\big|\, \|q_b\|_1 - 1 \,\big| \leq C_{32}. \tag{3.128}$$

By Lemma 30 and the Chebyshev inequality, the event corresponding to Equation (3.128) can have arbitrarily high probability when n is larger than a suitably chosen constant. Now, write $\mathbf{I}_0 = \{0\} \cup \{j \in \mathbb{N}^* : j \geq U\}$.

$$\begin{aligned} \|p_0 - p_b\|_t^t &= \sum_{j\in\mathbf{I}_0} b_j \int_{\widetilde{C}_j} \left| p_0 - \frac{p_0 + \gamma_j^{(\uparrow)}}{\|q_b\|_1} \right|^t + \sum_{j\in\mathbf{I}_0} (1-b_j) \int_{\widetilde{C}_j} \left| p_0 - \frac{p_0 + \gamma^{(\downarrow)}}{\|q_b\|_1} \right|^t \\ &\geq \sum_{j\in\mathbf{I}_0} b_j \int_{\widetilde{C}_j} \left| p_0 - \frac{p_0 + \gamma_j^{(\uparrow)}}{\|q_b\|_1} \right|^t = \sum_{j\in\mathbf{I}_0} b_j \left\| p_0 - \frac{p_0 + \gamma_j^{(\uparrow)}}{\|q_b\|_1} \right\|_{t,\widetilde{C}_j}^t \\ &\geq \sum_{j\in\mathbf{I}_0} b_j \left| \frac{\|\gamma_j^{(\uparrow)}\|_{t,\widetilde{C}_j}}{\|q_b\|_1} - \left\| p_0 \left(1 - \frac{1}{\|q_b\|_1}\right) \right\|_{t,\widetilde{C}_j} \right|^t \quad \text{by the reverse triangle inequality} \\ &\geq \sum_{j\in\mathbf{I}_0} b_j \left(\frac{\|\gamma_j^{(\uparrow)}\|_{t,\widetilde{C}_j}}{\|q_b\|_1} - \left(1 - \frac{1}{\|q_b\|_1}\right) \frac{1}{C_{32}} \|\gamma_j^{(\uparrow)}\|_{t,\widetilde{C}_j} \right)^t \quad \text{by Lemma 32 and Equation (3.128)} \\ &= \sum_{j\in\mathbf{I}_0} b_j \|\gamma_j^{(\uparrow)}\|_{t,\widetilde{C}_j}^t \left(\frac{1}{\|q_b\|_1} \left(1 + \frac{1}{C_{32}}\right) - \frac{1}{C_{32}} \right)^t \\ &\geq \sum_{j\in\mathbf{I}_0} b_j \|\gamma_j^{(\uparrow)}\|_{t,\widetilde{C}_j}^t \left(\frac{1}{2 + C_{32}} \right)^t \quad \text{by Equation (3.128)} \end{aligned}$$

$$\geq \sum_{j\geq U} b_j c_u c^{(\downarrow)} L^t h^{\alpha t+d} \left(\frac{1}{2+C_{32}}\right)^t := C_{\text{gap}} L^t h^{\alpha t+d} \sum_{j\geq U} b_j.$$

It now remains to prove that, *whp*, $L^t h^{\alpha t+d} \sum_{j\geq U} b_j \gtrsim {\rho^*_{tail}}^t$.

$$\mathbb{E}\Big[\sum_{j\geq U} b_j\Big] = \sum_{j\geq U} \pi_j = \frac{n^2\left(\sum_{j\geq U} p_j\right)^2}{2c_u} \geq \frac{D^2 c_{tail}^2}{2c_u} \quad \text{by Lemma 46.}$$

Moreover,

$$\mathbb{V}\Big[\sum_{j\geq U} b_j\Big] \leq \sum_{j\geq U} \pi_j = \mathbb{E}\Big[\sum_{j\geq U} b_j\Big]$$

We now consider the event

$$\sum_{j\geq U} b_j \geq \frac{1}{2}\mathbb{E}\Big[\sum_{j\geq U} b_j\Big]. \tag{3.129}$$

By the Chebyshev inequality, the probability of this event can be made arbitrarily large by choosing the constant c_{tail} small enough, since $\mathbb{V}\Big[\sum_{j\geq U} b_j\Big] = o\Big(\mathbb{E}^2\Big[\sum_{j\geq U} b_j\Big]\Big)$ as $c_{tail} \to +\infty$. Therefore, on the intersection of the events defined in Equations (3.128) and (3.129), we have that:

$$\begin{aligned}
\|p_0 - p_b\|_t^t &\geq C_{\text{gap}} L^t h^{\alpha t+d} \sum_{j\geq U} b_j \geq \frac{C_{\text{gap}}}{2} L^t h^{\alpha t+d} \sum_{j\geq U} \pi_j \\
&\geq \frac{C_{\text{gap}}}{2} L^t h^{\alpha t+d} D \int_{\mathcal{T}(u_B)} p_0 \qquad \text{by Lemma 46} \\
&\asymp {\rho^*_{tail}}^t.
\end{aligned}$$

□

3.F.4 Proof of Proposition 3.14

Proof of Proposition 3.14. We draw $\widetilde{n} \sim \text{Poi}(2n)$ and $\widetilde{n}'|b \sim \text{Poi}\big(2n\int_\Omega q_b\big)$ independent of $\widetilde{n}$, and we let $\mathcal{A}_1 = \{\widetilde{n} \geq n\}$ and $\mathcal{A}'_1 = \{\widetilde{n}' \geq n\}$. By Lemma 39, we can ensure $\mathbb{P}(\mathcal{A}_1), \mathbb{P}(\mathcal{A}'_1) \geq 1 - \eta/100$ for n larger than a constant. This condition will be assumed throughout the proof of Proposition 3.14. We will also slightly abuse notation and identify the probability densities with their associated probability measures. Moreover, we will use the notation $\widetilde{C}_0 = D(U)^c$ where we recall that $D(U) = \bigcup_{j\geq U} \widetilde{C}_j$. We recall the definition of $\bar{p}^{(n)}$ in (3.122) and introduce the *poissonized* probability measures $\bar{p}_b^{(\widetilde{n}')}$, $q_b^{(\widetilde{n})}$ and $p_0^{\otimes\widetilde{n}}$, defined over $\widetilde{\mathcal{X}} = \bigcup_{n\in\mathbb{N}} \Omega^n$. The core of the proof is to link our target quantity $d_{TV}(\bar{p}^{(n)}, p_0^{\otimes n}) = d_{TV}(\mathbb{P}_{p_0^{\otimes n}}, \mathbb{P}_{\bar{p}^{(n)}})$ (which we want to upper bound by a small constant), to the quantity $d_{TV}(q_b^{(\widetilde{n})}, p_0^{\otimes\widetilde{n}})$ which is easier to work with. For clarity, we give

the densities associated to each of the poissonized probability measures. For any $x \in \widetilde{\mathcal{X}}$, we denote by $\widetilde{n}(x) \in \mathbb{N}$ the unique integer such that $x = (x_1, \dots, x_{\widetilde{n}(x)})$.

$$p_0^{\otimes \widetilde{n}}(x) \Big| \{\widetilde{n} = \nu\} = \begin{cases} \text{if } \widetilde{n}(x) \neq \nu: & 0 \\ \text{otherwise:} & p_0^{\otimes \nu}(x). \end{cases}$$

$$\bar{p}_b^{(\widetilde{n}')}(x) \Big| \{\widetilde{n}' = \nu\} = \begin{cases} \text{if } \widetilde{n}'(x) \neq \nu: & 0 \\ \text{otherwise:} & \sum\limits_{(\beta_j)_{j \geq U} \in \{0,1\}} \prod\limits_{j \geq U} \pi_j^{\beta_j} (1-\pi_j)^{1-\beta_j} \left(\frac{q_\beta}{\|q_\beta\|_1} \right)^{\otimes \nu}(x). \end{cases}$$

$$q_b^{(\widetilde{n})}(x) \Big| \{\widetilde{n} = \nu\} = \begin{cases} \text{if } \widetilde{n}(x) \neq \nu: & 0 \\ \text{otherwise:} & \sum\limits_{(\beta_j)_{j \geq U} \in \{0,1\}} \prod\limits_{j \geq U} \pi_j^{\beta_j} (1-\pi_j)^{1-\beta_j} \; q_\beta^{\otimes \nu}(x). \end{cases}$$

Note that q_b is not a density. Therefore the term $q_b^{(\widetilde{n})}$ with $\widetilde{n} \sim \mathrm{Poi}(2n)$ denotes the mixture of inhomogeneous spatial Poisson processes with intensity functions $(2n\, q_b)_b$, over the realizations of $(b_j)_{j \in \mathbf{I}_0}$ where we recall that $\mathbf{I}_0 = \{0\} \cup \{j \in \mathbb{N}^* : j \geq U\}$. We define the histogram of x over the domain $D(U)$ by setting for all $j \geq U : \widetilde{N}_j(x) = \sum_{i=1}^{\widetilde{n}(x)} \mathbb{1}_{x_i \in \widetilde{C}_j}$.

We have $R_B^{tail} = 1 - d_{TV}(\mathbb{P}_{p_0}, \mathbb{P}_{\bar{p}})$, where $\mathbb{P}_{p_0}$ and $\mathbb{P}_{\bar{p}}$ are respectively the probability measures of the densities $p_0^{\otimes n}$ and $\bar{p}$. We therefore aim at proving $d_{TV}(\mathbb{P}_{p_0}, \mathbb{P}_{\bar{p}}) < 1 - \eta$. We will denote by $\mathrm{Poi}(f)$ the inhomogeneous spatial Poisson process with nonnegative intensity function f, by $f_{|\widetilde{C}_j}$ the restriction of f to the cell $\widetilde{C}_j$. Moreover, for any two probability measures P, Q over the same measurable space $(\mathcal{Y}, \mathcal{C})$, and for any event $A_0 \in \mathcal{C}$ we will denote by $d_{TV}^{A_0}(P, Q)$ the total variation restricted to A_0, defined as the quantity:

$$d_{TV}^{(A_0)}(P, Q) = \sup_{A \in \mathcal{C}} \Big| P(A \cap A_0) - Q(A \cap A_0) \Big|. \tag{3.130}$$

We prove the following lemmas concerning the total variation restricted to A_0:

Lemma 33. *For any two probability measures P, Q over the same measurable space $(\mathcal{Y}, \mathcal{C})$, and for any event $A_0 \in \mathcal{C}$, if $P, Q \ll \mu$ over $(\mathcal{Y}, \mathcal{C})$, i.e. $dP = p d\mu$ and $dQ = q d\mu$, it holds:*

$$d_{TV}^{(A_0)}(P, Q) = \frac{1}{2} \left[|P(A_0) - Q(A_0)| + \int_{A_0} |p - q| d\mu \right].$$

Lemma 34. *For any two probability measures P_1, Q_1 (resp. P_2, Q_2) over the same measurable space $(\mathcal{Y}_1, \mathcal{C}_1)$(resp. $(\mathcal{Y}_2, \mathcal{C}_2)$), for any event $A_0 = A_0^{(1)} \times A_0^{(2)}$ such that $A_0^{(1)} \in \mathcal{C}_1$ and $A_0^{(2)} \in \mathcal{C}_2$, if $P_1, Q_1 \ll \mu_1$ (resp. $P_2, Q_2 \ll \mu_2$) over $(\mathcal{Y}_1, \mathcal{C}_1)$ (resp. $(\mathcal{Y}_2, \mathcal{C}_2)$), i.e. $dP_j = p_j d\mu_j$ and $dQ_j = q_j d\mu_j$*

for $j = 1, 2$, then it holds:

$$d_{TV}^{(A_0)}(P_1 \otimes P_2, Q_1 \otimes Q_2) \leq d_{TV}^{(A_0^{(1)})}(P_1, Q_1) + d_{TV}^{(A_0^{(2)})}(P_2, Q_2).$$

We now come back to the proof of Proposition 3.14. We define $\forall j \geq U$:

$$\mathcal{A}^{(j)} = \Big\{ x \in \widetilde{\mathcal{X}} \;\Big|\; \widetilde{N}_j(x) \leq 1 \Big\}, \tag{3.131}$$

$$\mathcal{A}^{(0)} = \widetilde{\mathcal{X}}. \tag{3.132}$$

We also introduce

$$\mathcal{A} = \Big\{ x \in \widetilde{\mathcal{X}} \;\Big|\; \forall j \geq U : \widetilde{N}_j \leq 1 \Big\} = \bigcap_{j \geq U} \mathcal{A}^{(j)}, \tag{3.133}$$

the subset of $\widetilde{\mathcal{X}}$ of all the vectors of observations such that any cube $(\widetilde{C}_j)_{j \geq U}$ contains at most one observation. $\mathcal{A}$ will play an essential role, as it is the high probability event on which we will place ourselves to approximate the total variation $d_{TV}(p_0^{\otimes n}, \bar{p}^{(n)})$.
In Lemmas 35-41, we will successively use equivalence of models to formalize the following (informal) chain of approximations:

$$\begin{aligned} d_{TV}(\mathbb{P}_{p_0}, \mathbb{P}_{\bar{p}}) &\lesssim d_{TV}(\bar{p}^{(\widetilde{n}')}, p_0^{\otimes \widetilde{n}}) = d_{TV}(q_b^{(\widetilde{n})}, p_0^{\otimes \widetilde{n}}) \lesssim d_{TV}^{(\mathcal{A})}\Big(q_b^{(\widetilde{n})}, p_0^{\otimes \widetilde{n}}\Big) \\ &= d_{TV}^{(\mathcal{A})}\left(\bigotimes_{j \in \mathbf{I}_0} \mathrm{Poi}\Big(2n\, p_{0|\widetilde{C}_j}\Big), \bigotimes_{j \in \mathbf{I}_0} \mathrm{Poi}\Big(2n\, q_{b_j|\widetilde{C}_j}\Big) \right) \\ &\leq \sum_{j \geq U} d_{TV}^{(\mathcal{A}^{(j)})} \left\{ \mathrm{Poi}\Big(2n\, q_{b|\widetilde{C}_j}\Big),\ \mathrm{Poi}\Big(2n\, p_{0|\widetilde{C}_j}\Big) \right\}, \end{aligned}$$

At each step, we will control the approximation errors. We recall that $\mathbb{P}_{p_0^{\otimes n}}$ and $\mathbb{P}_{\bar{p}^{(n)}}$ are defined over Ω^n whereas $p_0^{\otimes \widetilde{n}}, \bar{p}_b^{(\widetilde{n}')}$ and $q_b^{(\widetilde{n})}$ are defined on $\widetilde{\mathcal{X}} = \bigcup_{n \in \mathbb{N}} \Omega^n$. More precisely we will prove the following lemmas:

Lemma 35. *It holds $d_{TV}(\bar{p}^{(n)}, p_0^{\otimes n}) \leq d_{TV}(\bar{p}^{(\widetilde{n}')}, p_0^{\otimes \widetilde{n}}) / \mathbb{P}(\mathcal{A}_1 \cap \mathcal{A}_1')$.*

Note that in the right hand side of Lemma 35, we have $\widetilde{n}'$ observations for $\bar{p}^{(\widetilde{n}')}$ and $\widetilde{n}$ for $p_0^{\otimes \widetilde{n}}$.

Lemma 36. *It holds $\overline{p}^{(\widetilde{n}')} = q_b^{(\widetilde{n})}$.*

Lemma 37. *It holds: $d_{TV}(q_b^{(\widetilde{n})}, p_0^{\otimes \widetilde{n}}) \leq d_{TV}^{(\mathcal{A})}\Big(q_b^{(\widetilde{n})}, p_0^{\otimes \widetilde{n}}\Big) + q_b^{(\widetilde{n})}(\mathcal{A}^c) + p_0^{\otimes \widetilde{n}}(\mathcal{A}^c)$.*

Lemma 38. *The following tensorization of the spatial Poisson processes holds:*

$Poi(2np_0) = \bigotimes_{j \in \mathbf{I}_0} Poi(2np_{0|\widetilde{C}_j})$ and $Poi(2nq_b) = \bigotimes_{j \in \mathbf{I}_0} Poi(2nq_{b_j|\widetilde{C}_j})$. Hence we have:

$$d_{TV}^{(\mathcal{A})}\Big(q_b^{(\widetilde{n})}, p_0^{\otimes \widetilde{n}}\Big) \leq \sum_{j \geq U} d_{TV}^{(\mathcal{A}^{(j)})} \left(Poi\Big(2n\, q_{b|\widetilde{C}_j}\Big),\ Poi\Big(2n\, p_{0|\widetilde{C}_j}\Big) \right).$$

Furthermore, we will prove the following lemmas controlling the error at each step:

Lemma 39. *There exists a constant $n_0 \in \mathbb{N}$ such that whenever $n \geq n_0$, it holds $\mathbb{P}(\mathcal{A}_1), \mathbb{P}(\mathcal{A}_1') \geq 1 - \eta/100$.*

Lemma 40. *It holds $p_0^{\otimes n}(\mathcal{A}^c) \leq A_{40}$ and $q_b^{(n)}(\mathcal{A}^c) \leq B_{40}$ where A_{40} and B_{40} are two constants which can be made arbitrarily small by choosing successively c_I, c_u and $c^{(\downarrow)}$ small enough.*

Finally we compute each of the terms in the last sum from Lemma 38:

Lemma 41. *For all $j \geq U$ it holds $d_{TV}^{(\mathcal{A}^{(j)})}\left(Poi\left(2n\, q_{b|\widetilde{C}_j}\right),\ Poi\left(2n\, p_{0|\widetilde{C}_j}\right)\right) \leq A_{41} n^2 p_j^2 + B_{41} \frac{p_j}{\sum_{l \geq U} p_l}$ where A_{41} and B_{41} are two constants and B_{41} can be made arbitrarily small by choosing c_u small enough.*

Bringing together Lemmas 35 - 41, we get:

$$\begin{aligned} d_{TV}(\bar{p}^{(n)}, p_0^{\otimes n}) &\leq \left[A_{40} + B_{40} + A_{41} \sum_{j \geq U} n^2 p_j^2 + B_{41}\right] / \mathbb{P}(\mathcal{A}_1 \cap \mathcal{A}_1') \\ &\leq \left[A_{40} + B_{40} + A_{41} c_u + B_{41}\right] / \mathbb{P}(\mathcal{A}_1 \cap \mathcal{A}_1'), \end{aligned}$$

where the right-hand side can be made arbitrarily small by choosing successively c_I, c_u and $c^{(\downarrow)}$ small enough, which ends the proof of Proposition 3.14.

We now prove Lemmas 33 - 41.

Proof of Lemma 33. Suppose by symmetry $P(A_0) \geq Q(A_0)$ and set $B_0 = \{x \in \mathcal{Y} : p(x) \geq q(x)\}$. Then we have:

$$\begin{aligned} |P(B_0 \cap A_0) - Q(B_0 \cap A_0)| &= P(B_0 \cap A_0) - Q(B_0 \cap A_0) = \int_{B_0 \cap A_0} |p - q| d\mu \\ &= P(A_0) - Q(A_0) + \int_{A_0 \setminus B_0} |p - q| d\mu, \\ \text{so that} \quad |P(B_0 \cap A_0) - Q(B_0 \cap A_0)| &= \frac{1}{2}\left[|P(A_0) - Q(A_0)| + \int_{A_0} |p - q| d\mu\right], \\ \text{which yields:} \quad d_{TV}^{(A_0)}(P, Q) &\geq \frac{1}{2}\left[|P(A_0) - Q(A_0)| + \int_{A_0} |p - q| d\mu\right]. \end{aligned}$$

Moreover, for any $B \in \mathcal{C}$, we consider $|P(B \cap A_0) - Q(B \cap A_0)|$. There are two cases.
First case: $P(B \cap A_0) \geq Q(B \cap A_0)$. Then we have:

$$\begin{aligned} |P(B \cap A_0) - Q(B \cap A_0)| &= P(B \cap A_0) - Q(B \cap A_0) \\ &= P(B \cap A_0 \cap B_0) - Q(B \cap A_0 \cap B_0) + \underbrace{P(B \cap A_0 \setminus B_0) - Q(B \cap A_0 \setminus B_0)}_{\leq 0} \end{aligned}$$

$$\leq P(B \cap A_0 \cap B_0) - Q(B \cap A_0 \cap B_0) \leq P(A_0 \cap B_0) - Q(A_0 \cap B_0)$$
$$= |P(A_0 \cap B_0) - Q(A_0 \cap B_0)|.$$

Second case: $Q(B \cap A_0) \geq P(B \cap A_0)$. Then we have:

$$\begin{aligned}
|P(B \cap A_0) - Q(B \cap A_0)| &= Q(B \cap A_0) - P(B \cap A_0)| \\
&= Q(B \cap A_0 \setminus B_0) - P(B \cap A_0 \setminus B_0) + \underbrace{Q(B \cap A_0 \cap B_0) - P(B \cap A_0 \cap B_0)}_{\leq 0} \\
&\leq Q(B \cap A_0 \setminus B_0) - P(B \cap A_0 \setminus B_0) \leq Q(A_0 \setminus B_0) - P(A_0 \setminus B_0) \\
&= \underbrace{Q(A_0) - P(A_0)}_{\leq 0} + P(A_0 \cap B_0) - Q(A_0 \cap B_0) \\
&\leq |P(A_0 \cap B_0) - Q(A_0 \cap B_0)|.
\end{aligned}$$

In both cases, the result is proven. □

Proof of Lemma 34. We have by Lemma 33:

$$\begin{aligned}
2d_{TV}^{(A_0)}(P_1 \otimes P_2, Q_1 \otimes Q_2) &= |P_1 \otimes P_2(A_0) - Q_1 \otimes Q_2(A_0)| + \int_{A_0} |p_1(x)p_2(y) - q_1(x)q_2(y)| d\mu_1(x) d\mu_2(y) \\
&\leq P_1(A_0^{(1)}) \left| P_2(A_0^{(2)}) - Q_2(A_0^{(2)}) \right| + Q_2(A_0^{(2)}) \left| P_1(A_0^{(1)}) - Q_1(A_0^{(1)}) \right| \\
&\quad + P_1(A_0^{(1)}) \int_{A_0^{(2)}} |p_2(y) - q_2(y)| d\mu_2(y) \\
&\quad + Q_2(A_0^{(2)}) \int_{A_0^{(1)}} |p_1(x) - q_1(x)| d\mu_1(x) \\
&= 2P_1(A_0^{(1)}) d_{TV}^{(A_0^{(2)})}(P_2, Q_2) + 2Q_2(A_0^{(2)}) d_{TV}^{(A_0^{(1)})}(P_1, Q_1) \\
&\leq 2d_{TV}^{(A_0^{(1)})}(P_1, Q_1) + 2d_{TV}^{(A_0^{(2)})}(P_2, Q_2).
\end{aligned}$$

□

Proof of Lemma 35. We have:

$$\begin{aligned}
d_{TV}\left(\bar{p}_b^{(\widetilde{n}')}, p_0^{\otimes \widetilde{n}}\right) &= \sup_{A \in \widetilde{\mathcal{X}}} \left| \bar{p}_b^{(\widetilde{n}')}(A) - p_0^{\otimes \widetilde{n}}(A) \right| \geq \sup_{A \in \widetilde{\mathcal{X}} \cap \mathcal{A}_1 \cap \mathcal{A}_1'} \left| \bar{p}_b^{(\widetilde{n}')}(A) - p_0^{\otimes \widetilde{n}}(A) \right| \\
&= \mathbb{P}(\mathcal{A}_1 \cap \mathcal{A}_1') \sup_{A \in \widetilde{\mathcal{X}}} \left| \bar{p}_{b,\mathcal{A}_1'}^{(\widetilde{n}')}(A) - p_{0,\mathcal{A}_1}^{\otimes \widetilde{n}}(A) \right| \quad \text{where} \begin{cases} \bar{p}_{b,\mathcal{A}_1'}^{(\widetilde{n}')} = \bar{p}_b^{(\widetilde{n}')}(\cdot | \mathcal{A}_1') \\ p_{0,\mathcal{A}_1}^{\otimes \widetilde{n}} = p_0^{\otimes \widetilde{n}}(\cdot | \mathcal{A}_1) \end{cases} \\
&= \mathbb{P}(\mathcal{A}_1 \cap \mathcal{A}_1') \, d_{TV}\left(\bar{p}_{b,\mathcal{A}_1'}^{(\widetilde{n}')}, p_{0,\mathcal{A}_1}^{\otimes \widetilde{n}}\right).
\end{aligned}$$

Furthermore, over $\mathcal{A}_1 \cap \mathcal{A}_1'$ it holds $\widetilde{n} \geq n$ and $\widetilde{n}' \geq n$, so that $d_{TV}\left(\bar{p}_{b,\mathcal{A}_1}^{(\widetilde{n})}, p_{0,\mathcal{A}_1}^{\otimes \widetilde{n}}\right) \geq d_{TV}\left(\bar{p}_b^{(n)}, p_0^{\otimes n}\right)$, which yields the result. □

Proof of Lemma 36. For fixed $b = (b_j)_{j\geq U}$ we have $\widetilde{n}' \sim \mathrm{Poi}(\|q_b\|_1)$ and $\overline{p}_b^{(\widetilde{n}')} = \mathrm{Poi}\Big(n\|q_b\|_1 \frac{q_b}{\|q_b\|_1}\Big) = \mathrm{Poi}(nq_b) = q_b^{\otimes\widetilde{n}}$ so that taking the mixture over all realizations of b yields that, unconditionally on b: $\overline{p}_b^{(\widetilde{n}')} = q_b^{(\widetilde{n})}$. □

Proof of Lemma 37. We have:

$$\begin{aligned} d_{TV}\Big(q_b^{(\widetilde{n})}, p_0^{\otimes\widetilde{n}}\Big) &\leq \sup_{A\in\widetilde{X}\cap\mathcal{A}} \Big|q_b^{(\widetilde{n})}(A) - p_0^{\otimes\widetilde{n}}(A)\Big| + q_b^{(\widetilde{n})}(\mathcal{A}^c) + p_0^{\otimes\widetilde{n}}(\mathcal{A}^c) \\ &= d_{TV}^{(\mathcal{A})}\Big(q_b^{(\widetilde{n})}, p_0^{\otimes\widetilde{n}}\Big) + q_b^{(\widetilde{n})}(\mathcal{A}^c) + p_0^{\otimes\widetilde{n}}(\mathcal{A}^c). \end{aligned}$$

Hence the result. □

Proof of Lemma 38. To further transform the last quantity $d_{TV}^{(\mathcal{A})}\Big(q_b^{(\widetilde{n})}, p_0^{\otimes\widetilde{n}}\Big)$ from Lemma 37, we recall that drawing an inhomogeneous spatial Poisson process with intensity function f, defined on $\bigcup_{j\in\mathbf{I}_0}\widetilde{C}_j$, is equivalent to drawing independently for each cell $\widetilde{C}_j, j\in\mathbf{I}_0$, one inhomogeneous spatial Poisson process with intensity $f_{|\widetilde{C}_j}$. For any non-negative function g, denote by $\mathrm{Poi}(g)$ the spatial Poisson process with intensity function g. We can therefore re-index the data generated from $p_0^{\otimes\widetilde{n}} = \mathrm{Poi}(np_0)$, as data generated by $\bigotimes_{j\in\mathbf{I}_0}\mathrm{Poi}\Big(np_{0|\widetilde{C}_j}\Big)$ - and respectively $q_b^{(\widetilde{n})}$ as $\bigotimes_{j\in\mathbf{I}_0}\mathrm{Poi}\Big(nq_{b|\widetilde{C}_j}\Big)$. Moreover, by independence of $(b_j)_{j\in\mathbf{I}_0}$, the events $(\mathcal{A}^{(j)})_{j\in\mathbf{I}_0}$ defined in (3.131) and (3.132) are independent under both $\mathrm{Poi}(2np_0)$ and $\mathrm{Poi}(2nq_b)$ so that Lemma 34 yields:

$$\begin{aligned} d_{TV}^{(\mathcal{A})}\Big(q_b^{(\widetilde{n})}, p_0^{\otimes\widetilde{n}}\Big) &\leq \sum_{j\in\mathbf{I}_0} d_{TV}^{(\mathcal{A}^{(j)})}\left(\mathrm{Poi}\Big(2n\,q_{b|\widetilde{C}_j}\Big),\ \mathrm{Poi}\Big(2n\,p_{0|\widetilde{C}_j}\Big)\right) \\ &= \sum_{j\geq U} d_{TV}^{(\mathcal{A}^{(j)})}\left(\mathrm{Poi}\Big(2n\,q_{b|\widetilde{C}_j}\Big),\ \mathrm{Poi}\Big(2n\,p_{0|\widetilde{C}_j}\Big)\right) \quad \text{since on } \widetilde{C}_0:\ p_0 = q_b \text{ for all } b. \end{aligned}$$

□

Proof of Lemma 39. By Chebyshev's inequality:

$$\mathbb{P}(\mathcal{A}_1^c) = \mathbb{P}(\mathrm{Poi}(2n) < n) \leq \mathbb{P}(|\mathrm{Poi}(2n) - 2n| > n) \leq \frac{2n}{n^2} = \frac{2}{n}$$

Moreover,

$$\begin{aligned} \mathbb{P}(\mathcal{A}_1'^c) = \mathbb{P}(\mathrm{Poi}(2n\|q_b\|_1) < n) &\leq \mathbb{P}\left(\mathrm{Poi}(2n\|q_b\|_1) < n \;\middle|\; \|q_b\|_1 \geq \frac{2}{3}\right) + \mathbb{P}\left(\|q_b\|_1 < \frac{2}{3}\right) \\ &\leq \mathbb{P}\left(\mathrm{Poi}\Big(\frac{4}{3}n\Big) < n\right) + \mathbb{P}\left(\Big|\|q_b\|_1 - 1\Big| > \frac{1}{3}\right) \leq \frac{12}{n} + \frac{9C_{30}}{n^2} \quad \text{by Lemma 30.} \end{aligned}$$

Choosing n_0 such that $\frac{12}{n_0} + \frac{9C_{30}}{n_0^2} \leq \eta/100$, we get the result. □

Proof of Lemma 40.

- For the first quantity:

$$p_0^{\otimes n}(\mathcal{A}^c) \leq \frac{1}{\mathbb{P}(\mathcal{A}_1)}\mathbb{E}_{\widetilde{n}}\left[\mathbb{P}_{p_0^{\otimes n}}\left(\exists j \geq U : \widetilde{N}_j \geq 2 \mid \widetilde{n}\right) \;\middle|\; \widetilde{n} \geq n\right]$$
$$\leq \frac{1}{\mathbb{P}(\mathcal{A}_1)}\mathbb{E}_{\widetilde{n}}\left[\sum_{j\geq U}\widetilde{n}^2 p_j^2 \;\middle|\; \widetilde{n} \geq n\right] = \frac{2n^2}{\mathbb{P}(\mathcal{A}_1)}\sum_{j\geq U} p_j^2$$
$$\leq \frac{2n^2}{\mathbb{P}(\mathcal{A}_1)}\sum_{j\geq U} h^d \int_{\widetilde{C}_j} p_0^2 \leq 3C_{48},$$

by the Cauchy-Schwarz inequality and Lemma 48, and taking $n \geq n_0$. Setting $A_{40} = 3C_{48}$ yields the result.

- For the second quantity:

$$q_b^{(\widetilde{n})}(\mathcal{A}^c) = \mathbb{E}_{b,\widetilde{n}}\left[\mathbb{P}_{q_b^{(\widetilde{n})}}\left(\exists j \geq U : \widetilde{N}_j \geq 2 \mid \widetilde{n}\right)\right]$$
$$\leq \sum_{j\geq U}\mathbb{E}_{b_j,\widetilde{n}}\left[\mathbb{P}_{q_{b_j|\widetilde{C}_j}^{(\widetilde{n})}}\left(\widetilde{N}_j \geq 2 \mid \widetilde{n}\right)\right] \leq \sum_{j\geq U}\mathbb{E}_{b_j,\widetilde{n}}\left[\widetilde{n}^2\left(\int_{\widetilde{C}_j} q_b\right)^2\right]$$
$$= 2n^2\left[\mathbb{V}\left[\|q_b\|_1\right] + \sum_{j\geq U} p_j^2\right] \leq 2C_{30} + 2C_{48} =: B_{40}.$$

Setting $B_{40} := 2C_{30} + 2C_{48}$ yields the result.

□

Proof of Lemma 41. We will use the notation

$$p_j^{(\uparrow)} = p_j + \Gamma_j^{(\uparrow)}, \tag{3.134}$$
$$p_j^{(\downarrow)} = p_j - \Gamma_j^{(\downarrow)}. \tag{3.135}$$

First, we show the following two facts:

Fact 1: For all $a, b \geq 0$ such that $a + b = 1$, and for all $x, y, z \in \mathbb{R}_+$ such that $x = ay + bz$, it holds

$$\left|e^{-x} - ae^{-y} - be^{-z}\right| \leq \frac{x^2}{2} + a\frac{y^2}{2} + b\frac{z^2}{2}.$$

The proof of Fact 1 is straightforward by the relation $1 - u \leq e^{-u} \leq 1 - u + \frac{u^2}{2}$ for all $u \geq 0$.

Fact 2: We have $\mathbb{P}_{p_0^{\otimes n}}\left(\widetilde{N}_j = 1\right) = 2np_je^{-2np_j}$ and moreover it holds:

$$0 \;\leq\; \mathbb{P}_{q_b^{(\widetilde{n})}}\left(\widetilde{N}_j = 1\right) - 2np_je^{-2np_j} \leq C_q n^2 p_j^2,$$

where C_q is a constant.

We now prove Fact 2. Under $p_0^{\otimes\widetilde{n}}$ we have that the number of observations is distributed as $\widetilde{n} \sim \mathrm{Poi}(2n)$ so that $\widetilde{N}_j \sim \mathrm{Poi}(2n\,p_j)$, hence $\mathbb{P}_{p_0^{\otimes\widetilde{n}}}\left(\widetilde{N}_j = 1\right) = 2np_je^{-2np_j}$.

Now, under $q_b^{(\widetilde{n})}$, it holds: $\widetilde{N}_j \sim \pi_j\mathrm{Poi}\left(2np_j^{(\uparrow)}\right) + (1-\pi_j)\mathrm{Poi}\left(2np_j^{(\downarrow)}\right)$. Therefore,

$$\mathbb{P}_{q_b^{(\widetilde{n})}}\left(\widetilde{N}_j = 1\right) = \pi_j 2np_j^{(\uparrow)}e^{-2np_j^{(\uparrow)}} + (1-\pi_j)2np_j^{(\downarrow)}e^{-2np_j^{(\downarrow)}} \tag{3.136}$$

$$= e^{-2np_j}\left[\pi_j 2np_j^{(\uparrow)}e^{-2n\Gamma_j^{(\uparrow)}} + (1-\pi_j)2np_j^{(\downarrow)}e^{2n\Gamma_j^{(\downarrow)}}\right], \tag{3.137}$$

hence $\mathbb{P}_{q_b^{(\widetilde{n})}}\left(\widetilde{N}_j = 1\right) \geq 2np_je^{-2np_j}$ using the inequality $e^x \geq 1+x$. We now prove $\mathbb{P}_{q_b^{(\widetilde{n})}}\left(\widetilde{N}_j = 1\right) \leq 2np_je^{-2np_j} + 8n^2p_j^2$. First, we have $2n\Gamma_j^{(\uparrow)} \leq 1$ since $2n\Gamma_j^{(\uparrow)} = 2n\frac{\|f\|_1 c_j^{(\uparrow)}}{n^2\int_{\mathcal{T}(u_B)}p_0} \leq 2\frac{\|f\|_1 c_u c^{(\downarrow)}}{c_{tail}} \leq 1$ by choosing c_{tail} large enough. Therefore, using the inequality $e^x \leq 1+2x$ for $0 \leq x \leq 1$, we get from Equation (3.137):

$$\begin{aligned}\mathbb{P}_{q_b^{(\widetilde{n})}}\left(\widetilde{N}_j = 1\right) &\leq e^{-2np_j}\left(\pi_j 2np_j^{(\uparrow)} + (1-\pi_j)2np_j^{(\downarrow)}\left(1+4n\Gamma_j^{(\downarrow)}\right)\right)\\ &\leq 2np_je^{-2np_j} + 8n^2p_j\Gamma_j^{(\downarrow)} \leq 2np_je^{-2np_j} + 8n^2p_j^2\\ &=: 2np_je^{-2np_j} + 8n^2p_j\Gamma_j^{(\downarrow)} \leq 2np_je^{-2np_j} + 8n^2p_j^2,\end{aligned}$$

which ends the proof of Fact 2.

Facts 1 and 2 being established, we can now compute $d_{TV}^{(\mathcal{A}^{(j)})}\left(\mathrm{Poi}\left(2n\,q_{b|\widetilde{C}_j}\right),\ \mathrm{Poi}\left(2n\,p_{0|\widetilde{C}_j}\right)\right)$ for fixed $j \geq U$. By Lemma 33 we have:

$$\begin{aligned}&d_{TV}^{(\mathcal{A}^{(j)})}\left(\mathrm{Poi}\left(2n\,q_{b|\widetilde{C}_j}\right),\ \mathrm{Poi}\left(2n\,p_{0|\widetilde{C}_j}\right)\right)\\ &\leq \frac{1}{2}\left[\mathrm{Poi}\left(2n\,q_{b|\widetilde{C}_j}\right)(\mathcal{A}^{(j)^c}) + \mathrm{Poi}\left(2n\,p_{0|\widetilde{C}_j}\right)(\mathcal{A}^{(j)^c}) + \int_{\mathcal{A}^{(j)}}\left|p_{0|\widetilde{C}_j}^{\otimes\widetilde{n}} - q_{b|\widetilde{C}_j}^{(\widetilde{n})}\right|\right]\end{aligned} \tag{3.138}$$

We now compute the term $\int_{\mathcal{A}^{(j)}}\left|p_{0|\widetilde{C}_j}^{\otimes\widetilde{n}} - q_{b|\widetilde{C}_j}^{(\widetilde{n})}\right|$.

$$\int_{\mathcal{A}^{(j)}}\left|p_{0|\widetilde{C}_j}^{\otimes\widetilde{n}} - q_{b|\widetilde{C}_j}^{(\widetilde{n})}\right| = \underbrace{\int_{\{\widetilde{N}_j=0\}}\left|p_{0|\widetilde{C}_j}^{\otimes\widetilde{n}} - q_{b|\widetilde{C}_j}^{(\widetilde{n})}\right|}_{\text{Term 1}} + \underbrace{\int_{\{\widetilde{N}_j=1\}}\left|p_{0|\widetilde{C}_j}^{\otimes\widetilde{n}} - q_{b|\widetilde{C}_j}^{(\widetilde{n})}\right|}_{\text{Term 2}}. \tag{3.139}$$

$$\begin{aligned}\text{Term 1} &= \left|\mathrm{Poi}\left(2n\,p_{0|\widetilde{C}_j}\right)(\widetilde{N}_j = 0) - \mathrm{Poi}\left(2n\,q_{b|\widetilde{C}_j}\right)(\widetilde{N}_j = 0)\right|\\ &= \left|e^{-2np_j} - \pi_j e^{-2np_j^{(\uparrow)}} - (1-\pi_j)e^{-2np_j^{(\downarrow)}}\right|\end{aligned}$$

$$\leq 2n^2p_j^2 + \pi_j 2n^2 {p_j^{(\uparrow)}}^2 + (1-\pi_j)2n^2{p_j^{(\downarrow)}}^2 \quad \text{using Fact 1}$$
$$\leq 4n^2p_j^2 + \pi_j 2n^2 {p_j^{(\uparrow)}}^2.$$

Moreover:

$$\pi_j 2n^2 {p_j^{(\uparrow)}}^2 = \left[\frac{1}{2c_u} p_j n^2 \sum_{l\geq U} p_l\right] 2n^2 \left[p_j + \frac{c_j^{(\uparrow)}\|f\|_1}{n^2 \int_{\mathcal{T}(u_B)} p_0}\right]^2$$
$$\leq \left[\frac{1}{2c_u} p_j\, n^2 \sum_{l\geq U} p_l\right] 2n^2 \left[\frac{1}{n^2 \sum_{l\geq U} p_l}\right]^2 \left[c_u + c_j^{(\uparrow)}\|f\|_1 C_{47}\right]^2 \text{ by Lemma 47}$$
$$\leq \frac{p_j}{\sum_{l\geq U} p_l} c_u \left[1 + 2c^{(\downarrow)}\|f\|_1 C_{47}\right]^2 =: C_\pi^{(\uparrow)} \frac{p_j}{\sum_{l\geq U} p_l}, \tag{3.140}$$

where $C_\pi^{(\uparrow)}$ can be made arbitrarily small by choosing c_u small enough. It follows that

$$\text{Term 1} \leq 4n^2p_j^2 + C_\pi^{(\uparrow)} \frac{p_j}{\sum_{l\geq U} p_l}. \tag{3.141}$$

We now consider Term 2. We set $p^{(\uparrow)}(x) = p_0(x) + \gamma_j^{(\uparrow)}(x)$ and $p^{(\downarrow)}(x) = p_0(x) - \gamma_j^{(\downarrow)}(x)$.

$$\text{Term 2} = \int_{\{\widetilde{N}_j=1\}} \left| p_{0|\widetilde{C}_j}^{\otimes \widetilde{n}} - q_{b|\widetilde{C}_j}^{(\widetilde{n})} \right| = \int_{\widetilde{C}_j} \left| p_0(x) \mathbb{P}_{p_0^{\otimes \widetilde{n}}}\left(\widetilde{N}_j = 1\right) - q_b(x) \mathbb{P}_{q_b^{(\widetilde{n})}}\left(\widetilde{N}_j = 1\right) \right| dx$$
$$= \int_{\widetilde{C}_j} \left| p_0(x) \mathbb{P}_{p_0^{\otimes \widetilde{n}}}\left(\widetilde{N}_j = 1\right) - \left(\pi_j p^{(\uparrow)}(x) + (1-\pi_j)p^{(\downarrow)}(x)\right) \mathbb{P}_{q_b^{(\widetilde{n})}}\left(\widetilde{N}_j = 1\right) \right| dx$$
$$\leq p_j \left| \mathbb{P}_{p_0^{\otimes \widetilde{n}}}\left(\widetilde{N}_j = 1\right) - \mathbb{P}_{q_b^{(\widetilde{n})}}\left(\widetilde{N}_j = 1\right) \right|$$
$$+ \mathbb{P}_{q_b^{(\widetilde{n})}}\left(\widetilde{N}_j = 1\right) \int_{\widetilde{C}_j} \left| \left(\pi_j \gamma_j^{(\uparrow)}(x) - (1-\pi_j)\gamma_j^{(\downarrow)}(x)\right) \right| dx$$
$$\leq C_q n^2 p_j^3 + \mathbb{P}_{q_b^{(\widetilde{n})}}\left(\widetilde{N}_j = 1\right) 2\Gamma_j^{(\downarrow)} \quad \text{by Fact 2 and recalling } \pi_j\Gamma_j^{(\uparrow)} = (1-\pi_j)\Gamma_j^{(\downarrow)}$$
$$\leq C_q n^2 p_j^3 + \left(2np_j + C_q n^2 p_j^2\right) 2p_j \quad \text{by Fact 2 and recalling } \Gamma_j^{(\downarrow)} \leq p_j$$
$$\leq (3C_q + 4) n^2 p_j^2. \tag{3.142}$$

We now control the term $\text{Poi}\left(2n\, q_{b|\widetilde{C}_j}\right)(\mathcal{A}^{(j)c}) + \text{Poi}\left(2n\, p_{0|\widetilde{C}_j}\right)(\mathcal{A}^{(j)c})$ from equation (3.138). We have:

$$\text{Poi}\left(2n\, p_{0|\widetilde{C}_j}\right)(\mathcal{A}^{(j)c}) = \mathbb{P}(\text{Poi}(2np_j) \geq 2) = 1 - e^{-2np_j}(1 + 2np_j)$$
$$\leq 1 - (1 - 2np_j)(1 + 2np_j) = 4n^2p_j^2. \tag{3.143}$$

$$\begin{aligned}
\text{Moreover we have:}\quad \mathrm{Poi}\Big(2n\, q_{b|\widetilde{C}_j}\Big)(\mathcal{A}^{(j)^c}) &= \mathbb{P}\left(\mathrm{Poi}(2n\int_{\widetilde{C}_j} q_b) \geq 2\right) \\
&= \pi_j \mathbb{P}(\mathrm{Poi}(2np_j^{(\uparrow)}) \geq 2) + (1-\pi_j)\mathbb{P}(\mathrm{Poi}(2np_j^{(\downarrow)}) \geq 2) \\
&\leq 4\pi_j n^2 {p_j^{(\uparrow)}}^2 + (1-\pi_j)4n^2 {p_j^{(\downarrow)}}^2 \text{ by (3.143)} \\
&\leq 2C_\pi^{(\uparrow)} \frac{p_j}{\sum_{l\geq U} p_l} + 4n^2 p_j^2 \quad \text{by (3.140).}
\end{aligned} \tag{3.144}$$

Bringing together equations (3.138), (3.139), (3.141), (3.142), (3.143) and (3.144), we get:

$$\begin{aligned}
&d_{TV}^{(\mathcal{A}^{(j)})}\left(\mathrm{Poi}\Big(2n\, q_{b|\widetilde{C}_j}\Big),\ \mathrm{Poi}\Big(2n\, p_{0|\widetilde{C}_j}\Big)\right) \leq \\
&\frac{1}{2}\left[4n^2p_j^2 + 2C_\pi^{(\uparrow)}\frac{p_j}{\sum_{l\geq U} p_l} + 4n^2p_j^2 + 4n^2p_j^2 + 2C_\pi^{(\uparrow)}\frac{p_j}{\sum_{l\geq U} p_l} + (3C_q+4)n^2p_j^2\right] \\
&= (8+\frac{3}{2}C_q)n^2p_j^2 + \frac{3}{2}C_\pi^{(\uparrow)}\frac{p_j}{\sum_{l\geq U} p_l} \\
&=: A_{41}\, n^2p_j^2 + B_{41}\frac{p_j}{\sum_{l\geq U} p_l}.
\end{aligned}$$

□

which ends the proof of Proposition 3.14.

□

Lemma 42. *Denote by $\mathbb{P}_b$ the probability distribution over the realizations of the random variable b. Set $\mathcal{A}_{sep} = \{b \text{ is such that } \|p_0 - p_b\|_t \geq C_{tail}^{LB}\rho_{tail}^*\}$ and $p_{b,cond} = \mathbb{E}(p_b^{(n)}|\mathcal{A}_{sep})$ where the expectation is taken according to the realizations of b. Suppose $\mathbb{P}_b(\mathcal{A}_{sep}^c) \leq \epsilon$. Then $d_{TV}(p_0^{\otimes n}, p_{b,cond}) \leq d_{TV}(p_0^{\otimes n}, \mathbb{E}_b(p_b^{(n)})) + 2\epsilon$.*

Proof of Lemma. We have:

$$\begin{aligned}
d_{TV}\Big(p_{b,cond}, \mathbb{E}_b(p_b^{(n)})\Big) &= \sup_A \Big|p_{b,cond}(A) - p_{b,cond}(A)\mathbb{P}(\mathcal{A}_{\text{sep}}) - \mathbb{E}\Big[p_b^{(n)}(A|\mathcal{A}_{\text{sep}}^c)\Big]\mathbb{P}(\mathcal{A}_{\text{sep}}^c)\Big| \\
&\leq \sup_A \Big|p_{b,cond}(A)(1-\mathbb{P}(\mathcal{A}_{\text{sep}}))\Big| + \mathbb{P}(\mathcal{A}_{\text{sep}}^c) \leq 2\epsilon,
\end{aligned}$$

so that: $d_{TV}\Big(p_0^{\otimes n}, p_{b,cond}\Big) \leq d_{TV}\Big(p_0^{\otimes n}, \mathbb{E}_b(p_b^{(n)})\Big) + d_{TV}\Big(\mathbb{E}_b(p_b^{(n)}), p_{b,cond}\Big) \leq d_{TV}\Big(p_0^{\otimes n}, \mathbb{E}_b(p_b^{(n)})\Big) + 2\epsilon.$

□

3.F.5 Technical results

Lemma 43. *It holds: $\int_{\mathcal{T}(2\bar{C}Lh^\alpha)} p_0 \geq \frac{1}{2}\int_{\mathcal{T}(u_B)} p_0$.*

Proof of Lemma 43. We have:

$$2\bar{C}Lh^\alpha \int_{\mathcal{T}(u_B)\setminus\mathcal{T}(2\bar{C}Lh^\alpha)} p_0 \leq \int_{\mathcal{T}(u_B)} p_0^2 \leq \frac{\bar{C}}{n^2h^d} \quad \text{by Lemma 5.}$$

Therefore:

$$\int_{\mathcal{T}(u_B)} p_0 - \int_{\mathcal{T}(2\bar{C}Lh^\alpha)} p_0 \leq \frac{1}{2}\int_{\mathcal{T}(u_B)} p_0.$$

□

Lemma 44. *It holds $p_U < C_{44}Lh^{\alpha+d}$ where $C_{44} = 2\bar{C}(1+c_\star) + \sqrt{d}^\alpha$ and $h = h_{tail}(u_B)$.*

Proof of Lemma 44. Fix $j \in \mathbb{N}^*$ such that $\widetilde{C}_j \cap \mathcal{T}(2\bar{C}Lh^{\alpha+d}) \neq \emptyset$ and $x \in \widetilde{C}_j$ such that $p_0(x) < 2\bar{C}Lh^{\alpha+d}$. Then by Assumption ($\star$), for all $y \in \widetilde{C}_j$

$$p_0(y) \leq (1+c_\star)p_0(x) + Lh^\alpha\sqrt{d}^\alpha < C_{44}Lh^\alpha$$

so that $p_j \leq C_{44}L^{\alpha+d}$. Therefore, if we had $p_U \geq C_{44}Lh^{\alpha+d}$, then necessarily, $\mathcal{T}(2\bar{C}Lh^\alpha) \subset \bigcup_{j\geq U} \widetilde{C}_j$, hence:

$$\frac{c_u}{n^2} \geq p_U \sum_{j\geq U} p_j \geq C_{44}Lh^{\alpha+d}\int_{\mathcal{T}(2\bar{C}Lh^\alpha)} p_0 \geq \frac{C_{44}}{2n^2} > \frac{c_u}{n^2}$$

for c_u small enough. Contradiction. □

Lemma 45. *Set $h = h_{tail}(u_B)$ and assume that the tail dominates i.e. $C_{BT}\rho^*_{bulk} \leq \rho^*_{tail}$. For all $j \in \mathbb{N}^*$, $j \geq 2$, if $p_j \leq C_{44}Lh^{\alpha+d}$ then $p_{j-1} \leq C_{45}Lh^{\alpha+d}$ where C_{45} is a constant depending only on C_{44}.*

Proof of Lemma 45. Let $j \in \mathbb{N}^*$, $j \geq 2$ and $z_j \in \widetilde{C}_j$ such that $p_0(z_j)h^d = p_j$ and assume $p_j < C_{44}Lh^{\alpha+d}$. Set $C'_{45} = 4(1+c_\star)C_{44} + \sqrt{d}^\alpha$.

Let $y \in \mathcal{T}(u_B)$ such that $p_0(y) = C'_{45}Lh^\alpha$. We can assume $C'_{45}Lh^\alpha < u_B$ by choosing C_{BT} large enough. Indeed, by Lemma 10, we have $u_B \geq c_A L \inf_{x\in\mathcal{B}} h_b^\alpha(x) \geq c_A L C_{BT}^{(2)\alpha} h_{tail}^\alpha(u_B)$ and choosing C_{BT} large enough ensures that $C_{BT}^{(2)\alpha}$ is large enough. Denote by l the index of the cube $\widetilde{C}_l$ containing y.

- First, $p_l > p_j$. Indeed, for all $z \in \widetilde{C}_l$ we have

  $$p_0(z) \geq (1-c_\star)C'_{45}Lh^\alpha - Lh^\alpha\sqrt{d}^\alpha \geq (3C_{44} - \sqrt{d}^\alpha)Lh^\alpha \geq 2C_{44}Lh^\alpha,$$

  hence $p_l = \int_{\widetilde{C}_j} p_0(z)dz \geq 2C_{44}Lh^{\alpha+d} > p_j$.

- Second, for all $z \in \widetilde{C}_l$ we have $p_0(z) \leq (1+c_\star)C'_{45}Lh^\alpha + \sqrt{d}^\alpha Lh^\alpha =: C_{45}Lh^\alpha$ hence $p_l \leq C_{45}Lh^\alpha$.

Since the $(p_l)_l$ are sorted in decreasing we also have $p_{j-1} \leq C_{45} L h^\alpha$. □

Lemma 46. *Suppose that the tail dominates, i.e.: $\rho^*_{tail} \geq C_{BT} \rho^*_{bulk}$. There exists a constant D such that whenever $\int_{\mathcal{T}(u_B)} p_0 \geq \frac{c_{tail}}{n}$, it holds:*

$$\sum_{j \geq U} p_j \geq D \int_{\mathcal{T}(u_B)} p_0.$$

Moreover, D can be made arbitrarily small by choosing c_u small enough and c_{tail} large enough, successively.

Proof of Lemma 46. Recall that by Lemma 44, $p_U \leq C_{44} L h^{\alpha+d}$. Therefore, we cannot have $U = 1$. Indeed, there always exists $j \in \mathbb{N}^*$ such that for some $x \in \widetilde{C}_1, p_0(x) = u_B$ and for this index j:

$$\forall y \in \widetilde{C}_j : p_0(y) \geq (1 - c_\star) u_B - L h^\alpha \sqrt{d}^\alpha > C_{44} L h^\alpha \tag{3.145}$$

for C_{BT} large enough, by Lemma 10 and recalling $u_B \geq c_A L \min_{\mathcal{B}} h_b^\alpha$. Therefore, $p_1 > C_{44} L h^{\alpha+d} \geq p_U$ hence $U \geq 2$. We can then write, by definition of U:

$$p_{U-1} \sum_{j \geq U-1} p_j > \frac{c_u}{n^2},$$

where $p_{U-1} \leq C_{45} L h^{\alpha+d}$ by Lemma 45. Therefore:

$$\begin{aligned}
\sum_{j \geq U} p_j &> \frac{c_u}{n^2 p_{U-1}} - p_{U-1} \geq \frac{c_u}{n^2 C_{45} L h^{\alpha+d}} - C_{45} L h^{\alpha+d} \\
&= \frac{c_u}{C_{45}} \int_{\mathcal{T}(u_B)} p_0 - \frac{C_{45}}{n^2 \int_{\mathcal{T}(u_B)} p_0} \\
&\geq \frac{c_u}{C_{45}} \int_{\mathcal{T}(u_B)} p_0 - \frac{C_{45}}{c_{tail}^2} \int_{\mathcal{T}(u_B)} p_0 \geq D \int_{\mathcal{T}(u_B)} p_0.
\end{aligned}$$

Choosing $c_{tail}^2 \geq \frac{C_{45}^2}{2 c_u}$ yields the result with $D = \frac{c_u}{C_{45}} - \frac{C_{45}}{c_{tail}^2}$.

□

Lemma 47. *In the case where the tail dominates, i.e. when $\rho^*_{tail} \geq C_{BT} \rho^*_{bulk}$, there exists a constant C_{47} such that $\sum_{j \geq U} p_j \leq C_{47} \int_{\mathcal{T}(u_B)} p_0$.*

Proof of Lemma 47. Set $h = h_{tail}(u_B)$. Proceeding like in equation (3.145), it is impossible that $\exists j \geq U,\ \exists x \in \widetilde{C}_j : p_0(x) = u_B$. For all $j \geq U$, we therefore have: $\sup_{\widetilde{C}_j} p_0 < v \leq u_B$, hence $\bigcup_{j \geq U} \widetilde{C}_j \subset \mathcal{T}(u_B)$, which yields $\int_{\bigcup_{j \geq U} \widetilde{C}_j} p_0 \leq \int_{\mathcal{T}(u_B)} p_0$. □

Lemma 48. *Whenever the tail dominates, i.e. when $\rho^*_{tail} \geq C_{BT} \rho^*_{bulk}$, there exists a constant C_{48} such that $\int_{\bigcup_{j \geq U} \widetilde{C}_j} p_0^2 \leq C_{48} \frac{1}{n^2 h^d}$.*

Proof of Lemma 48. Set $h = h_{tail}(u_B)$. Proceeding like in equation (3.145), we have: $\sup_{\widetilde{C}_j} p_0 < v \leq u_B$, hence $\bigcup_{j \geq U} \widetilde{C}_j \subset \mathcal{T}(u_B)$, which yields by Lemma 5: $\int_{\bigcup_{j\geq U} \widetilde{C}_j} p_0^2 \leq \frac{\tilde{C}}{n^2 h^d}$. □

Lemma 49. *Let $p \in \mathcal{P}_{\mathbb{R}^d}(\alpha, L, c_\star)$ and $\Omega' \subset \Omega$ a countable union of cubic domains of $\mathbb{R}^d$.*

1. *If $\int_{\Omega'} p \leq \frac{c}{n}$ for some constant $c > 0$, then*

$$\left(\int_{\Omega'} p^t\right)^{1/t} \leq A_{49}(c) \cdot \rho_r^*,$$

for $A_{49}(c)$ a constant depending only on c, η, d, α and t. Moreover, $A_{49}(c) \underset{c\to 0}{\to} 0$ and $A_{49}(c) \underset{c\to+\infty}{\to} +\infty$.

2. *There exists a constant B_{49} such that $\|p\|_t \leq B_{49} \cdot L^{\frac{d(t-1)}{t(\alpha+d)}}$.*

Proof of Lemma 49. Let $x \in \mathbb{R}^d \setminus \Omega'$ and let $h = \left(\frac{p(x)}{4L}\right)^{1/\alpha}$. By Assumption ($\star$), we have for all $y \in B(x,h)$:

$$p(y) \geq \frac{p_0(x)}{2} - L\left(\frac{p(x)}{4}\right) = \frac{p(x)}{4}.$$

Moreover, by assumption over Ω', $\mathrm{Vol}\left(B(x,h) \cap (\mathbb{R}^d \setminus \Omega')\right) \geq \frac{1}{2}\mathrm{Vol}(B(x,h))$. Therefore:

$$\int_{\mathbb{R}^d\setminus\Omega'} p \;\geq\; \frac{p(x)}{4}\frac{1}{2}\mathrm{Vol}(B(x,h)) \;=\; C_d \frac{p(x)^{\frac{\alpha+d}{\alpha}}}{L^{d/\alpha}}, \tag{3.146}$$

where $C_d = \frac{\mathrm{Vol}\left(B(0,1)\right)}{8\times 4^{d/\alpha}}$.

1. Therefore, if $\int_{\mathbb{R}^d\setminus\Omega'} p \leq \frac{c}{n}$, then $p(x) \leq \left(\frac{c}{C_d}\frac{L^d}{n^\alpha}\right)^{\frac{1}{\alpha+d}}$, which yields:

$$\int_{\mathbb{R}^d\setminus\Omega'} p^2 \leq \left(\frac{c}{C_d}\frac{L^d}{n^\alpha}\right)^{\frac{1}{\alpha+d}} \times \frac{c}{n} = c\left(\frac{c}{C_d}\right)^{\frac{1}{\alpha+d}} \frac{L^{\frac{d}{\alpha+d}}}{n^{\frac{2\alpha+d}{\alpha+d}}}.$$

Now, by Hölder's inequality, we have

$$\int_{\mathbb{R}^d\setminus\Omega'} p^t \leq \left(\int_{\mathbb{R}^d\setminus\Omega'} p\right)^{2-t} \left(\int_{\mathbb{R}^d\setminus\Omega'} p^2\right)^{t-1} =: A_{49}^t(c)\rho_r^{*t}$$

2. The proof of the second assertion follows the same lines. We have by Equation (3.146) that $\forall x \in \mathbb{R}^d : 1 = \int_{\mathbb{R}^d} p \geq C_d \frac{p(x)^{\frac{\alpha+d}{\alpha}}}{L^{d/\alpha}}$, hence $p(x) \leq \frac{1}{C_d^{\alpha/(\alpha+d)}} L^{\frac{d}{\alpha+d}}$ for all $x \in \mathbb{R}^d$. Therefore,

$\int_{\mathbb{R}^d} p^2 \leq \frac{1}{C_d^{\alpha/(\alpha+d)}} L^{\frac{d}{\alpha+d}} \int_{\mathbb{R}^d} p \leq \frac{1}{C_d^{\alpha/(\alpha+d)}} L^{\frac{d}{\alpha+d}}$, so that by Hölder's inequality:

$$\int_{\mathbb{R}^d} p^t \leq \left(\int_{\mathbb{R}^d} p\right)^{2-t} \left(\int_{\mathbb{R}^d} p^2\right)^{t-1} = C_d^{\frac{\alpha(1-t)}{\alpha+d}} L^{\frac{d(t-1)}{\alpha+d}} =: B_{49}^t L^{\frac{d(t-1)}{\alpha+d}}.$$

□

3.G Homogeneity and rescaling

Introduce, for any cubic domain $\Omega \subset \mathbb{R}^d$:

$$\mathcal{P}_\Omega(\alpha, L, c_\star) = \left\{ p \text{ density over } \Omega' \;\middle|\; p \in H(\alpha, L) \text{ and } p \text{ satisfies } (\star) \text{ over } \Omega' \right\}, \tag{3.147}$$

For $\lambda > 0$, define the rescaling operator:

$$\Phi_\lambda : \begin{cases} \mathcal{P}_{\lambda\Omega}(\alpha, L, c_\star) & \longrightarrow & \mathcal{P}_\Omega(\alpha, L\lambda^{\alpha+d}, c_\star) \\ p & \longmapsto & \lambda^d\, p(\lambda\,\cdot) \end{cases} \tag{3.148}$$

where $p(\lambda\,\cdot) : x \mapsto p(\lambda x)$ and $\lambda\Omega = \{\lambda x : x \in \Omega\}$. For any cubic domain $\Omega \subset \mathbb{R}^d$, we define $\rho_\Omega^*(p_0, \alpha, L, n)$ as the minimax separation radius for the following testing problem over Ω, upon observing $X_1, \ldots, X_n$ iid with density $p \in \mathcal{P}_\Omega(\alpha, L, c_\star)$

$$\begin{aligned} H_0 \quad &: p = p_0 \qquad \text{versus} \\ H_1^{(\Omega)}(\rho) &: p \in \mathcal{P}_\Omega(\alpha, L', c_\star') \;\text{ s.t. }\; \|p - p_0\|_t \geq \rho. \end{aligned} \tag{3.149}$$

Proposition 3.15. *(Rescaling) Let $\lambda > 0$ and let $p_0 \in \mathcal{P}_{\lambda\Omega}(\alpha,\, L,\, c_\star)$. It holds*

$$\rho_\Omega^*\left(\Phi_\lambda(p_0),\, \alpha,\, L\lambda^{\alpha+d},\, n\right) = \lambda^{d-d/t}\, \rho_{\lambda\Omega}^*\left(p_0, \alpha, L, n\right).$$

Proposition 3.16. *(Restriction of support) Let $p_0 \in \mathcal{P}_\Omega(\alpha, L, c_\star)$ and $\Omega' \subset \Omega$ another (possibly bounded) cubic domain of $\mathbb{R}^d$. Assume that the support of p_0 is included in Ω'. Then*

$$\rho_\Omega^*(p_0, n, \alpha, L) \asymp \rho_{\Omega'}^*(p_0, n, \alpha, L).$$

Proof of Proposition 3.15. It is direct to prove that $\forall \lambda > 0$, Φ_λ is well-defined and bijective. We can also immediately check that

$$\forall p, q \in \mathcal{P}_{\lambda\Omega}(\alpha, L, c_\star) : \|\Phi_\lambda(p) - \Phi_\lambda(q)\|_t = \lambda^{d-d/t} \|p - q\|_t.$$

Let ψ_λ^* be a test such that

$$\forall p \in \mathcal{P}_{\lambda\Omega}(\alpha, L, c_\star) : \|p - p_0\|_t \geq C\rho_{\lambda\Omega}^*\left(p_0, \alpha, L, n\right) \Longrightarrow \mathbb{P}_{p_0}(\psi_\lambda^* = 1) + \mathbb{P}_p(\psi_\lambda^* = 0) \leq \eta,$$

for some constant C. Now, let $\widetilde{p} \in \mathcal{P}_\Omega(\alpha, L\lambda^{\alpha+d}, c_\star)$ such that

$$\|\widetilde{p} - \Phi_\lambda(p_0)\|_t \geq C\lambda^{d-\frac{d}{t}} \rho^*_{\lambda\Omega}(p_0, \alpha, L, n).$$

It then follows that $\|\Phi_{\lambda^{-1}}(\widetilde{p}) - p_0\|_t \geq C\rho^*_{\lambda\Omega}(p_0, \alpha, L, n)$ hence $\mathbb{P}_{p_0}(\psi^*_\lambda = 1) + \mathbb{P}_{\Phi_{\lambda^{-1}}(\widetilde{p})}(\psi^*_\lambda = 0) \leq \eta$ i.e $\mathbb{P}_{\Phi_\lambda(p_0)}(\psi^* = 1) + \mathbb{P}_{\widetilde{p}}(\psi^* = 0) \leq \eta$ where $\psi^*(x_1, \ldots, x_n) = \psi^*_\lambda(\lambda x_1, \ldots, \lambda x_n)$. Therefore, $C\lambda^{d-d/t}\rho^*_{\lambda\Omega}(p_0, \alpha, L, n) \geq \rho^*_\Omega\left(\Phi(p_0), \alpha, L\lambda^{\alpha+d}, n\right)$ and the converse bound can be proved by symmetry using $\Phi_{\lambda^{-1}}$. □

Proof of Proposition 3.16. Clearly $\rho^*_\Omega(p_0) \geq \rho^*_{\Omega'}(p_0)$. For the converse bound, we define $\psi_{\text{out}} = \mathbb{1}\left\{\bigvee_{i=1}^n (X_i \notin \Omega')\right\}$ the test rejecting H_0 whenever one of the observations X_i belongs to $\Omega \setminus \Omega'$. Lemma 49 shows that, for c_{out} and C_{out} two large enough constants, if $p \in \mathcal{P}_\Omega(\alpha, L, c_\star)$ is such that $\int_{\Omega\setminus\Omega'} p^t \geq C_{\text{out}}\rho^{*t}_r$ then $\int_{\Omega\setminus\Omega'} p \geq \frac{c_{\text{out}}}{n}$, so that $\mathbb{P}_p(\psi_{\text{out}} = 1) > 1 - \eta/2$. Now, assume $\|p - p_0\|_t \geq C\rho^*_{\Omega'}(p_0)$ over Ω, for C a large enough constant, and let ψ^* be an optimal test over Ω', i.e. such that $\mathbb{P}_{p_{0,\Omega'}}(\psi^* = 1) + \mathbb{P}_{p_{\Omega'}}(\psi^* = 0) \leq \eta$ whenever $\|p - p_0\|_t \geq C'\rho^*_{\Omega'}(p_0)$. Then if $\int_{\Omega\setminus\Omega'} p^t \geq C_{\text{out}}\rho^{*t}_r$, $\mathbb{P}_{p_{\Omega'}}(\psi_{\text{out}} \vee \psi^* = 0) \leq \eta/2$. Otherwise, $\int_{\Omega'} |p - p_0|^t \geq C'\rho^*_\Omega(p_0) - C_{\text{out}}\rho^*_r \geq \frac{C'}{2}\rho^*_\Omega(p_0)$ so that $\mathbb{P}_{p_{\Omega'}}(\psi_{\text{out}} \vee \psi^* = 0) \leq \eta/2$ for C' large enough. Moreover, under H_0 , we clearly have $\mathbb{P}_{p_0}(\psi_{\text{out}} \vee \psi^* = 1) \leq \eta/2$. Hence the result. □

3.H Proofs of examples

3.H.1 Uniform distribution

See [183] for $\lambda = 1$ and use Proposition 3.15 for arbitrary $\lambda > 0$.

3.H.2 Arbitrary p_0 over $\Omega = [-1, 1]^d$ with $L = 1$

First, note that by equation (3.146), p_0 is upper bounded by a constant denoted by $C_{\max}$ since $L = 1$. For any small constant c, there exists a fixed constant $\delta > 0$ such that for all p_0 with support over $[-1, 1]^d$, the set $\{p_0 \geq c\}$ has Lebesgue measure at least δ. Fix such a c. Now, there exists a constant n_0 such that for all $n \geq n_0$, for all p_0, $u_{aux}(p_0) \leq c$. We then have

$$C^r_{\max} 2^d \geq \int_{\mathcal{B}(u_{aux})} p_0^r \geq c^r \delta \quad \text{which is a constant.}$$

Therefore, ${\rho^*_{bulk}}^t \asymp n^{-\frac{2\alpha t}{4\alpha+d}}$.

As for the tail, we have by the Cauchy-Schwarz inequality and Lemma 5

$$\left(\int_{\mathcal{T}(u_B)} p_0\right)^2 \leq \frac{\left(\int_{\mathcal{T}(u_B)} p_0\right)^2}{\left|\mathcal{T}(u_B) \cap [0,1]^d\right|} \leq \int_{\mathcal{T}(u_B)} p_0^2 \leq \bar{C}\widetilde{L}^{\frac{1}{\alpha+d}} \left(\int_{\mathcal{T}(u_B)} p_0\right)^{\frac{d}{\alpha+d}}$$

hence

$$p_0[\mathcal{T}] \leq \bar{C}^{\frac{\alpha+d}{2\alpha+d}} n^{-\frac{2\alpha}{2\alpha+d}}.$$

We can now immediately check that $\rho^*_{bulk} \gg \rho^*_{tail}$ and $\rho^*_{bulk} \gg \rho^*_r$ as $n \to \infty$. Since ρ^*_{bulk} is independent of p_0, the result is proven.

3.H.3 Spiky null

Set $\widetilde{p_0}(x) = \frac{f}{\|f\|_1}$ over $\mathbb{R}^d$. Since $\widetilde{p_0}$ takes nonzero values only over $[\pm\frac{1}{2}]^d$ we have $\rho^*(,\alpha,1,n) \asymp n^{-\frac{2\alpha}{4\alpha+d}}$ by the preceding case. Now, by homogeneity (see Proposition 3.15), we have $\rho^*(p_0,\alpha,L,n) = L^{\frac{d(t-1)}{t(\alpha+d)}}\rho^*(\widetilde{p_0},\alpha,1,n)$, which yields the result.

3.H.4 Gaussian null

Note that

$$\int_{\|x\|>b} p_0(x)dx = e^{-\frac{b^2}{2\sigma^2}(1+o(1))} \quad \text{when } b \to +\infty. \tag{3.150}$$

Therefore, noting b_I the unique value such that if $\|x\| = b_I$, then $p_0(x) = u_{aux}$, we have by the definition of u_{aux} that $b_I = \sigma^2\frac{4\alpha}{2\alpha+d}\log(n)(1+o(1))$ when $n \to +\infty$. By Lemma 6 and using (3.150), it holds $\int_{\mathcal{T}(u_B)} p_0 \asymp \int_{\widetilde{\mathcal{T}}(u_{aux})} p_0 = \int_{\mathcal{T}(u_{aux})} p_0 = n^{-\frac{2\alpha}{2\alpha+d}(1+o(1))} \gg \frac{1}{n}$, so that the tail rate writes

$$\rho^*_{tail} \asymp L^{\frac{d(t-1)}{t(\alpha+d)}} n^{-\frac{2\alpha}{2\alpha+d}(1+o(1))} \gg \rho^*_r.$$

Now, by direct calculation, $\rho^*_{bulk} \asymp \frac{L^{\frac{d}{4\alpha+d}}}{n^{\frac{2\alpha}{4\alpha+d}}}(\sigma^d)^{\frac{(4-3t)\alpha+d}{t(4\alpha+d)}}$ and we can immediately check that it is the dominant term.

3.H.5 Pareto null

Fix $d = t = 1$ and $\alpha \leq 1$. We let $q_{aux} > x_1$ denote the unique value such that $p_0(q_{aux}) = u_{aux}$. By the definition of u_{aux} and using simple algebra we get $q_{aux} \asymp \widetilde{L}^{-\frac{1}{3\beta+\alpha+1}}$. Moreover, we have by Lemma 6 that $\int_{\mathcal{T}(u_B)} p_0 \asymp \int_{\widetilde{\mathcal{T}}(u_{aux})} p_0 = \int_{\mathcal{T}(u_{aux})} p_0 = \widetilde{L}^{\frac{\beta}{3\beta+\alpha+1}}$ so that (by recalling $t = 1$): $\rho^*_{tail} \asymp \int_{\mathcal{T}(u_B)} p_0 \asymp \widetilde{L}^{\frac{\beta}{3\beta+\alpha+1}} \gg \rho^*_r$. Now, we can easily get $\rho^*_{bulk} \asymp \widetilde{L}^{\frac{1}{4\alpha+1}} \ll \rho^*_{tail}$ which ends the proof.

Part II

Estimation with learning constraints

Chapter 4

Robust learning under local differential privacy

This Chapter is based on the paper "Robust Estimation of Discrete Distributions under Local Differential Privacy" [169], by Julien Chhor and Flore Sentenac (arXiv:2202.06825).

Abstract

Although robust learning and local differential privacy are both widely studied fields of research, combining the two settings is just starting to be explored. We consider the problem of estimating a discrete distribution in total variation from n contaminated data batches under a local differential privacy constraint. A fraction $1-\epsilon$ of the batches contain k i.i.d. samples drawn from a discrete distribution p over d elements. To protect the users' privacy, each of the samples is privatized using an α-locally differentially private mechanism. The remaining ϵn batches are an adversarial contamination. The minimax rate of estimation under contamination alone, with no privacy, is known to be $\epsilon/\sqrt{k} + \sqrt{d/kn}$, up to a $\sqrt{\log(1/\epsilon)}$ factor. Under the privacy constraint alone, the minimax rate of estimation is $\sqrt{d^2/\alpha^2 kn}$. We show that combining the two constraints leads to a minimax estimation rate of $\epsilon\sqrt{d/\alpha^2 k} + \sqrt{d^2/\alpha^2 kn}$ up to a $\sqrt{\log(1/\epsilon)}$ factor, larger than the sum of the two separate rates. We provide a polynomial-time algorithm achieving this bound, as well as a matching information theoretic lower bound.

Keywords Privacy, Robustness, Adversarial Contamination, Multionmial Distributions, Statistical Optimality.

4.1 Introduction

In recent machine learning developments, the growing need to analyze potentially corrupted, biased or sensitive data has given rise to unprecedented challenges. To extract relevant information from today's data, studying algorithms under new learning constraints has emerged as a major necessity. To name a few, let's mention learning from incomplete data, transfer learning, fairness, robust learning or privacy. Although each one of them has been subject to intense progress in recent works, combining several learning constraints is not conventional. In this work, we propose

to study how to estimate discrete distributions under the constraint of both being robust to adversarial contamination and of ensuring local differential privacy.

On the one hand, robust learning has received considerable attention over the past decades. This recent research has been developing in two main directions. The first one deals with robustness to heavy tails, see [58], see also [113] for an excellent review. The second one explores robustness to outliers. It mainly considers two contamination models, which are the Huber contamination model [6], [21], [181], [49], where the outliers are iid with an unknown probability distribution, and the *adversarial* contamination, where the outliers are added by a malicious adversary who knows the estimation procedure, the underlying distribution and the data and seeks to deteriorate the procedure's estimation performance ([110, 55, 171, 141]).

On the other hand, preserving the privacy of individuals has emerged as a major concern, as more and more sensitive data are collected and processed. The most commonly used privatization framework is that of differential privacy ([43], [125], [174], [124], [156]). Both central and local models of privacy are considered in the field. In the centralized case, a global entity collects the data and analyzes it before releasing a privatized result, from which the original data should not be possible to infer. In local privacy, the data themselves are released and should remain private ([68]). The paper focuses on the latter notion. A vast line of work also studies private mechanism under communication constraints ([148], [117], [119]), which we do not consider here, but adding a communication constraint would be interesting future work.

Connections between robustness and *global* differential privacy have been recently discussed in ([115, 112], [145]). These papers show that the two notions rely on the same theoretical concepts, and that results in the two fields are related. In other words, robustness and *global* differential privacy work well together. Several papers developed algorithms under robustness and *global* differential privacy constraints ([163], [160], [164], [151]).

In this paper, we study how *local* differential privacy interacts with robustness. This interaction has been studied previously in [157], where the authors provide upper and lower bound for estimating discrete distributions under the two constraints. The lower bound was later tightened in [150]. The papers also study testing. We detail in Section 4.1.1 how our setting is a generalization of theirs. The work of [176] also considers this interaction. We explain in more details the differences between their setting and ours in Section 4.1.1.

In this paper, we study how to combine robust statistics with local differential privacy for estimating discrete distributions over finite domains. Assume that we want to gather information from n data centers (think of n hospitals for instance). For each of them, we collect k iid observations with unknown discrete distribution p to be estimated. To protect the users' privacy (patients data in the hospital example), each single one of the nk observations is privatized using an α-locally differentially private mechanism (see the formal definition of local differential privacy in Subsection 4.2.1). However, an ϵ-fraction of the data centers are untrustworthy and can send adversarially chosen data. The goal is to estimate p in total variation distance (or ℓ_1 distance) from these n

corrupted and privatized batches of size k. This setting is quite natural, as in many applications, the data are collected in batches, some of which may be untrustworthy or even adversarial.

4.1.1 Related work

With the local differential privacy constraint only (i.e. without contamination), the problem of estimating discrete distributions has been solved in ([38], [54])where the authors propose a polynomial-time and minimax optimal algorithm for estimation under ℓ_1 and ℓ_2 losses. Note that *without privacy and outliers*, the minimax estimation rate in ℓ_1 is known to be $\sqrt{\frac{d}{N}}$, where N is the number of iid samples with a discrete distribution over d elements (see, e.g. [73]). The paper [68] shows that under privacy alone, the ℓ_1 minimax rate scales as $\frac{d}{\alpha\sqrt{N}}$. We give an alternative proof of the lower bound of [68], in Appendix 4.D.

With the robustness constraint only (i.e. with n adversarially corrupted batches but without local differential privacy), the problem of estimating discrete distributions has been considered in [93]. For $k = 1$, it is well known that $\Omega(\epsilon)$ error is unavoidable. However, [93] surprisingly prove that the error can be reduced provided that k is large enough. More precisely, they show that with no privacy but under contamination, the minimax risk of estimation under ℓ_1 loss from n batches of size k and ϵ adversarial corruption on the batches scales as $\sqrt{\frac{d}{N}} + \frac{\epsilon}{\sqrt{k}}$, where $N = nk$. [93] both provide an information theoretic lower bound and a minimax optimal algorithm, unfortunately running in exponential time in either k or d. Polynomial-time algorithms were later proposed by [130], [138] and were shown to reach the information theoretic lower bound up to an extra $\sqrt{\log(\frac{1}{\epsilon})}$ factor. In this specific setting, it is not known if this extra factor represents a computational gap between polynomial-time and exponential-time algorithms. However, for the problem of robust mean estimation of *normal distributions*, some lower bounds suggest that this exact quantity cannot be removed from the rate of computationally tractable estimators (see [89]).

Closer to our setting, the papers by [157], [150] and [176] combine robustness with local differential privacy. The problem studied here is a generalisation of the first two papers where the authors consider un-batched data, which corresponds to $k = 1$ in our setting. The setting considered by [176] is not the same as ours, as do not consider discrete distributions and implicitly assume $k = 1$. More importantly, in their setting, contamination comes *before* privacy: some of the raw data $X_1, \ldots, X_n$ are outliers themselves, and the privacy mechanism is applied on each X_i. Conversely, in our work and in the previous two papers, contamination occurs *after* privacy: none of the raw data are outliers and the adversary is allowed to choose the contamination directly on the set of privatized data. As we will highlight below, this difference yields fundamentally different phenomena compared to the results in [176].

4.1.2 Summary of the contributions

In this paper, we study the interplay between local differential privacy and *adversarial* contamination, when the contamination comes *after* the data have been privatized. In this case, we prove that the resulting estimation rate is not merely the sum of the two estimation rates stated in [68] and [93] but is always slower. More specifically, the term due to the contamination in the bound

suffers a multiplicative inflation of $\sqrt{d}/\alpha$. This generalizes a phenomenon first observed in [157]. This phenomenon stands in contrast with [176], for which the resulting rate is exactly the sum of the rate with privacy but no contamination, plus the rate with contamination but no privacy. The reason is that in [176], contamination occurs *before* privacy. We provide an explicit algorithm that returns an estimator achieving the optimal bound up to a factor $\sqrt{\log(1/\epsilon)}$, and runs polynomially in all parameters. This algorithm is an adaptation to our setting of methods that were previously used for robust estimation of discrete distributions ([138, 161]). On a side note, the algorithms introduced in [157] and [150] require the use of a public coin. The proposed algorithm also holds in their setting and relieves this assumption.

4.2 Setting

4.2.1 Definitions

For any integer $d \geq 2$, denote by $\mathcal{P}_d = \left\{ p \in \mathbb{R}^d \;\middle|\; \forall j : p_j \geq 0 \text{ and } \sum_{j=1}^d p_j = 1 \right\}$ the set of probability vectors over $\{1, \dots, d\}$. For any $x \in \mathbb{R}^d$, we write $\|x\|_1 = \sum_{j \in [d]} |x_j|$ and $\|x\|_2^2 = \sum_{j \in [d]} x_j^2$. For any two probability distributions p, q over some measurable space $(\mathcal{X}, \mathcal{A})$, we denote by

$$TV(p, q) = \sup_{A \in \mathcal{A}} |p(A) - q(A)| \text{ the total variation between } p \text{ and } q.$$

Fix $\alpha \in (0, 1)$ and consider two measurable spaces $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Z}, \mathcal{B})$. A Markov transition kernel $Q : (\mathcal{X}, \mathcal{A}) \to (\mathcal{Z}, \mathcal{B})$ is said to be a (non-interactive) α-*locally differentially private* mechanism if it satisfies

$$\sup_{B \in \mathcal{B}} \sup_{x, x' \in \mathcal{X}} \frac{Q(B|x)}{Q(B|x')} \leq e^\alpha. \tag{4.1}$$

For any $x \in \mathcal{X}$, we say that the random variable Z is a privatized version of x if $Z \sim Q(\cdot|x)$. The measurable space $(\mathcal{Z}, \mathcal{B})$ is called the *image space* of Q. In what follows, we use the Landau notation O which hides an absolute constant, independent of $d, \epsilon, n, k, \alpha, Q, p$.

4.2.2 Model

We consider the problem of learning a discrete distribution p over a finite set $\{1, \dots, d\}$, $d \geq 3$ under two learning constraints: a) ensuring α-local differential privacy and b) being robust to adversarial contamination. To this end, we assume that the data are generated as follows. For some small enough absolute constant $c \in (0, \frac{1}{100})$ and for some known corruption level $\epsilon \in (0, c)$, we will use the notation $n' = n(1 - \epsilon)$ throughout and assume that $n' \in \mathbb{N}$.

1. First, n' iid *batches* of observations $X^1, \dots, X^{n'}$ are collected. More precisely, each batch X^b can be written as $X^b = (X_1^b, \dots, X_k^b)$ and consists of k iid random observations with an unknown discrete distribution $p \in \mathcal{P}_d$, i.e. $\forall (b, l, j) \in [n'] \times [k] \times [d] : \mathbb{P}(X_l^b = j) = p_j$.

2. Second, we privatize each of the $n'k$ observations using an α-LDP mechanism Q, yielding n' iid batches $Y^1, \dots, Y^{n'}$ such that $Y^b = (Y_1^b, \dots, Y_k^b)$ where $Y_l^b | X_l^b \sim Q(\cdot | X_l^b)$. We denote

by Qp the distribution of any random variable Y_l^b. We then have: $Qp(dz) = \sum_{j\in[d]} p_j Q(dz|j)$, where $Q(dz|j)$ is a shorthand for $Q(dz|X=j)$. The mechanism Q is chosen by the statistician in order to preserve statistical performance while ensuring privacy.

3. An adversary is allowed to build $n\epsilon$ batches $Y^{n'+1}, \dots, Y^n$ on which no restriction is imposed. Then, he shuffles the set of n batches $(Y_1, \dots, Y_n)$. The resulting set of observations, denoted as $B = (Z^1, \dots, Z^n)$, is referred to as the ϵ-corrupted family of batches.

The observed dataset therefore consists of $n = |B|$ batches of k samples each. Among these batches is an unknown collection of *good batches* $B_G \subset B$ of size $n(1-\epsilon)$, corresponding to the non-contaminated batches. The remaining set $B_A = B \setminus B_G$ of size $n\epsilon$, denotes the unknown set of adversarial batches.

The statistician never has access to the actual observations $X^1, \dots, X^{n'}$, but only to $Z^1, \dots, Z^n$ where $Z^b = (Z_1^b, \dots, Z_k^b)$. Each batch is assumed to be either entirely clean or adversarially corrupted. Note that observing n batches of size k encompasses the classical case where $k = 1$, for which the data consist of n iid and ϵ-corrupted *single observations* rather than batches. On top of being more general, the setting with general k allows us to derive faster rates for large k than for the classical case $k = 1$. Note also that in our setting, the contamination comes after the data have been privatized, which is one of the main differences with [176], where the authors assume that the Huber contamination comes before privacy. The examples considered by the authors are 1-dimensional mean estimation and density estimation without batches (i.e. for $k = 1$). In these settings, the authors surprisingly prove that the algorithm that would be used in absence of corruption is automatically robust to Huber contamination.

In our setting, we would like to answer the following questions:

1. When contamination comes *after* privacy, do we need to design robust procedures or would the private procedure be automatically robust like in [176]?
2. If Q_ϵ denotes the optimal privacy mechanism for ϵ-contamination, how does Q_ϵ depend on ϵ?

We answer these questions as follows:

1. With contamination *after* privacy, the procedure that we would use if there were no contamination is no longer robust and a new algorithm is needed.
2. The optimal privacy mechanism Q_ϵ does not depend on ϵ, whereas the optimal estimator does.

We introduce the minimax framework. An *estimator* $\widehat{p}$ is a measurable function of the data taking values in $\mathcal{P}_d$.

$$\widehat{p} : \mathcal{Z}^{nk} \longrightarrow \mathcal{P}_d.$$

For any set of n' clean batches $Y^1, \dots, Y^{n'}$ where $Y^b = (Y_1^b, \dots, Y_k^b)$ and $n' = n(1-\epsilon)$, we define the set of ϵ-contaminated families of n batches as

$$\mathcal{C}(Y^1, \dots, Y^{n'}) = \left\{ (Z^b)_{b=1}^n \,\middle|\, \exists J \subset [n] \text{ s.t. } |J| = n\epsilon \text{ and } \{Z^b\}_{b\notin J} = \{Y^1, \dots, Y^{n'}\} \right\}. \tag{4.2}$$

We are interested in estimating $p \in \mathcal{P}_d$ with guarantees in high probability. We therefore introduce the minimax estimation rate of p in high probability as follows.

Definition 4.1. *Given $\delta > 0$, the minimax rate of estimation rate of $p \in \mathcal{P}_d$ given the privatized and ϵ-corrupted batches $(Z^b)_{b=1}^n$ where $\forall i \in \{1, \dots, n\} : Z^b = (Z_1^b, \dots Z_k^b)$ is defined as the quantity $\psi_\delta^*(n, k, \alpha, d, \epsilon)$ satisfying*

$$\psi_\delta^*(n, k, \alpha, d, \epsilon) = \inf \left\{ \psi > 0 \;\middle|\; \inf_{\hat{p}, Q} \sup_{p \in \mathcal{P}_d} \mathbb{P}\Big(\sup_{z \in \mathcal{C}(Y)} \left\| \widehat{p}(z) - p \right\|_1 > \psi \Big) \leq \delta \right\}. \tag{4.3}$$

where the infimum is taken over all estimators $\widehat{p}$ and all α-LDP mechanisms Q, and the expectation is taken over all collections of n' clean batches $Y^1, \dots, Y^{n'}$ where $Y^b = (Y_1^b, \dots, Y_k^b)$ and $Y_l^b \overset{iid}{\sim} Qp$. Informally, ψ_δ^* represents the infimal distance such that there exists an estimator $\hat{p}$ able to estimate any $p \in \mathcal{P}_d$ within total variation ψ_δ^* with probability $\geq 1 - \delta$. The ℓ_1 norm is a natural metric for estimating discrete distributions since $TV(p, q) = \frac{1}{2}\|p - q\|_1$ for any $p, q \in \mathcal{P}_d$ (see [188]).

4.3 Results

We now state our main Theorem.

Theorem 4.1. *Assume $d \geq 3$. There exist absolute constants $c, C, C', C'' > 0$ such that for $\delta = C' e^{-d}$, we have:*

$$\psi_\delta^*(n, k, \alpha, \epsilon, d) \geq c \left\{ \left(\frac{d}{\alpha\sqrt{kn}} + \frac{\epsilon}{\alpha}\sqrt{\frac{d}{k}} \right) \wedge 1 \right\},$$

and if $n \geq C'' d$ then

$$\psi_\delta^*(n, k, \alpha, \epsilon, d) \leq C \left\{ \left(\frac{d}{\alpha\sqrt{kn}} + \frac{\epsilon\sqrt{\log(1/\epsilon)}}{\alpha}\sqrt{\frac{d}{k}} \right) \wedge 1 \right\}.$$

In short, we prove that with probability at least $1 - O(e^{-d})$, it is possible to estimate any $p \in \mathcal{P}_d$ within total variation of the order of $\left(\frac{d}{\alpha\sqrt{kn}} + \frac{\epsilon}{\alpha}\sqrt{\frac{d}{k}} \right) \wedge 1$ up to log factors and provided that $n \geq C'' d$. We can compare this rate with existing results in the literature.

- As shown in [68], the term $\frac{d}{\alpha\sqrt{kn}} \wedge 1$ corresponds to the estimation rate under privacy if there were no outliers, with a total number of observations of $N = nk$.
- The term $\frac{\epsilon\sqrt{d}}{\alpha\sqrt{k}} \wedge 1$ reveals an interesting interplay between contamination and privacy. In absence of privacy, [93] proved that the contribution of the contamination is of the order of $\frac{\epsilon}{\sqrt{k}} \wedge 1$. The effect of the corruption therefore becomes more dramatic when it occurs after privatization.
- Letting $k' = \frac{\alpha^2}{d} k$, our rate rewrites $\psi^*(n, k, \alpha, \epsilon, d) \asymp \left(\sqrt{\frac{d}{k'n}} + \frac{\epsilon}{\sqrt{k'}} \right) \wedge 1$. Noticeably, this rate exactly corresponds to the rate from [93] if we had an ϵ-corrupted family of n *non-privatized*

batches $X_1, \dots, X_n$, and if each batch contained k' observations. The quantity k' therefore acts as an effective sample size and the effect of privacy amounts to shrinking the number of observations by a factor α^2/d.

- For the upper bound, the assumption $n \geq C''d$ is classical in the robust statistics literature, even in the gaussian setting (see e.g. [171]).

4.3.1 Lower bound

The following Proposition yields an information theoretic lower bound on the best achievable estimation accuracy under local differential privacy and adversarial contamination.

Proposition 4.1. *Assume $d \geq 3$. There exist two absolute constants $C, c > 0$ such that for all $\epsilon \in (0, \frac{1}{2})$, for all estimator $\hat{p}$ and all α-LDP mechanism Q, there exists a probability vector $p \in \mathcal{P}_d$ satisfying*

$$\mathbb{P}_p \left[\sup_{z' \in \mathcal{C}(Y)} \left\| \widehat{p}(z') - p \right\|_1 \geq c \left\{ \left(\frac{d}{\alpha\sqrt{kn}} + \frac{\epsilon\sqrt{d}}{\alpha\sqrt{k}} \right) \wedge 1 \right\} \right] \geq Ce^{-d},$$

where the probability $\mathbb{P}_p$ is taken over all collections of n' clean batches $Y = (Y^1, \dots, Y^{n'})$ where $Y^b = (Y_1^b, \dots, Y_k^b)$ and $Y_l^b \overset{iid}{\sim} Qp$.

The proof is given in Appendix 4.C. At a high level, the term $\frac{d}{\alpha\sqrt{kn}} \wedge 1$ comes from the classical lower bound given in [68]. The proof of the second term $\frac{\epsilon\sqrt{d}}{\alpha\sqrt{k}} \wedge 1$ is new. It is based on the fact that for any α-LDP mechanism Q, it is possible to find two probability vectors $p, q \in \mathcal{P}_d$ such that $\|p - q\|_1 \gtrsim \frac{\epsilon\sqrt{d}}{\alpha\sqrt{k}} \wedge 1$ and $TV(Qp^{\otimes k}, Qq^{\otimes k}) \leq \epsilon$. In other words, we prove:

$$\inf_{Q} \sup_{\substack{(p,q) \in \mathcal{P}_d: \\ TV(Qp^{\otimes k}, Qq^{\otimes k}) \leq \epsilon}} \|p - q\|_1 \gtrsim \frac{\epsilon\sqrt{d}}{\alpha\sqrt{k}} \wedge 1.$$

In the proof, we argue that $TV(Qp^{\otimes k}, Qq^{\otimes k}) \leq \epsilon$ represents an indistinguishability condition under ϵ-contamination. Namely, it implies that, even if we had arbitrarily many clean batches drawn from p or q, the adversary could add $n\epsilon$ corrupted batches such that the resulting family of batches has the same distribution under p or q. By observing this limiting distribution, it is therefore impossible to recover the underlying probability distribution so that an error of $\|p - q\|_1/2$ is unavoidable.

To exhibit two vectors $p, q \in \mathcal{P}$ satisfying this, we restrict ourselves to vectors satisfying $\chi^2(Qp||Qq) \leq C\frac{\epsilon^2}{k}$ for some small enough absolute constant $C > 0$, which implies that $TV(Qp^{\otimes k}, Qq^{\otimes k}) \leq \epsilon$ (see [188] section 2.4). Noticeably, we prove the relation

$$\chi^2(p||q) = \Delta^T \Omega \Delta,$$

where $\Delta = p - q$ and $\Omega = \Omega(Q) = \left[\int_{\mathcal{Z}} \left(\frac{Q(z|i)}{Q(z|1)} - 1 \right) \left(\frac{Q(z|j)}{Q(z|1)} - 1 \right) Q(z|1) dz \right]_{i,j \in [d]}$ is a nonnegative symmetric matrix. The eigenvectors of Ω play an important role. Namely, we prove that we can

choose a vector Δ in the span of the first $\lceil \frac{2d}{3} \rceil$ eigenvectors of Ω such that $\Delta^T \Omega \Delta \leq C\frac{\epsilon^2}{k}$ and $\|\Delta\|_1 \gtrsim \frac{\epsilon\sqrt{d}}{\alpha\sqrt{k}} \wedge 1$. Defining the vectors $p = \left(\frac{|\Delta_j|}{\|\Delta\|_1}\right)_{j=1}^d \in \mathcal{P}_d$ and $q = p - \Delta$ ends the proof.

4.4 Upper bound

We now address the upper bound by proposing an α-LDP mechanism Q for privatizing the clean data $X^1, \ldots, X^{n'}$ as well as an algorithm $\widehat{p}$ for robustly estimating vector p given an ϵ-contaminated family of n batches $Z^1, \ldots, Z^n$.

Each non-private data point $X_i^b \in [d]$ is privatized using the RAPPOR algorithm introduced in ([68, 82]). In this procedure, the *privatization channel* Q randomly maps each point $X \in [d]$ to a point $Z \in \{0,1\}^d$ by flipping its coordinates independently at random with probability $\lambda = \frac{1}{e^{\alpha/2}+1}$:

$$\forall j \in [d]: \quad Z(j) = \begin{cases} \mathbb{1}_{X=j} & \text{with probability } 1-\lambda, \\ 1 - \mathbb{1}_{X=j} & \text{otherwise.} \end{cases}$$

We now derive a polynomial-time algorithm taking as input the ϵ-contaminated family of batches $(Z^b)_{b\in[n]}$ and returning an estimate $\hat{p}$ for p with the following properties.

Theorem 4.2 (Upper Bound)**.** *For any $\epsilon \in (0, 1/100]$, $\alpha \in (0,1]$, if $n \geq \frac{4d}{\epsilon^2 \ln(e/\epsilon)}$, Algorithm 4 runs in polynomial time in all parameters and its estimate $\widehat{p}$ satisfies $||\widehat{p} - p||_1 \lesssim \frac{\epsilon}{\alpha}\sqrt{\frac{d\ln(1/\epsilon)}{k}}$ w.p. at least $1 - O(e^{-d})$.*

If $n \geq O(d)$, then there exists $\epsilon' \in (0, 1/100]$ s.t. $n = \frac{4d}{(\epsilon')^2 \ln(1/\epsilon')}$. Running the algorithm with that parameter ϵ' rather than the true ϵ gives the following result.

Corollary 4.1. *If $n \geq O(d)$, then the algorithm's estimate satisfies $||\widehat{p} - p||_1 \lesssim \frac{d}{\alpha}\sqrt{\frac{e}{nk}}$ with probability at least $1 - O(e^{-d})$.*

Theorem 4.2 and Corollary 4.1 yield the upper bound. We have not seen the regime $d \leq n$ explored in the literature, even with robustness only. This would be an interesting research direction for future work. Note that for the estimate $\hat{p}$ given by Algorithm 4 we can have $\|\hat{p}\|_1 \neq 1$. The next corollary, proved in Appendix 4.B, states that normalizing $\hat{p}$ yields an estimator in $\mathcal{P}_d$ with the same estimation guarantees as in Theorem 4.2.

Corollary 4.2. *Let the assumptions of Theorem 4.2 be satisfied and let $\hat{p}$ denote the output of Algorithm 4. Define $\widehat{p}^* = \frac{\widehat{p}_+}{\|\widehat{p}_+\|_1}$ where $\widehat{p}_+(j) = 0 \vee \widehat{p}(j)$ for all $j \in [d]$, then $||\widehat{p}^* - p||_1 \leq 2||\widehat{p} - p||_1 \lesssim \frac{\epsilon}{\alpha}\sqrt{\frac{d\ln(1/\epsilon)}{k}}$ holds with probability at least $1 - O(e^{-d})$.*

4.4.1 Description of the algorithm

We now give a high level description of our algorithm. It is based on algorithms for robust discrete distribution estimation, [138, 161]. For each $S \subseteq [d]$, define $q(S) = \sum_{j\in S} q_j$ and $p(S) = \sum_{j\in S} p_j$.

The quantities $\widehat{q}$, $\widehat{p}$ will respectively denote the estimators of p and q. Recalling that $TV(p,\widehat{p}) = \sup_{S\subseteq[d]} \left|p(S) - \widehat{p}(S)\right|$, we aim at finding $\widehat{p}$ satisfying $\left|p(S) - \widehat{p}(S)\right| \lesssim \frac{\epsilon}{\alpha}\sqrt{\frac{d\ln(1/\epsilon)}{k}}$ for all $S \subseteq [d]$. To this end, it is natural to first estimate the auxiliary quantity

$$q(j) := \mathbb{E}_p\Big[Z(j)\Big|\ Z \text{ is a good sample}\Big] \quad \text{for all } j \in [d],$$

which is linked with $p(j)$ through the formula $p(j) = \frac{q(j)-1}{1-2\lambda}$. Our algorithm therefore first focuses on robustly estimating q and outputs $\widehat{p} = \frac{\widehat{q}-\mathbb{1}}{1-2\lambda}$. If there were no outliers, we would estimate $q(j)$ by $\frac{1}{nk}\sum_{b\in[n]}\sum_{l\in[k]} Z_l^b(j)$. In the presence of outliers, our algorithm iteratively deletes the batches that are likely to be contaminated, and returns the empirical mean of the remaining data. More precisely, at each iteration, the current collection of remaining batches B' is processed as follows:

1. Compute the *contamination rate* $\sqrt{\tau_{B'}}$ (defined in equation 4.9) of the collection B'. If $\sqrt{\tau_{B'}} \leq 200$, return the empirical mean of the elements in B'.
2. If $\sqrt{\tau_{B'}} \geq 200$, compute the *corruption score* ε_b (defined in equation 4.10) of each batch $b \in B'$. Select the subset B^o of the $n\epsilon$ batches of B' with top corruption scores. Iteratively delete one batch in B^o: at each step, choose a batch b with probability proportional to ε_b, until the sum of all ε_b in B^o has been halved.

At a high level, *contamination rate* $\tau_{B'}$ quantifies how many adversarial batches remain in the current collection B'. The *corruption score* ε_b quantifies how likely it is for batch b to be an outlier. Both the *contamination rate* and the *corruption scores* can be computed in polynomial time (see Remark 1). The algorithm therefore terminates in polynomial time, as it removes at least one batch per iteration. We give its pseudo-code below.

Algorithm 4: Robust Estimation Procedure

input: Corruption level ϵ, Batch collection B
$B' \leftarrow B$
while *contamination rate of* B', $\sqrt{\tau_{B'}} \geq 200$ **do**
  $\forall b \in B'$ compute corruption score ε_b
  $B^o \leftarrow \{\epsilon|B|$ Batches with top corruption scores$\}$
  $\epsilon_{\text{tot}} = \sum_{b\in B^o} \varepsilon_b$
  while $\sum_{b\in B^o} \varepsilon_b \geq \epsilon_{tot}/2$ **do**
    Delete a batch from B^o, picking batch b with probability proportional to ε_b
$\widehat{q}_{B'} = \frac{1}{|B'|}\sum_{b\in B'}\sum_{l=1}^{k} Z_l^b$ and $\hat{p} = \frac{\widehat{q}-\mathbb{1}}{1-2\lambda}$
output: Estimation $\hat{p}$

We now give a high level description of our algorithm's theoretical guarantees. Recall that B_G denotes the set of non-contaminated batches and B_A the set of adversarial batches. Throughout the paper, for any collection of batches $B' \subseteq [d]$, we will use the following shorthands:

$$B'_G = B' \cap B_G \text{ and } \ B'_A = B' \cap B_A.$$

Assume that $n \geq O\left(\frac{d}{\epsilon^2 \log(e/\epsilon)}\right)$.

- In Lemma 5, we show that each deletion step has a probability at least $3/4$ of removing an adversarial batch. By a direct Chernoff bound, there is only a probability $\leq O(e^{-\epsilon|B|}) \leq O(e^{-d})$ of removing more than $2\epsilon|B_G|$ clean batches before having removed all the corrupted batches. In other words, our algorithm keeps at least $(1-2\epsilon)n$ of the good batches with high probability.

- As proved in equations 4.11 and 4.12, as soon as a subset B' contains at least $(1-2\epsilon)n$ good batches, it holds with probability $\geq 1 - O(e^{-d})$ that for all $S \subseteq [d]$

$$\begin{cases} |\widehat{q}_{B'}(S) - q(S)| \lesssim \left(1 + \sqrt{\tau_{B'}}\right) \epsilon \sqrt{\frac{d \ln(e/\epsilon)}{k}}, & \text{(i)} \\ \sqrt{\tau_{B'_G}} \leq 200. & \text{(ii)} \end{cases} \tag{4.4}$$

  There are two cases. If the algorithm has eliminated all the outliers, then it has kept at least $(1-2\epsilon)n$ clean batches with probability $1 - O(e^{-d})$. Then condition (i) $\sqrt{\tau_{B'}} = \sqrt{\tau_{B'_G}} \leq 200$ ensures that the algorithm terminates. Otherwise, the algorithm stops before removing all of the outliers, but in this case, the termination condition guarantees that $\sqrt{\tau_{B'}} \leq 200$. In both cases, condition (ii) yields that the associated estimator $\widehat{q} := \widehat{q}_{B_{\mathbf{out}}}$ has an estimation error satisfying $\sup_{S \subseteq [d]} |\widehat{q}(S) - q(S)| \lesssim \epsilon \sqrt{\frac{d \ln(e/\epsilon)}{k}}$ with probability $\geq 1 - O(e^{-d})$.

- Finally, we link the estimation error of $\widehat{q}$ to that of $\hat{p}$

$$\begin{aligned} ||\widehat{p} - p||_1 &\leq 2 \max_{S \in [d]} |\widehat{p}(S) - p(S)| \qquad \text{(see Lemma 12)} \\ &\leq 2 \max_{S \subseteq [d]} \left| \sum_{j \in S} \frac{1}{1-2\lambda} (\hat{q}_j - 1) - \frac{1}{1-2\lambda} (q_j - 1) \right| \\ &\leq \frac{1}{1-2\lambda} \max_{S \in [d]} |\hat{q}(S) - q(S)| \leq \frac{5}{\alpha} \max_{S \in [d]} |\hat{q}(S) - q(S)| \\ &\lesssim \frac{\epsilon}{\alpha} \sqrt{\frac{d \ln(e/\epsilon)}{k}} \quad \text{with probability } \geq 1 - O(e^{-d}), \end{aligned}$$

  which yields the estimation guarantee over $\widehat{p}$ and proves Theorem 4.2.

We now move to the formal definitions of the quantities involved in the algorithm and state all the technical results mentioned.

4.4.2 Technical results

Wlog, assume that $6\epsilon\sqrt{\frac{d \ln(e/\epsilon)}{k}} \leq 1$. Otherwise the upper bound of the theorem is clear. For any set $S \subseteq [d]$ and any observation Z_i^b, we define the empirical weight of S in Z_i^b as $Z_i^b(S) := \sum_{j \in S} Z_i^b(j)$. This quantity is an estimator of $q(S)$. For each batch Z^b and each collection of batches $B' \subseteq B$,

we aggregate these estimators by building

$$\widehat{q}_b(S) := \frac{1}{k}\sum_{i=1}^{k} Z_i^b(S) \quad \text{and} \quad \widehat{q}_{B'}(S) := \frac{1}{|B'|}\sum_{b\in B'} \widehat{q}_b(S).$$

Our goal is to remove batches Z^b that do not satisfy some concentration properties verified by clean batches. To this end, we introduce empirical estimators of the second order moment:

$$\widehat{\mathrm{Cov}}_{S,S'}^{B'}(b) := \left[\widehat{q}_b(S) - \widehat{q}_{B'}(S)\right]\left[\widehat{q}_b(S') - \widehat{q}_{B'}(S')\right] \tag{4.5}$$

$$\widehat{\mathrm{Cov}}_{S,S'}\left(B'\right) := \frac{1}{|B'|}\sum_{b\in B'} \widehat{\mathrm{Cov}}_{S,S'}^{B'}(b)\,. \tag{4.6}$$

In Appendix 4.A.4, we give the expression of $\mathrm{Cov}_{S,S'}(q)$ s.t.

$$\mathrm{Cov}_{S,S'}(q) = \mathbb{E}\left[\widehat{\mathrm{Cov}}_{S,S'}\left(B'\right)\right].$$

We are now ready to define the essential concentration properties satisfied by the clean batches with high probability (see Lemma 1).

Definition 4.2 (Nice properties of good batches)**.** *1. For all $S \subseteq [d]$, all sub-collections $B'_G \subseteq B_G$ of good batches of size $|B'_G| \geq (1-2\epsilon)\,|B_G|$,*

$$\left|\widehat{q}_{B'_G}(S) - q(S)\right| \leq 6\epsilon\sqrt{\frac{d\ln(e/\epsilon)}{k}}, \tag{4.7}$$

$$\left|\widehat{Cov}_{S,S'}\left(B'_G\right) - Cov_{S,S'}(\widehat{q}_{B'_G})\right| \leq \frac{250d\epsilon\ln\left(\frac{e}{\epsilon}\right)}{k}. \tag{4.8}$$

2. For all $S, S' \subseteq [d]$, for any sub collection of good batches B''_G s.t. $|B''_G| \leq \epsilon|B_G|$,

$$\sum_{b\in B''_G}\left[\widehat{q}_b(S) - q(S)\right]\left[\widehat{q}_b(S') - q(S')\right] \leq \frac{33\epsilon d|B_G|\ln(e/\epsilon)}{k}.$$

Lemma 1 (Nice properties of good batches)**.** *If $|B_G| \geq \frac{3d}{\epsilon^2\ln(e/\epsilon)}$, the nice properties of the good batches hold with probability $1-10e^{-d}$.*

The proof is very similar to that of Lemma 3 in [138], and can be found in Appendix 4.A.2 where we clarify which technical elements change.

In the case where $S' = S$, we use the shorthands $\widehat{\mathrm{Cov}}_{S,S}(B') = \widehat{\mathbf{V}}_S(B')$ and $\mathrm{Cov}_{S,S}(B') = \mathbf{V}_S(B')$. The following Lemma states that the quality of estimator $\widehat{q}_{B'}$ is controlled by the concentration of $\left|\widehat{\mathbf{V}}_S(B') - \mathbf{V}_S(\widehat{q})\right|$.

Lemma 2 (Variance gap to estimation error)**.** *If conditions 1 and 2 hold and* $\max_{S\subseteq[d]} |\widehat{q}_{B'}(S) - q(S)| \leq 11$*, then for any subset* B' *s.t.* $|B'_G| \geq (1-2\epsilon)|B_G|$ *and for any* $S \subset [d]$*, we have:*

$$|\widehat{q}_{B'}(S) - q(S)| \leq 28\epsilon\sqrt{\frac{d\ln(6e/\epsilon)}{k}} + 2\sqrt{\epsilon\left|\widehat{\mathbf{V}}_S(B') - \mathbf{V}_S(\widehat{q})\right|}.$$

This Lemma is proved in Appendix 4.A.3. Together with equation (4.8), this Lemma ensures that removing enough outliers yields an estimator $\widehat{q}_{B'}$ with estimation guarantee $\sup_{S\subseteq[d]} |\widehat{q}_{B'}(S) - q(S)| \lesssim \epsilon\sqrt{\frac{d\ln(1/\epsilon)}{k}}$.

The adversarial batch deletion is achieved by identifying the batches Z^b for which $\widehat{\mathrm{Cov}}^{B'}_{S,S'}(b)$ (defined in equation (4.5)) is at odds with Definition 4.2 for some $S, S' \subset [d]$. Searching through all possible $S, S' \subseteq [d]$ would yield an exponential-time algorithm. A way around this is to introduce a semi-definite program that can be approximated in polynomial time. To this end, we prove the next Lemma, stating that the quantities $\widehat{\mathrm{Cov}}_{S,S'}(q)$ and $\mathrm{Cov}_{S,S'}(q)$ can be computed as scalar products of matrices.

Lemma 3 (Matrix expression)**.** *Denote by* $\mathbb{1}_S$ *the indicator vector of the elements in* S*. For each vector* q*, there exists a matrix* $\mathbf{C}(\widehat{q})$ *s.t. for any* $S, S' \subseteq [d]$*,*

$$Cov_{S,S'}(\widehat{q}) = \left\langle \mathbb{1}_S\mathbb{1}_{S'}^T, \mathbf{C}(\widehat{q})\right\rangle.$$

$$\widehat{Cov}^{B'}_{S,S'}(b) = \left\langle \mathbb{1}_S\mathbb{1}_{S'}^T, \widehat{\mathbf{C}}_{b,B'}\right\rangle \quad \text{and} \quad \widehat{Cov}_{S,S'}\left(B'\right) = \left\langle \mathbb{1}_S\mathbb{1}_{S'}^T, \widehat{\mathbf{C}}(B')\right\rangle,$$

with $\widehat{\mathbf{C}}(B') = \sum_{b\in B'} \widehat{\mathbf{C}}_{b,B'}$.

The proof of the Lemma and the precise expressions of the matrices can be found in Appendix 4.A.4. To define the semi-definite program, we introduce the following space of Gram matrices:

$$\mathcal{G} := \left\{ M \in \mathbb{R}^{d\times d}, M_{ij} = \langle u^{(i)}, v^{(j)}\rangle \;\middle|\; (u^{(i)})_{i=1}^d, (v^{(i)})_{j=1}^d \text{unit vectors in } (\mathbb{R}^d, \|\cdot\|_2) \right\}.$$

For a subset B', let us define $D_{B'} = \widehat{\mathbf{C}}(B') - \mathbf{C}(\widehat{q}_{B'})$, and define $M^*_{B'}$ as any matrix s.t.

$$\langle M^*_{B'}, D_{B'}\rangle \geq \max_{M\in\mathcal{G}} \langle M, D_{B'}\rangle - c\frac{\epsilon d\ln(e/\epsilon)}{k},$$

for some small enough absolute constant $c > 0$.

Remark 1. *Note that the quantity* $\max_{M\in\mathcal{G}}\langle M, D_{B'}\rangle$ *is an SDP. For all desired precision* $\delta > 0$*, it is possible to find the solution of this program up to an additive constant* δ *in polynomial time in all the parameters of the program and in* $\log(1/\delta)$*. Thus,* $M^*_{B'}$ *can be computed in polynomial time, as well as the contamination rate and the corruption score, defined below.*

Definition of the *contamination rate* and *corruption scores*. When $\widehat{q}(S) \gg \lambda|S|$ for some $S \subseteq [d]$, the *contamination rate* and *corruption scores* have special definitions. Formally, let $A = \left\{ j \in [d] \,\middle|\, \widehat{q}_{B'}(j) \geq \lambda \right\}$ and $S^* = \max_{S \subseteq [d]} |\widehat{q}_{B'}(S) - \lambda|S||$. We have $S^* = A$ or $S^* = [d] \setminus A$, which can be computed in polynomial time. In the special case where $|\widehat{q}_{B'}(S^*) - \lambda|S^*|| \geq 11$, the *contamination rate* $\sqrt{\tau_{B'}}$ of the collection B' is defined as $\tau_{B'} = \infty$ and the corruption score of a batch is defined as $\varepsilon_b(B') = |\widehat{q}_b(S^*) - \lambda|S^*||$.

Otherwise, the *contamination rate* $\sqrt{\tau_{B'}}$ of the collection B' is defined through the quantity satisfying

$$\langle M^*_{B'}, D_{B'} \rangle = \tau_{B'} \frac{\epsilon d \ln(e/\epsilon)}{k}. \tag{4.9}$$

Define the *corruption score* of a batch as

$$\varepsilon_b(B') = \langle M^*_{B'}, \widehat{\mathbf{C}}_{b,B'} \rangle. \tag{4.10}$$

The following Lemma guarantees that the quantity $\langle M^*_{B'}, D_{B'} \rangle$ is a good approximation of $\max_{S,S' \subseteq [d]} \left| \langle \mathbb{1}_S \mathbb{1}_{S'}^T, D_{B'} \rangle \right|$, with the advantage that it can be computed in polynomial time.

Lemma 4 (Grothendieck's inequality corollary)**.** *Assume $d \geq 3$. For all symmetric matrix $A \in \mathbb{R}^{d \times d}$, it holds*

$$\max_{S,S' \subseteq [d]} \left| \langle \mathbb{1}_S \mathbb{1}_{S'}^T, A \rangle \right| \leq \max_{M \in \mathcal{G}} \langle M, A \rangle \leq 8 \max_{S,S' \subseteq [d]} \left| \langle \mathbb{1}_S \mathbb{1}_{S'}^T, A \rangle \right|.$$

The proof of the Lemma can be found in Appendix 4.A.5. Together with Lemma 2, this Lemma implies that if conditions 1 and 2 hold, then for any subset B' s.t. $|B'_G| \geq (1 - 2\epsilon)|B_G|$ and for any $S \subset [d]$, we have:

$$|\widehat{q}_{B'}(S) - q(S)| \leq (30 + 2\sqrt{\tau_{B'}})\, \epsilon \sqrt{\frac{d \ln(e/\epsilon)}{k}}. \tag{4.11}$$

This Lemma implies that if equation (4.8) holds, then, for any B' s.t. $|B'_G| \geq (1 - 2\epsilon)|B_G|$

$$\sqrt{\tau_{B'_G}} \leq 200. \tag{4.12}$$

Lemma 5 (Score good vs. adversarial batches)**.** *If $\sqrt{\tau_{B'}} \geq 200$ and condition 1-2 hold, then for any collection of batches B' s.t. $|B' \cap B_G| \geq (1 - 2\epsilon)|B_G|$, for any sub-collection of good batches $B''_G \subseteq B$, $|B''_G| \leq \epsilon n$, we have:*

$$\sum_{b \in B''_G} \varepsilon_b(B') < \frac{1}{8} \sum_{b \in B'_A} \varepsilon_b(B').$$

This Lemma is proved in Appendix 4.A.6, where we argue that this Lemma ensures that each batch deletion has a probability at least $\frac{3}{4}$ of removing an adversarial batch.

4.5 Discussion and future work

We studied the problem of estimating discrete distributions in total variation, with both privacy and robustness constraints. We obtained an information theoretic lower bound of $\epsilon\sqrt{d/\alpha^2 k} + \sqrt{d^2/\alpha^2 kn}$. We proposed an algorithm running in polynomial time and returning an estimated parameter such that the estimation error is within $\sqrt{\log(1/\epsilon)}$ of the information theoretic lower bound. It would be interesting to explore if polynomial algorithms could achieve the optimal bound without this extra factor. We do not consider the adaptation to unknown contamination ϵ and leave it for future work. It would also be interesting to explore what happens if the contamination occurs before the privacy rather than after, like in [176]. Indeed, they do not consider batched data, and it would be interesting to check if their result holds in that case. Also, the upper bound holds only if $n \geq O(d)$. Exploring the regime $n \leq d$ would be an interesting research direction, which has not been done to our knowledge, even in the case of the sole robustness constraint. Finally, we could study the combination of the robustness and privacy constraints in other settings, such as density estimation.

Appendix

4.A Proofs

4.A.1 Proof of Lemma 6, Law of the sum

Lemma 6 (Law of the sum)**.** *For any subset $S \subseteq [d]$, we have:*

$$\sum_{j\in S} Z(j) \sim \sum_{j=1}^{|S|-1} b_j + b^S,$$

with the $(b_j)_{j=1}^{|S|-1}$ independent Bernoulli variables s.t. $\mathrm{P}(b_j = 1) = \lambda$ and b^S a Bernoulli independent of the others s.t.

$$\mathrm{P}(b^S = 1) = \lambda + (1-2\lambda)p_S.$$

For any $t \in [d]$,

$$\begin{aligned}
\mathbb{P}\left(\sum_{j\in S} Z(j) = t\right) &= \binom{|S|-1}{t-1}(1-\lambda)^{|S|-t+1}\lambda^{t-1}p(S) + (1-p(S))\binom{|S|}{t}(1-\lambda)^{|S|-t}\lambda^t \\
&+ \binom{|S|-1}{t}(1-\lambda)^{|S|-t-1}\lambda^{t+1}p(S) \\
&= \binom{|S|-1}{t-1}(1-\lambda)^{|S|-t}\lambda^{t-1}\left[(1-\lambda)p(S) + \lambda\left(1-p(S)\right)\right] \\
&+ \binom{|S|-1}{t}(1-\lambda)^{|S|-t-1}\lambda^{t}\left[(1-\lambda)\left(1-p(S)\right) + \lambda p(S)\right] \\
&= \binom{|S|-1}{t-1}(1-\lambda)^{|S|-t}\lambda^{t-1}\left[(1-2\lambda)p(S) + \lambda\right] \\
&+ \binom{|S|-1}{t}(1-\lambda)^{|S|-t-1}\lambda^{t}\left[1-\lambda-(1-2\lambda)p(S)\right].
\end{aligned}$$

Note that we have:

$$q(S) = (1-2\lambda)\,p(S) + \lambda|S|. \tag{4.13}$$

4.A.2 Proof of Lemma 1, Essential properties of good batches

We start with the following intermediary Lemma.

Lemma 7. *If $\left|B_G\right| \geq \frac{2d}{\epsilon^2 \ln(e/\epsilon)}$, then $\forall S \subseteq [d]$ and $\forall B'_G \subseteq B_G$ of size $\left|B'_G\right| \geq (1-2\epsilon)\left|B_G\right|$, with probability at least $1 - 4e^{-d}$,*

$$\left|\widehat{q}_{B'_G}(S) - q(S)\right| \leq 6\epsilon\sqrt{\frac{d\ln(e/\epsilon)}{k}}.$$

Proof. : The proof of this lemma is exactly part of that of lemma 11 in [138] with different constants, we repeat it for completeness. From Hoeffding's inequality, for any $S \subseteq [d]$,

$$\mathbb{P}\left[\left|B_G\right| \left|\widehat{q}_{B_G}(S) - q(S)\right| \geq \frac{\left|B_G\right|}{\sqrt{2}}\epsilon\sqrt{\frac{d\ln(e/\epsilon)}{k}}\right] \leq 2e^{-\epsilon^2\left|B_G\right|\ln(e/\epsilon)} \leq 2e^{-2d}.$$

Similarly, for a fixed sub-collection $U_G \subseteq B_G$ of size $1 \leq \left|U_G\right| \leq 2\epsilon\left|B_G\right|$,

$$\mathbb{P}\left[\left|U_G\right| \cdot \left|\widehat{q}_{U_G}(S) - q(S)\right| \geq 2\epsilon\left|B_G\right|\sqrt{\frac{d\ln(e/\epsilon)}{k}}\right] \leq 2e^{-8\frac{\epsilon^2|B_G|^2}{|U_G|}\ln(e/\epsilon)} \leq 2e^{-4\epsilon|B_G|\ln(e/\epsilon)}. \quad (4.14)$$

We now bound the number of subsets of cardinality smaller than $2\epsilon\left|B_G\right|$:

$$\begin{aligned}\sum_{j=1}^{\lfloor 2\epsilon|B_G|\rfloor}\binom{|B_G|}{j} &\leq 2\epsilon\left|B_G\right|\binom{|B_G|}{\lfloor 2\epsilon|B_G|\rfloor} \leq 2\epsilon\left|B_G\right|\left(\frac{e\left|B_G\right|}{2\epsilon\left|B_G\right|}\right)^{2\epsilon|B_G|} \\ &\leq e^{2\epsilon|B_G|\ln(e/\epsilon)+\ln(2\epsilon|B_G|)} < e^{3\epsilon|B_G|\ln(e/\epsilon)}. \end{aligned} \quad (4.15)$$

Thus, by union bound,

$$\mathbb{P}\left[\exists\left|U_G\right| \leq 2\epsilon\left|B_G\right| : \left|U_G\right|\left|\widehat{q}_{U_G}(S) - q(S)\right| \geq 2\epsilon\left|B_G\right|\sqrt{\frac{d\ln(e/2\epsilon)}{k}}\right] \leq 2e^{-\epsilon|B_G|\ln(e/\epsilon)} \leq 2e^{-2d}.$$

For any sub-collection $B'_G \subseteq B_G$ with $|B'_G| \geq (1-2\epsilon)\left|B_G\right|$,

$$\begin{aligned}\left|\sum_{b\in B'_G}\left[\widehat{q}_b(S) - q(S)\right]\right| &= \left|\sum_{b\in B_G}\left[\widehat{q}_b(S) - q(S)\right] - \sum_{b\in B_G\setminus B'_G}\left[\widehat{q}_b(S) - q(S)\right]\right| \\ &\leq \left|\sum_{b\in B_G}\left[\widehat{q}_b(S) - q(S)\right]\right| + \left|\sum_{b\in B_G\setminus B'_G}\left[\widehat{q}_b(S) - q(S)\right]\right| \\ &\leq \left|B_G\right| \times \left|\widehat{q}_{B_G}(S) - q(S)\right| + \max_{\substack{U_G \text{ s.t.} \\ \left|U_G\right|\leq 2\epsilon\left|B_G\right|}} \left|U_G\right| \times \left|\widehat{q}_{U_G}(S) - q(S)\right| \\ &\leq \left(2 + \frac{1}{\sqrt{2}}\right)\epsilon\left|B_G\right|\sqrt{\frac{d\ln(e/\epsilon)}{k}}.\end{aligned}$$

where the last inequality holds with probability at least $1 - 4e^{-2d}$. We conclude by using a union bound over the 2^d possible subsets and by noting that $(2 + \frac{1}{\sqrt{2}})\frac{\left|B_G\right|}{|B'_G|} \leq 6$. □

We now move to the following result.

Lemma 8. *If* $\left|B_G\right| \geq \frac{3d}{\epsilon^2 \ln(e/\epsilon)}$, *then* $\forall S, S' \subseteq [d]$ *and* $\forall B'_G \subseteq B_G$ *of size* $\left|B'_G\right| \geq (1-2\epsilon)\left|B_G\right|$, *with probability at least* $1 - 2e^{-d}$,

$$\left| \frac{1}{|B'_G|} \sum_{b \in B'_G} \left[\widehat{q}_b(S) - q(S)\right]\left[\widehat{q}_b(S') - q(S')\right] - Cov_{S,S'}(q) \right| \leq \frac{140 d\epsilon \ln\left(\frac{e}{\epsilon}\right)}{k}.$$

Proof. : Let $U_b(S,S') = \left(\frac{\widehat{q}_b(S)-q(S)}{d}\right)\left(\frac{\widehat{q}_b(S')-q(S')}{d}\right) - \frac{\mathrm{Cov}_{S,S'}(q)}{d^2}$. For $b \in B_G$, $\frac{\widehat{q}_b(S)-q(S)}{d} \sim \mathrm{subG}(1/4dk)$, therefore

$$\left(\frac{\widehat{q}_b(S) - q(S)}{d}\right)\left(\frac{\widehat{q}_b(S') - q(S')}{d}\right) - \mathbb{E}\left[\left(\frac{\widehat{q}_b(S) - q(S)}{d}\right)\left(\frac{\widehat{q}_b(S') - q(S')}{d}\right)\right] = Y_b \sim \mathrm{subE}\left(\frac{16}{4kd}\right).$$

Here subE is sub exponential distribution. For any $S, S' \subseteq [d]$, Bernstein's inequality gives:

$$\mathbb{P}\left[\left|\sum_{b \in B_G} U_b(S,S')\right| \geq 6\epsilon |B_G| \frac{\ln(e/\epsilon)}{kd}\right] \leq 2e^{-\epsilon^2 \left|B_G\right| \ln^2(e/\epsilon)} \leq 2e^{-3d}.$$

Next, for a fixed sub-collection $B''_G \subseteq B_G$ of size $1 \leq \left|B''_G\right| \leq \epsilon\left|B_G\right|$,

$$\begin{aligned}\Pr\left[\left|\sum_{b \in B''_G} U_b(S,S')\right| \geq 64\epsilon\left|B_G\right|\frac{\ln(e/\epsilon)}{n}\right] &\leq 2e^{-\frac{64\epsilon|B_G|\ln(e/\epsilon)}{2\times 2\times 4/n}} \\ &\leq 2e^{-4\epsilon|B_G|\ln(e/\epsilon)}.\end{aligned}$$

The same steps as the previous lemma terminate the proof, except that there are now 2^{2d} sets $S, S' \subseteq [d]$.

□

By Lemma 7 and 8, if $\left|B_G\right| \geq \frac{2d}{\epsilon^2 \ln(e/\epsilon)}$, then $\forall S \subseteq [d]$ and $\forall B'_G \subseteq B_G$ of size $\left|B'_G\right| \geq (1-2\epsilon)\left|B_G\right|$, with probability at least $1 - 8e^{-d}$:

$$\left|\widehat{q}_{B'_G}(S) - q(S)\right| \leq 6\epsilon\sqrt{\frac{d\ln(e/\epsilon)}{k}}$$

and

$$\left| \frac{1}{|B'_G|} \sum_{b \in B'_G} \left[\widehat{q}_b(S) - q(S)\right]\left[\widehat{q}_b(S') - q(S')\right] - \mathrm{Cov}_{S,S'}(q) \right| \leq \frac{140 d\epsilon \ln\left(\frac{6e}{\epsilon}\right)}{k}.$$

Additionally Lemma 11, this implies:

$$\left|\mathrm{Cov}_{S,S'}(q) - \mathrm{Cov}_{S,S'}(\widehat{q}_{B'_G})\right| \leq 66\epsilon\sqrt{\frac{d\ln(e/\epsilon)}{k}}.$$

Moreover:

$$\begin{aligned}
\frac{1}{|B'_G|}\sum_{b\in B'_G}\Big[\widehat{q}_b(S)-q(S)\Big]\Big[\widehat{q}_b(S')-q(S')\Big] = {} & \frac{1}{|B'_G|}\sum_{b\in B'_G}\Big[\widehat{q}_b(S)-\widehat{q}_{B'}(S)\Big]\Big[\widehat{q}_b(S')-\widehat{q}_{B'}(S')\Big] \\
& + \Big[q(S)-\widehat{q}_{B'}(S)\Big]\Big[q(S')-\widehat{q}_{B'}(S')\Big] \\
& + \frac{1}{|B'_G|}\sum_{b\in B'_G}\Big[\widehat{q}_b(S)-\widehat{q}_{B'}(S)\Big]\Big[q(S')-\widehat{q}_{B'}(S')\Big] \\
& + \frac{1}{|B'_G|}\sum_{b\in B'_G}\Big[q(S)-\widehat{q}_{B'}(S)\Big]\Big[\widehat{q}_b(S')-\widehat{q}_{B'}(S')\Big] \\
= {} & \frac{1}{|B'_G|}\sum_{b\in B'_G}\Big[\widehat{q}_b(S)-\widehat{q}_{B'}(S)\Big]\Big[\widehat{q}_b(S')-\widehat{q}_{B'}(S')\Big] \\
& + \Big[q(S)-\widehat{q}_{B'}(S)\Big]\Big[q(S')-\widehat{q}_{B'}(S')\Big].
\end{aligned}$$

Therefore:

$$\left|\widehat{\mathrm{Cov}}_{S,S'}\left(B'_G\right) - \mathrm{Cov}_{S,S'}(\widehat{q}_{B'_G})\right| \leq \frac{242d\epsilon\ln\left(\frac{e}{\epsilon}\right)}{k}.$$

Note that we also have:

$$\left|\widehat{\mathrm{Cov}}_{S,S'}\left(B'_G\right) - \mathrm{Cov}_{S,S'}(q)\right| \leq \frac{176d\epsilon\ln\left(\frac{e}{\epsilon}\right)}{k}. \tag{4.16}$$

The following Lemma gives condition 2.

Lemma 9. *If* $\left|B_G\right| \geq \frac{3d}{\epsilon^2\ln(e/\epsilon)}$, *then* $\forall S, S' \subseteq [d]$ *and* $\forall B''_G \subseteq B_G$ *of size* $\left|B''_G\right| \leq \epsilon\left|B_G\right|$, *with probability at least* $1-2e^{-d}$,

$$\left|\sum_{b\in B''_G}\Big[\widehat{q}_b(S)-q(S)\Big]\Big[\widehat{q}_b(S')-q(S')\Big]\right| \leq \frac{33\epsilon d\left|B_G\right|\ln(e/\epsilon)}{k}.$$

Proof. : For any $S, S' \subseteq [d]$ and any $B'_G \subseteq B_G$ Bernstein's inequality gives:

$$\mathbb{P}\left[\left|\sum_{b\in B''_G} U_b(S,S')\right| \geq 32\epsilon\left|B_G\right|\frac{\ln(e/\epsilon)}{kd}\right] \leq 2e^{-4\epsilon\left|B_G\right|\ln(e/\epsilon)}.$$

We have :

$$\left|\sum_{b\in B_G''}\Big[\widehat{q}_b(S)-q(S)\Big]\Big[\widehat{q}_b(S')-q(S')\Big]\right| = \left|\sum_{b\in B_G''} d^2U_b(S,S') + |B_G''|\mathrm{Cov}_{S,S'}(q)\right|$$
$$\leq \left|\sum_{b\in B_G''} d^2U_b(S,S')\right| + \epsilon\frac{d\Big|B_G\Big|}{k}.$$

A union bound over all the possible B_G'' and the 2^{2d} sets S, S' terminates the proof.

□

Combining the three Lemmas of the section gives Lemma 1.

4.A.3 Proof of Lemma 2, Variance gap to estimation error

Proof: By condition 1 and Cauchy-Schwartz:

$$\left|\widehat{q}_{B'}(S)-q(S)\right| \leq \frac{1}{|B'|}\left|\sum_{b\in B'_G}\widehat{q}_b(S)-q(S)\right| + \frac{1}{|B'|}\left|\sum_{b\in B'_A}\widehat{q}_b(S)-q(S)\right|$$
$$\leq 6\epsilon\sqrt{\frac{d\ln(e/\epsilon)}{k}} + \sqrt{\frac{|B'_A|}{|B'|}}\sqrt{\frac{1}{|B'|}\sum_{b\in B'_A}\Big[\widehat{q}_b(S)-q(S)\Big]^2}. \tag{4.17}$$

We can decompose the second term:

$$\frac{1}{|B'|}\sum_{b\in B'_A}\Big[\widehat{q}_b(S)-q(S)\Big]^2 = \frac{1}{|B'|}\sum_{b\in B'}\Big[\widehat{q}_b(S)-q(S)\Big]^2 - \frac{1}{|B'|}\sum_{b\in B'_G}\Big[\widehat{q}_b(S)-q(S)\Big]^2.$$

By Lemma 8,

$$\left|\frac{1}{|B'_G|}\sum_{b\in B'_G}\Big[\widehat{q}_b(S)-q(S)\Big]^2 - \mathbf{V}_S(q)\right| \leq 140\frac{\epsilon d\ln(e/\epsilon)}{k}.$$

Thus,

$$\frac{1}{|B'|}\sum_{b\in B'_G}\Big[\widehat{q}_b(S)-q(S)\Big]^2 = \frac{|B'_G|}{|B'|}\frac{1}{|B'_G|}\sum_{b\in B'_G}\Big[\widehat{q}_b(S)-q(S)\Big]^2$$
$$\geq (1-2\epsilon)\left(\mathbf{V}_S(q) - 140\frac{\epsilon d\ln(e/\epsilon)}{k}\right)$$
$$\geq \mathbf{V}_S(q) - 2\epsilon\mathbf{V}_S(q) - 140\frac{\epsilon d\ln(e/\epsilon)}{k}$$

$$\geq \mathbf{V}_S(\widehat{q}_{B'}) - 15\frac{\left|\widehat{q}_{B'}(S) - q(S)\right|}{k} - 142\frac{\epsilon d \ln(e/\epsilon)}{k},$$

where the last inequality comes from Lemma 11 and $\mathbf{V}_S(q) \leq d/k$. Now, we have

$$\begin{aligned}\left|\widehat{q}_{B'}(S) - \widehat{q}_{B'_G}(S)\right| &\leq \left|\left(\frac{1}{|B'_G|} - \frac{1}{|B'|}\right)\sum_{b\in B'_G}\widehat{q}_b(S)\right| + \left|\frac{1}{|B'|}\sum_{b\in B'\setminus B'_G}\widehat{q}_b(S)\right| \\ &\leq \frac{2d\epsilon}{1-\epsilon} \leq 3d\epsilon.\end{aligned}$$

Thus,

$$\begin{aligned}\frac{\left|\widehat{q}_{B'}(S) - q(S)\right|}{k} &\leq \frac{\left|\widehat{q}_{B'_G}(S) - q(S)\right|}{k} + \frac{\left|\widehat{q}_{B'}(S) - \widehat{q}_{B'_G}(S)\right|}{k} \\ &\leq \frac{6\epsilon}{k}\sqrt{\frac{d\ln(e/\epsilon)}{k}} + \frac{3d\epsilon}{k} \leq 3\frac{d\epsilon}{k}\ln(e/\epsilon).\end{aligned}$$

This implies

$$\frac{1}{|B'|}\sum_{b\in B'_G}\left[\widehat{q}_b(S) - q(S)\right]^2 \geq \mathbf{V}_S(\widehat{q}_{B'}) - 187\frac{\epsilon d\ln(e/\epsilon)}{k}. \tag{4.18}$$

On the other hand,

$$\begin{aligned}\frac{1}{|B'|}\sum_{b\in B'}\left[\widehat{q}_b(S) - q(S)\right]^2 &= \frac{1}{|B'|}\sum_{b\in B'}\left[\widehat{q}_b(S) - \widehat{q}_{B'}(S)\right]^2 \\ &\quad + \left[q(S) - \widehat{q}_{B'}(S)\right]^2 + 2\left[q(S) - \widehat{q}_{B'}(S)\right]\frac{1}{|B'|}\sum_{b\in B'}\left[\widehat{q}_b(S) - \widehat{q}_{B'}(S)\right] \\ &= \frac{1}{|B'|}\sum_{b\in B'}\left[\widehat{q}_b(S) - \widehat{q}_{B'}(S)\right]^2 + \left[q(S) - \widehat{q}_{B'}(S)\right]^2.\end{aligned}$$

Combining this equation with equations 4.18 and 4.17 gives

$$\begin{aligned}\left|\widehat{q}_{B'}(S) - q(S)\right| &\leq \frac{1}{|B'|}\left|\sum_{b\in B'_G}\widehat{q}_b(S) - q(S)\right| + \frac{1}{|B'|}\left|\sum_{b\in B'_A}\widehat{q}_b(S) - q(S)\right| \\ &\leq 6\epsilon\sqrt{\frac{d\ln(e/\epsilon)}{k}} + \sqrt{2\epsilon}\sqrt{\widehat{\mathbf{V}}_S(B') - \mathbf{V}_S(\widehat{q}_{B'}) + 187\frac{\epsilon d\ln(e/\epsilon)}{k} + \left[q(S) - \widehat{q}_{B'}(S)\right]^2} \\ &\leq 26\epsilon\sqrt{\frac{d\ln(e/\epsilon)}{k}} + \sqrt{2\epsilon\left|\widehat{\mathbf{V}}_S(B') - \mathbf{V}_S(\widehat{q}_{B'})\right|} + \sqrt{2\epsilon}\left|\widehat{q}_{B'}(S) - q(S)\right|.\end{aligned}$$

Noting that $2\epsilon \leq 1/8$ terminates the proof:

$$\left|\widehat{q}_{B'}(S) - q(S)\right| \leq 30\epsilon\sqrt{\frac{d\ln(e/\epsilon)}{k}} + 2\sqrt{\epsilon\left|\widehat{\mathbf{V}}_S(B') - \mathbf{V}_S(\widehat{q}_{B'})\right|}.$$

□

4.A.4 Proof of Lemma 3, Matrix expression

For each batch $b \in B$, define matrix $C^{EV}_{b,B'}$ as:

$$\widehat{\mathbf{C}}_{b,B'}(j,l) = \left[\widehat{q}_b(j) - \widehat{q}_{B'}(j)\right]\left[\widehat{q}_b(l) - \widehat{q}_{B'}(l)\right],\ \forall (j,l) \in [d]^2. \tag{4.19}$$

For each collection of batches B' define

$$\widehat{\mathbf{C}}(B') = \frac{1}{|B'|}\sum_{b\in B'}\widehat{\mathbf{C}}_b.$$

For a set $S \subseteq [d]$, define $\mathbb{1}_S$ as the indicator vector of the elements in S. For any $S, S' \subseteq [d]$

$$\begin{aligned}
\left\langle \widehat{\mathbf{C}}(B'), \mathbb{1}_S\mathbb{1}_{S'}^T \right\rangle &= \frac{1}{|B'|}\sum_{b\in B'}\sum_{j\in S}\sum_{l\in S'}\left[\widehat{q}_b(j) - \widehat{q}_{B'}(j)\right]\left[\widehat{q}_b(l) - \widehat{q}_{B'}(l)\right] \\
&= \frac{1}{|B'|}\sum_{b\in B'}\left(\sum_{j\in S}\widehat{q}_b(j) - \sum_{j\in S}\widehat{q}_{B'}(j)\right)\left(\sum_{l\in S'}\widehat{q}_b(l) - \sum_{l\in S'}\widehat{q}_{B'}(l)\right) \\
&= \widehat{\mathrm{Cov}}_{S,S'}\left(B'\right).
\end{aligned}$$

We can compute

$$\begin{aligned}
\mathbb{E}\left[\sum_{j\in S} Z(j)\middle| X\right] &= \lambda|S|\mathbb{1}_{X\notin S} + \left(\lambda(|S|-1) + 1 - \lambda\right)\mathbb{1}_{X\in S} \\
&= \lambda|S| + (1-2\lambda)\,\mathbb{1}_{X\in S}.
\end{aligned}$$

For a set S, let us define $Y_S = \left(\sum_{j\in S} Z(j)\right) - q(S)$ and $\Delta_S = \lambda|S| - q(S)$. For any sets $S, S' \subseteq [d]$ s.t. $S \cap S' = \emptyset$, we have:

$$\begin{aligned}
\mathbb{E}\left[Y_S Y_{S'}\right] &= \mathbb{E}\left[\mathbb{E}\left[Y_S\middle|X\right]\mathbb{E}\left[Y_{S'}\middle|X\right]\right] \\
&= \mathbb{E}\left[\left(\Delta_S + (1-2\lambda)\,\mathbb{1}_{X\in S}\right)\left(\Delta_{S'} + (1-2\lambda)\,\mathbb{1}_{X\in S'}\right)\right] \\
&= \Delta_{S'}\Delta_S + \Delta_S(1-2\lambda)p(S') + \Delta_{S'}(1-2\lambda)p(S) \quad \text{since } S\cap S' = \emptyset \\
&= -\Delta_{S'}\Delta_S \qquad \text{since by (4.13) we have } (1-2\lambda)p(S) = -\Delta_S.
\end{aligned}$$

On the other hand, using the notation from Lemma 6, we have:

$$\begin{aligned}\mathbb{E}\left[Y_S^2\right] &= \mathbb{E}\left[\left(\sum_{j=1}^{|S|-1}(b_j - \mathbb{E}b_j) + b^S - \mathbb{E}b^S\right)^2\right] = \sum_{j=1}^{|S|-1}\mathbb{V}[b_j] + \mathbb{V}[b^S] \\ &= (|S|-1)\lambda(1-\lambda) + (\lambda + (1-2\lambda)p(S))\,(1-\lambda-(1-2\lambda)p(S)) \\ &= (|S|-1)\lambda(1-\lambda) + (\lambda - \Delta_S)\,(1-\lambda+\Delta_S) \\ &= -\Delta_S^2 + |S|\lambda(1-\lambda) - (1-2\lambda)\Delta_S.\end{aligned}$$

For any $S, S' \subseteq [d]$, we thus have:

$$\begin{aligned}\mathbb{E}\left[Y_S Y_{S'}\right] &= \mathbb{E}\left[\left(Y_{(S\cap S')} + Y_{(S\setminus S')}\right)\left(Y_{(S\cap S')} + Y_{(S'\setminus S)}\right)\right] \\ &= \mathbb{E}\left[Y_{(S\cap S')}^2\right] + \mathbb{E}\left[Y_{(S\cap S')}Y_{(S\setminus S')}\right] + \mathbb{E}\left[Y_{(S\cap S')}Y_{(S'\setminus S)}\right] + \mathbb{E}\left[Y_{(S\setminus S')}Y_{(S'\setminus S)}\right] \\ &= -\left(\Delta_{(S\cap S')} + \Delta_{(S\setminus S')}\right)\left(\Delta_{(S\cap S')} + \Delta_{(S'\setminus S)}\right) + |S\cap S'|\lambda(1-\lambda) - (1-2\lambda)\Delta_{(S\cap S')} \\ &= -\Delta_S\Delta_{S'} + |S\cap S'|\lambda(1-\lambda) - (1-2\lambda)\Delta_{(S\cap S')}.\end{aligned}$$

For a vector q, define

$$k\mathbf{C}(q) = -\left(\lambda\mathbb{1} - q\right)\left(\lambda\mathbb{1} - q\right)^T + \lambda(1-\lambda)I_d - (1-2\lambda)\mathrm{Diag}(\lambda\mathbb{1} - q). \tag{4.20}$$

For any two sets $S, S' \subseteq [d]$, we have: $\mathbb{E}\left[Y_S Y_{S'}\right] = \mathbb{1}_S^T k\mathbf{C}(q)\mathbb{1}_{S'}$, so that $\mathbb{E}\left[\widehat{\mathrm{Cov}}_{S,S'}(B'_G)\right] = \mathbb{1}_S^T\mathbf{C}(q)\mathbb{1}_{S'}$. We now define:

$$\mathrm{Cov}_{S,S'}(B') := \mathbb{1}_S^T\mathbf{C}(\widehat{q}_{B'})\mathbb{1}_{S'}. \tag{4.21}$$

4.A.5 Proof of Lemma 4, Grothendieck's inequality corollary

Proof of Lemma 4.

- For the first inequality, fix any $x, y \in \{0,1\}^d$ and three orthonormal vectors $e_0, e_1, e_2 \in \mathbb{R}^d$. Define the following vectors:

$$\forall j \in \{1,\dots,d\} : u^{(j)} = \begin{cases} e_0 & \text{if } x_j = 1, \\ e_1 & \text{otherwise,}\end{cases} \quad \text{and} \quad v^{(j)} = \begin{cases} e_0 & \text{if } y_j = 1, \\ e_2 & \text{otherwise.}\end{cases}$$

  Then the matrix $M = \left[\langle u^{(i)}, v^{(j)}\rangle\right]_{ij}$ belongs to $\mathcal{G}$ and we have by construction $M = xy^T$ which proves the first inequality.

- For the second inequality, we have by Grothendieck's inequality

$$\max_{M\in\mathcal{G}}\langle M, A\rangle \le 2\max_{x,y\in\{\pm1\}^d}\langle xy^T, A\rangle.$$

  For all $a \in \mathbb{R}$, define $a^+ = a \vee 0$ and $a^- = (-a)\vee 0$ and for all vector $x \in \mathbb{R}^d$, define $x^+ = (x_j^+)_j$ and $x^- = (x_j^-)_j$. Note that if $x \in \{\pm1\}^d$, then $x^+, x^- \in \{0,1\}^d$. We therefore

have:

$$\max_{x,y\in\{\pm1\}^d}\langle xy^T,A\rangle = \max_{x,y\in\{\pm1\}^d}\Big|\langle x^+y^{+T},A\rangle-\langle x^-y^{+T},A\rangle-\langle x^+y^{-T},A\rangle+\langle x^-y^{-T},A\rangle\Big|$$
$$\leq 4\max_{a,b\in\{0,1\}^d}\Big|\langle ab^T,A\rangle\Big|,$$

which proves the second inequality.

□

4.A.6 Proof of Lemma 5, Score good vs. adversarial batches

We first note that the Lemma implies the desired property for the batches in B^o, namely that each batch deletion has a probability at least $\frac{3}{4}$ of removing an adversarial batch. Indeed, we have:

$$\sum_{b\in B^o}\varepsilon_b = \sum_{b\in B^o_G}\varepsilon_b + \sum_{b\in B^o_A}\varepsilon_b.$$

If we had $\sum_{b\in B^o_A}\varepsilon_b < 7\sum_{b\in B^o_G}\varepsilon_b$, this would imply:

$$\sum_{b\in B^o}\varepsilon_b < 8\sum_{b\in B^o_G}\varepsilon_b < \sum_{b\in B'_A}\varepsilon_b,$$

where the last inequality comes from the Lemma. However this is in contradiction with the definition of B^o, which is the sub-collection of $\epsilon|B|$ batches with top ε_b scores, since $|B'_A|\leq\epsilon|B|$. We therefore have that $\sum_{b\in B^o_A}\varepsilon_b \geq 7\sum_{b\in B^o_G}\varepsilon_b$ hence $\sum_{b\in B^o_G}\varepsilon_b \leq \frac{1}{8}\sum_{b\in B^o}\varepsilon_b$. Denote by $B^o(t)$ the current set obtained from B^o after having removed t batches (and before $\sum_{b\in B^o}\varepsilon_b$ has been halved). We keep deleting batches from B^o until $\sum_{b\in B^o(t)}\varepsilon_b \leq \frac{1}{2}\sum_{b\in B^o}\varepsilon_b$. At each step, we therefore have that $\sum_{b\in B^o_G(t)}\varepsilon_b \leq \frac{1}{4}\sum_{b\in B^o(t)}\varepsilon_b$ hence the probability of removing a good batch from $B^o(t)$ is always less than $\frac{3}{4}$.

Subcase 1 We first prove the Lemma in the case where $\max_{S\subseteq[d]}\left|\widehat{q}_{B'}(S)-\lambda|S|\right|\geq 11$.
We have:

$$|\widehat{q}_{B'}(S)-\lambda|S|| \leq \frac{1}{|B'|}\left|\sum_{b\in B'_G}\widehat{q}_b(S)-\lambda|S|\right| + \frac{1}{|B'|}\left|\sum_{b\in B'_A}\widehat{q}_b(S)-\lambda|S|\right|$$
$$\leq \frac{|B'_G|}{|B'|}\frac{1}{|B'_G|}\left|\sum_{b\in B'_G}\widehat{q}_b(S)-q(S)\right| + \frac{|B'_G|}{|B_G|}\left|\lambda|S|-q(S)\right| + \frac{1}{|B'|}\left|\sum_{b\in B'_A}\widehat{q}_b(S)-\lambda|S|\right|$$
$$\leq 6\epsilon\sqrt{\frac{d\ln(e/\epsilon)}{k}}+1+\frac{1}{|B'|}\left|\sum_{b\in B'_A}\widehat{q}_b(S)-\lambda|S|\right| \quad \text{by equation (4.7).}$$

Let $S^* = \arg\max_{S\subseteq[d]} |\widehat{q}_{B'}(S) - \lambda|S||$. We have:

$$\left|\sum_{b\in B'_A} \widehat{q}_b(S^*) - \lambda|S^*|\right| \geq 9(1-2\epsilon)\Big|B_G\Big|.$$

On the other hand, by equation 4.14, we have for any B''_G s.t. $|B''_G| \leq \epsilon\Big|B_G\Big|$:

$$\begin{aligned}\left|\sum_{b\in B''_G} \widehat{q}_b(S^*) - \lambda|S^*|\right| &\leq \left|\sum_{b\in B''_G} \widehat{q}_b(S^*) - q(S)\right| + |B''_G| \\ &\leq \left(\epsilon + 2\epsilon\sqrt{\frac{d\ln(e/\epsilon)}{k}}\right)\Big|B_G\Big| \\ &\leq \Big(1+\epsilon\Big)\Big|B_G\Big|.\end{aligned}$$

Thus we have:

$$\frac{\left|\sum_{b\in B'_A} \widehat{q}_b(S^*) - \lambda|S^*|\right|}{\left|\sum_{b\in B''_G} \widehat{q}_b(S^*) - \lambda|S^*|\right|} \geq \frac{9(1-2\epsilon)}{1+\epsilon} > 8.$$

Subcase 2 In the case where $\max_{S\subseteq[d]} \left|\widehat{q}_{B'}(S) - \lambda|S|\right| \leq 11$, the proof relies on the following intermediary Lemma 10.

Lemma 10. *If conditions 1 and 2 hold, then, for any $B' \subset [B]$, for any two sets S, S':*

$$\Big(\tau_{B'} - 11\sqrt{\tau_{B'}} - 1313\Big)\frac{\epsilon d\ln(e/\epsilon)}{k} \leq \frac{1}{|B'|}\sum_{b\in B'_A} \langle M^*, \widehat{\mathbf{C}}_{b,B'}\rangle.$$

Proof. : In this proof only, we use the shorthand:

$$\gamma := \frac{\epsilon d\ln(e/\epsilon)}{k}.$$

We have

$$\Big\langle M^*, D_{B'}\Big\rangle = \Big\langle M^*, \widehat{\mathbf{C}}(B') - \mathbf{C}(q)\Big\rangle + \Big\langle M^*, \mathbf{C}(q) - \mathbf{C}(\widehat{q}_{B'})\Big\rangle.$$

We analyse separately each term. For any S', S, according to Lemmas 11 and equation 4.11, we have:

$$\begin{aligned}\Big|\mathrm{Cov}_{S,S'}(\widehat{q}_{B'}) - \mathrm{Cov}_{S,S'}(q)\Big| &\leq \frac{11}{k}\max_{S''}\Big|\widehat{q}_{B'}(S'') - q(S'')\Big| \\ &\leq \frac{330 + 22\sqrt{\tau_{B'}}}{k}\epsilon\sqrt{\frac{d\ln(e/\epsilon)}{k}}\end{aligned}$$

$$\leq (330 + 22\sqrt{\tau_{B'}})\gamma\sqrt{\frac{1}{d\ln(e/\epsilon)k}}$$
$$\leq (96 + 7\sqrt{\tau_{B'}})\gamma.$$

Where the last line come from $d \geq 3$, $\epsilon \leq \frac{1}{20}$. Thus, by Lemma 4, we have:

$$\begin{aligned}\arg\max_{M\in\mathcal{G}}\langle M, \mathbf{C}(q) - \mathbf{C}(\widehat{q}_{B'})\rangle &\leq 8\max_{S,S'}\left|\mathrm{Cov}_{S,S'}(\widehat{q}_{B'}) - \mathrm{Cov}_{S,S'}(q)\right| \\ &\leq \frac{88}{k}\left(30 + 2\sqrt{\tau'_B}\right)\epsilon\sqrt{\frac{d\ln(e/\epsilon)}{k}} \\ &\leq (763 + 51\sqrt{\tau_{B'}})\gamma. \end{aligned} \tag{4.22}$$

On the other hand,

$$\begin{aligned}\widehat{\mathbf{C}}(B') - \mathbf{C}(q) &= \frac{1}{|B'|}\sum_{b\in B'}\widehat{\mathbf{C}}(b, B') - \mathbf{C}(q) \\ &= \frac{1}{|B'|}\sum_{b\in B'_G}\widehat{\mathbf{C}}(b, B') - \mathbf{C}(q) + \frac{1}{|B'|}\sum_{b\in B'_A}\widehat{\mathbf{C}}(b, B') - \mathbf{C}(q).\end{aligned}$$

From Lemma 4 and 8, we have:

$$\begin{aligned}\left|\left\langle M^*, \frac{1}{|B'_G|}\sum_{b\in B'_G}\widehat{\mathbf{C}}(b, B') - \mathbf{C}(q)\right\rangle\right| &\leq \left|\left\langle M^*, \frac{1}{|B'_G|}\sum_{b\in B'_G}\widehat{\mathbf{C}}(b, B'_G) - \widehat{\mathbf{C}}(b, B')\right\rangle\right| \\ &\quad + \left|\left\langle M^*, \frac{1}{|B'_G|}\sum_{b\in B'_G}\widehat{\mathbf{C}}(b, B'_G) - \mathbf{C}(q)\right\rangle\right|.\end{aligned}$$

We start by bounding the first term $A = \left|\left\langle M^*, \frac{1}{|B'_G|}\sum_{b\in B'_G}\widehat{\mathbf{C}}(b, B'_G) - \widehat{\mathbf{C}}(b, B')\right\rangle\right|$. By Lemma 4:

$$\begin{aligned}A &\leq \frac{8}{|B'_G|}\max_{S,S'\in[d]}\left|\sum_{b\in B'_G}\left[\widehat{q}_b(S) - \widehat{q}_{B'}(S)\right]\left[\widehat{q}_b(S') - \widehat{q}_{B'}(S')\right] - \left[\widehat{q}_b(S) - \widehat{q}_{B'_G}(S)\right]\left[\widehat{q}_b(S') - \widehat{q}_{B'_G}(S')\right]\right| \\ &= \frac{8}{|B'_G|}\max_{S,S'\in[d]}\left|\left[\widehat{q}_{B'_G}(S) - \widehat{q}_{B'}(S)\right]\left[\widehat{q}_{B'_G}(S') - \widehat{q}_{B'}(S')\right]\right|.\end{aligned}$$

By equation 4.4 and condition 1, for any $S \subseteq [d]$, we have:

$$\begin{aligned}\left|\widehat{q}_{B'_G}(S) - \widehat{q}_{B'}(S)\right| &\leq \left|\widehat{q}_{B'_G}(S) - q(S)\right| + \left|q(S) - \widehat{q}_{B'}(S)\right| \\ &\leq (36 + 2\sqrt{\tau_{B'}})\,\epsilon\sqrt{\frac{d\ln(e/\epsilon)}{k}}.\end{aligned}$$

Thus,

$$A \leq 8\left(36 + 2\sqrt{\tau_{B'}}\right)^2 \epsilon\gamma.$$

By equation 4.16 and Lemma 4, we have:

$$\left|\left\langle M^*, \frac{1}{|B'_G|}\sum_{b\in B'_G} \widehat{\mathbf{C}}(b, B'_G) - \mathbf{C}(q)\right\rangle\right| \leq 1408\gamma.$$

Thus:

$$\begin{aligned}\left|\left\langle M^*, \frac{1}{|B'|}\sum_{b\in B'_G} \widehat{\mathbf{C}}(b, B') - \mathbf{C}(q)\right\rangle\right| &= \frac{|B'_G|}{|B'|}\left|\left\langle M^*, \frac{1}{|B'_G|}\sum_{b\in B'_G} \widehat{\mathbf{C}}(b, B') - \mathbf{C}(q)\right\rangle\right| \\ &\leq 1408\gamma + 8\left(36 + 2\sqrt{\tau_{B'}}\right)^2 \epsilon\gamma.\end{aligned}$$

Finally, for any q, we have:

$$\langle M^*, C(q)\rangle \leq 8\max_{S,S'} \mathrm{Cov}_{S,S'}(q) \leq \frac{8d}{k}.$$

This gives:

$$\begin{aligned}\left|\left\langle M^*, \widehat{\mathbf{C}}(B') - \mathbf{C}(q)\right\rangle\right| &\leq \left|\left\langle M^*, \frac{1}{|B'|}\sum_{b\in B'_G} \widehat{\mathbf{C}}(b, B') - \mathbf{C}(q)\right\rangle\right| + \left|\left\langle M^*, \frac{1}{|B'|}\sum_{b\in B'_A} \widehat{\mathbf{C}}(b, B')\right\rangle\right| \\ &\quad + \frac{|B'_A|}{|B'|}\left|\left\langle M^*, \mathbf{C}(q)\right\rangle\right| \\ &\leq 1408\gamma + 8\left(36 + 2\sqrt{\tau_{B'}}\right)^2 \epsilon\gamma + \frac{\epsilon}{1-2\epsilon}\frac{8d}{k} + \left|\left\langle M^*, \frac{1}{|B'|}\sum_{b\in B'_A} \widehat{\mathbf{C}}(b, B')\right\rangle\right| \\ &\leq 1409\gamma + 8\left(36 + 2\sqrt{\tau_{B'}}\right)^2 \epsilon\gamma + \left|\left\langle M^*, \frac{1}{|B'|}\sum_{b\in B'_A} \widehat{\mathbf{C}}(b, B')\right\rangle\right|.\end{aligned}$$

We can now combine this with equations 4.22:

$$\begin{aligned}\tau_{B'}\gamma &= \langle M^*, D_{B'}\rangle \\ &\leq \left|\left\langle M^*, \widehat{\mathbf{C}}(B') - \mathbf{C}(q)\right\rangle\right| + \left|\left\langle M^*, \mathbf{C}(q) - \mathbf{C}(\widehat{q}_{B'})\right\rangle\right| \\ &\leq 2200\gamma + 8\left(36 + 2\sqrt{\tau_{B'}}\right)^2 \epsilon\gamma + 51\sqrt{\tau_{B'}}\gamma + \frac{1}{|B'|}\left|\left\langle M^*, \sum_{b\in B'_A} \widehat{\mathbf{C}}(b, B')\right\rangle\right|.\end{aligned}$$

Thus:

$$\left|\langle M^*, \sum_{b\in B'_A} \widehat{\mathbf{C}}(b, B')\rangle\right| \geq (1-2\epsilon)\left[(1-32\epsilon)\tau_{B'} - (\epsilon 1152+51)\sqrt{\tau_{B'}} - 2200 - 8*36^2\epsilon\right]\Big|B_G\Big|\gamma$$

With $\epsilon \leq 1/100$, we get:

$$\left|\langle M^*, \sum_{b\in B'_A} \widehat{\mathbf{C}}(b, B')\rangle\right| \geq (0.66\tau_{B'} - 62\sqrt{\tau_{B'}} - 2260)\Big|B_G\Big|\gamma$$

□

On the other hand, for any collection of good batches $B''_G \subseteq B'$ s.t. $|B''_G| \leq \epsilon\Big|B_G\Big|$, we have by Lemma 4:

$$\begin{aligned}\sum_{b\in B''_G} \langle M^*, \widehat{\mathbf{C}}_{b,B'}\rangle &\leq 8 \max_{S,S'\in[d]} \sum_{b\in B''_G} \langle \mathbb{1}_S \mathbb{1}_{S'}^T, \widehat{\mathbf{C}}_{b,B'}\rangle \\ &= 8 \max_{S,S'\in[d]} \sum_{b\in B''_G} \Big[\widehat{q}_b(S) - \widehat{q}_{B'}(S)\Big]\Big[\widehat{q}_b(S') - \widehat{q}_{B'}(S')\Big].\end{aligned}$$

We can decompose the terms in the sum:

$$\begin{aligned}\Big[\widehat{q}_b(S) - \widehat{q}_{B'}(S)\Big]\Big[\widehat{q}_b(S') - \widehat{q}_{B'}(S')\Big] &= \Big[\widehat{q}_b(S) - q(S)\Big]\Big[\widehat{q}_b(S') - q(S')\Big] + \Big[q(S) - \widehat{q}_{B'}(S)\Big]\Big[q(S') - \widehat{q}_{B'}(S')\Big] \\ &+ \Big[\widehat{q}_b(S) - q(S)\Big]\Big[q(S') - \widehat{q}_{B'}(S')\Big] + \Big[q(S) - \widehat{q}_{B'}(S)\Big]\Big[\widehat{q}_b(S') - q(S')\Big].\end{aligned}$$

By condition 1:

$$\max_{S,S'\in[d]} \sum_{b\in B''_G} \Big[\widehat{q}_b(S) - q(S)\Big]\Big[\widehat{q}_b(S') - q(S')\Big] \leq 33\Big|B_G\Big|\gamma.$$

By equation 4.11,

$$\max_{S,S'\in[d]} \sum_{b\in B''_G} \Big[q(S) - \widehat{q}_{B'}(S)\Big]\Big[q(S') - \widehat{q}_{B'}(S')\Big] \leq |B''_G|(33 + 2\sqrt{\tau_{B'}})^2\epsilon\gamma.$$

By equations 4.14 and 4.11,

$$\begin{aligned}\max_{S,S'\in[d]} \sum_{b\in B''_G} \Big[q(S) - \widehat{q}_{B'}(S)\Big]\Big[\widehat{q}_b(S') - q(S')\Big] &= \max_{S,S'\in[d]} |B''_G|\Big[\widehat{q}_{B''_G}(S') - q(S')\Big]\Big[q(S) - \widehat{q}_{B'}(S)\Big]. \\ &\leq 2(33 + 2\sqrt{\tau_{B'}})\Big|B_G\Big|\epsilon\gamma.\end{aligned}$$

Combining the three bounds we have:

$$\begin{aligned}\sum_{b\in B_G''}\langle M^*, \widehat{\mathbf{C}}_{b,B'}\rangle &\le 8\Big|B_G''\Big|(33+2\sqrt{\tau_{B'}})^2\epsilon\gamma + 32(33+2\sqrt{\tau_{B'}})\Big|B_G\Big|\epsilon\gamma + 264\Big|B_G\Big|\gamma \\ &\le 8\Big|B_G\Big|(33+2\sqrt{\tau_{B'}})^2\epsilon^2\gamma + 32(33+2\sqrt{\tau_{B'}})\Big|B_G\Big|\epsilon\gamma + 264\Big|B_G\Big|\gamma \\ &\le \left[32\tau_{B'}\epsilon^2 + (1056\epsilon^2+64)\sqrt{\tau_{B'}} + (1056\epsilon+264)\right]\Big|B_G\Big|\gamma.\end{aligned}$$

Which gives with $\epsilon \le 1/100$:

$$\sum_{b\in B_G''}\langle M^*, \widehat{\mathbf{C}}_{b,B'}\rangle \le (0.0032\tau_{B'} + 65\sqrt{\tau_{B'}} + 275)\Big|B_G\Big|\gamma.$$

Thus, we have:

$$\frac{\sum_{b\in B_A'}\langle M^*, \widehat{\mathbf{C}}_{b,B'}\rangle}{\sum_{b\in B_G''}\langle M^*, \widehat{\mathbf{C}}_{b,B'}\rangle} \ge \frac{0.66\tau_{B'} - 62\sqrt{\tau_{B'}} - 2260}{0.02\tau_{B'} + 65\sqrt{\tau_{B'}} + 275}.$$

With $\sqrt{\tau_{B'}} \ge 200$,

$$\frac{\sum_{b\in B_A'}\langle M^*, \widehat{\mathbf{C}}_{b,B'}\rangle}{\sum_{b\in B_G''}\langle M^*, \widehat{\mathbf{C}}_{b,B'}\rangle} \ge 8.$$

4.A.7 Auxiliary Lemmas

Lemma 11 (Covariance is Lipschitz)**.** *Let $q, q' \in \mathbb{R}^d$ and define $\epsilon = q' - q$. For any $S, S' \subset [d]$, if $\Big|\epsilon(S)\Big| \vee \Big|\epsilon(S')\Big| \le 12$, then*

$$\left|Cov_{S,S'}\left(q\right) - Cov_{S,S'}\left(q'\right)\right| \le \frac{15}{k}\max\left(\Big|\epsilon(S)\Big|, \Big|\epsilon(S')\Big|\right).$$

Proof of Lemma 11. By equation (4.13), we have $|\Delta_S| \le 1$ for all $S \subset [d]$. Therefore, by Lemma 3 and equation (4.20):

$$\begin{aligned}\left|\mathrm{Cov}_{S,S'}\left(q\right) - \mathrm{Cov}_{S,S'}\left(q'\right)\right| &= \left|\left\langle \mathbb{1}_S\mathbb{1}_{S'}^T, \mathbf{C}(q) - \mathbf{C}(q')\right\rangle\right| \\ &= \frac{1}{k}\left|\left\langle \mathbb{1}_S\mathbb{1}_{S'}^T, qq^T - (q+\epsilon)(q+\epsilon)^T + \lambda\mathbb{1}\epsilon^T + \lambda\epsilon\mathbb{1}^T + (1-2\lambda)\mathrm{Diag}(\epsilon)\right\rangle\right| \\ &= \frac{1}{k}\left|\epsilon(S)\Delta_{S'} + \epsilon(S')\Delta_S + (1-2\lambda)\epsilon(S\cap S') - \epsilon(S)\epsilon(S')\right| \\ &\le \frac{1}{k}\left(\Big|\epsilon(S)\Big| + \Big|\epsilon(S')\Big| + \Big|\epsilon(S)\Big| + 12\Big|\epsilon(S)\Big|\right) \\ &\le \frac{15}{k}\max\left(|\epsilon(S)|, |\epsilon(S')|\right).\end{aligned}$$

□

If $22\epsilon\sqrt{\frac{d\ln(e/\epsilon}{k}} \geq 1$, the proven bound for the algorithm is trivially true. Else, whenever condition 1 holds, we have for any $|B'_G| \geq (1-2\epsilon)\left|B_G\right|$:

$$\max \left|\widehat{q}_{B'_G}(S) - q(S)\right| \leq 1.$$

Thus, Lemma 11 may be applied to $\mathrm{Cov}_{S,S'}(\widehat{q}_{B'_G}) - \mathrm{Cov}_{S,S'}(q)$.

4.B Proof of Corollary 4.2

Lemma 12. *Let $p \in \mathcal{P}_d$ and $p' \in \mathbb{R}^d$. Then* $\sup_{S\subseteq[d} |p(S) - p'(S)| \leq \|p - p'\|_1 \leq 2 \sup_{S\subseteq[d} |p(S) - p'(S)|$.

Proof of Lemma 4.B. The first inequality follows from the triangle inequality. For the second one, letting $A = \{j \in [d] : p_j \geq p'_j\}$, we have: $\|p - p'\|_1 = p(A) - p'(A) + p'(A^c) - p(A^c) \leq 2 \sup_{S\subseteq[d} |p(S) - p'(S)|$. □

Proof of Corollary 4.2. Let $\widehat{p}$ be the output of Algorithm 4 and $\widehat{p}^* = \frac{\widehat{p}}{\|\widehat{p}\|_1}$. Then

$$\|p - \widehat{p}^*\|_1 \leq \|\widehat{p} - p\|_1 + \|\widehat{p} - \widehat{p}^*|_1 = \|\widehat{p} - p\|_1 + \left|\|\widehat{p}\|_1 - 1\right| \leq 2\|p - \widehat{p}\|_1.$$

□

4.C Lower bound: Proof of Proposition 4.1

For any two probability distributions p, q over some measurable space $(\mathcal{X}, \mathcal{A})$, we denote by

$$\chi^2(p||q) = \begin{cases} \int_{\mathcal{X}} \frac{p}{q}\, dp - 1 \text{ if } p \ll q \\ +\infty \text{ otherwise} \end{cases}$$

the χ^2 divergence between p and q. We start with the following Lemma.

Lemma 13. *Assume $d \geq 3$. There exists an absolute constant $c > 0$ such that for all estimator $\hat{p}$ and all α-LDP mechanism Q, there exists a probability vector $p \in \mathcal{P}_d$ satisfying*

$$\mathbb{E}\left[\sup_{z' \in \mathcal{C}(Z)} \left\|\hat{p}(z') - p\right\|_1\right] \geq c\left\{\left(\frac{d}{\alpha\sqrt{kn}} + \frac{\epsilon\sqrt{d}}{\alpha\sqrt{k}}\right) \wedge 1\right\},$$

where the expectation is taken over all collections of n' clean batches $Z^1, \ldots, Z^{n'}$ where $Z^b = (Z_1^b, \ldots, Z_k^b)$ and $Z_l^b \overset{iid}{\sim} Qp$.

This Lemma is the analog of Proposition 4.1 but with the guarantee in expectation rather than with high probability. We first prove this Lemma before moving to the proof of Proposition 4.1.

Proof of Lemma 13. We first show that $R^*_{n,k}(\alpha,\epsilon,d) \geq c\left(\frac{d}{\alpha\sqrt{kn}} \wedge 1\right)$ for some small enough absolute constant $c > 0$. Informally, this amounts to saying that the estimation problem under both contamination and privacy is more difficult than just under privacy. Formally:

$$R^*_{n,k}(\alpha,\epsilon,d) = \inf_{\hat{p},Q} \sup_{p\in\mathcal{P}_d} \mathbb{E}\left[\sup_{z'\in\mathcal{C}(Z))} \left\|\hat{p}(z') - p\right\|_1\right] \geq \inf_{\hat{p},Q} \sup_{p\in\mathcal{P}_d} \mathbb{E}\left[\left\|\hat{p} - p\right\|_1\right] \geq c\left(\frac{d}{\alpha\sqrt{kn}} \wedge 1\right),$$

where the last inequality follows from [68] Proposition 6. We also give a simpler proof of this fact in Appendix 4.D, using Assouad's lemma.

We now prove $R^*_{n,k}(\alpha,\epsilon,d) \geq c\left(\frac{\epsilon\sqrt{d}}{\alpha\sqrt{k}} \wedge 1\right)$. For any α-LDP mechanism Q and probability vector $p \in \mathcal{P}_d$, denote by Qp the density of the privatized random variable Z defined by $Z|X \sim Q(\cdot|X)$ and by $Qp^{\otimes k}$ the density of the joint distribution of k iid observations with distribution Qp. Define the set of pairs of probability vectors that are indistinguishable after privatization by Q and adversarial contamination

$$\mathcal{A}(Q) = \left\{(p,q) \in \mathcal{P}_d \;\middle|\; TV(Qp^{\otimes k}, Qq^{\otimes k}) \leq \epsilon\right\}. \tag{4.23}$$

To derive the adversarial rate, it suffices to prove

$$\inf_{Q} \sup_{p,q\in\mathcal{A}(Q)} \|p - q\|_1 \geq c\left\{\frac{\epsilon\sqrt{d}}{\alpha\sqrt{k}} \wedge 1\right\}. \tag{4.24}$$

To understand why (4.24) is a natural program to consider, fix an α-LDP mechanism Q and denote by $(\mathcal{Z},\mathcal{U},\nu)$ its image space. If $(p,q) \in \mathcal{A}(Q)$, then letting

$$A = \frac{Qp^{\otimes k} \vee Qq^{\otimes k}}{1 + TV(Qp^{\otimes k}, Qq^{\otimes k})}, \quad N^{(p)} = \frac{A - (1-\epsilon)Qp^{\otimes k}}{\epsilon}, \quad \text{and } N^{(q)} = \frac{A - (1-\epsilon)Qq^{\otimes k}}{\epsilon},$$

we can directly check that $A, N^{(p)}$ and $N^{(q)}$ are probability measures over $(\mathcal{Z},\mathcal{U})$ (for $N^{(p)}$ and $N^{(q)}$, we use the fact that $(p,q) \in \mathcal{A}(Q)$ to prove that $N^{(p)}(dz) \geq 0$ and $N^{(q)}(dz) \geq 0$). Moreover, it holds that $A = (1-\epsilon)Qp^{\otimes k} + \epsilon N^{(p)} = (1-\epsilon)Qq^{\otimes k} + \epsilon N^{(q)}$. This is exactly equivalent to saying that any clean family of n batches with distribution $Qp^{\otimes k}$ or $Qp^{\otimes k}$ can be transformed into a ϵ-contaminated family of n batches with distribution A through ϵ adversarial contamination. By observing such a contaminated family, it is therefore impossible to determine whether the underlying distribution is p or q, so that the quantity $\|p-q\|_1/2$ is a lower bound on the minimax estimation risk.

We now prove (4.24). For all $j \in \{1, \dots, d\}$ and $z \in \mathcal{Z}$, set

$$q_j(z) = \frac{Q(z|j)}{Q(z|1)} - 1, \qquad d\mu(z) = Q(z|1)d\nu(z), \tag{4.25}$$

and

$$\Omega_Q = \Big(\Omega_Q(j,j')\Big)_{jj'} = \left(\int_{\mathcal{Z}} q_j(z) q_{j'}(z) d\mu(z)\right)_{ij} \tag{4.26}$$

Given Q, we first prove that a sufficient condition for (p,q) to belong to $\mathcal{A}(Q)$ is that $(p-q)^T\Omega(p-q) \leq C\epsilon^2/k$ for some small enough absolute constant $C > 0$. Fix $p, q \in \mathcal{P}_d$ and define $\Delta = p - q$. By [188], Section 2.4, we have

$$TV(Qp^{\otimes k}, Qq^{\otimes k}) \leq \sqrt{-1 + (1 + \chi^2(Qp||Qq))^k}. \tag{4.27}$$

Now,

$$\begin{aligned}
\chi^2(Qp||Qq) &= \int_{\mathcal{Z}} \frac{(Qp(z) - Qq(z))^2}{Qq(z)} dz = \int_{\mathcal{Z}} \frac{\left(\sum_{j=1}^d Q(z|j)\ \Delta_j\right)^2}{\sum_{j=1}^d Q(z|j)\ q_j} dz \\
&= \int_{\mathcal{Z}} \frac{\left(\sum_{j=1}^d \left(\frac{Q(z|j)}{Q(z|1)} - 1\right) \Delta_j\right)^2}{\sum_{j=1}^d \frac{Q(z|j)}{Q(z|1)}\ q_j} Q(z|1) d\nu(z) \qquad \text{since } \sum_{j=1}^d \Delta_j = 0 \\
&\leq e^\alpha \int_{\mathcal{Z}} \sum_{j,j'=1}^d \Delta_j \Delta_{j'} q_j(z) q_{j'}(z) d\mu(z) \\
&= e^\alpha \Delta^T \Omega_Q \Delta.
\end{aligned}$$

Write $\Omega = \Omega_Q$ and assume that $\Delta^T\Omega_Q\Delta \leq C\epsilon^2/k$ for $C \leq e^{-2}$. Then equation (4.27) yields:

$$TV(Qp^{\otimes k}, Qq^{\otimes k}) \leq \sqrt{-1 + (1 + e^\alpha \Delta^T\Omega\Delta)^k} \leq \sqrt{-1 + \exp\left(e^\alpha k \Delta^T\Omega\Delta\right)} \leq \sqrt{-1 + \exp\left(Ce^\alpha\epsilon^2\right)} \leq \epsilon.$$

Defining

$$\mathcal{A}_{\chi^2}(Q) = \left\{(p,q) \in \mathcal{P} \;\middle|\; (p-q)^T\Omega(p-q) \leq \frac{C\epsilon^2}{k}\right\}, \tag{4.28}$$

it follows that $\mathcal{A}_{\chi^2}(Q) \subset \mathcal{A}(Q)$ for all Q, so that

$$\inf_Q \sup_{(p,q)\in\mathcal{A}(Q)} \|\Delta\|_1 \geq \inf_Q \sup_{(p,q)\in\mathcal{A}_{\chi^2}(Q)} \|\Delta\|_1.$$

Fix Q and note that Ω_Q is symmetric and nonnegative. We sort its eigenvalues as $\{\lambda_1 \leq \cdots \leq \lambda_d\}$ and denote by $v_1, \dots, v_d$ the associated eigenvectors. We also define $j_0 = \max\left\{j \in \{1, \dots, d\} : \lambda_j \leq 3e^2\alpha^2\right\}$. Noting that $\forall j : |q_j| \leq e\alpha$ and that μ is a probability measure, we get that $Tr(\Omega) = \sum_{j=1}^d \int_{\mathcal{Z}} q_j^2 d\mu \leq$

$de^2\alpha^2$, so that $(d-j_0)3e^2\alpha^2 \leq de^2\alpha^2$ hence $j_0 \geq 2d/3$.

Let $H = \left\{x \in \mathbb{R}^d : x^T\mathbb{1} = 0\right\}$, and note that $V := \mathrm{span}(v_j)_{j\leq j_0} \cap H$ is of dimension at least $m = \frac{2d}{3} - 1 \geq \frac{d}{3}$. Therefore by Lemma 14, there exists $\Delta \in V$ such that $\|\Delta\|_2^2 = \frac{C\epsilon^2}{2e^2\alpha^2 k} \wedge \frac{1}{d}$ and $\|\Delta\|_1 \geq C_{14}\sqrt{m}\|\Delta\|_2 \gtrsim \frac{\epsilon\sqrt{d}}{\alpha\sqrt{k}}$. Noting that over $\mathbb{R}^d$, $\|\cdot\|_1 \leq \sqrt{d}\|\cdot\|_2$, we also have $\|\Delta\|_1 \leq \frac{\epsilon}{\alpha\sqrt{k}} \wedge 1 \leq 1$.

This allows us to define the following vectors: $p = \left(\frac{|\Delta_j|}{\|\Delta\|_1}\right)_{j=1}^d \in \mathcal{P}_d$ and $q = p - \Delta$. To check that $q \in \mathcal{P}_d$, note that the condition $\Delta^T\mathbb{1} = 0$ ensures that $q^T\mathbb{1} = 1$. Moreover, for all $j \in \{1, \dots, d\}$ we have $q_j = \frac{|\Delta_j|}{\|\Delta\|_1} - \Delta_j \geq 0$ since $\|\Delta\|_1 \leq 1$.
Since by construction, we have $\Delta\Omega_Q\Delta \leq 2e^2\alpha^2\|\Delta\|_2^2 \leq \frac{C\epsilon^2}{k}$ and $p, q \in \mathcal{P}_d$, we have $(p,q) \in \mathcal{A}_{\chi^2}(Q)$. For all α-LDP mechanism Q, it therefore holds that $\sup_{(p,q)\in\mathcal{A}_{\chi^2}(Q)} \|\Delta\|_1 \gtrsim \frac{\epsilon\sqrt{d}}{\alpha\sqrt{k}} \wedge 1$. Taking the infimum over all Q, the result is proven. □

Lemma 14. *There exists an absolute constant C_{14} such that for all $m \in \{\lceil \frac{d}{3} \rceil, \dots, d\}$ and all linear subspace $V \subset \mathbb{R}^d$ of dimension m, it holds:*

$$\sup_{v\in V} \frac{\|v\|_1}{\|v\|_2} \geq C_{14}\sqrt{m}.$$

Proof of Lemma 14. Let V be a linear subspace of $\mathbb{R}^d$ of dimension m and denote by $\Pi_V := (\Pi_V(i,j))_{ij}$ the orthogonal projector onto V. Let $X \sim \mathcal{N}(0, \Pi_V)$. For some large enough absolute constant $C > 0$ we have:

$$\begin{aligned}\sup_{v\in V} \frac{\|v\|_1}{\|v\|_2} &\geq \mathbb{E}\left[\frac{\|X\|_1}{\|X\|_2}\right] \geq \mathbb{E}\left[\frac{\|X\|_1}{\|X\|_2}\mathbb{1}\left\{\|X\|_2 \leq C\sqrt{m}\right\}\right] \\ &\geq \underbrace{\frac{1}{C\sqrt{m}}\mathbb{E}\left[\|X\|_1\right]}_{\text{Principal term}} - \underbrace{\frac{1}{C\sqrt{m}}\mathbb{E}\left[\|X\|_1\mathbb{1}\left\{\|X\|_2 > C\sqrt{m}\right\}\right]}_{\text{Residual term}} \end{aligned} \tag{4.29}$$

We first analyze the principal term.

$$\mathbb{E}\|X\|_1 = \sum_{i,j=1}^d \mathbb{E}|X_{ij}| = \sqrt{\frac{2}{\pi}}\sum_{i,j=1}^d |\Pi_V(i,j)|^{1/2}$$

Note that $\forall i,j \in \{1,\dots,d\} : |\Pi_V(i,j)| \leq 1$ and that $\sum_{i,j=1}^d \Pi_V^2(i,j) = m$. Therefore:

$$\inf_{dim(V)=m} \sum_{i,j=1}^d |\Pi_V(i,j)|^{1/2} \geq \inf_{A\in\mathbb{R}^{d\times d}} \sum_{i,j=1}^d |a_{ij}|^{1/2} \text{ s.t. } \begin{cases}\|A\|_2^2 = m \\ \forall i,j : |a_{ij}| \leq 1.\end{cases}$$

$$= \inf_{a\in\mathbb{R}^{d\times d}} \sum_{i,j=1}^{d} a_{ij} \quad \text{s.t.} \quad \begin{cases} \sum_{i,j=1}^{d} a_{ij}^4 = m \\ \forall i,j : 0 \le a_{ij} \le 1. \end{cases} \tag{4.30}$$

The last optimization problem amounts to minimizing an affine function over a convex set, hence the solution, denoted by $(a^*_{ij})_{ij}$, is attained on the boundaries of the domain. Therefore, $\forall i,j \in \{1,\dots,d\} : a^*_{ij} \in \{0,1\}$. It follows from $\sum_{ij} a^4_{ij} = m$ that the family a^*_{ij} contains exactly m nonzero coefficients, which are all equal to 1. Therefore, the value of the last optimization problem is m, which yields that the principal term is lower bounded by $\frac{\sqrt{m}}{C}$.

We now move to the residual term. Writing $X = \sum_{j=1}^{m} x_j e_j$ where $(e_j)_{j=1}^m$ is an orthonormal basis of V, we have:

$$\begin{aligned} \mathbb{E}\left[\left\|X\right\|_1 \mathbb{1}\left\{\left\|X\right\|_2 > C\sqrt{m}\right\}\right] &\le \sqrt{d}\,\mathbb{E}\left[\left\|X\right\|_2 \mathbb{1}\left\{\left\|X\right\|_2 > C\sqrt{m}\right\}\right] \le \sqrt{d}\left\{\mathbb{E}\left[\left\|X\right\|_2^2 \mathbb{1}\left\{\left\|X\right\|_2^2 > C^2 m\right\}\right]\right\}^{1/2} \\ &\le \sqrt{d}\left\{m\,\mathbb{E}\left[x_1^2 \mathbb{1}\left\{\sum_{j=1}^{m} x_j^2 \ge C^2 m\right\}\right]\right\}^{1/2}. \end{aligned} \tag{4.31}$$

Moreover

$$\begin{aligned} \mathbb{E}\left[x_1^2 \mathbb{1}\left\{\sum_{j=1}^{m} x_j^2 \ge C^2 m\right\}\right] &\le \mathbb{E}\left[x_1^2\,\mathbb{1}\{x_1 \ge C\}\right] + \mathbb{E}\left[x_1^2\,\mathbb{1}\left\{\sum_{j=2}^{m} x_j^2 \ge C^2(m-1)\right\}\right] \\ &\le \mathbb{E}\left[x_1^2\,\mathbb{1}\{x_1 \ge C\}\right] + \mathbb{E}\left[x_1^2\right]\mathbb{P}\left(\left|\sum_{j=2}^{m} x_j^2 - \mathbb{E}x_1^2\right| \ge \left(C^2 - \mathbb{E}x_1^2\right)(m-1)\right) \end{aligned} \tag{4.32}$$

By the dominated convergence Theorem, $\lim_{C\to+\infty} \mathbb{E}\left[x_1^2\,\mathbb{1}\{x_1 \ge C\}\right] = 0$. Moreover, by Chebyshev's inequality:

$$\mathbb{P}\left(\left|\sum_{j=2}^{m} x_j^2 - \mathbb{E}x_1^2\right| \ge \left(C^2 - \mathbb{E}x_1^2\right)(m-1)\right) \le \frac{\mathbb{V}(x_1^2)}{\left(C^2 - \mathbb{E}x_1^2\right)^2 (m-1)} \underset{C\to+\infty}{\to} 0. \tag{4.33}$$

By (4.31), (4.32) and (4.33), we conclude that for all absolute constant $c > 0$, there exists a large enough absolute constant $C > 0$ such that the residual term is at most $\frac{c\sqrt{d}}{C}$. Take $c = \frac{1}{2}$ and $m \ge \frac{d}{3}$, then by equation (4.29) we get:

$$\sup_{v\in V} \frac{\|v\|_1}{\|v\|_2} \ge \frac{\sqrt{m}}{C} - \frac{c\sqrt{d}}{C} \ge \left(1 - \frac{\sqrt{3}}{2}\right)\frac{\sqrt{m}}{C} =: C_{14}\sqrt{m}.$$

□

Proof of Proposition 4.1. We distinguish between two cases.

1. **First case** If $\frac{d}{\alpha\sqrt{nk}} \leq \frac{\epsilon}{\alpha}\sqrt{\frac{d}{k}}$ i.e. if the dominating term comes from the contamination, taking $p, q \in \mathcal{P}_d$ like in the proof of Proposition 13 and $t \in \{p, q\}$ uniformly at random yields that

$$\inf_{\hat{p}} \sup_{p\in\mathcal{P}_d} \mathbb{P}\left(\sup_{Z\in\mathcal{C}(Y)} \|\hat{p}(Z) - p\|_1 \geq \|p-q\|_1/2\right)$$
$$\geq \inf_{\hat{p}} \mathbb{E}_{t\in\{p,q\}}\mathbb{P}_t\left(\sup_{Z\in\mathcal{C}(Y)} \|\hat{p}(Z) - p\|_1 \geq \|p-q\|_1/2\right) \geq \frac{1}{2} \geq O(e^{-d}),$$

   where $\|p-q\|_1 \gtrsim \frac{\epsilon}{\alpha}\sqrt{\frac{d}{k}} \wedge 1$.

2. **Second case** If $\frac{d}{\alpha\sqrt{nk}} \geq \frac{\epsilon}{\alpha}\sqrt{\frac{d}{k}}$ i.e. if the dominating term comes from the privacy constraint, then we set $N = nk$ and assume that we observe $Z_1, \ldots, Z_N$ iid with probability distribution $Z|X \sim Q(\cdot|X)$ such that X has a discrete distribution over $\{1, \ldots, d\}$. In other words, the random variables Z_i are no longer batches, but rather we have nk iid clean samples that are privatized versions of iid samples with distribution p. By section 4.D, it holds that

$$\inf_{\hat{p}} \sup_{p\in\mathcal{P}_d} \mathbb{E}\left[\sup_{\text{contamination}} \|\hat{p} - p\|_1\right] \geq \inf_{\hat{p}} \sup_{p\in\mathcal{P}_d} \mathbb{E}\|\hat{p} - p\|_1 \geq c\frac{d}{\alpha\sqrt{N}},$$

   for some small enough absolute constant $c > 0$. We use the definition of γ and of the cubic set of hypotheses $\mathcal{P}$ from (4.37). Let $\hat{p}$ be any estimator of the probability parameter and, for some small enough absolute constant $c > 0$, define

$$r = c\,\frac{d}{\sqrt{kn}}. \tag{4.34}$$

   We first justify that for this particular set of hypotheses, it is possible to assume *wlog* that

$$\|p - \hat{p}\|_1 \leq 6\gamma d \leq 6\,c_\gamma r. \tag{4.35}$$

   Indeed, define $u = \left(\frac{1}{d}\right)_{j=1}^{d}$. If for some observation $Z = (Z_1, \ldots Z_N)$ the estimate $\hat{p}(Z)$ satisfies $\|\hat{p}(Z) - u\|_1 > 4\gamma d$, then it is possible to improve $\hat{p}$ by replacing it with the estimator $\bar{p}$ satisfying $\|\bar{p}(Z) - p\|_1 \leq 6\gamma d$ and defined as:

$$\bar{p} := \hat{p}\mathbb{1}\left\{\left\|\hat{p} - u\right\|_1 \leq 4\gamma d\right\} + u\mathbb{1}\left\{\left\|\hat{p} - u\right\|_1 > 4\gamma d\right\}.$$

   Indeed, recalling that $\forall p \in \mathcal{P} : \left\|u - p\right\|_1 = 2\gamma d$, there are two cases.

   - If $\left\|\hat{p}(Z) - u\right\|_1 \leq 4\gamma d$, then $\hat{p} = \bar{p}$ so that $\left\|p - \bar{p}\right\|_1 \leq \left\|p - u\right\|_1 + \left\|u - \bar{p}\right\|_1 \leq 2\gamma d + 4\gamma d = 6\gamma d$.

- Otherwise, $\bar{p} = u$ and we get

$$\left\|\bar{p}(Z) - p\right\|_1 = 2\gamma d = 4\gamma d - 2\gamma d < \left\|\hat{p}(Z) - u\right\|_1 - \left\|u - p\right\|_1 \leq \left\|\hat{p}(Z) - p\right\|_1,$$

which proves that (4.35) can be assumed *wlog*. Now, from the proof of Lemma 15, we also have

$$\sup_{p\in\mathcal{P}} \mathbb{E}_p\left\|\hat{p} - p\right\|_1 \geq \frac{c_\gamma}{4} r =: Cr.$$

Fix any $p \in \mathcal{P}$ and write $\pi := \sup_{p\in\mathcal{P}} \mathbb{P}_p\left(\|\hat{p} - p\|_1 \geq cr\right)$ for $c = c_\gamma\left(\frac{1}{4} - 6\delta\right) > 0$ for $\delta < \frac{1}{24} =: c'$.

It follows that:

$$\begin{aligned}
Cr &\leq \sup_{p\in\mathcal{P}} \mathbb{E}_p\left\|\hat{p} - p\right\|_1 \\
&= \sup_{p\in\mathcal{P}} \left\{ \mathbb{E}_p\left[\left\|\hat{p} - p\right\|_1 \mathbb{1}\left\{\left\|\hat{p} - p\right\|_1 \geq cr\right\}\right] + \mathbb{E}_p\left[\left\|\hat{p} - p\right\|_1 \mathbb{1}\left\{\left\|\hat{p} - p\right\|_1 < cr\right\}\right] \right\} \\
&\leq 6c_\gamma r \cdot \pi + cr \quad \text{by equation (4.35), so that } \pi \geq \frac{C - c}{6c_\gamma} \geq \delta \geq O(e^{-d}).
\end{aligned}$$

□

4.D Simpler proof of the lower bound with privacy and no outliers

Here, we assume that $k = 1$ and that we observe $Z_1, \ldots, Z_n$ that are n iid with probability distribution $Z|X \sim Q(\cdot|X)$ and X has a discrete distribution over $\{1, \ldots, d\}$. We prove the following Lemma

Lemma 15. *In this setting, it holds*

$$\inf_{\hat{p}} \sup_{p\in\mathcal{P}_d} \mathbb{E}\|\hat{p} - p\|_1 \geq c\frac{d}{\alpha\sqrt{n}}, \tag{4.36}$$

for some small enough absolute constant $c > 0$.

For all $\epsilon \in \{\pm 1\}^{\lfloor d/2 \rfloor}$, define the probability vector $p_\epsilon \in \mathcal{P}_d$ such that

$$\forall j \in \{1, \ldots, d\} : p_\epsilon(j) = \begin{cases} \frac{1}{d} + \epsilon_j\, \gamma & \text{if } j \leq \frac{d}{2}, \\ \frac{1}{d} & \text{if } d \text{ is odd and } j = \frac{d+1}{2}, \\ \frac{1}{d} - \epsilon_{d-j+1}\gamma & \text{otherwise,} \end{cases} \tag{4.37}$$

where $\gamma = \frac{c_\gamma}{\alpha\sqrt{n}} \wedge \frac{c_\gamma}{d}$ and c_γ is a small enough absolute constant. Consider the cubic set of hypotheses

$$\mathcal{P} = \left\{ p_\epsilon \;\middle|\; \epsilon \in \{\pm 1\}^{\lfloor d/2 \rfloor} \right\}. \tag{4.38}$$

This set $\mathcal{P}$ consists of $M = 2^{\lfloor d/2 \rfloor}$ hypotheses. Over $\mathcal{P}$, the ℓ_1 distance simplifies as follows:

$$\forall \epsilon, \epsilon' \in \{\pm 1\}^{\lfloor d/2 \rfloor} : \|p_\epsilon - p_{\epsilon'}\|_1 = 4\gamma \; \rho(\epsilon, \epsilon'), \tag{4.39}$$

where $\rho(\epsilon, \epsilon') = \sum_{j=1}^{\lfloor d/2 \rfloor} \mathbb{1}_{\epsilon_j \neq \epsilon'_j}$ denotes the Hamming distance between ϵ and ϵ'.

To apply Assouad's Lemma (see e.g. [188] Theorem 2.12.(ii)), let $\epsilon, \epsilon' \in \{\pm 1\}^{\lfloor d/2 \rfloor}$ such that $\rho(\epsilon, \epsilon') = 1$. Recall that the observations $Z_1, \dots, Z_n$ are iid and follow the distribution $Z|X \sim Q(\cdot|X)$ where Q is an α-locally differentially private mechanism. Fix any such mechanism Q, and denote by q_ϵ and $q_{\epsilon'}$ the respective densities of Z when $X \sim p_\epsilon$ and $X \sim p_{\epsilon'}$. We therefore have $\forall z \in \mathcal{Z} : q_\epsilon(z) = \int Q(z|x) \; p_\epsilon(x) d\nu(x)$ where ν denotes the counting measure over $\{1, \dots, d\}$. For some probability distribution P, we also denote by $P^{\otimes n}$ the law of the probability vector $(X_1, \dots, X_n)$ when $X_i \overset{iid}{\sim} P$. Now, we have:

$$TV\left(q_\epsilon^{\otimes n}, q_{\epsilon'}^{\otimes n}\right) \leq \sqrt{\chi^2\left(q_\epsilon^{\otimes n} \;||\; q_{\epsilon'}^{\otimes n}\right)} = \sqrt{\left(1 + \chi^2\left(q_\epsilon \;||\; q_{\epsilon'}\right)\right)^n - 1}, \tag{4.40}$$

and defining $\Delta p(x) = p_\epsilon(x) - p_{\epsilon'}(x)$ for all $x \in \{1, \dots, d\}$, we can write:

$$\begin{aligned}
\chi^2\left(p_\epsilon \;||\; p_{\epsilon'}\right) &= \int_{\mathcal{Z}} \frac{\left(q_\epsilon(z) - q_{\epsilon'}(z)\right)^2}{q_{\epsilon'}(z)} dz = \int_{\mathcal{Z}} \frac{\left(\int Q(z|x) \; \Delta p(x) d\nu(x)\right)^2}{\int Q(z|x) \; p_{\epsilon'}(x) d\nu(x)} dz \\
&= \int_{\mathcal{Z}} Q(z|1) \; \frac{\left(\int_{\mathcal{X}} \left(\frac{Q(z|x)}{Q(z|1)} - 1\right) \Delta p(x) d\nu(x)\right)^2}{\int_{\mathcal{X}} \frac{Q(z|x)}{Q(z|1)} \; p_{\epsilon'}(x) d\nu(x)} dz \qquad \text{since } \int \Delta p(x) d\nu(x) = 0 \\
&\leq \int_{\mathcal{Z}} Q(z|1) \; \frac{\left(\int_{\mathcal{X}} \left|\frac{Q(z|x)}{Q(z|1)} - 1\right| \left|\Delta p(x)\right| d\nu(x)\right)^2}{\int_{\mathcal{X}} e^{-\alpha} \; p_{\epsilon'}(x) d\nu(x)} dz \\
&\leq \int_{\mathcal{Z}} Q(z|1) \; \frac{\left(C\alpha \; TV(p_\epsilon, p_{\epsilon'})\right)^2}{e^{-\alpha}} dz = e^\alpha C^2 \alpha^2 \left(2\gamma \; \rho(p_\epsilon, p_{\epsilon'})\right)^2 \\
&\leq 12 C^2 \alpha^2 \gamma^2 \leq \frac{12 C^2 c_\gamma^2}{n},
\end{aligned}$$

where $C > 0$ is an absolute constant such that for all $\alpha \in (0, 1)$ we have $e^\alpha - 1 \leq C\alpha$ and $1 - e^{-\alpha} \leq C\alpha$. Now by (4.40), we have:

$$TV\left(q_\epsilon^{\otimes n}, q_{\epsilon'}^{\otimes n}\right) \leq \sqrt{\left(1 + \chi^2\left(q_\epsilon \;||\; q_{\epsilon'}\right)\right)^n - 1} \leq \sqrt{\exp\left(12 C^2 c_\gamma^2\right) - 1}.$$

Choosing c_γ small enough therefore ensures that $TV\left(q_\epsilon^{\otimes n}, q_{\epsilon'}^{\otimes n}\right) \leq \frac{1}{2}$, so that by Assouad's lemma, the minimax risk is lower bounded as:

$$\inf_{\widehat{p}} \sup_{p \in \mathcal{P}_d} \mathbb{E}\|\widehat{p} - p\|_1 \geq \left\lfloor \frac{d}{2} \right\rfloor \frac{1}{2} 4\gamma \left(1 - \frac{1}{2}\right) \geq \frac{c_\gamma}{10} \left(\frac{d}{\alpha\sqrt{n}} \wedge 1\right).$$

Part III

Benign overfitting

Chapter 5

Benign overfitting in adaptive nonparametric regression

This chapter is based on the paper "benign overfitting and adaptive nonparametric regression" [170] with Suzanne Sigalla and Alexandre Tsybakov (arXiv:2206.13347).

Abstract

In the context of nonparametric regression with square loss, we construct an estimator that is a continuous function interpolating the data points with high probability, while being minimax optimal and adaptive to the unknown smoothness.

5.1 Introduction

Benign overfitting has attracted a great deal of attention in the recent years. It was initially motivated by the fact that deep neural networks have good predictive properties even when perfectly interpolating the training data [105], [98], [167], [154]. Such a behavior stands in strong contrast with the classical point of view that perfectly fitting the data points is not compatible with predicting well. With the aim of understanding this new phenomenon, a series of recent papers studied benign overfitting in linear regression setting, see [122], [146], [133], [144], [152], [175] and the references therein. The main conclusion for the linear model is that an unbalanced spectrum of the design matrix and over-parametrization, which in a sense approaches the model to non-parametric setting, are essential for benign overfitting to occur in linear regression. Extensions to kernel ridgeless regression were considered in [142] when the sample size n and the dimension d were assumed to satisfy $n \asymp d$, and in [143] for a more general case $d \asymp n^\alpha$ for $\alpha \in (0, 1)$. These papers give data-dependent upper bounds on the risk that can be small assuming favorable spectral properties of the data and the kernel matrix. On the other hand, if d is constant (independent of n) then the least-norm interpolating estimator with respect to the Laplace kernel is inconsistent [116].

In the line of work cited above, benign overfitting was understood as achieving simultaneously interpolation and prediction consistency, or possibly, consistency with some suboptimal rates. On the other hand, it was shown that, in non-parametric regression setting, interpolating estimators can attain minimax optimal rates [106]. Namely, it is proved in [106] that interpolation with minimax

optimal rates can be achieved by Nadaraya-Watson estimator with a singular kernel.

The idea of using singular kernels can be traced back to [7] giving start to popular techniques in image processing referred to as Shepard interpolation. In statistical language, Shepard interpolant is nothing else but the Nadaraya-Watson estimator with kernel $K(u) = 1/\|u\|^2$, where $\|\cdot\|$ denotes the Euclidean norm and $u \in \mathbf{R}^2$. Unaware of Shepard's work and its subsequent extensive use in image processing, [28] considered the same estimator in general dimension d, that is, with the kernel $K(u) = \|u\|^{-d}$ for $u \in \mathbf{R}^d$, and proved that the Nadaraya-Watson estimator with such a kernel is consistent in probability but fails to be pointwise almost surely consistent. However, this kernel is not integrable and has a peculiar property that the bandwidth cancels out from the definition of the estimator. Thus, the bias cannot be controlled and the bias-variance trade-off argument based on bandwidth selection does not apply. It remains unclear whether some rates of convergence can be achieved by such an estimator. Therefore, it was suggested in [106, 99] to localize and modify the kernel as $K(u) = \|u\|^{-a}\mathbf{1}(\|u\| \leq 1)$ where $0 < a < d/2$ rather than $a = d$ and $\mathbf{1}(\cdot)$ denotes the indicator function. The estimator with such a weaker type of singularity is also interpolating, and it was shown in [106, 99] that it achieves the minimax rates of convergence on the β-Hölder classes with $0 < \beta \leq 2$. Also, [97] proved a similar claim for the k nearest neighbor analog of this estimator with $0 < \beta \leq 1$. However, those results were restricted to functions with low smoothness β and the suggested estimators were not adaptive to β.

In this paper, we show that:

(i) interpolating estimators attaining minimax optimal rates on β-Hölder classes can be obtained for any smoothness $\beta > 0$,

(ii) estimators with such properties can be constructed adaptively to the unknown smoothness $\beta \in (0, \beta_{\max}]$, for any $\beta_{\max} > 0$.

The estimators that we consider to achieve (i) are local polynomial estimators (LPE) with singular kernels. In order to obtain adaptive estimators, we apply aggregation techniques to a family of LPE with singular kernels.
As a by-product, we obtain non-asymptotic bounds for the squared risk of LPE in classical setting with non-singular kernels. To the best of our knowledge, such bounds are missing in the existing literature on LPE that was mainly focused on asymptotic properties such as convergence in probability or pointwise asymptotic normality, cf. [11, 14, 18, 178].

Note that local polynomial method with singular kernels has been used as interpolation tool in numerical analysis, starting from [12]. It was also invoked in the context of non-parametric regression in [184]. However, [12, 184] only discussed functional properties, such as the smoothness of interpolants, rather than their statistical behavior.

5.2 Preliminaries

5.2.1 Notation

For any vector $x = (x_1, \ldots, x_d) \in \mathbf{R}^d$ and any multi-index $s = (s_1, \ldots, s_d) \in \mathbf{N}^d$, we define

$$|s| = \sum_{i=1}^{d} s_i, \qquad s! = s_1! \dots s_d!$$

$$x^s = x_1^{s_1} \dots x_d^{s_d} \qquad D^s = \frac{\partial^{s_1+\dots+s_d}}{\partial x_1^{s_1} \dots \partial x_d^{s_d}}.$$

We denote by $\|\cdot\|$ the Euclidean norm, and by $\mathrm{Card}(J)$ the cardinality of set J. For any integer $k \in \mathbf{N}^*$, we set $[k] = \{1, \dots, k\}$. For any $x \in \mathbf{R}^d$, $r > 0$, we denote by $\mathcal{B}_d(x, r)$ the closed Euclidean ball centered at x with radius r. We set for brevity $\mathcal{B}_d = \mathcal{B}_d(0, 1)$. For any $\beta > 0$, we denote by $\lfloor \beta \rfloor$ the maximal integer less than β, and by $\lceil \beta \rceil$ the minimal integer greater than β. We use symbols C, C' to denote positive constants that can vary from line to line.

For any $k > 0$, we denote by I_k the identity matrix of size k. For any square matrix M, the writing $M \succ 0$ means that M is positive definite. For any matrix M, we denote by M^+ its Moore-Penrose inverse, and by $\|M\|_\infty$ its spectral norm.

5.2.2 Model

Let (X, Y) be a pair of random variables in $\mathbf{R}^d \times \mathbf{R}$ with distribution P_{XY} and assume that we are given n i.i.d. observations $\mathcal{D} := \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ with distribution P_{XY}. We denote by P_X the marginal distribution of X and assume that it admits a density p with respect to the Lebesgue measure on the compact set $\mathrm{Supp}(p)$. We assume that for all $x \in \mathrm{Supp}(p)$, the regression function $f(x) = \mathbf{E}(Y|X = x)$ exists and is finite. Set $\xi(X) = Y - \mathbf{E}(Y|X)$. Equivalently, the model can be written as $Y_i = f(X_i) + \xi(X_i)$, where $\mathbf{E}(\xi(X_i)|X_i) = 0$. We make the following assumptions.

Assumption (A1). $\mathbf{E}(|\xi(X)|^{2+\delta}|X = x) \leq C$ for all $x \in \mathrm{Supp}(p)$, where δ and C are positive constants.

Assumption (A2). X is distributed with Lebesgue density $p(\cdot)$ such that $p \in [p_{\min}, p_{\max}]$ where $p_{\max} \geq p_{\min} > 0$. The support $\mathrm{Supp}(p)$ of p is a convex compact set contained in $\mathcal{B}_d$.

For any estimator f_n of f based on the sample $\mathcal{D}$, we consider the following L_2-loss :

$$\|f_n - f\|_{L_2}^2 = \mathbf{E}_X\left(\left[f_n(X) - f(X)\right]^2\right) = \int \left[f_n(x) - f(x)\right]^2 p(x)dx.$$

Here, $\mathbf{E}_X$ denotes the expectation with respect to P_X. Next, we define the expected risk as

$$\mathbf{E}\left[\|f_n - f\|_{L_2}^2\right],$$

where $\mathbf{E}$ denotes the expectation with respect to the distribution of $\mathcal{D}$.

Definition 1 (Interpolating estimator)**.** *An estimator f_n of f based on a sample $\mathcal{D} = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ is called interpolating over $\mathcal{D}$ if $f_n(X_i) = Y_i$ for $i = 1, \dots, n$.*

5.2.3 Hölder classes of functions

For any k-linear form $A : (\mathbf{R}^d)^k \longrightarrow \mathbf{R}$, we define its norm as follows

$$\|A\| := \sup\left\{\left|A[h_1, \dots, h_k]\right| : \|h_j\| \le 1, j \in [k]\right\}. \tag{5.1}$$

Given a k-times continuously differentiable function $f : \mathbf{R}^d \longrightarrow \mathbf{R}$ and $x \in \mathbf{R}^d$, we denote by $f^{(k)}(x) : (\mathbf{R}^d)^k \longrightarrow \mathbf{R}$ the following k-linear form

$$f^{(k)}(x)[h_1, \dots, h_k] = \sum_{|m_j|=1, \forall j \in [k]} D^{m_1+\dots+m_k} f(x) h_1^{m_1} \dots h_k^{m_k}, \quad \forall h_1, \dots, h_k \in \mathbf{R}^d,$$

where $m_1, \dots, m_k \in \mathbf{N}^d$ are multi-indices. Throughout the paper, we will consider the following Hölder class of functions.

Definition 2. *Let $\beta > 0$, $L > 0$, and let $f : \mathcal{B}_d \longrightarrow \mathbf{R}$ be a $\ell = \lfloor \beta \rfloor$ times continuously differentiable function. We denote by $\Sigma(\beta, L)$ the set of all functions f defined on $\mathcal{B}_d$ such that*

$$\max_{0 \le k \le \ell} \sup_{x \in \mathcal{B}_d} \|f^{(k)}(x)\| + \sup_{x, x' \in \mathcal{B}_d} \frac{\|f^{(\ell)}(x) - f^{(\ell)}(x')\|}{\|x - x'\|^{\beta - \ell}} \le L.$$

These classes of functions have nice embedding properties that will be needed to prove our result on adaptive estimation. For $\beta' \le \beta \le 1$, we clearly have $\Sigma(\beta, L) \subseteq \Sigma(\beta', L)$. Analogous embedding is valid for $\beta > 1$ as stated in the next lemma proved in the Appendix.

Lemma 1. *For any $0 < \beta' \le \beta$ and $L > 0$ we have $\Sigma(\beta, L) \subseteq \Sigma(\beta', 2L)$.*

The class $\Sigma(\beta, L)$ is closely related to several differently defined Hölder classes used in the literature. One of them is based on Taylor approximation, cf., for example, [11]. For any $x \in \mathbf{R}^d$ and any ℓ times continuously differentiable real-valued function f on $\mathbf{R}^d$, we denote by Tf_x its Taylor polynomial of degree ℓ at point x:

$$Tf_x(x') = \sum_{0 \le |s| \le \ell} \frac{(x - x')^s}{s!} D^s f(x').$$

Lemma 2. *Let $\beta > 0$, $L > 0$ and $f \in \Sigma(\beta, L)$. Then for all $x, y \in \mathcal{B}_d$, and $\ell = \lfloor \beta \rfloor$ it holds that*

$$\left|f(x) - Tf_y(x)\right| \le \frac{L}{\ell!} \|x - y\|^\beta.$$

Thus, we have $\Sigma(\beta, L) \subseteq \Sigma'(\beta, L/\lfloor \beta \rfloor!)$, where $\Sigma'(\beta, L')$ stands for the class of all functions f satisfying the relation $\left|f(x) - Tf_y(x)\right| \le L' \|x - y\|^\beta$.
Next, considering one more definition of Hölder class:

$$\widetilde{\Sigma}(\beta, L) = \left\{ f : \mathcal{B}_d \to \mathbf{R} : \sup_{x, x'} \frac{\|f^{(\ell)}(x) - f^{(\ell)}(x')\|}{\|x - x'\|^{\beta - \ell}} \le L \right\}$$

we also immediately have that $\Sigma(\beta, L) \subseteq \widetilde{\Sigma}(\beta, L)$. It follows from [14] that the minimax estimation rate on the class $\widetilde{\Sigma}(\beta, L)$ under the squared loss that we consider below is $n^{-\frac{2\beta}{2\beta+d}}$ up to constants depending only on β and d. Notice that the functions in $\widetilde{\Sigma}(\beta, L)$ used in the lower bound construction in [14] can be rescaled into functions in $\Sigma(\beta, L)$ by multiplying by a factor depending only on β and d. Hence, the lower bound construction in [14] remains valid for the class $\Sigma(\beta, L)$. It implies that the minimax rate of estimation on the class $\Sigma(\beta, L)$ is $n^{-\frac{2\beta}{2\beta+d}}$. In conclusion, though $\Sigma(\beta, L)$ is a subclass of suitable Hölder classes Σ' and $\widetilde{\Sigma}$ it is not substantially smaller, in the sense that estimation over these classes is essentially equally difficult.

5.3 Local polynomial estimators and interpolation

For $\ell \in \mathbf{N}$ let $C_{\ell,d} = \binom{\ell+d}{d}$ be the cardinality of the set of multi-indices $\{s = (s_1, \dots, s_d) \in \mathbb{N}^d, 0 \leq |s| \leq \ell\}$. We assume that the elements $s^{(1)}, \dots, s^{(C_{\ell,d})}$ of this set are ordered according to the increasing values of $|s|$, and in an arbitrary way for equal values of $|s|$. In particular, $s^{(1)} = (0, \dots, 0)$. For any $u \in \mathbf{R}^d$, define the vector $U(u) \in \mathbf{R}^{C_{\ell,d}}$ as follows:

$$U(u) := \left(\frac{u^s}{s!}\right)_{|s| \leq \ell},$$

where the components of $U(u)$ are ordered in the same way as $s^{(i)}$'s. In particular, the first component of $U(u)$ is 1 for any u.

The definition of local polynomial estimator usually given in the literature is as follows, cf., e.g., [188]. Let $K : \mathbf{R}^d \to \mathbf{R}_+$ be a kernel, $h > 0$ be a bandwith and $\ell \geq 0$ be an integer. Consider a vector $\hat{\theta}_n(x) \in \mathbf{R}^{C_{\ell,d}}$ such that

$$\hat{\theta}_n(x) \in \operatorname*{argmin}_{\theta \in \mathbf{R}^{C_{\ell,d}}} \sum_{i=1}^{n} \left[Y_i - \theta^\top U\left(\frac{X_i - x}{h}\right) \right]^2 K\left(\frac{X_i - x}{h}\right) \tag{5.2}$$

Then

$$f_n(x) = U^\top(0)\hat{\theta}_n(x) \tag{5.3}$$

is called a local polynomial estimator of order ℓ of $f(x)$. Note that $f_n(x)$ is the first component of $\hat{\theta}_n(x)$.

However, this definition is not convenient for our purposes. First, $\hat{\theta}_n(x)$ is not uniquely defined for such $x \in \mathbf{R}^d$ that the matrix

$$B_{nx} := \frac{1}{nh^d} \sum_{i=1}^{n} U\left(\frac{X_i - x}{h}\right) U^\top\left(\frac{X_i - x}{h}\right) K\left(\frac{X_i - x}{h}\right) \in \mathbf{R}^{C_{\ell,d} \times C_{\ell,d}}$$

is degenerate. Furthermore, $\hat{\theta}_n(x)$ is not defined for $x = X_i$ if the kernel K has a singularity at 0, which will be the main case of interest in what follows. Therefore, we adopt the following slightly different definition.

Definition 3 (Local polynomial estimator)**.** *If the kernel K is bounded then the local polynomial estimator of order ℓ (or shortly, LP(ℓ) estimator) of $f(x)$ at point x is defined as*

$$f_n(x) = \sum_{i=1}^{n} Y_i W_{ni}(x), \tag{5.4}$$

where, for $i = 1, \dots, n$, the weights $W_{ni}(x)$ are given by

$$W_{ni}(x) = \frac{U^\top(0)}{nh^d} B_{nx}^{+} U\left(\frac{X_i - x}{h}\right) K\left(\frac{X_i - x}{h}\right). \tag{5.5}$$

If the kernel K has a singularity at 0, that is, $\lim_{u\to 0} K(u) = +\infty$, then the LP($\ell$) estimator of $f(x)$ at point $x \notin \{X_1, \dots, X_n\}$ is still defined by (5.4) while we set, for $j = 1, \dots, n$,

$$f_n(X_j) = \limsup_{z \to X_j} f_n(z). \tag{5.6}$$

The purpose of (5.6) is to provide a valid definition for kernels with singularity at 0. We introduce lim sup in (5.6) for formal reasons. In the cases of our interest described in the next lemma there exists an exact limit in (5.6): $\lim_{x\to X_j} f_n(x) = Y_j$ for all $j \in [n]$, which means that the estimator f_n is interpolating.

Lemma 3. *[Interpolation property of LPE] Let f_n be an LP(ℓ) estimator with kernel $K : \mathbf{R}^d \to \mathbf{R}_+$ having a singularity at 0, that is, $\lim_{u\to 0} K(u) = +\infty$, and continuous on $\mathbf{R}^d \setminus \{0\}$. In particular, there exist $c_0 > 0$ and $\Delta > 0$ such that*

$$K(u) \geq c_0 \mathbf{1}(\|u\| \leq \Delta), \quad \forall u \in \mathbf{R}^d. \tag{5.7}$$

Assume that $X_1, \dots, X_n$ are distinct points in $\mathbf{R}^d$ and there exists a constant $\lambda_1 > 0$ such that

$$\sum_{j=1}^{n} U\left(\frac{X_j - x}{h}\right) U^\top\left(\frac{X_j - x}{h}\right) \mathbf{1}\left(\left\|\frac{X_j - x}{h}\right\| \leq \Delta\right) \succ \lambda_1 I_{C_{\ell,d}} \tag{5.8}$$

for all x in some neighborhood of X_i, where $I_{C_{\ell,d}}$ denotes the identity matrix. Then $f_n(X_i) = Y_i$.

For $\ell = 0$ (corresponding to the Nadaraya-Watson estimator) condition (5.8) is trivially satisfied since the expression on the left hand side is a positive scalar for any x in a neighborhood of X_i. For general ℓ, this condition is satisfied with high probability if X_j's are distributed with a density bounded away from zero on its support. Indeed, we have the following result. For $\Delta > 0$ consider the matrix

$$\overline{B}_{nx} := \frac{1}{nh^d} \sum_{i=1}^{n} U\left(\frac{X_i - x}{h}\right) U^\top\left(\frac{X_i - x}{h}\right) \mathbf{1}\left(\left\|\frac{X_i - x}{h}\right\| \leq \Delta\right) \in \mathbf{R}^{C_{\ell,d} \times C_{\ell,d}}.$$

Lemma 4. *Let $h \le \alpha$, where $\alpha > 0$. Let Assumption (A2) be satisfied. Then, the following holds.*
(i) For any $\Delta > 0$ there exist constants $\lambda_0(\ell) > 0$, $c > 0$ independent of n and x and depending only on $\ell, \alpha, \Delta, d, p(\cdot)$ such that

$$\mathbf{P}\Big(\inf_{x \in \mathrm{Supp}(p)} \lambda_{\min}(\overline{B}_{nx}) \ge \lambda_0(\ell) \Big) \ge 1 - c(h^{-d^2-d} e^{-nh^d/c} + e^{-n^3 h^{2d}/c}),$$

where $\lambda_{\min}(\overline{B}_{nx})$ is the minimal eigenvalue of $\overline{B}_{nx}$. Moreover, $\lambda_0(\ell) \ge \lambda_0(\ell')$ if $\ell \le \ell'$.
(ii) If K is a kernel satisfying (5.7) then there exist constants $\lambda_0'(\ell) > 0$, $c' > 0$ independent of n and x and depending only on $\ell, \alpha, \Delta, d, p(\cdot)$ such that

$$\mathbf{P}\Big(\inf_{x \in \mathrm{Supp}(p)} \lambda_{\min}(B_{nx}) \ge \lambda_0'(\ell) \Big) \ge 1 - c'(h^{-d^2-d} e^{-nh^d/c'} + e^{-n^3 h^{2d}/c'}).$$

Note that part (ii) of Lemma 4 is an immediate consequence of its part (i) and the fact that $B_{nx} \succ c_0 \overline{B}_{nx}$ if (5.7) holds. Also, the next corollary follows immediately from Lemmas 3 and 4.

Corollary 1. *Let f_n be an LP(ℓ) with kernel $K : \mathbf{R}^d \to \mathbf{R}_+$ having a singularity at 0, that is, $\lim_{u \to 0} K(u) = +\infty$, and continuous on $\mathbf{R}^d \setminus \{0\}$. Let $h = \alpha n^{-\frac{1}{2\beta+d}}$, where $\alpha, \beta > 0$ and let Assumption (A2) be satisfied. Then, there exists a constant $c' > 0$ such that, with probability at least $1 - c' e^{-A_n/c'}$, where $A_n = n^{\frac{2\beta}{2\beta+d}}$, the LPE f_n is interpolating, that is, $f_n(X_i) = Y_i$ for $i = 1, \dots, n$. Furthermore, the LP(0) estimator is interpolating with probability 1.*

Note that the kernels $K(u) = \|u\|^{-a} \mathbb{1}(\|u\| \le 1)$ with $a \in (0, d/2)$ considered in [99, 106] are not continuous on $\mathbf{R}^d \setminus \{0\}$ and thus do not satisfy the conditions of Lemma 3 and Corollary 1. On the other hand, these conditions are met, for example, for the kernels $K(u) = \|u\|^{-a} \cos^2(\pi \|u\|/2) \mathbb{1}(\|u\| \le 1)$ or $K(u) = \|u\|^{-a}(1 - \|u\|)_+$ with $a > 0$.

5.4 Minimax optimal interpolating estimator

In this section, we show that for any $\beta > 0$, one can construct an interpolating local polynomial estimator reaching the minimax rate $n^{-\frac{2\beta}{2\beta+d}}$ on the Hölder class $\Sigma(\beta, L)$.
In what follows, we assume that we know a constant L_0 such that $|f(x)| \le L_0$ for all $x \in \mathrm{Supp}(p)$. We denote the class of all such functions f by $\mathcal{F}_0$. This assumption is not crucial and can be avoided at the expense of a more cumbersome estimator construction (see Remark 1 below).
Let f_n be an LP(ℓ) estimator of order $\ell = \lfloor \beta \rfloor$. Set $\mu := L_0 \vee \max_{1 \le i \le n} |Y_i|$ and consider the truncated estimator

$$\bar{f}_n(x) = \Big[f_n(x) \Big]_{-\mu}^{\mu}, \tag{5.9}$$

where for all $y \in \mathbf{R}$ and $a \le b$ the truncation of y between a and b is defined as $[y]_a^b := (y \vee a) \wedge b$.

Theorem 1. *Let Assumptions (A1) and (A2) be satisfied. Let $f \in \Sigma(\beta, L)$ for $\beta > 0, L > 0$, and $|f(x)| \le L_0$ for all $x \in \mathrm{Supp}(p)$ and a constant $L_0 > 0$. Consider the estimator $\bar{f}_n$ defined in (5.9), where f_n is the LP(ℓ) estimator with $\ell = \lfloor \beta \rfloor$, $h = \alpha n^{-\frac{1}{2\beta+d}}$, for some $\alpha > 0$, and kernel K.*

(i) If K is a compactly supported kernel satisfying (5.7) and $\int K^2(u)du < \infty$ then

$$\mathbf{E}\left([\bar{f}_n(x) - f(x)]^2\right) \le Cn^{-\frac{2\beta}{2\beta+d}}, \quad \forall x \in \mathrm{Supp}(p), \tag{5.10}$$

$$\mathbf{E}\left(\|\bar{f}_n - f\|_{L_2}^2\right) \le Cn^{-\frac{2\beta}{2\beta+d}}, \tag{5.11}$$

where $C > 0$ is a constant independent of x and n.
(ii) If, in addition, $\lim_{u\to 0} K(u) = +\infty$ and K is continuous on $\mathbf{R}^d \setminus \{0\}$, then there exists a constant $c' > 0$ such that, with probability at least $1 - c'e^{-A_n/c'}$, where $A_n = n^{\frac{2\beta}{2\beta+d}}$, the estimator $\bar{f}_n$ is interpolating, that is, $\bar{f}_n(X_i) = Y_i$ for $i = 1, \dots, n$.

Note that, for the examples of singular kernels given at the end of the previous section, we need $a \in (0, d/2)$ to grant the condition $\int K^2(u)du < \infty$ required in Theorem 1. Moreover, Shepard kernel $K(u) = \|u\|^{-d}$ does not satisfy the assumptions of Theorem 1.

Remark 1. By modifying the estimator we can drop the assumption that $|f(x)| \le L_0$ for all $x \in \mathrm{Supp}(p)$. Indeed one can estimate the value $\max\limits_{x\in\mathrm{Supp}(p)} |f(x)|$ by $\max\limits_{x\in\mathrm{Supp}(p)} |\hat{f}(x)|$, where $\hat{f}$ is any estimator of f converging in sup-norm with some rate (for example, with the optimal rate $(\log(n)/n)^{\beta/(2\beta+d)}$). Given such an estimator, it is not hard to check that the argument in the proof goes through if we replace L_0 by the data-driven quantity $\hat{L}_0 = 2\max\limits_{x\in\mathrm{Supp}(p)} |\hat{f}(x)|$.

Remark 2. Theorem 1 completes the existing literature on LPE in classical setting with non-singular kernels. To the best of our knowledge, non-asymptotic bounds on the mean squared error of LPE are missing in the existing literature. The previous work was mainly focused on asymptotic properties such as convergence in probability or pointwise asymptotic normality, cf. [11, 14, 18, 178]. For binary $Y \in \{0, 1\}$ specific to classification setting, non-asymptotic deviation bounds for LPE were obtained in [45]. However, the techniques of [45] cannot be extended beyond the case of bounded Y.

Remark 3. The value $\max_{1\le i\le n} |Y_i|$ is introduced in the threshold μ only with the aim to preserve the interpolation property. Inspection of the proof shows that Theorem 1(i) remains valid when $\max_{1\le i\le n} |Y_i|$ is dropped from the definition of μ, so that $\mu = L_0$. Also, Theorem 1(i) remains valid if we truncate at $\hat{L}_0$ and not at L_0 as explained in Remark 1.

Remark 4. Inspection of the proof shows that Theorem 1 extends to kernels K that are not necessarily compactly supported. It suffices to assume that the integrals $\int(1 + \|u\|^\beta)K(u)du$ and $\int(1 + \|u\|^{2\beta})K^2(u)du$ are finite.

Remark 5. If kernel K is continuous on $\mathbf{R}^d$ the LPE f_n and $\bar{f}_n$ are continuous functions on $\mathrm{Supp}(p)$.

5.5 Adaptive interpolating estimator

In this section, we will use the following assumption on the noises $\xi(X_i)$.

Assumption (A3). Conditionally on $X = x$, the random variable $\xi(X)$ is a zero-mean σ_ξ-subgaussian random variable for all $x \in \mathrm{Supp}(p)$.

We propose an adaptive estimator that does not need the knowledge of β, L, achieves the minimax L_2 rate on convergence on classes $\Sigma(\beta, L)$ for all $L > 0$ and $\beta \in (0, \beta_{\max}]$, where $\beta_{\max} > 0$ is an arbitrary given value, and is interpolating with high probability. Our adaptive estimator is based on exponentially weighted aggregation, cf. [47]. Consider the sample $\mathcal{D} = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$. Assuming without loss of generality that n is even we split $\mathcal{D}$ into two independent subsamples $\mathcal{D}_1 = \Big\{(X_1, Y_1), \dots, (X_{\frac{n}{2}}, Y_{\frac{n}{2}})\Big\}$ and $\mathcal{D}_2 = \Big\{(X_{\frac{n}{2}+1}, Y_{\frac{n}{2}+1}), \dots, (X_n, Y_n)\Big\}$. We proceed in two steps.

1. Choose a finite grid $(\beta_j)_{j \in J}$ on the values of β. Let $f_{n,j}$ denote a $\mathrm{LP}(\ell_j)$ estimator (with $\ell_j = \lfloor \beta_j \rfloor$) based on the subsample $\mathcal{D}_1$ with bandwidth $h = \alpha n^{-\frac{1}{2\beta_j + d}}$, $\alpha > 0$, and kernel K satisfying the assumptions of Theorem 1. Construct $|J|$ truncated local polynomial estimators:

$$\bar{f}_{n,j}(x) = \Big[f_{n,j}(x)\Big]_{-\mu}^{\mu}, \qquad j \in J. \tag{5.12}$$

   By Theorem 1, each estimator $\bar{f}_{n,j}$ is interpolating over $\mathcal{D}_1$ with high probability, and satisfies

$$\sup_{f \in \Sigma(\beta_j, L) \cap \mathcal{F}_0} \mathbf{E}_1 \left[\|\bar{f}_{n,j} - f\|_{L_2}^2\right] \le C n^{-\frac{2\beta_j}{2\beta_j + d}}, \tag{5.13}$$

   where $\mathbf{E}_1$ denotes the expectation with respect to the distribution of $\mathcal{D}_1$.

2. Aggregate the estimators $(\bar{f}_{n,j})_{j \in J}$ using an exponentially weighted procedure and the second subsample $\mathcal{D}_2$. Set $k = \frac{n}{2}$. For $t = 1, \dots, k$, consider the vectors of weights $\theta_t^{EW} = (\theta_t^{EW}(j), j \in J)$ with components $\theta_1^{EW}(j) = 1/|J|, \forall j \in J$, and

$$\theta_t^{EW}(j) = \frac{\exp\left(-\eta \sum_{s=k+1}^{k+t-1} \left[\bar{f}_{n,j}(X_s) - Y_s\right]^2\right)}{\sum_{j=1}^{M} \exp\left(-\eta \sum_{s=k+1}^{k+t-1} \left[\bar{f}_{n,j}(X_s) - Y_s\right]^2\right)}, \quad j \in J,\ t \ge 2,$$

   where $\eta = \left[2\sigma_\xi^2 + 2\Big(L_0 + L_0 \vee \max_{i=1,\dots,k} |Y_i|\Big)^2\right]^{-1}$. For each t, we define an estimator, which is a convex combination of $(\bar{f}_{n,j})_{j \in J}$ weighted by θ_t^{EW}:

$$\hat{f}_t^{EW} = \sum_{j \in J} \theta_t^{EW}(j) \bar{f}_{n,j},$$

   and consider the averaged aggregate

$$\widetilde{f}_k = \frac{1}{k} \sum_{t=1}^{k} \hat{f}_t^{EW}. \tag{5.14}$$

As a convex combination of estimators $(\bar{f}_{n,j})_{j \in J}$ interpolating over $\mathcal{D}_1$, the estimator $\widetilde{f}_k$ is also interpolating over $\mathcal{D}_1$, but not over $\mathcal{D}_2$. We therefore introduce the estimator $\widetilde{g}_k$ obtained in the same way as $\widetilde{f}_k$ by interchanging $\mathcal{D}_1$ and $\mathcal{D}_2$. Thus, $\widetilde{g}_k$ is interpolating over $\mathcal{D}_2$. Next, we define an estimator interpolating over $\mathcal{D}_1 \cup \mathcal{D}_2$ by combining $\widetilde{f}_k$ and $\widetilde{g}_k$ as follows.

For any $x \in \mathbf{R}^d$ and any set $A \subseteq \mathbf{R}^d$, denote by $d(x, A) = \inf_{y \in A} \|x - y\|$ the distance between x and A. Let $\lambda : \mathbf{R}^d \to [0, 1]$ be any continuous function such that $\lambda(x) \to 0$ as $d(x, \mathcal{D}_2) \to 0$ and $\lambda(x) \to 1$ as $d(x, \mathcal{D}_1) \to 0$. For example, take $\lambda(x) = \frac{2}{\pi} \arctan\left(\frac{d(x,\mathcal{D}_2)}{d(x,\mathcal{D}_1)}\right)$ with $\frac{1}{0} = \infty$ and $\arctan(\infty) = 1$ by convention. We define our final estimator as

$$\hat{\mathsf{f}}_n(x) = \lambda(x)\widetilde{f}_k(x) + (1 - \lambda(x))\widetilde{g}_k(x). \tag{5.15}$$

Theorem 2. *Let $n \geq 3$, $\beta_{\max} > 1$. Consider the grid points β_j defined as follows:*

$$\beta_j = \left(1 + \frac{1}{\log n}\right)^j, \quad j = -M, \dots, M_{\max},$$

where $M = 2\lfloor \log(n) \log\log(n) \rfloor$ and $M_{\max} = M \wedge \lfloor \log(n) \log(\beta_{\max}) \rfloor$. Let Assumptions (A1) and (A3) be satisfied. If kernel K satisfies the assumptions of Theorem 1(i), then for any $\beta \in (0, \beta_{\max}]$ and $L > 0$ for the estimator $\hat{\mathsf{f}}_n$ defined by (5.15) we have

$$\sup_{f \in \Sigma(\beta, L) \cap \mathcal{F}_0} \mathbf{E}\left[\|\hat{\mathsf{f}}_n - f\|_{L_2}^2\right] \leq C n^{-\frac{2\beta}{2\beta+d}}, \tag{5.16}$$

where $C > 0$ is a positive constant depending only on $\beta, L, L_0, d, \beta_{\max}, \sigma_\xi, K, p_{\max}, p_{\min}$ and α.
If, in addition, kernel K satisfies the assumptions of Theorem 1(ii), then the estimator $\hat{\mathsf{f}}_n$ is interpolating with probability at least $1 - c'' \exp(-n^{\frac{2}{2+d}}/c'')$, where c'' is a positive constant depending only on $L, L_0, d, \beta_{\max}, K, p_{\max}, p_{\min}$ and α.

Note that one can get rid of the dependence of the estimators on L_0 and drop the intersection with $\mathcal{F}_0$ in (5.16). It can be achieved by using, for each estimator $\bar{f}_{n,j}$, a data-driven threshold instead of L_0 as explained in Remark 1.
If kernel K is continuous on $\mathbf{R}^d$ the adaptive estimator $\hat{\mathsf{f}}_n$ is a continuous function on $\mathrm{Supp}(p)$.

Appendix

Proof of Lemma 1. The result is straightforward if there exists an integer $\ell \geq 0$ such that $\ell < \beta' \leq \beta \leq \ell + 1$. Indeed, for any integer $\ell \geq 0$,

$$\ell < \beta' \leq \beta \leq \ell + 1 \quad \Longrightarrow \quad \Sigma(\beta, L) \subseteq \Sigma(\beta', L). \tag{17}$$

Thus, it remains to consider the case $\ell < \beta' \leq \ell + 1 < \beta$ for an integer ℓ. Handling this case will be based on the following embedding:

$$\Sigma(\beta, L) \subseteq \Sigma(\ell', 2L), \quad \forall \ell' \in \mathbf{N} \text{ such that } \ell' < \beta. \tag{18}$$

We now prove (18). Indeed, let $f \in \Sigma(\beta, L)$ and let ℓ' be an integer less than β. Then, in particular, $\max\limits_{0 \leq s \leq \ell'} \sup\limits_{x \in \mathcal{B}_d} \|f^{(s)}(x)\| \leq L$. Consider $x, y \in \mathcal{B}_d$ and $h = y - x$. Denote by h_i the ith component of h and by e_i the ith canonical basis vector in $\mathbf{R}^d$. Set $k = \ell' - 1$. Then for any multi-indices $m_1, \dots, m_k \in \mathbf{N}^d$ we have

$$\begin{aligned} D^{m_1+\dots+m_k} f(y) - D^{m_1+\dots+m_k} f(x) &= \int_0^1 \left\langle \nabla D^{m_1+\dots+m_k} f(x+th), h \right\rangle dt \\ &= \int_0^1 \sum_{i=1}^d D^{m_1+\dots+m_k+e_i} f(x+th) h_i dt \\ &= \sum_{i=1}^d \int_0^1 D^{m_1+\dots+m_k+e_i} f(x+th) dt \, h^{e_i}. \end{aligned}$$

Writing for brevity $G_{m_1,\dots,m_k,e_i}(x,h) = \int_0^1 D^{m_1+\dots+m_k+e_i} f(x+th) dt$ we obtain

$$\begin{aligned} \|f^{(k)}(y) - f^{(k)}(x)\| &= \sup_{\substack{\|u_j\| \leq 1, \\ j \in [k]}} \left| \sum_{|m_j|=1, \forall j \in [k]} \sum_{i=1}^d G_{m_1,\dots,m_k,e_i}(x,h) \, h^{e_i} u_1^{m_1} \dots u_k^{m_k} \right| \\ &= \|h\| \sup_{\substack{\|u_j\| \leq 1, \\ j \in [k]}} \left| \sum_{|m_j|=1, \forall j \in [k]} \sum_{i=1}^d G_{m_1,\dots,m_k,e_i}(x,h) \left(\frac{h}{\|h\|} \right)^{e_i} u_1^{m_1} \dots u_k^{m_k} \right| \\ &\leq \|h\| \sup_{\substack{\|u_j\| \leq 1, \\ j \in [k+1]}} \left| \sum_{|m_j|=1, \forall j \in [k+1]} \int_0^1 D^{m_1+\dots+m_{k+1}} f(x+th) dt \, u_1^{m_1} \dots u_{k+1}^{m_{k+1}} \right| \\ &\leq \|h\| \int_0^1 \sup_{\substack{\|u_j\| \leq 1, \\ j \in [k+1]}} \left| f^{(k+1)}(x+th)[u_1, \dots, u_{k+1}] \right| dt \\ &\leq \|h\| \sup_{z \in \mathcal{B}_d} \|f^{(k+1)}(z)\|_* \leq L \|x - y\|, \end{aligned}$$

which, together with bound $\max\limits_{0 \leq s \leq \ell' - 1} \sup\limits_{x \in \mathcal{B}_d} \|f^{(s)}(x)\| \leq L$ implies that $f \in \Sigma(\ell', 2L)$. Thus, we have proved (18).

It follows from (18) that if $\ell < \beta' \le \ell + 1 < \beta$ for an integer ℓ then $\Sigma(\beta, L) \subseteq \Sigma(\ell + 1, 2L)$, while taking $\beta = \ell + 1$ in (17) implies that $\Sigma(\ell + 1, 2L) \subseteq \Sigma(\beta', 2L)$. This proves the lemma when $\ell < \beta' \le \ell + 1 < \beta$ for an integer ℓ. □

Proof of Lemma 2. The result is clear for $\beta \le 1$. Assume that $\beta > 1$ and fix some $x, y \in \mathcal{B}_d$. By Taylor expansion, there exists $c \in (0, 1)$ such that

$$f(x) = \sum_{0 \le |k| \le \ell - 1} \frac{1}{k!} D^k f(y)(x-y)^k + \sum_{|k| = \ell} \frac{1}{k!} D^k f(y + c(x-y))(x-y)^k,$$

and

$$\left| f(x) - \sum_{|k| \le \ell} \frac{1}{k!} D^k f(y)(x-y)^k \right| = \left| \sum_{|k| = \ell} \frac{1}{k!} \Big[D^k f\Big(y + c(x-y)\Big) - D^k f(y) \Big](x-y)^k \right|.$$

By a standard combinatorial argument, it is not hard to check that, for any $h, z \in \mathbf{R}^d$,

$$f^{(k)}(z)[h]^k := \sum_{|m_1| = \ldots = |m_\ell| = 1} D^{m_1 + \ldots + m_\ell} f(z) h^{m_1 + \ldots + m_\ell} = \sum_{|k| = \ell} \frac{\ell!}{k!} D^k f(z) h^k \ .$$

It follows that

$$\begin{aligned} &\left| \sum_{|k| = \ell} \frac{1}{k!} \Big[D^k f\Big(y + c(x-y)\Big) - D^k f(y) \Big](x-y)^k \right| \\ &\quad = \frac{1}{\ell!} \left| f^{(\ell)}\Big(y + c(x-y)\Big)[x-y]^\ell - f^{(\ell)}(y)[x-y]^\ell \right| \\ &\quad \le \frac{1}{\ell!} \left\| f^{(\ell)}\Big(y + c(x-y)\Big) - f^{(\ell)}(y) \right\|_* \|x-y\|^\ell \\ &\quad \le \frac{L}{\ell!} \|x-y\|^\ell \|c(x-y)\|^{\beta - l} \le \frac{L}{\ell!} \|x-y\|^\beta. \end{aligned} \tag{19}$$

□

Proof of Lemma 3. In this proof, we fix $i \in [n]$, and our aim is to prove that $\lim_{x \to X_i} f_n(x) = Y_i$. Let $\mathcal{V}$ be the neighborhood of X_i where (5.8) holds. Since $X_1, \ldots, X_n$ are distinct, we assume w.l.o.g. that $\mathcal{V}$ does not contain $(X_j)_{j \ne i}$. Due to conditions (5.7) and (5.8), we have that $B_{nx} \succ 0$ for all x in $\mathcal{V}_- := \mathcal{V} \setminus \{X_i\}$. Thus, for all $x \in \mathcal{V}_-$ the vector $\hat{\theta}_n(x)$ is the unique solution of (5.2), and $f_n(x)$ is given by (5.3):

$$\begin{aligned} \hat{\theta}_n(x) &= \operatorname*{argmin}_{\theta \in \mathbf{R}^{C_{\ell,d}}} \sum_{i=1}^n \left[Y_i - \theta^\top U\left(\frac{X_i - x}{h} \right) \right]^2 K\left(\frac{X_i - x}{h} \right), \\ f_n(x) &= U^\top(0) \hat{\theta}_n(x). \end{aligned}$$

Define $g_i(x) = \left(Y_i - \hat{\theta}_n(x)^\top U\left(\frac{X_i - x}{h}\right)\right)^2$. First, we prove by contradiction that $\lim\limits_{x \to X_i} g_i(x) = 0$ for any $i \in [n]$. Indeed, suppose that $\lim_{x \to X_i} g_i(x) \neq 0$. Then, there is a sequence $(x_k)_k$ in $\mathbf{R}^d$ converging to X_i as $k \to \infty$ such that $\lim_{k\to\infty} g_i(x_k) = +\infty$ or $\lim_{k\to\infty} g_i(x_k) = \text{const} > 0$. In both cases,

$$\lim_{k\to\infty} \sum_{j=1}^n g_j(x_k) K\left(\frac{X_j - x_k}{h}\right) = +\infty \tag{20}$$

since the kernel K has a singularity at 0. On the other hand, the definition of $\hat{\theta}_n(x_k)$ implies that, for any k and any $\theta_* \in \mathbf{R}^{C_{\ell,d}}$,

$$\sum_{j=1}^n g_j(x_k) K\left(\frac{X_j - x_k}{h}\right) \le \sum_{j=1}^n \left(Y_j - \theta_*^\top U\left(\frac{X_j - x_k}{h}\right)\right)^2 K\left(\frac{X_j - x_k}{h}\right).$$

In particular, for $\theta_*^\top = (Y_i \quad 0 \dots 0)$ we have

$$\begin{aligned}
\sum_{j=1}^n \left(Y_j - \theta_*^\top U\left(\frac{X_j - x_k}{h}\right)\right)^2 K\left(\frac{X_j - x_k}{h}\right) &= \sum_{j=1}^n (Y_j - Y_i)^2 K\left(\frac{X_j - x_k}{h}\right) \\
&= \sum_{j\neq i} (Y_j - Y_i)^2 K\left(\frac{X_j - x_k}{h}\right) \\
&\underset{k\to+\infty}{\to} \sum_{j\neq i} (Y_j - Y_i)^2 K\left(\frac{X_j - X_i}{h}\right) < +\infty,
\end{aligned}$$

which is in contradiction with (20). Therefore, for any $i \in [n]$ we have $\lim\limits_{x \to X_i} g_i(x) = 0$.
A similar argument yields that $\limsup\limits_{x \to X_i} g_j(x) < +\infty$ for any $j \neq i$. Indeed, if for some $j \neq i$ this relation does not hold then there is a sequence $(x_k)_k$ in $\mathbf{R}^d$ converging to X_i as $k \to \infty$ such that $\lim_{k\to\infty} g_j(x_k) = +\infty$. It implies (20), which is not possible as shown above.
Next, we prove that $\|\hat{\theta}_n(x)\|$ is bounded for all x in a neighborhood of X_i. Since $\lim\limits_{x \to X_i} g_i(x) = 0$, and for any $j \neq i$ we have $\limsup\limits_{x \to X_i} g_j(x) < +\infty$ the values $g_j(x)$ are bounded for all $j \in [n]$ and all x in a neighborhood of X_i. We will further denote this neighborhood by $\mathcal{V}'$. It follows that $\varphi_j(x) = \hat{\theta}_n(x)^\top U\left(\frac{X_j - x}{h}\right)$, $j = 1, \dots, n$, are bounded for $x \in \mathcal{V}'$ and thus the sum $\sum_{j=1}^n \varphi_j^2(x)$ is bounded as well. On the other hand, by assumption (5.8), for all $x \in \mathcal{V}_-$,

$$\begin{aligned}
\sum_{j=1}^n \varphi_j^2(x) &\ge \sum_{j=1}^n \hat{\theta}_n(x)^\top U\left(\frac{X_j - x}{h}\right) U^\top\left(\frac{X_j - x}{h}\right) \mathbf{1}\left(\left\|\frac{X_j - x}{h}\right\| \le \Delta\right) \hat{\theta}_n(x) \\
&\ge \lambda_1 \|\hat{\theta}_n(x)\|^2,
\end{aligned}$$

where $\lambda_1 > 0$. It follows that $\|\hat{\theta}_n(x)\|$ is bounded for all $x \in \mathcal{V}' \cap \mathcal{V}_-$.

Let $\hat{\theta}_{n,(1)}(x) = f_n(x)$ denote the first component of $\hat{\theta}_n(x)$ and $\hat{\theta}_{n,(2)}(x)$ the vector of its remaining $C_{\ell,d} - 1$ components, so that $\hat{\theta}_n(x)^\top = \left(\hat{\theta}_{n,(1)}(x), \hat{\theta}_{n,(2)}(x)^\top\right)$. Recall that the first component of $U(u)$ is equal to 1 for all $u \in \mathbf{R}^d$. Denote by $U_{(2)}(u)$ the vector of its remaining $C_{\ell,d} - 1$ components, so that $U(u)^\top = \left(1, U_{(2)}(u)^\top\right)$. With this notation, the relation $\lim_{x \to X_i} g_i(x) = 0$ proved above can be written as:

$$g_i(x) = \left(Y_i - \hat{\theta}_{n,(1)}(x) - \hat{\theta}_{n,(2)}(x)^\top U_{(2)}\left(\frac{X_i - x}{h}\right)\right)^2 \underset{x \to X_i}{\to} 0.$$

Since $\|\hat{\theta}_n(x)\|$ is bounded for $x \in \mathcal{V}' \cap \mathcal{V}_-$ we get that $|\hat{\theta}_{n,(1)}(x)|$ and $\|\hat{\theta}_{n,(2)}(x)\|$ are also bounded for $x \in \mathcal{V}' \cap \mathcal{V}_-$. The definition of $U(u)$ implies the convergence $\lim_{x \to X_i} \|U_{(2)}\left(\frac{X_i - x}{h}\right)\| = 0$. It follows that

$$\hat{\theta}_{n,(2)}(x)^\top U_{(2)}\left(\frac{X_i - x}{h}\right) \underset{x \to X_i}{\to} 0$$

and therefore

$$\hat{\theta}_{n,(1)}(x) \underset{x \to X_i}{\to} Y_i,$$

which concludes the proof since $\hat{\theta}_{n,(1)}(x) = f_n(x)$. □

Proof of Lemma 4. We prove only part (i) of the lemma since part (ii) is its immediate consequence. We have

$$\overline{B}_{nx} = \frac{1}{nh^d} \sum_{i=1}^n U\left(\frac{X_i - x}{h}\right) U^\top\left(\frac{X_i - x}{h}\right) \mathbf{1}\left(\frac{\|X_i - x\|}{\Delta} \le h\right)$$

and, for any $\lambda_0 > 0$,

$$\begin{aligned}
&\mathbf{P}\left(\inf_{x \in \mathrm{Supp}(p)} \lambda_{\min}(\overline{B}_{nx}) < \lambda_0\right) = \mathbf{P}\left(\inf_{x \in \mathrm{Supp}(p)} \inf_{\|v\|=1} v^\top \overline{B}_{nx} v < \lambda_0\right) \\
&\le \mathbf{P}\left(\inf_{x \in \mathrm{Supp}(p)} \inf_{\|v\|=1} v^\top \overline{B}(x) v - \sup_{x \in \mathrm{Supp}(p)} \|\overline{B}_{nx} - \overline{B}(x)\|_\infty < \lambda_0\right) \qquad (21)
\end{aligned}$$

where $\overline{B}(x) := \mathbf{E}(\overline{B}_{nx})$. Set $S(x, h, \Delta) = \{u \in \mathcal{B}_d(0, \Delta) : x + uh \in \mathrm{Supp}(p)\}$. Then we have

$$\begin{aligned}
v^\top \overline{B}(x) v &= \frac{1}{h^d} \int \left[v^\top U\left(\frac{z - x}{h}\right)\right]^2 \mathbf{1}\left(\left\|\frac{z - x}{h}\right\| \le \Delta\right) p(z) \mathrm{d}z \\
&\ge p_{\min} v^\top \left[\int_{S(x,h,\Delta)} U(u) U(u)^\top du\right] v
\end{aligned}$$

$$\geq p_{\min} v^\top \left[\int_{S(x,\alpha,\Delta)} U(u)U(u)^\top du \right] v,$$

where for the last inequality we used the fact that $S(x,\alpha,\Delta) \subset S(x,h,\Delta)$ since $h \leq \alpha$ and $\mathrm{Supp}(p)$ is a convex set. Notice that $S(x,\alpha,\Delta)$ is also a convex set and it is not reduced to one point x as $\mathrm{Supp}(p)$ is a convex set with positive Lebesgue measure. Thus, $S(x,\alpha,\Delta)$ is of infinite cardinality for any $x \in \mathrm{Supp}(p)$.
Denote by $S_d(0,1)$ the unit sphere in $\mathbf{R}^d$ centered at 0. Note that, for fixed Δ and α, the function

$$\begin{cases} \mathrm{Supp}\ p \times S_d(0,1) & \longrightarrow & \mathbf{R} \\ (x,v) & \mapsto & v^\top \left[\int_{S(x,\alpha,\Delta)} U(u)U(u)^\top du \right] v \end{cases}$$

is continuous and defined on a compact set. Therefore, it attains its minimum at some (x_0, v_0), where $x_0 \in \mathrm{Supp}(p)$ and $\|v_0\| = 1$. We argue now that the value of this minimum is positive. Indeed, it is clearly non-negative, and if it were 0 we would have:

$$0 = v_0^\top U(u) = \sum_{|k| \leq \ell} v_0(k) \frac{u^k}{k!}, \quad \forall u \in S(x_0,\alpha,\Delta). \tag{22}$$

As observed above, $S(x_0,\alpha,\Delta)$ is a set of infinite cardinality. On the other hand, the expression in (22) is a polynomial in u, so that for $v_0 \neq 0$ it can vanish only in a finite number of points. Thus, (22) is impossible. It follows that

$$\lambda_1(\ell) := \min_{v \in S_d(0,1), x \in \mathrm{Supp}(p)} v^\top \left[\int_{S(x,\alpha,\Delta)} U(u)U(u)^\top du \right] v > 0.$$

Next, note that the vector $U(u) = U_\ell(u)$ depends on ℓ, and that for $\ell \leq \ell'$ and any fixed x, the matrix $\int_{S(x,\alpha,\Delta)} U_\ell(u)U_\ell(u)^\top du$ is an extraction of the matrix $\int_{S(x,\alpha,\Delta)} U_{\ell'}(u)U_{\ell'}(u)^\top du$. Hence, the smallest eigenvalue of the former matrix is necessarily not less than that of the latter. Thus, $\lambda_1(\ell) \geq \lambda_1(\ell')$ for $\ell \leq \ell'$.
Setting $\lambda_0 = \lambda_0(\ell) := p_{\min}\lambda_1(\ell)/2$ and using (21) we find:

$$\mathbf{P}\left(\inf_{x \in \mathrm{Supp}(p)} \lambda_{\min}(\overline{B}_{nx}) < \lambda_0 \right) \leq \mathbf{P}\left(\sup_{x \in \mathrm{Supp}(p)} \|\overline{B}_{nx} - \overline{B}(x)\|_\infty > \lambda_0 \right). \tag{23}$$

It remains now to bound the probability on the right hand side of (23).
By Assumption (A2), the convex compact set $\mathrm{Supp}(p)$ is included in $\mathcal{B}_d = \mathcal{B}_d(0,1)$. For $\varepsilon > 0$, let $\{x_1, \ldots, x_N\} \subset \mathcal{B}_d^N$ be the minimal ε-net on $\mathcal{B}_d$ in the Euclidean metric. Then we have:

$$\begin{aligned} \sup_{x \in \mathrm{Supp}(p)} \|\overline{B}(x) - \overline{B}_{nx}\|_\infty \leq & \sup_{x \in \mathcal{B}_d} \min_{1 \leq k \leq N} \|\overline{B}(x) - \overline{B}(x_k)\|_\infty \\ & + \max_{1 \leq k \leq N} \|\overline{B}(x_k) - \overline{B}_{nx_k}\|_\infty + \sup_{\substack{x,x' \in \mathcal{B}_d, \\ \|x-x'\| \leq \varepsilon}} \|\overline{B}_{nx} - \overline{B}_{nx'}\|_\infty. \end{aligned}$$

Thus,

$$\mathbf{P}\left(\sup_{x\in\mathrm{Supp}(p)}\|\overline{B}_{nx}-\overline{B}(x)\|_\infty>\lambda_0\right)\le P_1+P_2+P_3,\quad\text{where}\tag{24}$$

$$P_1=\mathbf{P}\left(\sup_{x\in\mathcal{B}_d}\min_{1\le k\le N}\|\overline{B}(x)-\overline{B}(x_k)\|_\infty>\frac{\lambda_0}{3}\right),$$

$$P_2=\mathbf{P}\left(\max_{1\le k\le N}\|\overline{B}(x_k)-\overline{B}_{nx_k}\|_\infty>\frac{\lambda_0}{3}\right),$$

$$P_3=\mathbf{P}\left(\sup_{\substack{x,x'\in\mathcal{B}_d:\\ \|x-x'\|\le\varepsilon}}\|\overline{B}_{nx}-\overline{B}_{nx'}\|_\infty>\frac{\lambda_0}{3}\right).$$

In the rest of the proof, we control the terms P_1, P_2, P_3.
Control of P_2. Since all norms in the space of $C_{\ell,d}\times C_{\ell,d}$ matrices are equivalent there exists a constant $c_1>0$ depending only on ℓ, d such that, for all $k\in\{1,\dots,N\}$,

$$\|\overline{B}(x_k)-\overline{B}_{nx_k}\|_\infty\le c_1\max_{1\le i,j\le C_{\ell,d}}|b_{nx_k}(i,j)-b_{x_k}(i,j)|$$

where $b_{nx_k}(i,j)$ and $b_{x_k}(i,j)$ are the elements of $\overline{B}_{nx_k}$ and $\overline{B}(x_k)$, respectively. Then, for any $k\in\{1,\dots,N\}$,

$$\mathbf{P}\left(\|\overline{B}(x_k)-\overline{B}_{nx_k}\|_\infty>\frac{\lambda_0}{3}\right)\le C_{\ell,d}^2\max_{1\le i,j\le C_{\ell,d}}\mathbf{P}\left(|b_{nx_k}(i,j)-b_{x_k}(i,j)|>\frac{\lambda_0}{3c_1}\right).$$

We recall that $b_{x_k}(i,j)=\mathbf{E}\left[b_{nx_k}(i,j)\right]$. Setting $s=s^{(i)}$ and $r=s^{(j)}$ we have

$$b_{nx_k}(i,j)=\frac{1}{nh^d}\sum_{m=1}^n\frac{(X_m-x_k)^s}{h^s s!}\frac{(X_m-x_k)^r}{h^r r!}\mathbf{1}\left(\left\|\frac{X_m-x_k}{h}\right\|\le\Delta\right).$$

This is a sum of n i.i.d. random variables, each of which is bounded in absolute value by $\frac{C}{nh^d}$ and has variance not exceeding $\frac{C}{n^2h^d}$, where $C>0$ is a constant depending only on ℓ, d, Δ. By Bernstein's inequality,

$$\mathbf{P}\left(|b_{nx_k}(i,j)-b_{x_k}(i,j)|>\frac{\lambda_0}{3c_1}\right)\le 2\exp(-c_2nh^d),$$

where $c_2>0$ only depends on ℓ, d, Δ and not on n, k, i, j. It follows from the above inequalities and the union bound that

$$P_2\le 2NC_{\ell,d}^2\exp(-c_2nh^d).\tag{25}$$

Control of P_3. For any $x, x' \in \mathcal{B}_d$,

$$\overline{B}_{nx} - \overline{B}_{nx'} = \frac{1}{nh^d}\sum_{i=1}^{n}\Big[U\left(\frac{X_i - x}{h}\right)U^\top\left(\frac{X_i - x}{h}\right)\mathbf{1}\left(\left\|\frac{X_i - x}{h}\right\| \le \Delta\right) - U\left(\frac{X_i - x'}{h}\right)U^\top\left(\frac{X_i - x'}{h}\right)\mathbf{1}\left(\left\|\frac{X_i - x'}{h}\right\| \le \Delta\right)\Big].$$

For any $u \in \mathbf{R}^d$ consider the matrix

$$V(u) = U(u)U^\top(u)\mathbf{1}\{\|u\| \le \Delta\}. \tag{26}$$

Notice that $U(u) \in \mathbf{R}^{C_{\ell,d}}$ is Lipschitz continuous in u on the ball $\mathcal{B}_d(0,\Delta)$ since the components of vector $U(u)$ are polynomials in u. Thus, there exists a constant $\widetilde{L} > 0$ depending only on ℓ and d such that for any $u, u' \in \mathbf{R}^d$, if either $\|u\| \le \Delta, \|u'\| \le \Delta$ or $\|u\| > \Delta, \|u'\| > \Delta$, then

$$\Big\|V(u) - V(u')\Big\|_\infty \le \widetilde{L}\|u - u'\|,$$

and if (u, u') belongs to the set

$$\widetilde{\Delta} := \{(u,u') : \|u\| \le \Delta, \|u'\| > \Delta\} \cup \{(u,u') : \|u\| > \Delta, \|u'\| \le \Delta\}$$

then

$$\Big\|V(u) - V(u')\Big\|_\infty \le \widetilde{L},$$

taking $\widetilde{L} \ge \max\limits_{\|u\|\le\Delta}\Big\|U(u)U(u)^\top\Big\|_\infty$. It follows that

$$\Big\|V(u) - V(u')\Big\|_\infty \le \widetilde{L}\Big\{\|u - u'\| + \mathbf{1}((u,u') \in \widetilde{\Delta})\Big\}, \tag{27}$$

which implies the bound

$$\|\overline{B}_{nx} - \overline{B}_{nx'}\|_\infty \le \frac{\widetilde{L}}{h^{d+1}}\|x - x'\| + \frac{\widetilde{L}}{nh^d}\mathrm{Card}\Big\{i \in [n] : X_i \in \widetilde{\Delta}(x, x', h\Delta)\Big\},$$

where we denote by $\widetilde{\Delta}(x, x', h\Delta)$ the symmetric difference $\mathcal{B}_d(x, h\Delta) \triangle \mathcal{B}_d(x', h\Delta)$. Thus,

$$\sup_{\substack{x,x' \in \mathcal{B}_d: \\ \|x - x'\| \le \varepsilon}} \|\overline{B}_{nx} - \overline{B}_{nx'}\|_\infty \le \frac{\widetilde{L}\varepsilon}{h^{d+1}} + \frac{\widetilde{L}}{nh^d}\sup_{\substack{x,x' \in \mathcal{B}_d: \\ \|x - x'\| \le \varepsilon}}\sum_{i=1}^{n}\mathbf{1}\Big(X_i \in \widetilde{\Delta}(x, x', h\Delta)\Big), \tag{28}$$

If $\|x - x'\| \le \varepsilon$ then

$$\widetilde{\Delta}(x, x', h\Delta) \subseteq \{z : h\Delta < \|z - x\| \le h\Delta + \varepsilon\} \cup \{z : h\Delta < \|z - x'\| \le h\Delta + \varepsilon\}.$$

Therefore, for $\|x-x'\| \le \varepsilon$ we have $\left|\widetilde{\Delta}(x,x',h\Delta)\right| \le C_* h^{d-1}\varepsilon$, where we denote by $|S|$ the Lebesgue measure of a measurable set $S \subset \mathbf{R}^d$, and $C_* > 0$ is a constant depending only on Δ and d. Set $\varepsilon = c_0 h^{d+1}$, where the constant c_0 satisfies $0 < c_0 \le \dfrac{\lambda_0}{6\widetilde{L}}$. Then for $\|x-x'\| \le \varepsilon$ we get $\mathbf{P}(X_1 \in \Delta(x,x',h\Delta)) \le p_{\max} C_* c_0 h^{2d}$. Choose c_0 small enough (and depending only on $\ell, d, p_{\min}p_{\max}, \Delta$) to satisfy $p_{\max} C_* c_0 \alpha^d \le \dfrac{\lambda_0}{12\widetilde{L}}$. Consider the random event

$$\mathcal{A} = \Big\{ \sup_{\substack{x,x' \in \mathcal{B}_d: \\ \|x-x'\| \le \varepsilon}} \sum_{i=1}^n \mathbf{1}\Big(X_i \in \widetilde{\Delta}(x,x',h\Delta)\Big) \le A \Big\},$$

where $A = \dfrac{\lambda_0}{6\widetilde{L}} n h^d$. Due to the choice of c_0 and the fact that $h \le \alpha$ the bound $\mathbf{P}(X_1 \in \Delta(x,x',h\Delta)) \le A/2$ holds whenever $\|x-x'\| \le \varepsilon$. Hence,

$$\mathbf{P}\Big(\overline{\mathcal{A}}\Big) \le \mathbf{P}\Big\{ \sup_{\substack{x,x' \in \mathcal{B}_d: \\ \|x-x'\| \le \varepsilon}} \Big| \frac{1}{n} \sum_{i=1}^n \mathbf{1}\Big(X_i \in \widetilde{\Delta}(x,x',h\Delta)\Big) - \mathbf{P}\Big(X_1 \in \Delta(x,x',h\Delta)\Big)\Big| \ge A/2 \Big\}. \tag{29}$$

The class of all balls in $\mathbf{R}^d$ has a VC-dimension at most $d+2$, cf. Corollary 13.2 in [177]. Consequently, the class of all intersections of two balls in $\mathbf{R}^d$ has a VC-dimension at most Cd where $C > 0$ is an absolute constant [50]. This allows us to apply the Vapnik-Chervonenkis inequality to bound the probability in (29). Indeed, we can use the decomposition

$$\begin{aligned} \mathbf{1}\Big(X_i \in \widetilde{\Delta}(x,x',h\Delta)\Big) &= \mathbf{1}\Big(X_i \in \mathcal{B}_d(x,h\Delta)\Big) + \mathbf{1}\Big(X_i \in \mathcal{B}_d(x',h\Delta)\Big) \\ &\quad - 2 \cdot \mathbf{1}\Big(X_i \in \mathcal{B}_d(x,h\Delta) \cap \mathcal{B}_d(x',h\Delta)\Big) \end{aligned} \tag{30}$$

and bound from above the probability in (29) by the three probabilities corresponding to the three terms on the right hand side of (30). Applying the Vapnik-Chervonenkis inequality [177, Theorem 12.5] to each of these probabilities we get

$$\mathbf{P}\Big(\overline{\mathcal{A}}\Big) \le c_3 n^{c_3} \exp(-nA^2/128) \le c_3 n^{c_3} \exp(-c_4 n^3 h^{2d}),$$

where $c_3 > 0, c_4 > 0$ are constants depending only on $d, \ell, p(\cdot), \Delta$. On the other hand, due to (28) and the definitions of ε and A, on the event $\mathcal{A}$ we have

$$\sup_{\substack{x,x' \in \mathcal{B}_d: \\ \|x-x'\| \le \varepsilon}} \|\overline{B}_{nx} - \overline{B}_{nx'}\|_\infty \le \frac{\lambda_0}{3}.$$

Thus, we have proved that

$$P_3 \le c_3 n^{c_3} \exp(-c_4 n^3 h^{2d}). \tag{31}$$

Control of P_1. Fix $x \in \mathcal{B}_d$ and let $k \in \{1, \dots, N\}$ be such that $\|x - x_k\| \le \varepsilon$. Using (27) we obtain

$$\begin{aligned}
\|\overline{B}(x) - \overline{B}(x_k)\|_\infty &\le \frac{1}{h^d} \int_{\mathbf{R}^d} \left\| V\left(\frac{z-x}{h}\right) - V\left(\frac{z-x_k}{h}\right) \right\|_\infty p(z)dz \\
&\le \frac{\widetilde{L}}{h^d} \int_{\mathbf{R}^d} \left[\frac{\epsilon}{h} + \mathbf{1}(z \in \widetilde{\Delta}\big(x, x_k, h\Delta\big)\Big)\right] p(z)dz \\
&\le \widetilde{L}\varepsilon \left(\frac{1}{h^{d+1}} + \frac{C_* p_{\max}}{h}\right) \quad (\text{since } \left|\widetilde{\Delta}(x, x_k, h\Delta)\right| \le C_* h^{d-1}\varepsilon) \\
&= \widetilde{L}c_0 \left(1 + C_* p_{\max} h^d\right) \le \widetilde{L}c_0 \left(1 + C_* p_{\max} \alpha^d\right) < \frac{\lambda_0}{3}
\end{aligned}$$

provided that c_0 is chosen small enough (depending only on $\ell, d, p(\cdot), \Delta, \alpha$). Thus, $P_1 = 0$ under this choice of c_0. Combining this remark with (23), (25) and (31) we conclude that

$$\mathbf{P}\left(\inf_{x \in \mathrm{Supp}(p)} \lambda_{\min}(\overline{B}_{nx}) < \lambda_0\right) \le 2NC_{\ell,d}^2 \exp(-c_2 n h^d) + c_3 n^{c_3} \exp(-c_4 n^3 h^{2d}).$$

Recall that the cardinality N of the minimal ε-net on the ball $\mathcal{B}_d = \mathcal{B}_d(0,1)$ satisfies $N \le \left(\frac{2}{\varepsilon} + 1\right)^d$. The result of the lemma now follows by observing that under our choice of ε we have $N \le C h^{-d^2-d}$, where the constant $C > 0$ depends only on $\ell, d, p(\cdot), \Delta, \alpha$. □

In the proof of Theorem 1 below, we will use the fact that an LP(ℓ) estimator reproduces the polynomials of degree $\le \ell$ for all $x \in \mathbf{R}^d$ such that $B_{nx} \succ 0$. We state this property in the next proposition. The proof is omitted. It follows the same lines as the proof of Proposition 1.12 in [188] dealing with the case $d = 1$.

Prop .1. *Let $x \in \mathbf{R}^d$ such that $B_{nx} \succ 0$ and let Q be a polynomial of degree $\le \ell$. Then the LP(ℓ) weights W_{ni} are such that*

$$\sum_{i=1}^n Q(X_i) W_{ni}(x) = Q(x).$$

In particular,

$$\sum_{i=1}^n W_{ni}(x) = 1 \text{ and } \sum_{i=1}^n (X_i - x)^k W_{ni}(x) = 0 \text{ for } |k| \le \ell. \tag{32}$$

Proof of Theorem 1. Part (ii) of the theorem follows from Corollary 1. Also, note that (5.11) is an immediate consequence of (5.10) and Assumption (A2). Therefore, we need only to prove (5.10). Fix $x \in \mathrm{Supp}(p)$ and define the random events $\mathcal{E}_0 = \Big\{x \notin \{X_1, \dots, X_n\}\Big\}$ and

$$\mathcal{E} = \Big\{\lambda_{\min}(B_{nx}) \ge \lambda_0'\Big\} \cap \mathcal{E}_0,$$

where $\lambda_0' = \lambda_0'(\ell)$ is a constant from Lemma 4 that does not depend on n and x. From Assumption (A2) we get that $\mathbf{P}(\mathcal{E}_0) = 1$. This and Lemma 4 with our choice of h yield:

$$\mathbf{P}(\overline{\mathcal{E}}) \le c' e^{-A_n/c'}, \tag{33}$$

where $A_n = n^{\frac{2\beta}{2\beta+d}}$ and $c' > 0$ does not depend on x and n.
Since $|\bar{f}_n(x)| \le \mu = \max_{1\le i\le n} |Y_i| \vee L_0$ we obtain

$$\begin{aligned}\mathbf{E}\left(\left[\bar{f}_n(x) - f(x)\right]^2\right) &\le \mathbf{E}\Big(\left[\bar{f}_n(x) - f(x)\right]^2 \mathbf{1}(\mathcal{E})\Big) + \mathbf{E}\Big([L_0+\mu]^2 \mathbf{1}(\overline{\mathcal{E}})\Big) \\ &\le \mathbf{E}\Big(\left[f_n(x) - f(x)\right]^2 \mathbf{1}(\mathcal{E})\Big) + \mathbf{E}\Big([L_0+\mu]^{2+\delta}\Big)^{\frac{2}{2+\delta}} \mathbf{P}(\overline{\mathcal{E}})^{\frac{\delta}{2+\delta}}\end{aligned}$$

where we have used Hölder inequality and the fact that $|\bar{f}_n(x) - f(x)| \le |f_n(x) - f(x)|$ for all $x \in \mathrm{Supp}(p)$. Next,

$$\mathbf{E}\Big([L_0+\mu]^{2+\delta}\Big) \le \mathbf{E}\Big([2L_0 + \max_{1\le i\le n} |\xi(X_i)|]^{2+\delta}\Big) \le C\Big[1 + n\mathbf{E}\Big(|\xi(X_1)|^{2+\delta}\Big)\Big].$$

Using this inequality and Assumption (A1) we get

$$\mathbf{E}\left(\left[\bar{f}_n(x) - f(x)\right]^2\right) \le \mathbf{E}\Big(\left[f_n(x) - f(x)\right]^2 \mathbf{1}(\mathcal{E})\Big) + Cn^{\frac{2}{2+\delta}} \mathbf{P}(\overline{\mathcal{E}})^{\frac{\delta}{2+\delta}}. \tag{34}$$

We now bound the main term $\mathbf{E}\Big(\left[f_n(x) - f(x)\right]^2 \mathbf{1}(\mathcal{E})\Big)$ on the right hand side of (34). Writing for brevity $\mathbf{E}[\cdot|X_1,\dots,X_n] = \widetilde{\mathbf{E}}[\cdot]$ we have

$$\begin{aligned}\mathbf{E}\Big(\left[f_n(x) - f(x)\right]^2 \mathbf{1}(\mathcal{E})\Big) &\le 2\mathbf{E}\left(\Big(f_n(x) - \widetilde{\mathbf{E}}[f_n(x)]\Big)^2 \mathbf{1}(\mathcal{E})\right) \\ &\quad + 2\mathbf{E}\left(\Big(\widetilde{\mathbf{E}}[f_n(x)] - f(x)\Big)^2 \mathbf{1}(\mathcal{E})\right).\end{aligned} \tag{35}$$

We analyze separately the two terms (bias and variance terms) on the right hand side of (35).
Bound on the variance term. On the event $\mathcal{E}$ we have

$$\widetilde{\mathbf{E}}[f_n(x)] = \sum_{i=1}^n f(X_i) W_{ni}(x),$$

where

$$W_{ni}(x) = \frac{1}{nh^d} U^\top(0) B_{nx}^{-1} U\left(\frac{X_i - x}{h}\right) K\left(\frac{X_i - x}{h}\right).$$

Thus, using Assumption (A1) the variance term can be bounded as follows:

$$\mathbf{E}\left(\Big(f_n(x) - \widetilde{\mathbf{E}}[f_n(x)]\Big)^2 \mathbf{1}(\mathcal{E})\right) = \mathbf{E}\left(\left(\sum_{i=1}^n \xi(X_i) W_{ni}(x)\right)^2 \mathbf{1}(\mathcal{E})\right)$$

$$= \mathbf{E}\left(\sum_{i=1}^{n} \mathbf{E}\left[\xi^2(X_i)\Big|X_i\right] W_{ni}^2(x)\mathbf{1}(\mathcal{E})\right) \le C\sigma^2(x),$$

where

$$\sigma^2(x) = \mathbf{E}\left(\sum_{i=1}^{n} W_{ni}^2(x)\mathbf{1}(\mathcal{E})\right).$$

In what follows, we assume w.l.o.g. that $\mathrm{Supp}(K) \subseteq \mathcal{B}_d$. On the event $\mathcal{E}$, we have $\|B_{nx}^{-1}v\| \le \|v\|/\lambda_0'$ for any $v \in \mathbf{R}^{C_{\ell,d}}$. This inequality and the fact that $\|U(0)\| = 1$ imply

$$\begin{aligned}
|W_{ni}(x)| &\le \frac{1}{nh^d}\left\|B_{nx}^{-1}U\left(\frac{X_i - x}{h}\right)K\left(\frac{X_i - x}{h}\right)\right\| \\
&\le \frac{1}{nh^d\lambda_0'}\left\|U\left(\frac{X_i - x}{h}\right)\right\| K\left(\frac{X_i - x}{h}\right) \\
&\le \frac{1}{nh^d\lambda_0'}K\left(\frac{X_i - x}{h}\right)\sqrt{\sum_{0\le|s|\le\ell}\frac{1}{(s!)^2}} \qquad \text{(since } \mathrm{Supp}(K) \subseteq \mathcal{B}_d\text{)} \\
&\le \frac{c_5}{nh^d}K\left(\frac{X_i - x}{h}\right) =: \zeta_i,
\end{aligned}$$

where $c_5 > 0$ is a constant that does not depend on n and x. Using Assumption (A2) and the compactness of the support of K we get

$$\mathbf{E}(\zeta_1^2) \le \frac{c_5^2 p_{\max}}{n^2h^d}\int K^2(u)\mathrm{d}u \le \frac{C}{n^2h^d}, \tag{36}$$

$$\mathbf{E}(\zeta_1) \le \frac{c_5 p_{\max}}{n}\int K(u)\mathrm{d}u \le \frac{C}{n}\left(\int K^2(u)\mathrm{d}u\right)^{1/2} \le \frac{C}{n}. \tag{37}$$

It follows that

$$\sigma^2(x) \le \mathbf{E}\left(\sum_{i=1}^{n}\zeta_i^2\right) \le \frac{C}{nh^d}$$

and

$$\mathbf{E}\left(\left(f_n(x) - \widetilde{\mathbf{E}}[f_n(x)]\right)^2\mathbf{1}(\mathcal{E})\right) \le \frac{C}{nh^d}. \tag{38}$$

Bound on the bias term. On the event $\mathcal{E}$ we have

$$\widetilde{\mathbf{E}}[f_n(x)] - f(x) = \sum_{i=1}^{n} f(X_i)W_{ni}(x) - f(x)$$

$$= \sum_{i=1}^{n} [f(X_i) - f(x)] W_{ni}(x),$$

so that the bias term in (35) can be written as

$$\mathbf{E}\left(\left(\widetilde{\mathbf{E}}[f_n(x)] - f(x)\right)^2 \mathbf{1}(\mathcal{E})\right) = \mathbf{E}\left(\left[\sum_{i=1}^{n} [f(X_i) - f(x)] W_{ni}(x)\right]^2 \mathbf{1}(\mathcal{E})\right) =: b^2(x).$$

Using (32) and the Taylor expansion of f we get that for some $\tau_i \in [0,1]$,

$$\begin{aligned}\sum_{i=1}^{n} [f(X_i) - f(x)] W_{ni}(x) &= \sum_{i=1}^{n} \sum_{|k|=\ell} \frac{D^k f(x + \tau_i(X_i - x))}{k!} (X_i - x)^k W_{ni}(x) \\ &= \sum_{i=1}^{n} \sum_{|k|=\ell} \frac{(D^k f(x + \tau_i(X_i - x)) - D^k f(x)}{k!} (X_i - x)^k W_{ni}(x).\end{aligned}$$

Since f belongs to $\Sigma(\beta, L)$ we can apply (19), which yields

$$\begin{aligned} b^2(x) &\le \mathbf{E}\left[\left(\sum_{i=1}^{n} \frac{L}{\ell!} \|X_i - x\|^\beta |W_{ni}(x)|\right)^2 \mathbf{1}(\mathcal{E})\right] \\ &= \mathbf{E}\left[\left(\sum_{i=1}^{n} \frac{L}{\ell!} \|X_i - x\|^\beta |W_{ni}(x)| \, \mathbf{1}(\|X_i - x\| \le h)\right)^2 \mathbf{1}(\mathcal{E})\right] \quad (\text{as } \mathrm{supp}(K) \subset \mathcal{B}_d) \\ &\le \mathbf{E}\left[\left(\sum_{i=1}^{n} \frac{L}{\ell!} h^\beta |W_{ni}(x)|\right)^2 \mathbf{1}(\mathcal{E})\right]. \end{aligned}$$

As $|W_{ni}(x)| \le \zeta_i$ we further get

$$\begin{aligned} b^2(x) &\le C h^{2\beta} \mathbf{E}\left[\left(\sum_{i=1}^{n} \zeta_i\right)^2\right] = C h^{2\beta} \left[\sum_{i=1}^{n} \mathbf{E}(\zeta_i^2) + \sum_{i \ne j}^{n} \mathbf{E}(\zeta_i)\mathbf{E}(\zeta_j)\right] \\ &= C h^{2\beta} \left[n\mathbf{E}(\zeta_1^2) + n(n-1)\mathbf{E}(\zeta_1)^2\right] \le C h^{2\beta}, \end{aligned}$$

where the last inequality follows from (36), (37) and the fact that $h = \alpha n^{-\frac{1}{2\beta+d}}$. Combining this bound on $b^2(x)$ with (33), (34), (35) and (38) we finally obtain

$$\mathbf{E}\left(\left[\bar{f}_n(x) - f(x)\right]^2\right) \le C\left(\frac{1}{nh^d} + h^{2\beta} + n^{\frac{2}{2+\delta}} e^{-n^a/C}\right),$$

where $a = \frac{2\beta}{2\beta+d}$. Since $h = \alpha n^{-\frac{1}{2\beta+d}}$ the desired bound (5.10) follows. □

Proof of Theorem 2. If K satisfies the assumptions of Theorem 1(ii) then each estimator $\bar{f}_{n,j}$ is interpolating on $\mathcal{D}_1$ with probability at least

$$1 - C\exp(-n^{-2\beta_j/(2\beta_j+d)}/C) \geq 1 - C\exp(-n^{-\frac{2}{2+d}}/C)$$

if $\beta_j > 1$, and with probability 1 if $0 < \beta_j \leq 1$. Hence all of them are simultaneously interpolating with probability at least

$$1 - CM_{\max}\exp(-n^{-\frac{2}{2+d}}/C) \geq 1 - C'\exp(-n^{-\frac{2}{2+d}}/C'),$$

and the same holds true for their convex combination $\widetilde{f}_k$. Analogously, the estimator $\widetilde{g}_k$ is interpolating on $\mathcal{D}_2$ with the same probability. These remarks and the definition of $\hat{\mathsf{f}}_n$ in (5.15) ensure that $\hat{\mathsf{f}}_n$ is interpolating on the whole sample $\mathcal{D}$ with probability at least $1 - 2C'\exp(-n^{-\frac{2}{2+d}}/C')$. We now prove the bound (5.16). First, we show that such a bound holds for the estimator $\widetilde{f}_k$. Corollary 5.5 in [47] with $b_0 = +\infty$, $\widetilde{L} = L_0 + L_0 \vee \max\limits_{i=1,\dots,k} |Y_i|$, and $\eta = \left[2\sigma_\xi^2 + 2\widetilde{L}^2\right]^{-1}$ implies that

$$\mathbf{E}_2\left[\|\widetilde{f}_k - f\|_{L_2}^2\right] \leq \min_{-M\leq j\leq M_{\max}} \|\bar{f}_{n,j} - f\|_{L_2}^2 + C(1 + \max_{i=1,\dots,k} Y_i^2)\frac{\log(2M+1)}{n},$$

where we denote by $\mathbf{E}_2$ the expectation over the distribution of the sample $\mathcal{D}_2$ and we have used the fact that $M_{\max} \leq M$. Taking the expectations over $\mathcal{D}_1$ on both sides and using the fact that the noise is subgaussian we further get

$$\mathbf{E}_1\mathbf{E}_2\left[\|\widetilde{f}_k - f\|_{L_2}^2\right] \leq \min_{-M\leq j\leq M_{\max}} \mathbf{E}_1\left[\|\bar{f}_{n,j} - f\|_{L_2}^2\right] + C\frac{(\log n)^2 \log\log n}{n}. \tag{39}$$

Assume now that $\beta \in [\beta_j, \beta_{j+1}]$ for some $j \in \{-M, \dots, M_{\max} - 1\}$. Lemma 1 implies that $\Sigma(\beta, L) \subseteq \Sigma(\beta_j, 2L)$. Hence, using (5.13), we obtain:

$$\sup_{f\in\Sigma(\beta,L)\cap\mathcal{F}_0} \mathbf{E}_1\left[\|\bar{f}_{n,j} - f\|_{L_2}^2\right] \leq \sup_{f\in\Sigma(\beta_j,2L)\cap\mathcal{F}_0} \mathbf{E}_1\left[\|\bar{f}_{n,j} - f\|_{L_2}^2\right] \leq Cn^{-\frac{2\beta_j}{2\beta_j+d}}. \tag{40}$$

Combining (39) and (40) we get that, for $\beta \in [\beta_j, \beta_{j+1}]$,

$$\sup_{f\in\Sigma(\beta,L)\cap\mathcal{F}_0} \mathbf{E}_1\mathbf{E}_2\left[\|\widetilde{f}_k - f\|_{L_2}^2\right] \leq Cn^{-\frac{2\beta_j}{2\beta_j+d}}. \tag{41}$$

Notice that if $\beta \in [\beta_j, \beta_{j+1}]$ for some $j \in \{-M, \dots, M_{\max} - 1\}$ then

$$n^{-\frac{2\beta_j}{2\beta_j+d}} \leq en^{-\frac{2\beta}{2\beta+d}}.$$

Indeed,

$$\frac{\beta}{2\beta+d}-\frac{\beta_j}{2\beta_j+d}\le\frac{\beta_{j+1}-\beta_j}{(2\beta+d)(2\beta_j+d)}=\frac{\beta_j}{(2\beta_j+d)(2\beta+d)\log n}$$
$$\le\frac{\beta}{(2\beta_j+d)(2\beta+d)\log n}\le\frac{1}{2\log n}.$$

The case $\beta\in[\beta_{M_{\max}},\beta_{\max}]$ is treated analogously. Therefore, by equation (41), for each $\beta\in[\beta_{-M},\beta_{\max}]$ there exists a constant $C>0$ such that

$$\sup_{f\in\Sigma(\beta,L)\cap\mathcal{F}_0}\mathbf{E}\left[\|\widetilde{f}_k-f\|_{L_2}^2\right]\le Cn^{-\frac{2\beta}{2\beta+d}}. \tag{42}$$

Next, note that for any fixed $\beta>0$ it is only possible to have $\beta<\beta_{-M}$ for a finite number $n_0(\beta)$ of integers n ($n\le n_0(\beta)$). For such values of n the estimation error of $\widetilde{f}_k$ is bounded by a constant depending only on β, d and L_0:

$$\mathbf{E}\Big[\|\widetilde{f}_k-f\|_{L_2}^2\Big]\le 4\mathbf{E}_1\Big[\max_{i=1,\dots,n_0(\beta)/2}Y_i^2\Big]+2L_0^2\le C(\log(n_0(\beta))+L_0^2).$$

Consequently, (42) also holds for $0<\beta<\beta_{-M}$ (and thus for all $\beta\in(0,\beta_{\max}]$) if we take the constant $C>0$ in (42) large enough.
By the same argument, we deduce that the bound (42) holds for the estimator $\widetilde{g}_k$. Combining both bounds and using the fact that function $\lambda(\cdot)$ appearing in (5.15) takes values in $[0,1]$ we get the desired bound (5.16) for the final estimator $\hat{\mathsf{f}}_n$. □

Bibliography

[1] Jerzy Neyman and Egon S Pearson. IX. On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. R. Soc. Lond. A*, 231(694-706):289–337, 1933.

[2] John W Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, pages 448–485, 1960.

[3] Peter J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101, 1964.

[4] Wassily Hoeffding. Asymptotically optimal tests for multinomial distributions. *The Annals of Mathematical Statistics*, pages 369–401, 1965.

[5] Peter J Huber. A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, pages 1753–1758, 1965.

[6] Peter J Huber. Robust confidence limits. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 10(4):269–278, 1968.

[7] Donald Shepard. A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM national conference*, pages 517–524. ACM, 1968.

[8] Frank R Hampel. Robust estimation: A condensed partial survey. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 27(2), 1973.

[9] Lucien LeCam et al. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1(1):38–53, 1973.

[10] Stephen E Fienberg. The use of chi-squared statistics for categorical data problems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(1):54–64, 1979.

[11] Charles J Stone. Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8:1348–1360, 1980.

[12] P Lancaster and K Salkauskas. Surfaces generated by moving least squares methods. *Mathematics of Computation*, 37(155):141–158, 1981.

[13] Yuri Izmailovich Ingster. On the minimax nonparametric detection of signals in white gaussian noise. *Problemy Peredachi Informatsii*, 18(2):61–73, 1982.

[14] Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, pages 1040–1053, 1982.

[15] IA Ibragimov and RZ Khas' minskii. More on the estimation of distribution densities. *Journal of Soviet Mathematics*, 25(3):1155–1165, 1984.

[16] Yu 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)*, 136:74, 1984.

[17] Yuri Izmailovich Ingster. The minimax test of nonparametric hypothesis on a distribution density in metrics L_p. *Teoriya Veroyatnostei i ee Primeneniya*, 31(2):384–389, 1986.

[18] A B Tsybakov. Robust reconstruction of functions by the local-approximation method. *Problems of Information Transmission*, 22:133–146, 1986.

[19] Andrew R Barron. Uniformly powerful goodness of fit tests. *The Annals of Statistics*, pages 107–124, 1989.

[20] Michael Sergeevich Ermakov. Minimax detection of a signal in a gaussian white noise. *Theory of Probability & Its Applications*, 35(4):667–679, 1991.

[21] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer, 1992.

[22] Yuri I Ingster. Asymptotically minimax hypothesis testing for nonparametric alternatives. i, ii, iii. *Math. Methods Statist*, 2(2):85–114, 1993.

[23] Michael Sergeevich Ermakov. Minimax nonparametric testing of hypotheses on the distribution density. *Theory of Probability & Its Applications*, 39(3):396–416, 1995.

[24] Béatrice Laurent. Efficient estimation of integral functionals of a density. *The Annals of Statistics*, 24(2):659–681, 1996.

[25] Ronitt Rubinfeld and Madhu Sudan. Robust characterizations of polynomials with applications to program testing. *SIAM Journal on Computing*, 25(2):252–271, 1996.

[26] Vladimir G Spokoiny. Adaptive hypothesis testing using wavelets. *The Annals of Statistics*, 24(6):2477–2498, 1996.

[27] Béatrice Laurent. Estimation of integral functionals of a density and its derivatives. *Bernoulli*, pages 181–211, 1997.

[28] Luc Devroye, Laszlo Györfi, and Adam Krzyżak. The Hilbert kernel regression estimate. *Journal of Multivariate Analysis*, 65(2):209–227, 1998.

[29] Yanqin Fan. Goodness-of-fit tests based on kernel density estimators with fixed smoothing parameters. *Econometric Theory*, 14(5):604–621, 1998.

[30] Oded Goldreich, Shari Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. *Journal of the ACM (JACM)*, 45(4):653–750, 1998.

[31] Oleg V Lepski and Vladimir G Spokoiny. Minimax nonparametric hypothesis testing: the case of an inhomogeneous alternative. *Bernoulli*, pages 333–358, 1999.

[32] Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D Smith, and Patrick White. Testing that distributions are close. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 259–269. IEEE, 2000.

[33] Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

[34] Latanya Sweeney. Simple demographics often identify people uniquely. *Health (San Francisco)*, 671(2000):1–34, 2000.

[35] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[36] Yannick Baraud. Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8(5):577–606, 2002.

[37] Alexandre Evfimievski, Johannes Gehrke, and Ramakrishnan Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 211–222, 2003.

[38] Alexandre Evfimievski, Johannes Gehrke, and Ramakrishnan Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '03, page 211–222, New York, NY, USA, 2003. Association for Computing Machinery.

[39] Anatoli Juditsky and Sophie Lambert-Lacroix. On minimax density estimation on $\mathbb{R}$. *Bernoulli*, 10(2):187–220, 2004.

[40] Yannick Baraud, Sylvie Huet, and Béatrice Laurent. Testing convex hypotheses on the mean of a gaussian vector. application to testing qualitative hypotheses on a regression function. *The Annals of Statistics*, 33(1):214–257, 2005.

[41] Noam Berger, Christian Borgs, Jennifer T Chayes, and Amin Saberi. On the spread of viruses on the internet. In *Soda*, volume 5, pages 301–310, 2005.

[42] Béatrice Laurent. Adaptive estimation of a quadratic functional of a density by model selection. *ESAIM: Probability and Statistics*, 9:1–18, 2005.

[43] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[44] Magalie Fromont and Béatrice Laurent. Adaptive goodness-of-fit tests in a density model. *The Annals of Statistics*, 34(2):680–720, 2006.

[45] Jean-Yves Audibert and Alexandre B Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007.

[46] Cynthia Dwork. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer, 2008.

[47] Anatoli Juditsky, Philippe Rigollet, and Alexandre B Tsybakov. Learning by mirror averaging. *The Annals of Statistics*, 36(5):2183–2206, 2008.

[48] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008.

[49] Peter J Huber and EM Ronchetti. Robust statistics. 2nd john wiley & sons. *Hoboken, NJ*, 2, 2009.

[50] A van der Vaart and J A Wellner. A note on bounds for VC dimensions. In *High Dimensional Probability*, volume 5, pages 103–107. IMS Collections, 2009.

[51] Nicolas Verzelen and Fanny Villers. Goodness-of-fit tests for high-dimensional gaussian linear models. *The Annals of Statistics*, 38(2):704–752, 2010.

[52] Darren P Croft, Joah R Madden, Daniel W Franks, and Richard James. Hypothesis testing in animal social networks. *Trends in ecology & evolution*, 26(10):502–507, 2011.

[53] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106, 2011.

[54] Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

[55] Peter J Rousseeuw and Mia Hubert. Robust statistics for outlier detection. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 1(1):73–79, 2011.

[56] Nino Shervashidze, Pascal Schweitzer, Erik Jan Van Leeuwen, Kurt Mehlhorn, and Karsten M Borgwardt. Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research*, 12(9), 2011.

[57] Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Shengjun Pan, and Ananda Suresh. Competitive classification and closeness testing. In *Conference on Learning Theory*, pages 22–1, 2012.

[58] Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pages 1148–1185, 2012.

[59] Nicolas Verzelen. Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electronic Journal of Statistics*, 6:38–90, 2012.

[60] Quentin Berthet and Philippe Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *Conference on learning theory*, pages 1046–1066. PMLR, 2013.

[61] Quentin Berthet and Philippe Rigollet. Optimal detection of sparse principal components in high dimension. *The Annals of Statistics*, 41(4):1780–1815, 2013.

[62] John Duchi, Martin J Wainwright, and Michael I Jordan. Local privacy and minimax bounds: Sharp rates for probability estimation. *Advances in Neural Information Processing Systems*, 26, 2013.

[63] Daniel R Hyduke, Nathan E Lewis, and Bernhard Ø Palsson. Analysis of omics data with genome-scale models of metabolism. *Molecular BioSystems*, 9(2):167–174, 2013.

[64] Sebastian Moreno and Jennifer Neville. Network hypothesis testing using mixed kronecker product graph models. In *2013 IEEE 13th International Conference on Data Mining*, pages 1163–1168. IEEE, 2013.

[65] Ery Arias-Castro and Nicolas Verzelen. Community detection in dense random networks. *The Annals of Statistics*, 42(3):940–969, 2014.

[66] Siu-On Chan, Ilias Diakonikolas, Paul Valiant, and Gregory Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 1193–1203. SIAM, 2014.

[67] Ilias Diakonikolas, Daniel M Kane, and Vladimir Nikishkin. Testing identity of structured distributions. In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*, pages 1841–1854. SIAM, 2014.

[68] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Local privacy, data processing inequalities, and statistical minimax rates, 2014.

[69] Alexander Goldenshluger and Oleg Lepski. On adaptive minimax density estimation on $\mathbb{R}^d$. *Probability Theory and Related Fields*, 159(3):479–543, 2014.

[70] Jing Qian and Venkatesh Saligrama. Efficient minimax signal detection on graphs. *Advances in Neural Information Processing Systems*, 27:2708–2716, 2014.

[71] Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. *Advances in Neural Information Processing Systems*, 28:3591–3599, 2015.

[72] Bhaswar Bhattacharya and Gregory Valiant. Testing closeness with unequal sized samples. In *Advances in Neural Information Processing Systems*, pages 2611–2619, 2015.

[73] Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Minimax estimation of discrete distributions under ℓ_1 loss. *IEEE Transactions on Information Theory*, 61(11):6343–6354, 2015.

[74] Nicolas Verzelen and Arias-Castro. Community detection in sparse random networks. *The Annals of Applied Probability*, 25(6):3465–3510, 2015.

[75] Bo Waggoner. l^p testing and learning of discrete distributions. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, pages 347–356, 2015.

[76] Meng Wang, Chaokun Wang, Jeffrey Xu Yu, and Jun Zhang. Community detection in social networks: an in-depth benchmarking study with a procedure-oriented framework. *Proceedings of the VLDB Endowment*, 8(10):998–1009, 2015.

[77] Emmanuel Abbe and Colin Sandon. Achieving the KS threshold in the general stochastic block model with linearized acyclic belief propagation. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 1342–1350. Citeseer, 2016.

[78] Punam Bedi and Chhavi Sharma. Community detection in social networks. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 6(3):115–135, 2016.

[79] Sébastien Bubeck, Jian Ding, Ronen Eldan, and Miklós Z Rácz. Testing for high-dimensional geometry in random graphs. *Random Structures & Algorithms*, 49(3):503–532, 2016.

[80] Mengjie Chen, Chao Gao, and Zhao Ren. A general decision theory for huber's ϵ-contamination model. *Electronic Journal of Statistics*, 10(2):3752–3774, 2016.

[81] Ilias Diakonikolas and Daniel M Kane. A new approach for testing properties of discrete distributions. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 685–694. IEEE, 2016.

[82] Peter Kairouz, Keith Bonawitz, and Daniel Ramage. Discrete distribution estimation under local privacy. In *International Conference on Machine Learning*, pages 2436–2444. PMLR, 2016.

[83] Jing Lei. A goodness-of-fit test for stochastic block models. *The Annals of Statistics*, 44(1):401–424, 2016.

[84] Tengyao Wang, Quentin Berthet, and Richard J Samworth. Statistical and computational trade-offs in estimation of sparse principal components. *The Annals of Statistics*, 44(5):1896–1930, 2016.

[85] Emmanuel Abbe. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

[86] Yannick Baraud, Lucien Birgé, and Mathieu Sart. A new method for estimation and model selection: ρ-estimation. *Inventiones mathematicae*, 207(2):425–517, 2017.

[87] Olivier Collier, Laëtitia Comminges, and Alexandre B Tsybakov. Minimax estimation of linear and quadratic functionals on sparsity classes. *The Annals of Statistics*, 45(3):923–958, 2017.

[88] Ilias Diakonikolas, Daniel M Kane, and Vladimir Nikishkin. Near-optimal closeness testing of discrete histogram distributions. *arXiv preprint arXiv:1703.01913*, 2017.

[89] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 73–84. IEEE, 2017.

[90] Chao Gao and John Lafferty. Testing network structure using relations between small subgraph probabilities. *arXiv preprint arXiv:1704.06742*, 2017.

[91] Debarghya Ghoshdastidar, Maurilio Gutzeit, Alexandra Carpentier, and Ulrike von Luxburg. Two-sample tests for large random graphs using network statistics. In *Conference on Learning Theory*, pages 954–977. PMLR, 2017.

[92] Cedric E Ginestet, Jun Li, Prakash Balachandran, Steven Rosenberg, and Eric D Kolaczyk. Hypothesis testing for network data in functional neuroimaging. *The Annals of Applied Statistics*, 11(2):725–750, 2017.

[93] Mingda Qiao and Gregory Valiant. Learning discrete distributions from untrusted batches. *arXiv preprint arXiv:1711.08113*, 2017.

[94] Minh Tang, Avanti Athreya, Daniel L Sussman, Vince Lyzinski, Youngser Park, and Carey E Priebe. A semiparametric two-sample hypothesis testing problem for random graphs. *Journal of Computational and Graphical Statistics*, 26(2):344–354, 2017.

[95] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46(1):429–455, 2017.

[96] Sivaraman Balakrishnan and Larry Wasserman. Hypothesis testing for high-dimensional multinomials: A selective review. *The Annals of Applied Statistics*, 12(2):727–749, 2018.

[97] Mikhail Belkin, Daniel J Hsu, and Partha Mitra. Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate. *Advances in Neural Information Processing systems*, 31, 2018.

[98] Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pages 541–549. PMLR, 2018.

[99] Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? *Oberwolfach Reports*, 15(2):1776–1779, 2018.

[100] Clément L Canonne, Ilias Diakonikolas, Themis Gouleakis, and Ronitt Rubinfeld. Testing shape restrictions of discrete distributions. *Theory of Computing Systems*, 62(1):4–62, 2018.

[101] Mengjie Chen, Chao Gao, and Zhao Ren. Robust covariance and scatter matrix estimation under Huber's contamination model. *The Annals of Statistics*, 46(5):1932–1960, 2018.

[102] Debarghya Ghoshdastidar and Ulrike von Luxburg. Practical methods for graph two-sample testing. *Advances in Neural Information Processing Systems*, 31:3019–3028, 2018.

[103] Costin Bădescu, Ryan O'Donnell, and John Wright. Quantum state certification. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 503–514, 2019.

[104] Sivaraman Balakrishnan, Larry Wasserman, et al. Hypothesis testing for densities and high-dimensional multinomials: Sharp local minimax rates. *Annals of Statistics*, 47(4):1893–1927, 2019.

[105] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[106] Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? In *Proceedings of AISTATS-2019*, volume 89, pages 1611–1619. PMLR, 2019.

[107] Eric Blais, Clément L Canonne, and Tom Gur. Distribution testing lower bounds via reductions from communication complexity. *ACM Transactions on Computation Theory (TOCT)*, 11(2):1–37, 2019.

[108] Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L Bartlett. Fast mean estimation with sub-gaussian rates. In *Conference on Learning Theory*, pages 786–806. PMLR, 2019.

[109] Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. Testing ising models. *IEEE Transactions on Information Theory*, 65(11):6829–6852, 2019.

[110] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.

[111] Maurilio Gutzeit. Topics in statistical minimax hypothesis testing. 2019.

[112] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 656–672. IEEE, 2019.

[113] Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

[114] Gabor Lugosi and Shahar Mendelson. Risk minimization by median-of-means tournaments. *Journal of the European Mathematical Society*, 22(3):925–965, 2019.

[115] Rafael Pinot, Florian Yger, Cédric Gouy-Pailler, and Jamal Atif. A unified view on differential privacy and robustness to adversarial examples. *arXiv preprint arXiv:1906.07982*, 2019.

[116] Alexander Rakhlin and Xiyu Zhai. Consistency of interpolation with laplace kernels is a high-dimensional phenomenon. In *Conference on Learning Theory*, pages 2595–2623. PMLR, 2019.

[117] Jayadev Acharya, Clément L Canonne, Ziteng Sun, and Himanshu Tyagi. Unified lower bounds for interactive high-dimensional estimation under information constraints. *arXiv preprint arXiv:2010.06562*, 2020.

[118] Jayadev Acharya, Clément L. Canonne, and Himanshu Tyagi. Inference under information constraints ii: Communication constraints and shared randomness. *IEEE Transactions on Information Theory*, 66(12):7856–7877, 2020.

[119] Jayadev Acharya, Peter Kairouz, Yuhan Liu, and Ziteng Sun. Estimating sparse discrete distributions under local privacy and communication constraints. *arXiv preprint arXiv:2011.00083*, 2020.

[120] Marco Avella-Medina. The role of robust statistics in private data analysis. *CHANCE*, 33(4):37–42, 2020.

[121] Yannick Baraud and Lucien Birgé. Robust bayes-like estimation: Rho-bayes estimation. *The Annals of Statistics*, 48(6):3699–3720, 2020.

[122] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

[123] Amir-Hossein Bateni and Arnak S Dalalyan. Confidence regions and minimax rates in outlier-robust estimation on the probability simplex. *Electronic Journal of Statistics*, 14(2):2653–2677, 2020.

[124] Thomas Berrett and Cristina Butucea. Locally private non-asymptotic testing of discrete distributions is faster using interactive mechanisms. *Advances in Neural Information Processing Systems*, 33:3164–3173, 2020.

[125] Cristina Butucea, Amandine Dubois, Martin Kroll, and Adrien Saumard. Local differential privacy: Elbow effect in optimal density estimation and adaptation over Besov ellipsoids. *Bernoulli*, 26(3):1727–1764, 2020.

[126] Cristina Butucea, Angelika Rohde, and Lukas Steinberger. Interactive versus non-interactive locally differentially private estimation: Two elbows for the quadratic functional. *arXiv preprint arXiv:2003.04773*, 2020.

[127] Timothy I Cannings, Yingying Fan, and Richard J Samworth. Classification with imperfect training labels. *Biometrika*, 107(2):311–330, 2020.

[128] Clément L Canonne. A survey on distribution testing: Your data is big. But is it blue? *Theory of Computing*, pages 1–100, 2020.

[129] Clément L Canonne, Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Testing bayesian networks. *IEEE Transactions on Information Theory*, 66(5):3132–3170, 2020.

[130] Sitan Chen, Jerry Li, and Ankur Moitra. Efficiently learning structured distributions from untrusted batches. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 960–973, 2020.

[131] Sitan Chen, Jerry Li, and Ankur Moitra. Learning structured distributions from untrusted batches: Faster and simpler. *Advances in Neural Information Processing Systems*, 33:4512–4523, 2020.

[132] Julien Chhor and Alexandra Carpentier. Sharp local minimax rates for goodness-of-fit testing in large random graphs, multivariate poisson families and multinomials. *arXiv preprint arXiv:2012.13766*, 2020.

[133] Geoffrey Chinot and Matthieu Lerasle. On the robustness of the minimum ℓ_2 interpolator. *arXiv preprint arXiv:2003.05838*, 2020.

[134] Soham Dan and Bhaswar B Bhattacharya. Goodness-of-fit tests for inhomogeneous random graphs. In *International Conference on Machine Learning*, pages 2335–2344. PMLR, 2020.

[135] Debarghya Ghoshdastidar, Maurilio Gutzeit, Alexandra Carpentier, and Ulrike Von Luxburg. Two-sample hypothesis testing for inhomogeneous random graphs. *Annals of Statistics*, 48(4):2208–2229, 2020.

[136] Samuel B Hopkins. Mean estimation with sub-gaussian rates in polynomial time. *The Annals of Statistics*, 48(2):1193–1213, 2020.

[137] Ayush Jain and Alon Orlitsky. A general method for robust learning from batches. *Advances in Neural Information Processing Systems*, 33:21775–21785, 2020.

[138] Ayush Jain and Alon Orlitsky. Optimal robust learning of discrete distributions from batches. In *International Conference on Machine Learning*, pages 4651–4660. PMLR, 2020.

[139] Ilmun Kim, Sivaraman Balakrishnan, and Larry Wasserman. Robust multivariate nonparametric tests via projection averaging. *The Annals of Statistics*, 48(6):3417–3441, 2020.

[140] Dmitry Kobak, Jonathan Lomond, and Benoit Sanchez. The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *J. Mach. Learn. Res.*, 21:169–1, 2020.

[141] Guillaume Lecué, Matthieu Lerasle, and Timlothée Mathieu. Robust classification via mom minimization. *Machine Learning*, 109(8):1635–1665, 2020.

[142] Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):1329–1347, 2020.

[143] Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory*, pages 2683–2711. PMLR, 2020.

[144] Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1):67–83, 2020.

[145] Mohammad Naseri, Jamie Hayes, and Emiliano De Cristofaro. Toward robustness and privacy in federated learning: Experimenting with local and central differential privacy. *arXiv e-prints*, pages arXiv–2009, 2020.

[146] Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*, 2020.

[147] Denny Wu and Ji Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. *Advances in Neural Information Processing Systems*, 33:10112–10123, 2020.

[148] Jayadev Acharya, Clément L Canonne, Yuhan Liu, Ziteng Sun, and Himanshu Tyagi. Interactive inference under information constraints. *IEEE Transactions on Information Theory*, 68(1):502–516, 2021.

[149] Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Robust testing and estimation under manipulation attacks. In *International Conference on Machine Learning*, pages 43–53. PMLR, 2021.

[150] Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Robust testing and estimation under manipulation attacks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 43–53. PMLR, 18–24 Jul 2021.

[151] Hassan Ashtiani and Christopher Liaw. Private and polynomial time algorithms for learning gaussians and beyond. *arXiv preprint arXiv:2111.11320*, 2021.

[152] Peter L Bartlett and Philip M Long. Failures of model-dependent generalization bounds for least-norm interpolation. *Journal of Machine Learning Research*, 22(204):1–15, 2021.

[153] Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta numerica*, 30:87–201, 2021.

[154] Mikhail Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248, 2021.

[155] Thomas B Berrett, Ioannis Kontoyiannis, and Richard J Samworth. Optimal rates for independence testing via u-statistic permutation tests. *The Annals of Statistics*, 49(5):2457–2490, 2021.

[156] T Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5):2825–2850, 2021.

[157] Albert Cheu, Adam Smith, and Jonathan Ullman. Manipulation attacks in local differential privacy. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 883–900. IEEE, 2021.

[158] Julien Chhor and Alexandra Carpentier. Goodness-of-fit testing for h\" older-continuous densities: Sharp local minimax rates. *arXiv preprint arXiv:2109.04346*, 2021.

[159] Amandine Dubois, Thomas Berrett, and Cristina Butucea. Goodness-of-fit testing for Hölder continuous densities under local differential privacy. *arXiv preprint arXiv:2107.02439*, 2021.

[160] Samuel B. Hopkins, Gautam Kamath, and Mahbod Majid. Efficient mean estimation with pure differential privacy via a sum-of-squares exponential mechanism, 2021.

[161] Ayush Jain and Alon Orlitsky. Robust density estimation from batches: The best things in life are (nearly) free. In *International Conference on Machine Learning*, pages 4698–4708. PMLR, 2021.

[162] Subhodh Kotekal and Chao Gao. Minimax rates for sparse signal detection under correlation. *arXiv preprint arXiv:2110.12966*, 2021.

[163] Xiyang Liu, Weihao Kong, Sham Kakade, and Sewoong Oh. Robust and differentially private mean estimation. *Advances in Neural Information Processing Systems*, 34, 2021.

[164] Xiyang Liu, Weihao Kong, and Sewoong Oh. Differential privacy and robust statistics in high dimensions, 2021.

[165] Dominic Richards, Jaouad Mourtada, and Lorenzo Rosasco. Asymptotics of ridge (less) regression under general source condition. In *International Conference on Artificial Intelligence and Statistics*, pages 3889–3897. PMLR, 2021.

[166] Antonin Schrab, Ilmun Kim, Mélisande Albert, Béatrice Laurent, Benjamin Guedj, and Arthur Gretton. Mmd aggregated two-sample test. *arXiv preprint arXiv:2110.15073*, 2021.

[167] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

[168] Clément L. Canonne. Topics and Techniques in Distribution Testing: A Biased but Representative Sample. March 2022.

[169] Julien Chhor and Flore Sentenac. Robust estimation of discrete distributions under local differential privacy. *arXiv preprint arXiv:2202.06825*, 2022.

[170] Julien Chhor, Suzanne Sigalla, and Alexandre B Tsybakov. Benign overfitting and adaptive nonparametric regression. *arXiv preprint arXiv:2206.13347*, 2022.

[171] Arnak S Dalalyan and Arshak Minasyan. All-in-one robust estimator of the gaussian mean. *The Annals of Statistics*, 50(2):1193–1219, 2022.

[172] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

[173] Joseph Lam-Weil, Alexandra Carpentier, and Bharath K Sriperumbudur. Local minimax rates for closeness testing of discrete distributions. *Bernoulli*, 28(2):1179–1197, 2022.

[174] Joseph Lam-Weil, Béatrice Laurent, and Jean-Michel Loubes. Minimax optimal goodness-of-fit testing for densities and multinomials under a local differential privacy constraint. *Bernoulli*, 28(1):579–600, 2022.

[175] Guillaume Lecué and Zong Shang. A geometrical viewpoint on the benign overfitting property of the minimum ℓ_2-norm interpolant estimator. *arXiv preprint arXiv:2203.05873*, 2022.

[176] Mengchu Li, Thomas B Berrett, and Yi Yu. On robustness and local differential privacy. *arXiv preprint arXiv:2201.00751*, 2022.

[177] L Devroye, L Györfi, and G Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, NY e.a., 1996.

[178] J Fan and I Gijbels. *Local Polynomial Modelling and its Applications*. Chapman and Hall, NY., 1996.

[179] Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge university press, 2021.

[180] Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. *Robust statistics: the approach based on influence functions*, volume 196. John Wiley & Sons, 2011.

[181] Peter J Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004.

[182] Jurij I Ingster. *Adaptation in minimax nonparametric hypothesis testing for ellipsoids and Besov bodies*. WIAS, 1998.

[183] Yuri Ingster and Irina A Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*, volume 169. Springer Science & Business Media, 2012.

[184] V Ya Katkovnik. *Nonparametric identification and data smoothing (in Russian)*. Nauka, Moscow, 1985.

[185] László Lovász. *Large networks and graph limits*, volume 60. American Mathematical Soc., 2012.

[186] Ricardo A Maronna, R Douglas Martin, Victor J Yohai, and Matías Salibián-Barrera. *Robust statistics: theory and methods (with R)*. John Wiley & Sons, 2019.

[187] Elvezio M Ronchetti and Peter J Huber. *Robust statistics*. John Wiley & Sons, 2009.

[188] Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

Titre : Problèmes d'inférence non-paramétrique et en grande dimension

Mots clés : Tests minimax, statistiques robustes, confidentialité, régression non-paramétrique, statistiques en grande dimension, distributions discrètes

Résumé : Dans cette thèse, nous traitons les sujets suivants: tests minimax locaux, estimation sous contraintes combinées de robustesse et de confidentialité locale différentielle, estimation adaptative en régression non-paramétrique avec benign overfitting. Nous étudions en premier lieu le problème de test minimax d'adéquation pour des lois discrètes et des lois à densité Hölder-régulières. Le problème consiste à tester l'égalité à une loi connue contre une alternative composée de distributions séparées de l'hypothèse nulle au sens d'une certaine métrique. Nous identifions les vitesses locales non-asymptotiques sur la séparation nécessaire pour assurer l'existence d'un test uniformément consistant et nous donnons leur dépendance précise par rapport à l'hypothèse nulle pour différentes distances de séparation. Nous identifions également les tests locaux optimaux correspondants. Nous étudions le problème d'estimation de lois discrètes avec contrainte de confidentialité differentielle locale en supposant de plus que les données sont issues d'un modèle de contamination adversariale. Nous proposons un algorithme robuste aux outliers et adapté à la confidentialité, dont nous montrons l'optimalité statistique ainsi que l'efficacité en temps de calcul. Enfin, dans le cadre de la régression nonparamétrique, nous exhibons un estimateur adaptatif à la régularité, capable d'interpoler tous les points de données avec grande probabilité tout en atteignant l'optimalité statistique - un phenomène connu sous le nom de "benign overfitting".

Title : Topics in high-dimensional and non-parametric inference

Keywords : Minimax testing, robust statistics, local differential privacy, nonparametric regression, high-dimensional statistics, discrete distributions

Abstract : In this thesis, we consider the following topics: Local minimax testing, estimation under the combined constraints of robustness and local differential privacy, estimation of discrete distributions under low-rank assumptions, adaptive estimation in nonparametric regression with benign overfitting. A first theme addressed in this thesis is the local minimax goodness-of-fit testing problem for high-dimensional discrete distributions and Hölder-smooth densities. The problem consists in testing equality to a given distribution when observing iid data, against an alternative separated from the null distribution in some given metric. We identify the sharp non-asymptotic local minimax rates on the separation needed to ensure the existence of a uniformly consistent test and give its precise dependency on the null distribution for a variety of separation distances. We also derive the corresponding local minimax tests. In the second part of the thesis, we study an estimation problem with learning constraints. We consider the problem of estimating discrete distributions under local differential privacy, assuming also that the data follow an adversarial contamination model. We propose a locally differentially private algorithm that is robust to adversarially chosen outliers. We prove its near statistical optimality and show that it has polynomial time complexity. In the third part of the thesis, we consider the nonparametric regression setting, and we show that local polynomial estimators with singular kernel can be minimax optimal and adaptive to unknown smoothness, while interpolating all the datapoints with high probability – a phenomenon known as "benign overfitting".

Institut Polytechnique de Paris
91120 Palaiseau, France