Testing of distributions, minimax optimality and extensions

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium
(Dr. rer. nat.)

von M.Sc. Joseph Lam

geb. am 4. September 1994 in Strassburg, Frankreich

genehmigt durch die Fakultät für Mathematik
der Otto-von-Guericke-Universität Magdeburg

Gutachter: Prof. Dr. Alexandra Carpentier
Prof. Dr. Cristina Butucea

eingereicht am: 29. März 2021

Verteidigung am: 14. Juni 2021

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG

Abstract

Institut für Mathematische Stochastik
Fakultät für Mathematik

Testing of distributions, minimax optimality and extensions

by Joseph Lam

The subject of this thesis is a minimax optimal study of statistical inference, with a focus on hypothesis testing. This will give us the opportunity of reviewing tools and ideas from the literature of estimation and testing of discrete and continuous distributions. In the course of our analysis, we will explore several extensions of classical minimax statistical problems. The first one is a local refinement of the minimax framework and we will contribute by obtaining local minimax optimal rates for closeness testing. The second is the study of minimax optimal methods while preserving the privacy of data sets. Our contribution in that area will be minimax rates for identity testing under local differential privacy. Finally, we extend the scope of our study to a sequential setting, where we will employ techniques from bandit theory in order to obtain the first minimax rate for adaptive rejection sampling.

Gegenstand dieser Dissertation ist die Untersuchung inferenzstatistischer Methoden - insbesondere von Hypothesentests - auf Minimax-Optimalität. Wir werden Ideen und Techniken untersuchen, die bereits in der Literatur zum Schätzen und Testen für diskrete oder auch stetige Verteilungen vorhanden sind. Im Verlauf unserer Analyse werden wir verschiedene Erweiterungen klassischer statistischer Minimax-Probleme untersuchen. Die Erste ist eine lokale Verfeinerung des Minimax-Frameworks und wir zeigen neue lokale Minimax-optimale Raten für Closeness Testing. Die Zweite ist die Untersuchung der optimalen Minimax-Methoden unter Wahrung der Privacy von Datensätzen. Unser Beitrag in diesem Bereich sind Minimax-Raten für Anpassungstests unter lokaler differenzieller Privatsphäre. Anschließend erweitern wir unsere Resultate auf eine sequentielle Umgebung, wobei wir Techniken aus der Banditentheorie anwenden, um erstmalig eine Minimax-Rate für eine solche adaptive Testmethode zu erhalten - bisher gab es in der Literatur keine Resultate dafür.

Contents

List of Figures

List of Tables

Chapter 1

Preface

1.1 Notations

Let $(\Omega, \mathcal{A})$ be a measure space and $(\mathbb{P}_\theta)_{\theta\in\Theta}$ a family of probability measures on $(\Omega, \mathcal{A})$. We assume that X is a random variable in the statistical model $(\Omega, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta})$. Expectation and variance with respect to the measure $\mathbb{P}_\theta$ are denoted as $\mathbb{E}_\theta$ and $\mathbb{V}_\theta$. For any $x \in \mathbb{R}^d$, we have $\|x\|_{\ell_u} = (\sum_{i=1}^d |x_i|^u)^{1/u}$. For any set S, we define $\mathbb{L}_u(S)$ as the set of functions f with support in S such that$|f|^u$ is integrable, and for any $f \in \mathbb{L}_u(S)$, we write $\|f\|_{\mathbb{L}_u} = (\int |f|^u)^{1/u}$. For any positive integer d, $\mathbf{P}_d = \{p \in \mathbb{R}^d : \forall i, p_i \geq 0, \sum_{j=1}^d p_j = 1\}$ denotes the set of d-dimensional probability vectors. For any countable set S, $|S|$ denotes the number of elements of S.

1.2 Introduction to estimation

Roll a die and you can end up with either one of six possible results. In this experiment, most mechanisms involved in delivering the final result are out of our grasp. So we end up considering it as random instead. The field of probability is tasked with modeling this random experiment with structural assumptions. So in the example of the dice roll, you would model each possible result with some probability to be determined. And at the heart of statistics is statistical inference, the process of deducing intrinsic properties, like the probability of each result, just from the observations. Key questions are when such a task is possible and how to accomplish this the most efficiently.

An important problem in statistical inference is the estimation of a distribution. Given a specified class of distributions, the goal of estimation is to identify some characteristics of the distribution of the observed random variables.

Definition 1. *Let $X_1, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables and $\theta \in \Theta$ be a parameter associated with the distribution of X_1.*

- *A statistic is a random variable $\varphi(X_1, ..., X_n)$ such that φ is a measurable function.*
- *An estimator of $\theta \in \Theta$ is a statistic ranging in Θ.*

Now this definition is very unrestrictive and allows for the definition of meaningless estimators. Historically estimators have been studied according to various performance criteria that we will develop in the sections to come. This chapter will review basic ideas from the field of statistics building up to minimax statistical inference.

1.2.1 Asymptotic evaluation of the quality of an estimator

We begin with the study of asymptotic criteria related to stochastic convergence, for which a complete survey can be found in Van der Vaart (2000). Those hold when considering a number n of random variables growing to infinity. Of course, such properties are not as appealing in practice as non-asymptotic properties. They are nonetheless interesting initial results and failing to fulfill simple asymptotic properties is often a red flag for considering an estimator. We now present a few asymptotic properties associated with the study of estimators: consistency and asymptotic distribution. We illustrate both using theorems fundamental to the study of statistics and probability: the law of large numbers and the central limit theorem.

Definition 2.
- *A sequence of random variables (Y_n) converges in probability to a constant θ, if for any $\varepsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}(\|Y_n - \theta\|_2 \geq \varepsilon) = 0.$$

- *An estimator is consistent in Θ, if for any $\theta \in \Theta$, the estimator converges in probability to θ.*

Remark 1. *In the definition of consistency, the condition is enforced to hold for any $\theta \in \Theta$. Indeed, even a broken clock is correct twice a day. For example, if $\Theta = \mathbb{R}$, then the constant equal to 0 is a trivial estimator which converges in probability to θ if $\theta = 0$.*

We illustrate this definition with the following theorem justifying the use of the sample average for estimating the mean of a distribution.

Theorem 1 (The weak law of large numbers.)**.** *Let $X_1, \ldots, X_n$ be i.i.d. random variables with $\mathbb{E}(X_1) = \mu$. Then $\sum_{i=1}^n X_i/n$ converges in probability to μ.*

In the following definition, we introduce the concept of asymptotic distribution.

Definition 3. *A sequence of random variables (Y_n) converges in distribution to Y, if*

$$\mathbb{E}[g(Y_n)] \to \mathbb{E}[g(Y)],$$

for any bounded, uniformly continuous g. And the distribution associated with Y is the asymptotic distribution of Y_n.

When an estimator is consistent, its asymptotic distribution is a Dirac distribution at θ. However, in an analog way to the study of convergence of deterministic sequences, there exists finer convergence results describing the behaviour of the estimator as it converges to θ.

Theorem 2 (The central limit theorem of Lindeberg-Levy.)**.** *Let $X_1, \ldots, X_n$ be i.i.d. random variables with $\mathbb{E}(X_1) = \mu$ and $\mathbb{V}(X_1) = \sigma^2 < \infty$. Then $\sqrt{n}(\sum_{i=1}^n X_i/n - \mu)$ converges in distribution to $\mathcal{N}(0, \sigma^2)$.*

Knowing the asymptotic distribution has applications in building exact asymptotic confidence intervals and the asymptotic distribution can also serve the purpose of measuring the performance of an estimator in the following way. Let L be a loss function, so it is nonnegative. Given the asymptotic distribution Q_θ of $\hat{\theta} - \theta$, a good estimator will obtain low values of the following mean loss.

$$\int L(t) dQ_\theta(t).$$

Example 1. *A fundamental example is the mean squared error, taking* $L : t \mapsto t^2$, *that we will discuss in more detail in the next section in a nonasymptotic context.*

1.2.2 Nonasymptotic evaluation of the quality of an estimator

We now turn to the study of nonasymptotic criteria, which will be the focus of the results presented in this dissertation. Ideas presented here are very much related to asymptotic properties. We start with the definition of unbiasedness.

Definition 4. *An estimator is unbiased, if for any* $\theta \in \Theta$, $\mathbb{E}_\theta(\varphi(X_1, \dots, X_n)) = \theta$.

Unbiasedness is a criterion relying on the first moment of the estimator. However, one might consider the variance as well, that is to say, the second moment. An unbiased estimator is considered good if it has low variance. The low variance criterion is not sufficient by itself. Again, a constant estimator has 0 variance, but one would not consider it a good estimator.

So we might be considering this as an optimization problem, minimizing the variance under the constraint of having no bias over all estimators of $\theta \in \Theta$. Now there are two issues with the formulation of this problem.

1. The class of all estimators might be too large to optimize over.
2. Enforcing absolutely no bias puts a disproportionate emphasis on bias in comparison with the variance.

For the first issue, one can reduce the class of all possible estimators. For example, one can consider linear estimators only and look for the best linear unbiased estimators. However, this is a very restrictive class of estimators and it turns out that this first issue can be solved for a lot of problems using information-theoretic bounds, while keeping a very large class of estimators.

For the second issue, in the case $\Theta \subset \mathbb{R}$, one could consider minimizing over all estimators $\hat{\theta}$,

$$\mathbb{V}_\theta(\hat{\theta}) + cB_\theta(\hat{\theta})^2,$$

where $B_\theta(\hat{\theta}) = \mathbb{E}_\theta(\hat{\theta} - \theta)$ is the bias of $\hat{\theta}$ and c is some positive constant.

Taking the squared value keeps the term associated with the bias non-negative, and the definition 'physically' homogeneous. In fact, we highlight the connection with the mean squared error, defined as

$$\mathbb{E}_\theta[\|\hat{\theta} - \theta\|_2^2].$$

But a more practical form of the mean squared error is given by the bias-variance decomposition in the following proposition.

Proposition 1. *If* $\Theta \subset \mathbb{R}^d$, *where* $d > 0$, *then we have for any* $\theta \in \Theta$ *and estimators* $\hat{\theta}$,

$$\mathbb{E}_\theta[\|\hat{\theta} - \theta\|_2^2] = tr[\mathbb{V}_\theta(\hat{\theta})] + \|B_\theta(\hat{\theta})\|_2^2,$$

where tr is the trace function and $\mathbb{V}$ *is the covariance matrix.*

Note that the mean squared error depends heavily on the number of samples and one can easily generalize the mean squared error to an $\mathbb{L}_u$-estimation risk for any $u \in (0, \infty]$ by

$$\mathbb{E}_\theta[\|\hat{\theta} - \theta\|_u^u],$$

where one simply ignores the power if $u = \infty$.

The estimation risk gives an explicit measure of the performance of an estimator $\hat{\theta}$ for a given θ. We will now study the question of comparing estimators over Θ.

1.2.3 Comparison of estimators

Comparing the performance of estimators is like comparing real-valued functions. If a function dominates another, that is, if it is larger than the other everywhere, then its performance is obviously better. In other cases however, pointwise comparison is not trivially summarized into a general comparison.

A possibility is the Bayesian perspective, where the comparison is done on average. Another is the minimax framework, where one compares the performance of estimators at their worst.

Definition 5. *Let $n > 0$ be some fixed integer and $\mathcal{E}_n$ the set of estimators of $\theta \in \Theta$ using n observations. Let $u \in (0, \infty]$, and if $u = \infty$, one simply ignores the power in the following definitions.*

- *Let ν be a distribution with support in Θ. Then the $\mathbb{L}_u$-Bayesian estimation risk associated with ν is*
$$\inf_{\hat{\theta} \in \mathcal{E}_n} \mathbb{E}_{\theta \sim \nu} \left(\mathbb{E} \left[\|\hat{\theta} - \theta\|_u^u \middle| \theta \right] \right).$$
- *And the $\mathbb{L}_u$-minimax estimation risk is*
$$\inf_{\hat{\theta} \in \mathcal{E}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta [\|\hat{\theta} - \theta\|_u^u].$$

Remark 2. *The definition of the Bayesian risk relies on a probability distribution ν to be chosen. An example is choosing ν such that any $\theta \in \Theta$ is given equal weight. However, leaving the discussion on the choice of ν open reduces the comparability of the performance of estimators.*

We will focus throughout this work on the study of minimax results and a good introduction to this type of analysis can be found in Tsybakov (2008). Now, finding the exact minimax risk is extremely complicated for most problems. Instead, the interest will lie on finding a function ϕ such that

$$c\phi(n, d) \leq \inf_{\hat{\theta} \in \mathcal{E}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta [\|\hat{\theta} - \theta\|_u^u] \leq C\phi(n, d),$$

where c and C are constants independent of n, or scale at most polylogarithmically with n.

Then ϕ characterizes the minimax risk of a problem, making for a quantitative analysis of a problem in a nonasymptotic way. So the minimax framework allows for an easy comparison of the efficiency of methods and, in a dual way, the difficulty of problems.

1.3 Minimax estimation

Having introduced both the estimation problem and the minimax framework, we will be tackling the problem of estimation of distributions within the minimax framework. We will first discuss in Section 1.3.1 the discrete case with multinomial and Poisson distributions. Then the continuous case will be partitioned into the study of Hölder densities in Section 1.3.2 and Besov densities in Section 1.3.3.

A lot of sections from this thesis will resonate with one another and we hint at a few of their connections here. Section 1.4 provides the testing counterparts to the estimation problems that we will study here. In Section 1.5, we will twist the study of

minimax estimation with the added constraint of privacy preservation. Finally, Hölder estimation with kernel methods will be useful for the problem of adaptive rejection sampling presented in Section 1.6.4.

1.3.1 Minimax estimation of discrete distributions

We will present minimax estimation results for Poisson distributions which can be translated into results for multinomial distributions. Similar considerations of independent Poisson samples in order to simplify the proofs are made in Chan et al. (2014) and Valiant and Valiant (2017).

We define the Poisson distribution as follows.

Definition 6. *Let $\lambda \in (0, \infty)$. A random variable $X \sim \mathcal{P}(\lambda)$ with respect to probability measure $\mathbb{P}$, if for any $k \in \mathbb{N}$,*

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

We define the multinomial distribution as well.

Definition 7. *Let $n \in \mathbb{N}$ and $q \in \mathbf{P}_d$. A random variable $\xi \sim \mathcal{M}(n, q)$ with respect to $\mathbb{P}$, if for any $i \in \{1, \dots, n\}$, $k \in \{1, \dots, d\}$,*

$$\mathbb{P}(\xi_i = k) = q_k.$$

We provide an important building block in connecting results with Poisson distributions to results with multinomial distributions. It is an equivalence result between the samples from a multinomial distribution with a random number of trials and samples from independent Poisson distributions.

Proposition 2. *Let $n \in \mathbb{R}^+$, $q \in \mathbf{P}_d$. Let $\hat{n} \sim \mathcal{P}(n)$. Let the conditional distribution of ξ be $\mathcal{M}(\hat{n}, q)$, conditionally on $\hat{n}$. For any $i \leq d$, we have $X_i = \sum_{j=1}^{\hat{n}} \mathbb{1}\{\xi_j = i\}$. Then we have independent*

$$X_i \sim \mathcal{P}(nq_i).$$

The proof of this proposition is in Section 1.7.1.

Remark 3. *Note that for any $\lambda_1 > 0$, there exists $n \in \mathbb{R}^+$ and $q_1 \geq 0$ such that $\lambda_1 = nq_1$. And for any $c > 0$, we have $\lambda_1 = (cn)(q_1/c)$. In particular, there exists c such that $\sum_{i=1}^{d} q_i/c \leq 1$.*

We provide more details on obtaining Poisson samples with large probability from multinomial samples, as well as bounding the total variation distance between multinomial distributions from a bound between Poisson distributions in Sections 2.2, 2.6.1 and 2.7.3.

We now provide upper bounds on the ℓ_u-minimax estimation risk for $0 < u \leq 2$ of the probability vector $q \in \mathbf{P}_d$ from Poisson observations. For any $i \leq d$, let $X_i \sim \mathcal{P}(nq_i)$ be independent random variables. We show that X/n is a minimax optimal estimator of q. Discussions on such a problem can be found in Berend and Kontorovich (2013) and Han, Jiao, and Weissman (2015). We provide the following upper bound on the minimax estimation risk.

Theorem 3. *Let $\mathcal{E}_d$ be the set of estimators of q using X_i for all $i \leq d$. Let $u \in (0, 2]$. We have the following upper bound on the ℓ_u-minimax estimation risk*

$$\inf_{\hat{\theta} \in \mathcal{E}_d} \sup_{\theta \in \Theta} \mathbb{E}(\|\hat{\theta} - \theta\|_{\ell_u}^u) \leq d^{(1-u/2)} n^{-u/2}.$$

We provide the proof of this Theorem in Section 1.7.2 and one can obtain the following matching lower bound.

Theorem 4. *We have the following lower bound on the ℓ_u minimax estimation risk*

$$\inf_{\hat{\theta}\in\mathcal{E}_n}\sup_{\theta\in\Theta}\mathbb{E}(\|\hat{\theta}-\theta\|_{\ell_u}^u)\geq cd^{(1-u/2)}n^{-u/2}.$$

The proof relies on reducing the estimation problem to a testing problem and finding a lower bound for that testing problem. Details on such a method can be found in Assouad (1996), Yang and Barron (1999), Tsybakov (2008), and Duchi, Jordan, and Wainwright (2013c), and a detailed lower bound proof for testing will also be provided in Section 1.4.2.

We have tackled discrete estimation, and we go on with presenting this problem in the continuous case as in Tsybakov (2008) and Giné and Nickl (2021). We will be considering continuous distributions characterized by their probability density functions. In comparison with discrete distributions, where one considers the individual probability of each category, there exists an immense diversity of probability distribution functions associated with continuous distributions to be studied. We will restrict the probability density functions to classes of regular functions. Now, the regularity of a function can be characterized globally, as with Besov functions, where there are bounds on the norm of any number of derivatives of the function. It can also be defined locally, as with Hölder functions, where the bounds are for local variations of f.

1.3.2 Hölder functions and kernel density estimation

Hölder spaces are traditional smoothness classes defined in the following way.

Definition 8. *Let $0<s\leq 2$ and $H\geq 0$. Let $u\in(0,\infty]$. A function $f:[0,1]^d\to\mathbb{R}$ is (s,H)-Hölder with respect to the $\mathbb{L}_u$-norm, if for any $(x,y)\in([0,1]^d)^2$,*

$$|f(x)-f(y)-\langle\nabla f(x),(x-y)\rangle\mathbb{1}\{s>1\}|\leq H\|x-y\|_{\mathbb{L}_u}^s.$$

Assume $X_1,\ldots,X_n$ are independent random variables with common marginal (s,H)-Hölder density f.

We follow in the steps of Tsybakov (2008) and define a kernel.

Definition 9. *A kernel $K:\mathbb{R}\to\mathbb{R}$ of order l is a function such that $u\mapsto u^jK(u)$ is integrable for any $0\leq j\leq l$ and satisfying $\int K(u)du=1$, and $\int u^jK(u)du=0$ for any $1\leq j\leq l$.*

Let $\mathbb{K}:\mathbb{R}^d\to\mathbb{R}$ be such that $\mathbb{K}=\prod_{i=1}^d K$, where K is a kernel. Then we define a kernel density estimator, the Parzen-Rosenblatt estimator as

$$\hat{f}(t)=\frac{1}{nh^d}\sum_{i=1}^n\mathbb{K}\left(\frac{X_i-t}{h}\right),$$

where $h>0$ is a bandwidth to be chosen. If $X_1,\ldots,X_n$ are fixed and $K\geq 0$, then $\hat{f}$ is a probability density function.

Remark 4. *In particular, taking $K:u\mapsto\mathbb{1}\{-1<u\leq 1\}/2$, we have $\hat{f}_n(t)=(F_n(x+h)-F_n(x-h))/2h$, where $F_n(t)=\frac{1}{n}\sum_{i=1}^n\mathbb{1}\{X_i\leq t\}$. Now, F_n is a consistent estimator of F by the law of large numbers and $\lim_{h\to 0}\frac{F(t+h)-F(t-h)}{2h}=f(t)$. This justifies the use of $\hat{f}_n$ as an estimator of f.*

The following propositions are extensions to the multidimensional case of results from Tsybakov (2008).

Proposition 3. *Assume f is a density such that $f \leq C_f \in \mathbb{R}$. Let $K : \mathbb{R} \to \mathbb{R}$ be a function such that $\int K^2(u)du < \infty$. Then the variance of the estimator is upper bounded by $C_1/(nh^d)$, for any choice of $h > 0$, where $C_1 = C_f \int \mathbb{K}^2(u)du$.*

Proposition 4. *Assume f is (s, H)-Hölder and let K be a kernel of order $\lfloor s \rfloor$ such that*

$$\int |u|^s |K(u)|du < \infty.$$

Then the squared bias is upper bounded by $C_2 h^{2s}$ for any choice of $h > 0$, where

$$C_2 = \left(\frac{L}{\lfloor s \rfloor} \int |u|^s |\mathbb{K}(u)|du \right)^2.$$

Combining both propositions, we obtain the following theorem.

Theorem 5. *If the assumptions from Propositions 3 and 4 hold, then the mean squared error is upper bounded by*

$$C_2 h^{2s} + \frac{C_1}{nh^d},$$

for any choice of $h > 0$.

So for the right choice of bandwidth for the kernel, we deduce the following upper bound.

Corollary 1. *Taking $h = \left(\frac{C_1}{2sC_2} \right)^{1/(2s+d)} n^{-1/(2s+d)}$, one obtains the following upper bound on the mean squared error*

$$Cn^{-2s/(2s+d)},$$

where C is a constant large enough.

Remark 5. *This rate is known as the nonparametric rate of estimation, in contrast with n^{-1} from the parametric rate.*

Hölder densities have been tackled in other problems as well. For details on identity testing in Hölder sets, we refer to Ingster (1987) and Arias-Castro, Pelletier, and Saligrama (2018). We now present regression, a problem related to estimation, tackled in Tsybakov (2008) and Raskutti, Wainwright, and Yu (2011).

The regression problem. The Parzen-Rosenblatt estimator only relies on the values of the random variables X_i. One can however also consider kernel methods for the regression problem where one observes $(X_1, Y_1), \ldots, (X_n, Y_n)$ such that any $i \leq n$

$$Y_i = g(X_i) + \xi_i, \tag{1.1}$$

where ξ_i are independent random variables with $\mathbb{E}(\xi_i) = 0$ and $g : [0, 1]^d \to \mathbb{R}$ is an unknown function. One particular case of interest is when $g = f$ is the density of X_i. A kernel estimator associated with this problem is the Nadaraya-Watson estimator,

$$f_n^{NW}(t) = \sum_{i=1}^{n} Y_i W_{ni}^{NW}(t),$$

where

$$W_{ni}^{NW}(t) = \frac{\mathbb{K}(\frac{X_i - t}{h})}{\sum_{j=1}^{n} \mathbb{K}(\frac{X_j - t}{h})} \mathbb{1}\left\{ \sum_{j=1}^{n} \mathbb{K}(\frac{X_j - t}{h}) \neq 0 \right\}.$$

It turns out that the minimax estimation rate for such a regression problem coincides with the minimax estimation rate of density estimation for Hölder densities.

Besides, in the particular case where f is the uniform distribution on $[0,1]^d$, the Nadaraya-Watson estimator and the Parzen-Rosenblatt estimator coincide.

Finally, the noiseless regression problem, where $\xi_i = 0$ in Equation (1.1), is an important case that has not been as extensively studied as the classical regression problem. This corresponds to estimating f from noiseless observations of f, which can also be seen as an interpolation problem with random design. Such a framework has been studied in Kohler and Krzyżak (2013), Kohler (2014), Bauer et al. (2017), and Berthier, Bach, and Gaillard (2020), and a minimax optimal method for noiseless regression of (H, s)-Hölder functions when $s \leq 1$, is proved to be the nearest neighbor estimator, defined as

$$f_n(t) = f(X_j) \mathbb{1}\left\{ j = \arg\min_l \|t - X_l\|_2 \right\}, \tag{1.2}$$

whose behavior is also studied in Kpotufe (2011), Shalev-Shwartz and Ben-David (2014), Chaudhuri and Dasgupta (2014), and Reeve and Brown (2017). For such a problem, one can reach a minimax mean squared error of $n^{-2s/d}$, which converges faster than the minimax mean squared error $n^{-2s/2s+d}$ from noisy regression. This faster convergence will be discussed in Section 1.6.6 and it will motivate our analysis of adaptive rejection sampling in Chapter 4.

1.3.3 Besov spaces and wavelet density estimation

We will now describe minimax estimation results in Besov spaces, which have been widely used in statistics since the seminal paper by Donoho et al. (1996). Indeed, thanks to interesting properties of Besov spaces from approximation theory, a large variety of signals can be dealt with, especially those built using wavelet bases. As explained in Giné and Nickl (2021), there exist multiple constructions for Besov spaces, and we will focus on one revolving around wavelet bases. Similar constructions to ours can be found in Kerkyacharian and Picard (1992), Kerkyacharian and Picard (1993), Donoho et al. (1996), Donoho and Johnstone (1998), Meyer (1990), Juditsky and Lambert-Lacroix (2004), Härdle et al. (2012), Goldenshluger and Lepski (2014), and Butucea et al. (2020).

Definition of wavelet bases. We provide the following assumptions in order to define wavelet bases.

Assumption 1. *Let ϕ be a compactly supported scaling function such that*

1. *$\{\phi(x-k), k \in \mathbb{Z}\}$ is an orthonormal family of $\mathbb{L}_2(\mathbb{R})$. Let V_0 be the subspace spanned by this basis.*
2. *For any $j \in \mathbb{Z}$, $V_j \subset V_{j+1}$, where V_j denotes the space spanned by $\{\phi_{j,k}, k \in \mathbb{Z}\}$.*
3. *We also assume the same regularity assumptions as in Donoho et al. (1996), where for some integer κ large enough, ϕ is of class $\mathcal{C}^\kappa$, ϕ and every derivative up to order κ is rapidly decreasing.*

Let W_j be defined such that $V_{j+1} = V_j \oplus W_j$. Then we define ψ, a mother wavelet such that

1. *$\{\psi(x-k), k \in \mathbb{Z}\}$ is an orthonormal basis of W_0.*
2. *$\{\psi_{j,k}, k \in \mathbb{Z}, j \in \mathbb{Z}\}$ is an orthonormal basis of $\mathbb{L}_2(\mathbb{R})$.*
3. *ψ is of class $\mathcal{C}^\kappa$, ψ and every derivative up to order κ is rapidly decreasing.*

Remark 6. *Note that $\cap_{j\in\mathbb{Z}} V_j = \{0\}$. Besides, if $\phi \in \mathbb{L}_2(\mathbb{R})$ and $\int \phi = 1$, then $\mathbb{L}_2(\mathbb{R}) = \cup_{j\in\mathbb{Z}} V_j$.*

We now present a useful proposition from Meyer (1990) for the representation of any $f \in \mathbb{L}_2(\mathbb{R})$ with wavelet bases.

Proposition 5. *If Assumption 1 holds, then for any $J \in \mathbb{Z}$*

$$\mathbb{L}_2(\mathbb{R}) = V_J \oplus W_J \oplus W_{J+1} \oplus \dots,$$

that is, for any $f \in \mathbb{L}_2(\mathbb{R})$, there exists a unique family $(\phi_J, \psi_J, \psi_{J+1}, \dots) \in V_J \times \prod_{j\geq J} W_j$, such that

$$f = \phi_J + \sum_{j\geq J} \psi_j.$$

Fix $J \in \mathbb{Z}$, we will consider

$$\{\phi_{J,k} = 2^{J/2}\phi(2^J x - k); k \in \mathbb{Z}\}, \quad \{\psi_{j,k} = 2^{j/2}\phi(2^j x - k); k \in \mathbb{Z}, j \geq J\},$$

and we denote for all $j \geq J, k \in \mathbb{Z}$,

$$\alpha_{J,k}(f) = \int 2^{J/2} f\phi(2^J(\cdot) - k), \quad \beta_{j,k}(f) = \int 2^{j/2} f\psi(2^j(\cdot) - k).$$

Example 2. *The Haar basis is defined with $\phi = \mathbb{1}[0,1]$ and $\psi = \mathbb{1}[1/2,1] - \mathbb{1}[0,1/2]$. In particular, for any $k \in \Lambda(j) = \{0, 1, \dots 2^j - 1\}$, $\phi_{j,k}$ and $\psi_{j,k}$ are supported in the dyadic interval $[k/2^j, (k+1)/2^j]$. So if $f \in \mathbb{L}_2([0,1])$, then for any $j \geq J$ and $k \notin \Lambda(j)$, $\alpha_{j,k} = \beta_{j,k} = 0$. So we will only consider in that case,*

$$\{\phi_{J,k}; k \in \Lambda(J)\}, \quad \{\psi_{j,k}; k \in \Lambda(j), j \geq J\}.$$

Definition of a Besov space. Now we will define Besov spaces with a condition on the coefficients of f in the wavelet basis.

Definition 10. *Let E_j be the associated projection operator onto V_j and $D_j = E_{j+1} - E_j$. Let $s > 0$, $1 \leq u \leq \infty$, $1 \leq h \leq \infty$, we define the Besov space as*

$$B_{s,u,h} = \left\{ f; \|E_J(f)\|_{\mathbb{L}_u} + \left(\sum_{j\geq J} [2^{js}\|D_j f\|_{\mathbb{L}_u}]^h \right)^{1/h} < \infty \right\}.$$

Then we write the following lemma from Meyer (1990) and Donoho et al. (1996) connecting the norm of f with the norm of its wavelet coefficients.

Lemma 1. *Let g be a function such that $\{g(x-k), k \in \mathbb{Z}\}$ is an orthonormal family of $\mathbb{L}_2(\mathbb{R})$ and also satisfying the third point from Assumption 1. Let $f(x) = \sum \lambda_k 2^{j/2} g(2^j x - k)$. Then there exists constants c and C such that*

$$c2^{j(1/2-1/u)}\|\lambda\|_{\ell_u} \leq \|f\|_{\mathbb{L}_u} \leq C2^{j(1/2-1/u)}\|\lambda\|_{\ell_u}.$$

This lemma will be useful in order to define the Besov space in terms of coefficients of f. We consider the projections of f, for any $j \geq J$,

$$E_J f = \sum_{k\in\mathbb{Z}} \alpha_{J,k}\phi_{0,k}, \quad D_j f = \sum_{k\in\mathbb{Z}} \beta_{j,k}\psi_{j,k}.$$

Then, by Lemma 1, we obtain the following alternative equivalent definition of a Besov space.

Definition 11. *Let $1 \leq u \leq \infty$, $1 \leq h \leq \infty$. A Besov space is defined as*

$$\left\{ f; \|\alpha_{0\cdot}\|_{\ell_u} + \left(\sum_{j\geq 0} [2^{j(s+1/2-1/u)} \|\beta_{j\cdot}\|_{\ell_u}]^h \right)^{1/h} < \infty \right\},$$

where for any $j \geq 0$, $\|\beta_{j\cdot}\|_{\ell_u} = (\sum_{k\in\mathbb{Z}} |\beta_{j,k}|^u)^{1/u}$.

And we will focus on f with a bounded support and Besov balls with a fixed radius in particular. Let $R' > 0$. Then we consider

$$\left\{ f \in \mathbb{L}_u([0,1]); \|E_0(f)\|_{\mathbb{L}_u} + \left(\sum_{j\geq 0} [2^{js} \|D_j f\|_{\mathbb{L}_u}]^h \right)^{1/h} \leq R' \right\},$$

that is, for $R > 0$

$$\widetilde{B}_{s,u,h}(R) = \{ f \in \mathbb{L}_u([0,1]); \|\alpha_{0\cdot}\|_{\ell_u} + (\sum_{j\geq 0} [2^{j(s+1/2-1/u)} \|\beta_{j\cdot}\|_{\ell_u}]^h)^{1/h} \leq R \}.$$

Example 3. *We provide a particular example of Besov ball which will be of interest in Chapter 3. For $R > 0$ and $s > 0$, the Besov ball $\widetilde{B}_{s,2,\infty}(R)$ with radius R associated with the Haar basis is defined as*

$$\widetilde{B}_{s,2,\infty}(R) = \left\{ f \in \mathbb{L}_2([0,1]); \forall j \geq 0, \sum_{k\in\Lambda(j)} \beta_{jk}^2(f) \leq R^2 2^{-2js} \right\}.$$

Now note that, if $s \leq 1$, then there is an equivalence between the definition of $\widetilde{B}_{s,2,\infty}(R)$ and the definition of the corresponding Besov space using moduli of smoothness – see e.g. Theorem 4.3.2 in Giné and Nickl (2021). And for larger s, Besov spaces defined with Daubechies wavelets satisfy this equivalence property. Further discussion on the relevance of Besov spaces to density estimation can be found in Donoho et al. (1996).

Density estimation in a Besov ball. We consider probability density functions

$$f = \sum_{k\in\mathbb{Z}} \alpha_{J,k}\phi_{J,k} + \sum_{j\geq J} \sum_{k\in\mathbb{Z}} \beta_{jk}\psi_{jk}.$$

So estimating f amounts to estimating $\alpha_{J,k}$ and β_{jk} for any k and $J \leq j \leq j_1$, where j_1 is chosen appropriately as detailed in Donoho et al. (1996). And we end up with the following estimators.

$$\hat{\alpha}_{J,k} = n^{-1} \sum_{i=1}^{n} \phi_{J,k}(X_i), \qquad \hat{\beta}_{j,k} = n^{-1} \sum_{i=1}^{n} \psi_{j,k}(X_i).$$

So one obtains the following minimax result, as detailed in Donoho et al. (1996), Härdle et al. (2012), Butucea et al. (2020), and Giné and Nickl (2021).

Theorem 6. *Let $1 \leq u \leq \infty$, $1 \leq h \leq \infty$, $s > 0$. We have the following sharp lower bound up to a logarithmic factor on the minimax $\mathbb{L}_r$-risk over $\tilde{B}_{s,u,h}(R)$,*

$$\inf_{\hat{\theta} \in \mathcal{E}_n} \sup_{\theta \in \tilde{B}_{s,u,h}(R)} \mathbb{E}_\theta[\|\hat{\theta} - \theta\|_r^r] \geq cn^{(-rs+r/u-1)/(1+2(s-1/u))} \vee (n/\log n)^{-rs/(2s+1)}.$$

Remark 7. *Note that Besov sets are function classes parametrized by smoothness parameters and, as illustrated here, the minimax rates depend exclusively on those parameters in a lot of problems.*

Then in the particular case, where $r = u$, we have the minimax rate up to a logarithmic factor,

$$n^{-us/(2s+1)}.$$

1.4 Minimax and local minimax testing of distributions

A problem related to estimation is testing the equality of two densities f and f_0, expressed in the following way in statistics.

$$\mathcal{H}_0 : f = f_0, \qquad \text{versus} \qquad \mathcal{H}_1 : f \neq f_0. \tag{1.3}$$

$\mathcal{H}_0$ is the null hypothesis and $\mathcal{H}_1$ is the alternative hypothesis. Now such a formulation of a testing problem is rather imprecise, that is the reason why, we will instead define the problem using three sets:

- Two sets of densities associated with each hypothesis.
- A set of statistics ranging in $\{0, 1\}$.

Remark 8. *One can equivalently define a statistic ranging in $\{0, 1\}$ as a test function evaluated at the observed random variables. The interpretation of the test taking value 1 is the null hypothesis being rejected. If the value is 0, then the null hypothesis is not rejected. So the set of statistics ranging in $\{0, 1\}$ characterize all possible tests built from the observations. Thus, assumptions on the observations are encoded in that set. From now on, the set of test statistics will refer to the set of statistics ranging in $\{0, 1\}$.*

The first testing problem we will consider is the classical problem of identity testing, also known as goodness-of-fit testing, or one-sample testing, where one would like to determine whether the observed random variables follow one fixed specified distribution or not. Let f_0 be a fixed probability density function. Let $\mathcal{C}$ be a set of probability density functions and dist some distance function. We now give the formulation of an identity testing problem with the following two sets associated with $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively.

$$H_0 = \{(f_0, f); f = f_0\}, \quad H_1(\rho, \text{dist}) = \{(f_0, f); f \in \mathcal{C}, \text{dist}(f_0, f) \geq \rho\},$$

where $\rho > 0$.

Remark 9. *Since f_0 is a fixed density, it is known and $\mathcal{H}_0$ is associated with a set of densities with a single element. In that case, we say that the hypothesis $\mathcal{H}_0$ is simple. On the other hand, $\mathcal{H}_1$ is composite, because it has multiple elements. So in the case that f_0 is fixed, we are considering a simple-composite testing problem.*

Let independent $X_1, \dots, X_n \sim f$ with respect to probability measure $\mathbb{P}_f$. Let $\mathcal{X}$ be the range of X_1. Let $\Phi = \{\varphi : \mathcal{X}^n \to \{0,1\}\}$. So we define the set of test statistics for this hypothesis testing problem as

$$\mathcal{T}_n = \{\hat{\theta}; \hat{\theta} = \varphi(X_1, \dots, X_n), \varphi \in \Phi\}.$$

In the same way as in estimation, the quality of a test is measured according to some criterion. We define the type I and type II error risks.

- Type I error: $\varphi(X_1, \dots, X_n) = 1$, when $f_0 = f$.
- Type II error: $\varphi(X_1, \dots, X_n) = 0$, when $f_0 \neq f$.

The significance level of φ is defined as a lower bound on $1 - \mathbb{P}_f(\varphi(X_1, \dots, X_n) = 1)$ if $f = f_0$. The power is defined as $1 - \mathbb{P}_f(\varphi(X_1, \dots, X_n) = 0)$ if $f \neq f_0$.

In practical applications, one fixes the significance level of the test, and maximizes the power then. However, both values are real numbers between 0 and 1, depending on φ, f, f_0 and the number of observations. Instead, we will define a criterion which is comparable to the mean squared error from estimation. Let H_0 be a set of (f_0, f) such that $\mathcal{H}_0$ holds. Let $H_1(\rho, \text{dist})$ be a set (f_0, f) such that f and f_0 are ρ apart according to some distance dist. Let the testing error risk be

$$\begin{aligned} &R(H_0, H_1(\rho, \text{dist}), \varphi(X_1, \dots, X_n)) \\ &= \sup_{(f_0,f)\in H_0} \mathbb{P}_f(\varphi(X_1, \dots, X_n) = 1) + \sup_{(f_0,f)\in H_1(\rho,\text{dist})} \mathbb{P}_f(\varphi(X_1, \dots, X_n) = 0). \end{aligned}$$

Remark 10. *Now in the alternative hypothesis, f and f_0 are ρ apart instead of simply being different from one another. Indeed, if f is allowed to get arbitrarily close to f_0, then the testing error risk will be 1.*

We define the separation distance as

$$\begin{aligned} &\rho_\gamma(H_0, H_1, \varphi(X_1, \dots, X_n); \text{dist}) \\ &= \inf\{\rho > 0; R(H_0, H_1(\rho, \text{dist}), \varphi(X_1, \dots, X_n)) \leq \gamma\}. \end{aligned}$$

This definition extends the notion of critical radius introduced in Ingster (1993) to the non-asymptotic framework. And we consider the minimax separation distance.

$$\rho^*_\gamma(H_0, H_1, \mathcal{T}_n; \text{dist}) = \inf_{\hat{\theta}\in\mathcal{T}_n} \rho_\gamma(H_0, H_1, \hat{\theta}; \text{dist}).$$

In the sections to come, we will be making comparisons between testing and estimation results. Note however, that the definition of minimax separation distance that we just provided and of minimax risk from Definition 5 are not homogeneous. Instead, the comparison will be made with $\inf_{\hat{\theta}\in\mathcal{E}_n} \sup_{\theta\in\Theta} \mathbb{E}_\theta[\|\hat{\theta} - \theta\|_u^u]^{1/u}$. Then, showing that testing is simpler than estimation amounts to showing that the rate associated with estimation is at least as large as the rate of $\rho^*_\gamma(H_0, H_1, \mathcal{T}_n; \text{dist})$. This is the expected result, because testing whether $f = f_0$ can be solved by estimation of f. We refer to Balakrishnan and Wasserman (2018) for a survey on hypothesis testing.

This chapter will focus on minimax testing and a local refinement of the concept of minimaxity, mostly with discrete distributions. We will first tackle upper and lower bounds on the minimax separation distance for identity testing in Sections 1.4.1 and 1.4.2. Then we will treat the related problem of minimax closeness testing in Section 1.4.3. After this, we will take a more local point of view on minimax identity testing in Section 1.4.4 and consider an upper bound for local minimax closeness

testing in Section 1.4.5. This will lead us in Section 1.4.6 to discuss the motivation of working on finding sharp local minimax rates for closeness testing of discrete distributions in Chapter 2. Finally, we will present results for minimax identity testing for continuous distributions with Besov densities in Section 1.4.7. This will give us the chance to discuss adaptivity to the unknown smoothness parameter of a Besov density. Section 1.4.7 will be useful for the study of identity testing under local differential privacy in Chapter 3.

1.4.1 An illustrating example: χ^2-test for discrete distributions

We start with the study of identity testing for discrete distributions, which has also been discussed in Valiant and Valiant (2017), Balakrishnan and Wasserman (2017a), Balakrishnan and Wasserman (2018), and Kim (2020). In this section and the next one, we will provide matching minimax upper and lower bounds focusing on uniformity testing, that is, testing whether the observed random variables are uniformly distributed. Minimax uniform testing is related to minimax identity testing, because the bounds will coincide with the ℓ_1-minimax separation distance for identity testing from Paninski (2008), as explained in Remark 11. So one can conclude that the uniform distribution is the hardest distribution to test in identity testing with ℓ_1-norm. Note however that it might not be the case for other norms, a typical example being the ℓ_2-norm. We consider a uniform probability vector p over d classes. Let $u \in (0, 2]$. Let $\Phi = \{\varphi : \mathbb{N}^d \to \{0, 1\}\}$. We define the following sets for the definition of the hypothesis testing problem.

$$H_0 = \{(p, q); p_i = 1/d, i \leq d,\ p = q\},$$

$$H_1(\rho, \|\cdot - \cdot\|_{\ell_u}) = \{(p, q); p_i = 1/d, i \leq d,\ q \in \mathbf{P}_d, \|p - q\|_{\ell_u} \geq \rho\},$$

$$\mathcal{T}_d = \{\hat{\theta}; \hat{\theta} = \varphi(X_1, \ldots, X_d), \varphi \in \Phi\},$$

where $X_i \sim \mathcal{P}(nq_i)$ is an independent random variable with respect to probability measure $\mathbb{P}_q$ for any $i \leq d$. Note that we provide results for Poisson distributions, which are connected to multinomial distributions, as hinted at in Section 1.3.1. We define the χ^2-statistic as follows

$$T = \sum_{i=1}^{d} [(X_i - np_i)^2 - X_i]. \tag{1.4}$$

The associated threshold is

$$cn\|p\|_2 = cnd^{-1/2},$$

where c is a constant depending on γ. So the test will be defined as $\mathbb{1}\{T \geq cnd^{-1/2}\}$. Note that the usual χ^2-statistic renormalizes each term by dividing by p_i and we neglect this, because here p is uniform. We will discuss this further in Section 1.4.4 tackling a local refinement to the minimax framework in identity testing.

We now present the following theorem providing an upper bound on the minimax separation distance for testing uniformity for Poisson distributions.

Theorem 7. *We have the following upper bound on the minimax ℓ_u-separation distance*

$$\rho^*_\gamma(H_0, H_1, \mathcal{T}_d; \|\cdot - \cdot\|_{\ell_u}) \leq n^{-1/2} d^{(1/u - 1/2)} (\sqrt{2} d^{-1/4} + \sqrt{68} n^{-1/2}) / \sqrt{(\gamma - 3/4)}.$$

We provide the proof of this theorem in Section 1.7.3.

Corollary 2. *If $n \geq d^{1/2}$, then*

$$\rho_\gamma^*(H_0, H_1, \mathcal{T}_d; \|\cdot - \cdot\|_{\ell_u}) \leq C d^{(1/u-3/4)} n^{-1/2},$$

where C is a constant depending on γ.

Remark 11. *Now we provide results focusing on p being the uniform distribution, but it coincides with the global ℓ_1-minimax separation distance $d^{1/4}/\sqrt{n}$ in the worst case of p presented in Paninski (2008). So the uniform distribution is indeed the worst case for identity testing.*

We compare the testing rate $d^{(1/u-3/4)}n^{-1/2}$ with the estimation rate $d^{(1/u-1/2)}n^{-1/2}$ and, as expected, this identity testing problem is simpler than estimating the density. One concludes that using the test (1.4) is more efficient than estimating q and plugging this estimate into $\|p-q\|_{\ell_u}$ in order to make a comparison with p.

1.4.2 Lower bound for identity testing of a uniform probability vector

We will provide a detailed proof on finding a lower bound for the problem of identity testing, when p is a uniform probability vector. The technique that we will describe is also discussed in Assouad (1996), Yang and Barron (1999), Baraud (2002a), Tsybakov (2008), Duchi, Jordan, and Wainwright (2013c), and Arias-Castro, Pelletier, and Saligrama (2018).

Theorem 8. *We have the following lower bound on the minimax ℓ_u-separation distance*

$$\rho_\gamma^*(H_0, H_1, \mathcal{T}_d; \|\cdot - \cdot\|_{\ell_u}) \geq (1-\gamma)^{1/2}[(24^{-1/4}n^{-1/2}d^{1/u-3/4}) \wedge (4^{-1/4}n^{-1/4}d^{1/u-1})].$$

Corollary 3. *If $n \geq d$, we have*

$$\rho_\gamma^*(H_0, H_1, \mathcal{T}_d; \|\cdot - \cdot\|_{\ell_u}) \geq c n^{-1/2} d^{1/u-3/4},$$

where c is constant depending on γ.

Presentation of a lower bound technique. We define the following distances.

Definition 12. *Let ν_0, ν_1 be probability measures.*

- *Then the total variation distance between ν_0, ν_1 is*

$$d_{TV}(\nu_0, \nu_1) = \sup_A |\nu_0(A) - \nu_1(A)|.$$

- *Besides, if ν_1 is absolutely continuous with respect to ν_0, then the chi-squared distance between ν_0 and ν_1 is defined as*

$$\chi^2(\nu_0, \nu_1) = \mathbb{E}_{\nu_0}\left[\left(\frac{d\nu_1 - d\nu_0}{d\nu_0}\right)^2\right].$$

Our lower bound technique relies on a Bayesian approach and the connection is made with the following lemma.

Lemma 2. *Let $\mathcal{C}$ be a set of densities. Let $\rho > 0$. Let $(\delta_0, \delta_1) \in [0,1)^2$ and $\gamma \in (0,1)$ such that $\gamma + \delta_0 + \delta_1 < 1$. Let $H_0 = \{(f_0, f) \in \mathcal{C}^2; f = f_0\}$, $H_1(\rho, \|\cdot - \cdot\|_{\mathbb{L}_u}) = \{(f_0, f) \in \mathcal{C}^2 : \|f - f_0\|_{\mathbb{L}_u} \geq \rho\}$ for any $\rho > 0$ and $\mathcal{T}$ a set of $\hat{\theta} = \varphi(\mathcal{X})$ such that*

$\mathcal{X} \sim f$ with respect to probability measure $\mathbb{P}_f$. Let ν_0 and $\nu_{1,\rho}$ be probability measures such that $\nu_0(H_0) \geq 1 - \delta_0$ and $\nu_{1,\rho}(H_1(\rho, dist)) \geq 1 - \delta_1$ for any $\rho > 0$. If, for some $\rho > 0$,

$$d_{TV}(\mathbb{P}_{\nu_{1,\rho}}, \mathbb{P}_{\nu_0}) < 1 - \gamma - \delta_0 - \delta_1,$$

then we have

$$\rho^*_\gamma(H_0, H_1, \mathcal{T}; dist) \geq \rho.$$

We prove this lemma in Section 1.7.4. Thus, the lower bound that is obtained heavily relies on the choice of prior distributions ν_0 and $\nu_{1,\rho}$ on (f_0, f). We prove this result now.

Besides, we connect total variation distance and chi-squared distance in the following lemma that we also prove in Section.

Lemma 3. *For two distributions ν_1, ν_0 such that ν_1 is absolutely continuous with respect to ν_0, we have*

$$d_{TV}(\nu_0, \nu_1) \leq \sqrt{\chi^2(\nu_0, \nu_1)}.$$

An important property of the chi-squared distance is its tensorization behaviour depicted in the following lemma.

Lemma 4. *For two distributions ν_1, ν_0, we have*

$$\sqrt{\chi^2(\nu_0^{\otimes d}, \nu_1^{\otimes d})} = \sqrt{(1 + \chi^2(\nu_0, \nu_1))^d - 1}.$$

Definition of a prior distribution. Let $m = n \vee d$. Assume $\Delta \leq 1/m$. Let $\eta \in \{-1, 1\}^{\lfloor d/2 \rfloor}$. For any $1 \leq i \leq d$, let

$$q_i = \begin{cases} 1/d + \eta_{\lceil i/2 \rceil}\Delta, \text{ if } i/2 \notin \mathbb{N} \text{ and } i \leq 2\lfloor d/2 \rfloor, \\ 1/d - \eta_{\lceil i/2 \rceil}\Delta, \text{ if } i/2 \in \mathbb{N} \text{ and } i \leq 2\lfloor d/2 \rfloor, \\ 1/d, \text{ if } 2\lfloor d/2 \rfloor < i \leq d, \end{cases}$$

since $p_i = 1/d$ for any $i \leq d$. Then q is a probability vector. So $\mathcal{P}(n/d)^{\otimes d}$ produces a sample set corresponding to H_0, and $\bigotimes_{i=1}^d \mathcal{P}(nq_i)$ corresponding to H_1.

So one can define the associated joint distributions Λ_0 and Λ_1. Firstly, let

$$Q_0 = \mathcal{P}(n/d)^{\otimes 2},$$

and

$$Q_1 = \frac{\mathcal{P}(n(1/d + \Delta)) \otimes \mathcal{P}(n(1/d - \Delta)) + \mathcal{P}(n(1/d - \Delta)) \otimes \mathcal{P}(n(1/d + \Delta))}{2}.$$

That is, for any $(j, k) \in \mathbb{N}^2$

$$Q_0(j, k) = (n/d)^{j+k} \frac{e^{-2n/d}}{j!k!}$$

and

$$Q_1(j, k) = [(1/d + \Delta)^j (1/d - \Delta)^k + (1/d - \Delta)^j (1/d + \Delta)^k] \frac{n^{j+k} e^{-2n/d}}{2(j!)(k!)}.$$

Then writing $\nu_0 = \mathcal{P}(n/d)$, we define

$$\Lambda_0 = \nu_0^{\otimes d} = \begin{cases} Q_0^{\otimes \lfloor d/2 \rfloor}, \text{ if } d/2 \in \mathbb{N}, \\ Q_0^{\otimes \lfloor d/2 \rfloor} \otimes \nu_0, \end{cases}$$

and

$$\Lambda_1 = \begin{cases} Q_1^{\otimes \lfloor d/2 \rfloor}, \text{ if } d/2 \in \mathbb{N}. \\ Q_1^{\otimes \lfloor d/2 \rfloor} \otimes \nu_0. \end{cases}$$

Note that Λ_0 and Λ_1 correspond to the joint distributions of a d-dimensional sample associated with the vector q under H_0 and H_1, respectively.

Information bound We have by Lemma 3 and 4,

$$d_{TV}(\Lambda_0, \Lambda_1) \leq \sqrt{\chi^2(\Lambda_0, \Lambda_1)} \leq \sqrt{(1 + \chi^2(Q_0, Q_1))^{\lfloor d/2 \rfloor} - 1}.$$

Now,

$$(1 + \chi^2(Q_0, Q_1))^{\lfloor d/2 \rfloor} - 1 \leq \exp(\lfloor d/2 \rfloor \chi^2(Q_0, Q_1)) - 1 \leq d\chi^2(Q_0, Q_1),$$

if $\chi^2(\nu_0, \nu_1) \leq 1/d$. And

$$\begin{aligned}
&\mathbb{E}_{(X_1,X_2)\sim Q_0}[(\frac{dQ_1}{dQ_0}(X_1, X_2) - 1)^2] = \sum_{j,k} \frac{[Q_1(j,k) - Q_0(j,k)]^2}{Q_0(j,k)} \\
&= \sum_{j,k} \frac{(n/p_1)^{j+k} e^{-2np_1}}{4j!k!} [(p_1 + \Delta)^j (p_1 - \Delta)^k + (p_1 - \Delta)^j (p_1 + \Delta)^k - 2p_1^{j+k}]^2 \\
&= \sum_{j,k} \frac{(np_1)^{j+k} e^{-2np_1}}{4j!k!} [(1 + \Delta/p_1)^j (1 - \Delta/p_1)^k + (1 - \Delta/p_1)^j (1 + \Delta/p_1)^k - 2]^2 \\
&\leq \sum_{j,k} \frac{(np_1)^{j+k} e^{-2np_1}}{4j!k!} [\exp((j-k)\Delta/p_1) + \exp((k-j)\Delta/p_1) - 2]^2 \\
&\leq \sum_{j,k} \frac{(np_1)^{j+k} e^{-2np_1}}{4j!k!} [2 + 2[(k-j)\Delta/p_1]^2 - 2]^2 \\
&\leq \sum_{j,k} \frac{[(k-j)\Delta/p_1]^4 (np_1)^{j+k} e^{-2np_1}}{j!k!} = [\Delta/p_1]^4 \mathbb{E}_{(X,Y)\sim \mathcal{P}(np_1)^{\otimes 2}}[(X-Y)^4].
\end{aligned}$$

We have

$$\mathbb{E}_{(X,Y)\sim \mathcal{P}(np_1)^{\otimes 2}}[(X-Y)^4] = 2\mathbb{E}(X^4) + 6\mathbb{E}(X^2)^2 - 8\mathbb{E}(X^3)\mathbb{E}(X).$$

So by Lemma 5,

$$\mathbb{E}_{(X,Y)\sim \mathcal{P}(np_1)^{\otimes 2}}[(X-Y)^4] = 12(np_i)^2 + 2np_i.$$

So

$$\begin{aligned}
\chi^2(Q_0, Q_1) = \mathbb{E}_{(X_1,X_2)\sim Q_0}[(\frac{dQ_1}{dQ_0}(X_1, X_2) - 1)^2] &\leq [\Delta/p_1]^4 (12(np_i)^2 + 2(np_i)) \quad (1.5) \\
&= \Delta^4 (12n^2d^2 + 2nd^3). \quad (1.6)
\end{aligned}$$

So $\chi^2(Q_0, Q_1) < (1-\gamma)^2/d$, if

$$\Delta^4 < (1-\gamma)^2[(24n^2d^3)^{-1} \wedge (4nd^4)^{-1}],$$

that is

$$\Delta < (1-\gamma)^{1/2}[(24^{-1/4}n^{-1/2}d^{-3/4}) \wedge (4^{-1/4}n^{-1/4}d^{-1})].$$

So

$$d_{TV}(\Lambda_0, \Lambda_1) < 1-\gamma,$$

with

$$\|p-q\|_{\ell_u} \geq (1-\gamma)^{1/2}[(24^{-1/4}n^{-1/2}d^{1/u-3/4}) \wedge (4^{-1/4}n^{-1/4}d^{1/u-1})].$$

We conclude with Lemma 2.

1.4.3 Closeness testing of discrete distributions

The problem of closeness testing is very much related to identity testing, but this time, both q and p are unknown probability vectors. It is also known as two-sample testing, homogeneity testing or equivalence testing. For simpler derivations of the upper bound, we assume that we observe two sets of samples from each distribution. Such a consideration is akin to sample splitting, that we detail further in Section 2.2. Let $\Phi = \{\varphi : \mathbb{N}^{4d} \to \{0,1\}\}$. We define the following sets for closeness testing of discrete distributions.

$$H_0 = \{(p,q) \in (\mathbf{P}_d)^2; p = q\}, \quad H_1(\rho, \|\cdot - \cdot\|_{\ell_u}) = \{(p,q) \in (\mathbf{P}_d)^2; \|p-q\|_{\ell_u} \geq \rho\}.$$

$$\mathcal{T}_d = \{\hat{\theta}; \hat{\theta} = \varphi(X_1^{(1)}, Y_1^{(1)}, X_1^{(2)}, Y_1^{(2)}, \ldots, X_d^{(1)}, Y_d^{(1)}, X_d^{(2)}, Y_d^{(2)}), \varphi \in \Phi\},$$

where for any $l \in \{1,2\}$, $i \leq d$, we consider independent $X_i^{(l)} \sim \mathcal{P}(nq_i)$, $Y_i^{(l)} \sim \mathcal{P}(np_i)$ with respect to some probability measure $\mathbb{P}_{p,q}$.

Remark 12. *Closeness testing is a composite-composite testing problem, in contrast to identity testing which has a simple null hypothesis.*

We provide the minimax rate for this problem proved in Chan et al. (2014).

Theorem 9. *The minimax ℓ_1-separation distance $\rho^*_\gamma(H_0, H_1, \mathcal{T}_d; \|\cdot - \cdot\|_1)$ has the following rate,*

$$(d^{1/2}n^{-3/4}) \vee (d^{1/4}n^{-1/2}).$$

Remark 13. *The minimax separation distance for closeness testing is worse than the one for identity testing of $d^{1/4}n^{-1/2}$, but still better than the minimax rate for estimation $d^{1/2}n^{-1/2}$. Now, both minimax rates of testing coincide, if $n \geq d$.*

We will provide a test inspired by the one presented in Section 1.4.1, but it will not rely on the knowledge of p and have the same performance if p is a uniform vector. We consider the following test statistic,

$$T = \sum_{i=1}^{d} (X_i^{(1)} - Y_i^{(1)})(X_i^{(2)} - Y_i^{(2)}), \tag{1.7}$$

together with the threshold,

$$\hat{t} = c(\sqrt{\sum Y_i^{(1)} Y_i^{(2)}} + 1),$$

where c is a constant depending on γ. So we build the following test,

$$\hat{\theta} = \mathbb{1}\{T \geq \hat{t}\}.$$

Note that the threshold is empirical, because p is unknown. Similar considerations will be made in Chapter 2.

Theorem 10. *We have the following ℓ_u-separation distance associated with the test defined above, when p is a uniform probability vector.*

$$\rho_\gamma(H_0, H_1, \hat{\theta}; \|\cdot - \cdot\|_{\ell_u}) \leq C[(n^{-1}d^{1/u-1/2}) \vee (n^{-1/2}d^{1/u-3/4})],$$

where C is a constant depending on γ.

We provide the proof of this theorem in Section 1.7.5.

Remark 14. *When p is uniform, the performance of a test built on the statistic from Equation (1.7) coincides with the minimax separation distance for identity testing presented in Theorem 7. So we manage to build a test capable of distinguishing whether p and q are uniform vectors only from comparing the samples from both distributions and is as efficient as comparing one sample set with a uniform vector. However, taking $u = 1$, the separation distance rate $(d^{1/2}n^{-1}) \vee (d^{1/4}n^{-1/2})$ from Theorem 10 is smaller than the minimax separation distance for closeness testing $(d^{1/2}n^{-3/4}) \vee (d^{1/4}n^{-1/2})$ presented in Theorem 9, if $d > n$. This demonstrates that the uniform probability vector is not the hardest case for the ℓ_1-closeness testing problem, if $d > n$. More on the global worst case study can be found in Chan et al. (2014) and an analysis of the local difficulty corresponding to any probability vector will be made in Chapter 2.*

1.4.4 Local minimax identity testing

Up until now, we have been providing minimax separation distances, either for a fixed vector p or in the worst case of p. However, it might differ greatly for other vectors p. For instance, if $p_1 = 1$ and $p_i = 0$ for any $i \in \{2, \dots, d\}$, the problem becomes much easier and we will give the minimax optimal separation distance for this case in Example 4.

Now, having a test guaranteed to be minimax optimal only for a uniform p is not very satisfying when tackling other problems. Instead, one would like minimax optimal guarantees specific to the problem at hand, that is, specific to p. Hence, the motivation for local minimax identity testing in Valiant and Valiant (2017) and Balakrishnan and Wasserman (2017a).

We present the local minimax testing problem, which has been tackled in ℓ_1 distance in the literature. Let $\Phi = \{\varphi : \mathbb{N}^d \to \{0, 1\}\}$. Let $\pi \in \mathbf{P}_d$. We define the following sets for the definition of the hypothesis testing problem.

$$H_{0,\pi} = \{(p, q); p = \pi = q\}, \quad H_{1,\pi}(\rho, \|\cdot - \cdot\|_1) = \{(p, q); p = \pi, q \in \mathbf{P}_d, \|p - q\|_{\ell_u} \geq \rho\},$$

$$\mathcal{T}_d = \{\hat{\theta}; \hat{\theta} = \varphi(X_1, \dots, X_d), \varphi \in \Phi\},$$

where $X_i \sim \mathcal{P}(nq_i)$ is an independent random variable with respect to probability measure $\mathbb{P}_{p,q}$ for any $i \leq d$.

This time, π might not be uniform and the magnitude of the probability of each class can vary a lot. We will partition the distribution into two distinct parts: the bulk and the tail.

Definition 13. *Assume without loss of generality that the probability vector π is ordered: $\pi_1 \geq \pi_2 \geq \ldots \geq \pi_d$. Let $0 \leq \kappa \leq 1$*

- *The κ-tail of π is*
$$\mathcal{D}_\kappa(\pi) = \{i; \sum_{j=i}^{d} \pi_j \leq \kappa\}.$$
- *The κ-bulk of π is*
$$\mathcal{B}_\kappa(\pi) = \{i > 1; i \notin \mathcal{D}_\kappa(\pi)\}.$$

The tests that will be proposed will be focused on either the bulk or the tail of the distributions.

Tail test. Let us first define a statistic which will be used for a test focused on the distribution tail.
$$T_{\text{tail}}(\kappa) = \sum_{j \in \mathcal{D}_\kappa(\pi)} (X_j - n\pi_j).$$
The tail test is then defined as
$$\varphi_{\text{tail}}(\sigma, \delta) = \mathbb{1}\{T_{\text{tail}}(\kappa) > \sqrt{n\pi_{\mathcal{D}_\kappa(\pi)}/\delta}\}.$$

Remark 15.
- *The tail corresponds to very low probabilities. So observing a negative term in the statistic is what one would expect for such probabilities. Conversely, a deviation from what one would expect can only correspond to $X_j \geq 1$. That is the reason why there is no need for an absolute value in the statistic.*
- *Taking the squared value like in a χ^2 test accentuates how small $n\pi_j$ is with respect to observations $X_j \geq 1$.*

We will now present two tests tackling the bulk of the distribution in different but related ways. Both rely on the squared deviation from the mean, like the tests we presented in Sections 1.4.1 and 1.4.3.

2/3-test. We now define the 2/3-test originating from Valiant and Valiant (2017).
$$\varphi_{2/3}(\kappa, \delta) = \mathbb{1}\left\{ \sum_{i \in \mathcal{B}_\kappa(\pi)} [(X_i - n\pi_i)^2 - X_i]\pi_i^{-2/3} > n \left(2 \sum_{i \in \mathcal{B}_\kappa(\pi)} \pi_i^{2/3}/\delta \right)^{1/2} \right\}.$$
It is related to a χ^2-test with a renormalization tweaked for local minimax optimal separation distance. Such a renormalization is important when summing the squared deviation for varied ranges of probabilities. Indeed, the usual χ^2-test which divides each deviation term by p places too much weight on deviations corresponding to probabilities in the lower range of the bulk, as illustrated in Example 5 from Valiant and Valiant (2017).

Max test. Another way to tackle varied ranges of probabilities is illustrated in the max test. Indeed, the usual χ^2-test is optimal when p is close to uniform. The max test from Diakonikolas and Kane (2016) and Balakrishnan and Wasserman (2017a) relies on this idea. We partition the probability vector into nearly uniform groups and apply the χ^2-test on each element of the partition.

For $j \geq 1$, let the sets $\mathcal{S}_j$ form a partition of B_σ, where

$$\mathcal{S}_j = \{t; 2^{-j}\pi_2 < \pi_t \leq 2^{1-j}\pi_2\}.$$

Let $k_{\mathcal{S}} = |\{j; \mathcal{S}_j \neq \emptyset\}|$ A test statistic is defined for each $j \geq 1$ as

$$T_j = \sum_{t \in \mathcal{S}_j} [(X_t - n\pi_t)^2 - X_t].$$

Then one can summarize the max test as

$$\varphi_{\max}(\kappa, \delta) = \bigvee_j \mathbb{1}\left\{ T_j > \sqrt{2k_S n^2 \sum_{t \in \mathcal{S}_j} \pi_t^2/\delta} \right\}.$$

The following theorem from Balakrishnan and Wasserman (2017a) gives nearly optimal bounds on the local minimax separation distance for identity testing of discrete distributions.

Theorem 11. *Let*

$$l_n(\pi) = \frac{1}{n} \vee \frac{(\sum_{i \in \mathcal{B}_{l_n(\pi)}(\pi)} \pi_i^{2/3})^{3/4}}{\sqrt{n}}, \qquad u_n(\pi) = \frac{1}{n} \vee \frac{(\sum_{i \in \mathcal{B}_{u_n(\pi)/16}(\pi)} \pi_i^{2/3})^{3/4}}{\sqrt{n}}.$$

Then the local minimax ℓ_1-separation distance for identity testing is bounded as follows,

$$c l_n(\pi) \leq \rho_\gamma^*(H_{0,\pi}, H_{1,\pi}; \|\cdot - \cdot\|_1) \leq C u_n(\pi),$$

where c and C are constants depending on γ.

This rate is attained with $\varphi_{\text{tail}} \vee \varphi_{2/3}$ as described in Valiant and Valiant (2017), as well as with $\varphi_{\text{tail}} \vee \varphi_{\max}$ from Balakrishnan and Wasserman (2017a).

Example 4. *If $\pi_1 = 1$ and $\pi_i = 0$ for any $i \in \{2, \ldots, d\}$, the local minimax separation distance is n^{-1}, which is clearly faster than the global minimax rate $d^{1/4}n^{-1/2}$ reached with π uniform.*

1.4.5 An adaptive upper bound on closeness testing

Diakonikolas and Kane (2016) present a test for the closeness testing problem, as well as an upper bound guarantee which depends on p. However, the problem is not trivially formalized in order to highlight a dependence on a fixed vector π, while keeping a composite null hypothesis. This section will focus on the test presented in Diakonikolas and Kane (2016) and the associated separation distance. Further discussion on a formalization of a local minimax optimal closeness testing problem is postponed to Chapter 2.

The construction from Diakonikolas and Kane (2016) is related to the max test presented in Section 1.4.4, but the partition of the distribution has to be done without the knowledge of p. Their test is a combination of a few different tests, based on either the squared deviation between samples like in Equation (1.7), or the absolute deviation restricted to each element of the empirical partition. In Diakonikolas and Kane (2016), the authors take a point of view that is dual to ours, where they fix the separation distance between p and q in the alternative hypothesis and look for the smallest number of samples necessary in order to be able to test between both hypotheses. We discuss in further detail the connection between both points of view

in Chapter 2. Now, at each step of their algorithm, they make careful considerations stopping the production of samples from p and q depending on whether they can conclude on whether p and q are different early on. They provide an upper bound which reaches the minimax optimal rate, when p is uniform. And it also coincides with the local minimax identity testing rate for some values of p. However, the question of whether the upper bound of Diakonikolas and Kane (2016) is local minimax optimal still remains.

1.4.6 Motivation for Chapter 2

Local minimax optimality results, where the rate depends optimally on the distribution tested, are a great step forward in comparison with global minimax results. They have already been presented for identity testing as explained in Section 1.4.4. Now, our interest will lie in proving similar results for closeness testing, where the goal is to distinguish whether two samples are drawn from the same unspecified distribution or not. Efforts have been made in Diakonikolas and Kane (2016) towards obtaining a local minimax optimal test of closeness, that we describe in Section 1.4.5. However, only an upper bound that will turn out to be suboptimal is provided. This problem is harder to grasp than local minimax identity testing, because one would like the separation distance to optimally depend on an unknown distribution. We describe the idea behind our formalization of this composite-composite hypothesis testing problem here. We consider a fixed vector $\pi \in \mathbf{P}$, on which the local minimax separation distance will depend. In order for the null hypothesis to be composite, p will not simply be taken equal to π. Instead, we will consider probability vector p's with similar profiles to π. In particular, this will include permutations of the elements of π, as well as vectors with level sets a multiplicative constant away from those of π. Having π fixed means that there exists test functions φ exploiting the information of π, which might have an impact on the local minimax separation distance. However, we provide a test which does not rely on the knowledge of π and its associated separation distance matches the local minimax rate. So the exploitation of the knowledge of π does not make the problem significantly simpler and this amounts to finding the local minimax rate of a hypothesis testing problem which relies on the observations without a priori knowledge on any vector. Hence the connection with closeness testing which, in view of the local minimax separation distance, turns out to be substantially harder than the related one-sample testing problem over a wide range of cases. We sum up our contributions presented in Chapter 2 as

- providing a lower bound on the local minimax separation distance for closeness testing.
- proposing a test providing an upper bound that nearly matches the obtained lower bound.

1.4.7 Minimax identity testing in Besov balls

In this section, we will dwell on the classical problem of identity testing of continuous distributions that has first been studied under the lens of minimax optimality in the seminal work by Ingster (1987) and Ingster (1993). This section's analysis will be a useful reference for Chapter 3 treating identity testing in Besov balls under local differential privacy. Let f_0 be the uniform density in $[0,1]$. For any $s > 0$ and $R > 0$,

we define the set $\mathcal{B}_{s,u,h}(R)$ as follows

$$\mathcal{B}_{s,u,h}(R) = \left\{ f \in \mathbb{L}_u([0,1]), f - f_0 \in \widetilde{B}_{s,u,h}(R) \right\}. \tag{1.8}$$

Let $\Phi = \{\varphi : [0,1]^n \to \{0,1\}\}$. We define the following sets describing the testing problem.

$$H_0 = \{(f_0, f); f_0 = f\},$$

$$H_1(\rho, \|\cdot - \cdot\|_{\mathbb{L}_u}) = \{(f_0, f); f \in \mathcal{B}_{s,u,h}(R), \int f = 1, \|f_0 - f\|_{\mathbb{L}_u} \geq \rho\},$$

and

$$\mathcal{T}_n = \{\hat{\theta}; \hat{\theta} = \varphi(X_1, \ldots, X_n), \varphi \in \Phi\},$$

where $X_i \sim f$ is an independent random variable with respect to probability measure $\mathbb{P}_f$ for any $i \leq n$.

We now describe the minimax optimal testing method described in Ingster (2000a) and Fromont and Laurent (2006a). One partitions $(0,1]$ uniformly and counts the number of points falling in each interval as follows. Let $L = n^{2/(4s+1)}$. Let $I_j = ((j-1)/L, j/L]$ for any $j \in \{1, \ldots, L\}$, and $n_j = \sum_{l \leq n} \mathbb{1}\{X_l \in I_j\}$. Then this problem reduces to identity testing of a discrete distribution, very much related to the one presented in Section 1.4.1. We then obtain the following result.

Theorem 12. *The minimax $\mathbb{L}_u$-separation distance is upper bounded by*

$$\rho^*_\gamma(H_0, H_1, \mathcal{T}_n; \|\cdot - \cdot\|_{\mathbb{L}_u}) \leq C n^{-s/(4s+1)},$$

where C is a constant depending on γ.

Remark 16. *This upper bound turns out to be sharp. Such a reduction of the problem of testing in Besov balls to testing discrete distributions will be useful in Chapter 3.*

Now, this method relies on the a priori knowledge of s. The question of whether one can test minimax optimally without the knowledge of s, and otherwise what is the cost of adaptivity. A similar study of adaptivity will be made in the context of differential privacy in Chapter 3.

Adaptive identity testing to the smoothness parameter s. The method presented in Ingster (2000a) and Fromont and Laurent (2006a) relies on the aggregation of multiple tests over a range of possible values of s. Now, the set of possible smoothness values is an interval. So instead of considering an uncountable set of possible values, one uses firstly the fact that the knowledge of s is only used for the number of sets in the partition of $(0,1]$. Besides, one relies on the fact that a partition with a number of sets a multiplicative constant away from the minimax optimal partition will also be minimax optimal.

So one defines

$$\mathcal{J} = \{J \in \mathbb{N}; 2^J \leq n^2\},$$

and applies the procedure previously described for any $L \in \mathcal{J}$. Now, $|\mathcal{J}| \leq 2\log_2(n)+1$. As shown in Ingster (2000a) and Fromont and Laurent (2006a), one can produce an adaptive test reaching the same performance as a test given the knowledge of s up to a multiplicative logarithmic factor in the minimax rate. A similar reasoning will be made for building adaptive tests under local differential privacy in Chapter 3. As for a continuous version of minimax closeness testing as well as adaptivity in that context

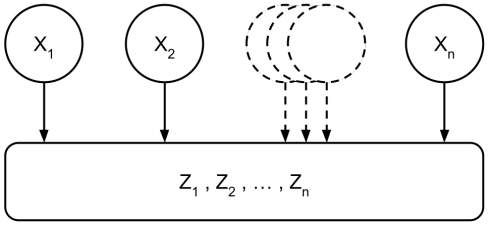

FIGURE 1.1: Graphical structure of the random variables X_i, Z_i for any $i \in \{1, \dots, n\}$ under *global* differential privacy, where X_i is unobserved and Z_i is observed.

is out of the scope of this dissertation, and we refer to Butucea and Tribouley, 2006; Fromont, Laurent, and Reynaud-Bouret, 2011 for details on that topic.

1.5 Minimax inference under local differential privacy

The concept of minimax optimality can be extended such that efficiency is not the only aspect to be considered. In particular, keeping the methods efficient while limiting the disclosure of the original data set has recently sparked a lot of interest across research communities. Various methods have been employed in preserving the information contained in a dataset and we refer to Wasserman and Zhou (2010) for a large list of methods and references in that regard. We will focus on one framework in particular, differential privacy formalized in Dwork et al. (2006b) and Dwork et al. (2006a). After formalizing differential privacy in Section 1.5.1 with a focus on non-interactive privacy, we will present some classical methods for satisfying this condition in Sections 1.5.2 and 1.5.3. Then we will consider estimation of discrete and continuous distributions under local differential privacy in Sections 1.5.4 and 1.5.5 leading up in Section 1.5.6 to the motivation for our study of identity testing under non-interactive privacy in Chapter 3.

1.5.1 Differential privacy

We begin with defining differential privacy summed up as the following condition: altering a single data point of the training set only affects the probability of an outcome to a limited degree. One main advantage of such a definition of privacy is that it can be parametrized by some positive parameter α, where α close to 0 corresponds to a more restrictive privacy condition. Let n be some positive integer and $\alpha > 0$. Let d_H be the Hamming distance defined for any $(x, x') \in (R^d)^n \times (R^d)^n$ as $d_H(x, x') = |\{i \leq n; x_i \neq x'_i\}|$. Let $X_1, \dots, X_n$ be i.i.d. random variables with respect to $\Omega, \mathcal{A}$ and taking value in $\mathbb{R}^d$. Let $Z_1, \dots, Z_n$ be random variables with respect to $\tilde{\Omega}, \tilde{\mathcal{A}}$ described by the following by some Markov kernel $Q : \tilde{\mathcal{A}}^{\otimes n} \times \Omega^n \to [0, 1]$. Q is an α-global differentially private channel with respect to $X_1, \dots, X_n$ if

$$\sup_{S \in \mathcal{Z}, (x,x') \in ((\mathbb{R}^d)^n)^2, d_H(x,x')=1} \frac{Q((Z_1, \dots, Z_n) \in S | X_1 = x_1, \dots, X_n = x_n)}{Q((Z_1, \dots, Z_n) \in S | X_1 = x'_1, \dots, X_n = x'_n)} \leq \exp(\alpha).$$

Our definition focuses on privacy mechanisms producing n random variables, which will then be used for statistical inference. And other definitions of global differential privacy can be found in the literature, for example in Dwork and Roth (2014).

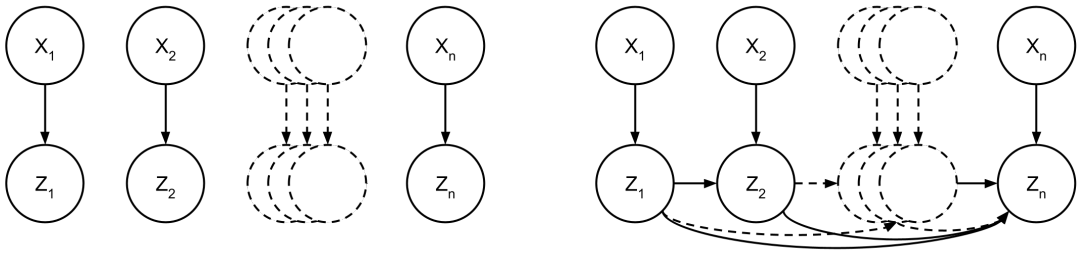

FIGURE 1.2: Graphical structures of the random variables X_i, Z_i for any $i \in \{1, \dots, n\}$ under *local* differential privacy, where X_i is unobserved and Z_i is observed. Left: non-interactive case. Right: sequentially interactive case. A similar figure can be found in Duchi, Jordan, and Wainwright (2013c).

The graphical structure associated with global differential privacy is depicted in Figure 1.1. The construction of Z_i for any $i \leq n$ can be made using all the original observations $X_1, \dots, X_n$. So this allows the use of a centralized machine handling all the original observations. This definition treats privacy in a global way with respect to the original dataset, in contrast with the privacy constraint that follows. We now define stronger assumptions corresponding to the case where one does not trust a single machine or person to handle the complete original data set. We provide the following definitions.

Definition 14. *Let $\alpha > 0$. Then*

- *Q satisfies the α-sequentially interactive local differential privacy condition, if for any $i \leq n$ the Markov kernel*

$$Q_i : \tilde{\mathcal{A}} \times \Omega \times \tilde{\Omega}^{i-1} \to [0, 1]$$

is defined such that

$$\sup_{S \in \mathcal{Z}_i, z_1, \dots, z_{i-1}, (x, x') \in (\mathbb{R}^d)^2} \frac{Q_i(Z_i \in S | X_i = x, Z_1 = z_1, \dots, Z_{i-1} = z_{i-1})}{Q_i(Z_i \in S | X_i = x', Z_1 = z_1, \dots, Z_{i-1} = z_{i-1})} \leq \exp(\alpha). \tag{1.9}$$

- *Q satisfies the α-non-interactive local differential privacy condition, if for any $i \leq n$ the Markov kernel*

$$Q_i : \mathcal{Z}_i \times \Omega \to [0, 1]$$

is defined such that

$$\sup_{S \in \mathcal{Z}_i, (x, x') \in (\mathbb{R}^d)^2} \frac{Q_i(Z_i \in S | X_i = x)}{Q_i(Z_i \in S | X_i = x')} \leq \exp(\alpha). \tag{1.10}$$

Remark 17. *Non-interactive privacy is strictly stronger assumption than sequentially interactive privacy which is itself strictly stronger than global privacy.*

We illustrate the graphical structures associated with both local differential privacy conditions in Figure 1.2. Depending on the problem, one might find different minimax results under non-interactive and sequentially interactive local differential privacy, as seen in Butucea, Rohde, and Steinberger (2020) and Berrett and Butucea (2020).

Having defined differential privacy, we provide the following proposition from Wasserman and Zhou (2010), which formally motivates differential privacy through a concept that one might describe as plausible deniability.

Proposition 6. *Suppose that Z is a private view of X. Let $X_i \sim f$ be an independent random variable for any $i \leq n$ with respect to probability measure $\mathbb{P}_f$. Let $\mathcal{Z}^{(\alpha)}(X_i)$ be the set of random variables Z_i ranging in $\mathcal{Z}_i$ such that Z_i is an α private transformation of X_i. Let $\Phi(Z_i) = \{\varphi : \mathcal{Z}_i \to \{0,1\}\}$. Let $(s,t) \in [0,1]^2$ and $i \in \{1,\dots,d\}$. We define $H_0 = \{X_i = s\}$, $H_1 = \{X_i = t\}$ and $\mathcal{T}_n^{(\alpha)} = \{\hat{\theta}; \hat{\theta} = \varphi(Z_1,\dots,Z_n), \varphi \in \Phi(Z_i), Z_i \in \mathcal{Z}^{(\alpha)}(X_i), i \leq n\}$. Let $\delta \in (0,1)$. For any $x \in [0,1]$, we write the conditional distribution of Z given $X = x$ as $Q(\cdot|x)$. Then for any test $\hat{\theta} \in \mathcal{T}_n^{(\alpha)}$ such that*

$$Q(\hat{\theta} = 1|s) \leq \delta,$$

we have

$$Q(\hat{\theta} = 0|t) \geq \delta \exp(\alpha).$$

Our results in Chapter 3 focus on non-interactive local differential privacy and we give examples of non-interactive privacy mechanisms in a few of the following sections.

1.5.2 Example of privacy mechanism: Laplace perturbation

The first privacy mechanism we present relies on the idea of adding correctly scaled noise to the original observations. We define the Laplace distribution as follows.

Definition 15. *The density of the Laplace distribution with mean μ and variance σ^2 is*

$$f(x) = \exp(-\sqrt{2}|x-\mu|/\sigma)/(\sqrt{2}\sigma).$$

The Laplace perturbation mechanism is described by Z_i for any $i \leq n$ such that

$$Z_i = X_i + \sigma W_i,$$

where W_i is an independent random variable with Laplace distribution with variance 1, and σ is a constant chosen appropriately depending on the range of values of X. Note that the distribution of Z_i will be continuous, even if X_i is discrete.

Adding independent Laplace noise is a classical privacy mechanism – see Dwork and Roth (2014). However, applying it to the correct basis with the corresponding scaling is critical in finding minimax optimal results as we will see in Section 1.5.5.

Remark 18.

- *Adding Laplace noise to the observations is part of the larger framework of exponential mechanisms presented in Wasserman and Zhou (2010). However, one notes that the Laplace perturbation mechanism fits naturally with the definition of differential privacy. For example, with Gaussian noise one can only satisfy a weaker constraint, approximate differential privacy, and the choice of an adequate variance in that case is a wide topic tackled in Dwork and Roth (2014) and Zhao et al. (2019).*

- *The definition of differential privacy mechanisms heavily relies on a bound on $|X|$, but loosening the constraint to one of approximate differential privacy, one can tackle distributions with unbounded support as well.*

1.5.3 Randomized-response-based privacy mechanisms

We will present methods based on randomized response, a classical privacy mechanism presented in the seminal paper, Warner (1965). For any $i \in \{1, \dots, n\}$, assume X_i takes value in $\mathcal{X}$ composed of d categories.

d-ary randomized response. We start with a method discussed in Kairouz, Bonawitz, and Ramage (2016). Denoting $2^{\mathcal{X}}$ as the power set of $\mathcal{X}$, one defines the d-ary randomized response mechanism as the Markov kernel $Q_i : 2^{\mathcal{X}} \times \mathcal{X}$, such that for any $(x, z) \in \mathcal{X}^2$

$$Q_i(\{z\}|x) = \frac{1}{d - 1 + \exp(\alpha)} \begin{cases} \exp(\alpha), \text{ if } z = x, \\ 1, \text{ if } z \neq x. \end{cases}$$

This mechanism keeps the values of the privatized random variables in the original finite set $\mathcal{X}$ and it can also be represented by the following matrix

$$\frac{1}{d - 1 + \exp(\alpha)} \begin{bmatrix} \exp(\alpha) & 1 & 1 & \dots & 1 \\ 1 & \exp(\alpha) & 1 & \dots & 1 \\ \dots & \dots & \dots & \dots & \dots \\ 1 & 1 & 1 & \dots & \exp(\alpha) \end{bmatrix}.$$

It amounts to lying on the value of X_i with some fixed probability depending on the parameter α of the condition privacy.

Remark 19. *If $d = 2$, this reduces to a simple randomized response method from Warner (1965).*

Randomized Aggregatable Privacy-Preserving Ordinal Response (RAPPOR). An alternative privacy mechanism based on randomized response is RAPPOR presented in Erlingsson, Pihur, and Korolova (2014). For any $i \leq n$ and $l \in \mathcal{X}$, write $\tilde{X}_{i,l} = \mathbb{1}\{X_i = l\}$. Then one defines RAPPOR as the Markov kernel $Q_i : 2^{\{0,1\}^d} \times \{0, 1\}^d$, such that for any $(x, z) \in (\{0, 1\}^d)^2$

$$Q_i(\{z\}|x) = \frac{1}{[1 + \exp(\alpha/2)]^d} \prod_{j \leq d} [\exp(\alpha/2)\mathbb{1}\{z_j = x_j\} + \mathbb{1}\{z_j \neq x_j\}].$$

That is for any $i \leq n$ one obtains the privatized random variable Z_i such that for any $(l, k) \in \mathcal{X}^2$, $Z_{i,l}$ is independent from $Z_{i,k}$ conditionally on X_i. Such a privacy mechanism can be more practically useful when d is large.

Theorem 13. *d-ary randomized response and RAPPOR are both α-non-interactive privacy mechanisms.*

The proof of this theorem can be found in Erlingsson, Pihur, and Korolova (2014) and Kairouz, Bonawitz, and Ramage (2016).

1.5.4 Multinomial estimation under local differential privacy

The mechanisms provided in Sections 1.5.2 and 1.5.3 can be applied in order to obtain minimax optimal results in various problems. We will illustrate this here with results from Duchi, Wainwright, and Jordan (2013) for estimating multinomial distributions under local differential privacy.

We now formalize the problem of minimax optimal estimation of multinomial distributions under local differential privacy. Let $\alpha \in \mathbb{R}$ such that $\alpha > 0$ and $n > 0$ be some fixed integer. Let $X_i \sim (n, q)$ be an independent random variable for any $i \leq n$ with respect to some measure $\mathbb{P}_q$. Let $\mathcal{Z}^{(\alpha)}(X_i)$ be the set of random variables Z_i ranging in $\mathcal{Z}_i$ such that Z_i is an α private transformation of X_i. Let $\Phi(Z_i) = \{\varphi : \mathcal{Z}_i \to \Theta\}$ and

$$\mathcal{E}_n^{(\alpha)} = \{\hat{\theta}; \hat{\theta} = \varphi(Z_1, \ldots, Z_n), \varphi \in \Phi(Z_i), Z_i \in \mathcal{Z}^{(\alpha)}(X_i), i \leq n\}.$$

Then we will consider the α-private minimax mean squared error,

$$\inf_{\hat{\theta} \in \mathcal{E}_n^{(\alpha)}} \sup_{q \in \mathbf{P}} \mathbb{E}_q[\|\hat{\theta} - q\|_2^2],$$

where $\mathbb{E}_q$ denotes the expectation with respect to $\mathbb{P}_q$.

In Duchi, Jordan, and Wainwright (2013a), the authors find the following bounds on the minimax mean squared error of estimating multinomial distributions under α-local differential privacy, for $\alpha \in [0, 1/4]$,

$$c\left(1 \wedge \frac{1}{\sqrt{n\alpha^2}} \wedge \frac{d}{n\alpha^2}\right) \leq \inf_{\hat{\theta} \in \mathcal{E}_n^{(\alpha)}} \sup_{q \in \mathbf{P}} \mathbb{E}_q[\|\hat{\theta} - q\|_2^2] \leq C\left(1 \wedge \frac{d}{n\alpha^2}\right),$$

where c and C are constants. Their upper bound can be obtained using the Laplace perturbation mechanism presented in Section 1.5.2 or the d-ary randomized response mechanism presented in Section 1.5.3.

Remark 20. *Comparing this rate with the non-private estimation rate presented in Section 1.3.1, we see a simple multiplicative degradation of the rate by the constant α^2.*

We now present a result of Duchi, Jordan, and Wainwright (2013c), which is key in obtaining the lower bound presented above. The following theorem provides a quantitative bound between the symmetrized Kullback-Leibler divergence of the privatized marginals and the total variation distance of the original distributions.

Theorem 14. *Let $\alpha \geq 0$ and Q an α-differential private channel. Let P_j be a distribution and $M_j(S) = \int_{\mathcal{X}} Q(S|x)dP_j(x)$, for $j \in \{1, 2\}$. Then*

$$D_{KL}(M_1, M_2) + D_{KL}(M_2, M_1) \leq (4 \wedge \exp(2\alpha))(\exp(\alpha) - 1)^2 \|P_1 - P_2\|_{TV}^2,$$

where D_{KL} is the Kullback-Leibler divergence.

The above theorem quantifies how much information is lost when transforming the data with an α-differential privacy channel, which contracts the space of probability measures. This strong data processing inequality can be used for providing lower bounds, combined with the technique provided in Section 1.4.2. For example, in Duchi, Jordan, and Wainwright (2013c), this technique is used for providing a sharp lower bound on the minimax risk of estimating a one-dimensional mean. It can also be used in order to obtain a *suboptimal lower bound* for the problem of identity testing under local differential privacy, as explained in Section 3.8.

1.5.5 Density estimation over Besov balls under local differential privacy

We illustrate the application of the Laplace perturbation mechanism from Section 1.5.2 for minimax estimation of a Besov density under local differential privacy, tackled in Butucea et al. (2020).

We now formalize the problem of density estimation in Besov balls under local differential privacy. Let $\alpha \in \mathbb{R}$ such that $\alpha > 0$ and $n > 0$ be some fixed integer. Let $X_i \sim f$ be an independent random variable for any $i \leq n$. Let $\mathcal{Z}^{(\alpha)}(X_i)$ be the set of random variables Z_i ranging in $\mathcal{Z}_i$ such that Z_i is an α private transformation of X_i. Let $\Phi(Z_i) = \{\varphi : \mathcal{Z}_i \to \Theta\}$ and

$$\mathcal{E}_n^{(\alpha)} = \{\hat{\theta}; \hat{\theta} = \varphi(Z_1, \dots, Z_n), \varphi \in \Phi(Z_i), Z_i \in \mathcal{Z}^{(\alpha)}(X_i), i \leq n\}.$$

So we will be considering the following minimax $\mathbb{L}_r$-risk

$$\inf_{\hat{f} \in \mathcal{E}_n^{(\alpha)}} \sup_{f \in \widetilde{B}_{s,u,h}(R)} \mathbb{E}_f[\|\hat{\theta} - q\|_r^r],$$

where $\mathbb{E}_f$ denotes the expectation with respect to $\mathbb{P}_f$.

Theorem 15. *We have the following lower bound,*

$$\inf_{\hat{f} \in \mathcal{E}_n^{(\alpha)}} \sup_{f \in \widetilde{B}_{s,u,h}(R)} \mathbb{E}_f[\|\hat{\theta} - q\|_r^r]$$
$$\geq c \left[(n(e^\alpha - 1)^2)^{-rs/(2s+2)} \vee \left(\frac{n(e^\alpha - 1)^2}{\log((n(e^\alpha - 1)^2))} \right)^{-r(s-1/u+1/r)/(2s-2/u+2)} \right],$$

where c is a constant.

The upper bound from Butucea et al. (2020) nearly matching this lower bound relies on the construction of Besov spaces using wavelets, as described in Section 1.3.3. Indeed their privacy mechanism amounts to adding Laplace noise to the coefficients of X in a wavelet basis. Using non-linear estimators, Butucea et al. (2020) find a sharp upper bound, but we will restrict our presentation to their linear estimator.

Privacy mechanism associated with linear wavelet estimators. We remind the following decomposition

$$f = \sum_{k \in \mathbb{Z}} \alpha_{J,k} \phi_{J,k} + \sum_{j \geq J} \sum_{k \in \mathbb{Z}} \beta_{j,k} \psi_{j,k}.$$

Let J and j_1 be integers to be determined later on. For $i \in \{1, \dots, n\}$, $j \in \{J - 1, \dots, j_1\}$, we denote by $\Lambda(j)$ the set of coefficients $\beta_{j,k}$ which are potentially different from 0, and we define

$$Z_{i,j,k} = \begin{cases} \phi_{J,k}(X_i) + \sigma_{J-1} W_{i,j,k}, \text{ if } j = J - 1, k \in \Lambda(j) \\ \psi_{j,k}(X_i) + \tilde{\sigma}_j W_{i,j,k}, \text{ if } j \in \{J, \dots, j_1\}, k \in \Lambda(j). \end{cases}$$

where $W_{i,j,k}$ are independent Laplace distributed random variables with variance 1 and

$$\sigma_{J-1} = 4 c_A \|\phi\|_\infty 2^{J/2} / \alpha,$$
$$\tilde{\sigma}_j = 4\sqrt{2} c_A \|\psi\|_\infty 2^{j_1/2} / (\alpha(\sqrt{2} - 1)),$$

where $j \in \{J, \dots, j_1\}$, $c_A = 2\lceil A \rceil + 1$, assuming that the support of ϕ and ψ is in $[-A, A]$. The proof that one indeed defines an α-private channel can be found in Butucea et al. (2020), and we detail it as well for the Haar basis in Section 3.4.1.

The decomposition of f suggests taking in the case of linear estimators, $J = 0$, and $\psi_{-1,k} = \phi_{0,k}$, then

$$\hat{f} = \sum_{j=-1}^{j_1} \sum_{k \in \Lambda(j)} \hat{\beta}_{jk} \psi_{jk},$$

where $\hat{\beta}_{jk} = \sum_{i=1}^{n} Z_{ijk}$. Butucea et al. (2020) obtain the following upper bound

$$\inf_{\hat{f} \in \mathcal{E}_n^{(\alpha)}} \sup_{f \in \widetilde{B}_{s,u,h}(R)} \mathbb{E}_f[\|\hat{\theta} - q\|_u^u] \leq C \left[(n\alpha^2)^{-us/(2s+2)} \vee n^{-us/(2s+1)} \right],$$

with the Laplace perturbation mechanism and wavelet estimators, choosing j_1 such that 2^{j_1} is of order $(n\alpha^2)^{1/(2s+2)} \wedge n^{1/(2s+1)}$.

Remark 21. *If $\alpha^2 \geq n^{1/(2s+1)}$, then the estimator using privatized observations is nonetheless able to reach the non-private minimax optimal rate. But if $\alpha^2 < n^{1/(2s+1)}$, we notice a polynomial degradation of the rate in n from the non-private minimax optimal rate.*

1.5.6 Motivation for Chapter 3

Local differential privacy and its consequences have been the topic of a lot of recent research in statistics and machine learning. Such a quantitative formulation of the privacy constraint is interesting, because one can then measure the theoretical impact of differential privacy on the minimax rates depending on some constant α parametrizing the privacy constraint. Results have been found for the problem of minimax estimation under local differential privacy already for multinomial distributions in Duchi, Wainwright, and Jordan (2013) presented in Section 1.5.4 and for Besov densities in Butucea et al. (2020) that we describe in 1.5.5. And we will be interested in minimax testing under local differential privacy, which around the time of the publication of Lam-Weil, Laurent, and Loubes (2020) was still largely uncharted territory. A few related works like Berrett and Butucea (2020) have since been released. We will examine in Chapter 3 the impact of non-interactive privacy on identity testing, i.e. the statistical problem assessing whether sample points are generated from a fixed density f_0, or not. But the observations are kept hidden and replaced by a stochastic transformation satisfying the local differential privacy constraint. In this setting, we propose a testing procedure which is based on an estimation of the quadratic distance between the density f of the unobserved samples and f_0. This is related to the test presented in Section 1.4.1. We will then use the Laplace perturbation mechanism presented in Section 1.5.2, where noise is added to the projection onto a wavelet basis in a similar way to Butucea et al. (2020) detailed in Section 1.5.5. Finally, using Theorem 14 from Duchi, Jordan, and Wainwright (2013c) discussed in Section 1.5.4, one can obtain a suboptimal lower bound for identity testing under local differential privacy. But we will make a new construction for a lower bound under local differential privacy, and instead rely on the technique presented in Section 1.4.2 with careful consideration of how much information is lost in the best case under non-interactive privacy, specifically in the problem considered. Such a result is useful in proving that there is a gap in the identity testing problem between non-interactive and sequentially interactive mechanisms that cannot be bridged, as explained in Butucea, Rohde, and Steinberger (2020) where the authors discuss their results in relation to ours in Lam-Weil, Laurent, and Loubes (2020). We now summarize our contributions.

- We provide the first minimax lower bound for the problem of identity testing under non-interactive privacy constraint over Besov balls.
- We present the first minimax optimal test with the associated local differentially private channel in this continuous setting.
- The test is made adaptive to the Besov smoothness parameter of the unknown density up to a logarithmic term.
- A minimax optimal test under local privacy can be derived for multinomial distributions as well.

1.6 Bandit theory and adaptive rejection sampling

We will be studying estimation problems, where the observations $X_1, \dots, X_n$ are sequentially observed. We first present the framework of stochastic multi-armed bandits for which a very complete presentation can also be found in Lattimore and Szepesvári (2020). This section will be organised as follows. After introducing the bandit framework in Section 1.6.1, we will present some fundamental methods in Sections 1.6.2 and 1.6.3. Now, techniques from bandit theory can prove useful in numerous sequential problems. We will direct our attention to the problem of adaptive rejection sampling that we introduce in Section 1.6.4. The adaptive rejection sampling method from Erraqabi et al. (2016) will be presented in Section 1.6.5. However, our discussion in Section 1.6.6 points out the flaws in the existing results motivating our minimax study of adaptive rejection sampling in Chapter 4.

1.6.1 Stochastic multi-armed bandit

A stochastic multi-armed bandit problem is a collection of real-valued distributions indexed by a finite set I, each index corresponding to an arm of the bandit problem. Let T be a positive integer. At each timestep $t \leq T$, one chooses an index $A_t = i \in I$ and one observes a sample X_t from the distribution indexed by i. X_t is interpreted as the reward obtained at time step t and one would like to maximize the cumulative reward at T. Let μ_i be the expected value of the distribution associated with the i-th arm, that is $\mu_i = \mathbb{E}(X_t | A_t = i)$ for any $t \leq T$.

Then bandit algorithms are strategies determining the choice of arm at every time step, and one measures the quality of one such algorithm by its mean reward put in perspective with the mean of the best arm. We formalize this idea with the cumulative regret

$$R_T = \sum_{t \leq T} (\sup_{i \in I} \mu_i - \mathbb{E}(X_t))$$

that we wish to minimize. So summing all instantaneous regrets, instead of considering the simple regret at the last step, means that a balance has to be struck between exploring to find the best arm and exploiting the accumulated knowledge about the arms.

The following proposition reformulates the cumulative regret and reduces the problem of bounding the regret to bounding the number of times a suboptimal arm is picked.

Proposition 7. *Let* $\Delta_j = \sup_{i \in I} \mu_i - \mu_j$. *Then*

$$R_T = \sum_{j \in I} \Delta_j \sum_{t \leq T} \mathbb{E}(\mathbb{1}\{A_t = j\}).$$

The proof of this proposition is given in Section 1.7.6.

In the next few sections, we will present some fundamental bandit algorithms. For each algorithm, we pull all the arms once following the index order. Assume from now on that $I = \{1, 2\}$, so this is a bandit problem with 2 arms, with the following distributions $\text{Ber}(\mu_1)$ and $\text{Ber}(\mu_2)$. Let $\mu_1 = 1/2$ and $\mu_2 = 1/2 - \Delta < 1/2$.

1.6.2 The ε-greedy bandit algorithm

We start with the ε-greedy bandit algorithm, for some $\varepsilon \in [0, 1]$. Its choice of arm at every time step t is described by the following mixture of distributions. For any $j \in \{1, 2\}$, we define the probability of picking arm j at time step $t > 2$,

$$\mathbb{P}(A_t = j) = \varepsilon/2 + (1 - \varepsilon)\mathbb{1}\left\{ j = \arg\max_{i \in I} \frac{\sum_{l \leq t} X_l \mathbb{1}\{i = A_t\}}{\sum_{l \leq t} \mathbb{1}\{i = A_t\}} \right\}.$$

The intuition is that with probability ε, we pick an arm at random, exploring in order to discover which arm has a higher mean. With probability $1 - \varepsilon$, we pick the arm with the best mean estimate.

In particular, taking $\varepsilon = 0$ corresponds to a fully greedy algorithm, also known as follow-the-leader. It relies on the fact that a good estimate of the mean is the empirical average. So one could draw the arm with the best average in order to maximally exploit the knowledge available. However, one might miss out on other better arms for which we never gain the information that they are better. Indeed, as explained in Theorem 16, we end up with a cumulative regret linear in T, that is, the worst rate possible.

On the other hand, the best way to know which arm has the largest mean is to explore the arms as much as possible. This corresponds to $\varepsilon = 1$. In fact, this would be a very viable strategy in a simple regret setting, where one is judged only on outputting the best arm after T time steps. So in the simple regret setting, there is no cost for exploring at time step $t < T$. But for a cumulative regret setting, one also obtains a linear regret. And actually, the following theorem will prove that setting ε to any constant in $[0, 1]$ leads to a linear cumulative regret.

Theorem 16. *Set $0 \leq \varepsilon \leq 1$ constant. Then, for some constant c depending on Δ and ε,*

$$R_T \geq cT.$$

The proof of this theorem is given in Section 1.7.7.

As explained in Auer, Cesa-Bianchi, and Fischer (2002) and Lattimore and Szepesvári (2020), this algorithm can still reach sublinear regret, if one takes ε decreasing over time, typically with a rate of order t^{-1}. The intuition is that, as we gain more information on all the arms, we explore less in favor of exploiting the accumulated knowledge more.

Remark 22. *The average budget allocated to complete exploration is εT. So one could consider the ε-greedy algorithm as a randomized version of the explore-and-commit strategy, where the complete exploration stage is done for a fixed number of initial timesteps, before pulling the arm with the best average all the time afterwards. Such a strategy is connected to the adaptive rejection sampling method from Erraqabi et al. (2016), that we will present in Section 1.6.5.*

1.6.3 The upper confidence bound algorithm

We now present the upper confidence bound algorithm, also described in Lattimore and Szepesvári (2020). Let $1/\delta = 1 + t\log(t)^2$. Then at every time step t, the algorithm deterministically chooses

$$A_t = \arg\max_{i \in I} \left\{ \frac{\sum_{j=1}^{t-1} X_j \mathbb{1}\{A_j = i\}}{\sum_{j=1}^{t-1} \mathbb{1}\{A_j = i\}} + \sqrt{\frac{2\sigma^2 \log(1/\delta)}{\sum_{j=1}^{t-1} \mathbb{1}\{A_j = i\}}} \right\}.$$

This algorithm picks the arm with the highest upper confidence bound, instead of the highest average. So it favours arms with high empirical means or only sampled a few times. One can describe the upper confidence bound algorithm as optimism in the face of uncertainty. Indeed, the more uncertain a direction, the higher the potential reward and so the greater the incentive to explore in that direction. This makes for a very organic exploration-exploitation tradeoff.

And using such a strategy one obtains the following results.

Theorem 17. *Following the upper confidence bound strategy, we have the following upper bound on the expected number of times a suboptimal arm will be pulled,*

$$\mathbb{E}(T_2(T)) \leq 8\alpha \log T/\Delta^2 + \alpha/(\alpha - 2),$$

for any $\alpha > 2$.

This result is translated to a regret in the following corollary.

Corollary 4. *The regret is then upper bounded by*

$$4\sqrt{2\alpha T \log(T)} + \alpha\Delta/(\alpha - 2),$$

for any $\alpha > 2$.

Our presentation of the proof of Theorem 17 will rely on the one from Orabona (2019). The proofs for both the theorem and its corollary can be found in Section 1.7.8.

Remark 23. *This upper bound is minimax optimal, as seen in Lattimore and Szepesvári (2020) and Orabona (2019).*

This concludes our quick overview of bandit theory, which will help contextualize its application in minimax adaptive rejection sampling. A lot of variations of the bandit problem have been studied and we refer to Lattimore and Szepesvári (2020) for more on that topic. We will now apply the bandit framework to the problem of adaptive rejection sampling.

1.6.4 Definition of minimax adaptive rejection sampling

We motivate the sampling problem as follows. We consider a target density f we wish to generate independent samples from. However, f is a density we cannot directly sample from. We also assume that we can evaluate f everywhere, but it is costly enough to be the computational bottleneck. So the number of evaluations of f would have to be minimized while maximizing the number of independent samples produced.

Rejection sampling. A classical method for solving this problem is rejection sampling. Assume that you know a constant M and a density g easy to sample from such that $f \leq Mg$. Then one can generate independent samples from f repeating

Algorithm 1 Rejection Sampling Step with (f, g, M) : RSS(f, g, M)

Input: target density f, proposal density g, rejection constant M.
Output: Either a sample X from f, or nothing.
Sample $X \sim g$ and $U \sim \mathcal{U}_{[0,1]}$.
if $U \leq \frac{f(X)}{Mg(X)}$ **then**
  output X.
else
  output $\emptyset$.
end if

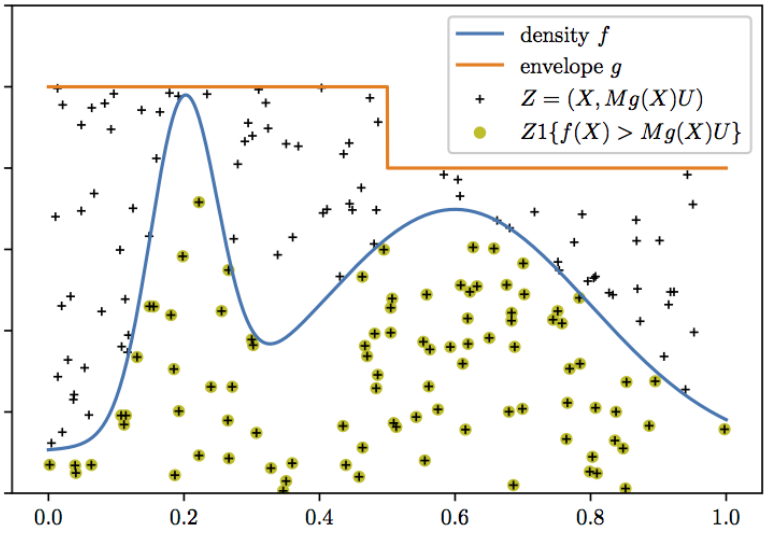

FIGURE 1.3: Geometrical interpretation of Rejection Sampling

the Rejection Sampling Step from Algorithm 1 n times. Then all accepted samples are independent samples from f. Indeed, consider the variable $Z = (X, Mg(X)U)$, where X, M, g and U are defined as in Algorithm 1. As shown in Figure 1.3, Z has a uniform distribution on the region under the graph of Mg, and the sample is accepted if it falls into the region under the graph of f. Conditional to acceptance, Z is then drawn uniformly from the area under the graph of f. Thus X is drawn from the distribution with density f. The acceptance probability is the ratio of the two areas, $1/M$. This means that the closer g is to f and M to 1, the more often samples are accepted. The goal is hence to find a good envelope of f in order to obtain a number of rejected samples as small as possible. In the absence of prior knowledge on the target density f, the proposal is typically the uniform density on a set including the support of f (here assumed to be compact), and the rejection constant M is set as an upper bound on f. Consequently, this method leads to rejecting many samples in most cases and f is evaluated many times uselessly. We presented rejection sampling with the same envelope Mg at every step. However, one could change the envelope over time. Besides, past evaluations of f can be used to define a better envelope over time. This motivates the study of adaptive rejection sampling.

Adaptive Rejection Sampling (ARS). The framework that we consider is *sequential and adaptive rejection sampling*. Set $\mathcal{S} = \emptyset$ and let n be the budget. An ARS method sequentially performs n steps At each step $t \leq n$, the samples $\{X_1, \dots, X_{t-1}\}$ collected until t, each in $[0,1]^d$, are known to the learner, as well as their images by f. The learner A chooses a positive constant M_t and a density g_t defined on $[0,1]^d$ that both depend on the previous samples and on the evaluations of f at these points $\{(X_1, f(X_1)), \dots, (X_{t-1}, f(X_{t-1}))\}$. Then the learner A performs a rejection sampling step with the proposal and rejection constant (g_t, M_t), as depicted in Algorithm 1. It generates a point X_t from g and a variable U_t that is independent from every other variable and drawn uniformly from $[0,1]$. X_t is accepted as a sample from f if $U_t \leq \frac{f(X_t)}{M_t g_t(X_t)}$ and rejected otherwise. If it is accepted, the output is X_t, otherwise the output is $\emptyset$. Once the rejection sampling step is complete, the learner adds the output of this rejection sampling step to $\mathcal{S}$. The learner iterates until the budget n of evaluations of f has been spent.

Definition 16. (Class of Adaptive Rejection Sampling (ARS) Algorithms) *An algorithm A is an ARS algorithm if, given f and n, at each step $t \in \{1 \dots n\}$:*

- *A chooses a density g_t, and a positive constant M_t, depending on $\Big\{(X_1, f(X_1)), \dots, (X_{t-1}, f(X_{t-1}))\Big\}$.*
- *A performs a Rejection Sampling Step with (f, g_t, M_t).*

The objective of an ARS algorithm is to sample as many i.i.d. points according to f as possible.

Theorem 18. *Given access to a positive, bounded density f defined on $[0,1]^d$, any Adaptive Rejection Sampling algorithm (as described above) satisfies:*
if $\forall t \leq n,\ \forall x \in [0,1]^d,\ f(x) \leq M_t g_t(x)$, the output $\mathcal{S}$ contains i.i.d. samples drawn according to f.

The proof of this theorem is given in Section 1.7.9. It gives a sufficient condition under which an adaptive rejection sampling algorithm is a *perfect sampler*, that is, it outputs i.i.d. samples. In contrast to this, the popular Markov Chain Monte Carlo class of methods relies on the construction of a Markov chain in order to produce samples, which are thus not independent.

Definition of the loss. Let us define the number of samples which are known to be independent and sampled according to f based on Theorem 18: $\hat{n} = \#\mathcal{S} \times \mathbf{1}\{\forall t \leq n :\ f \leq M_t g_t\}$. We define the loss of the learner as $L_n = n - \hat{n}$. This is justified by considering two complementary events. In the first, the rejection sampling procedure is correct at all steps, that is to say all proposed envelopes bound f from above; and in the second, there exists a step where the procedure is not correct. In the first case, the sampler will output i.i.d. samples drawn from the density f. So the loss of the learner L_n is the number of samples rejected by the sampler. In the second case, the sampler is not trusted to produce correct samples. So the loss becomes n,. Finally, we note that the rejection rate is L_n/n.

Remark on the loss. Let $\mathcal{A}$ be the set of ARS algorithms defined in Definition 16. Note that for any algorithm $A \in \mathcal{A}$, the loss $L_n(A)$ is related to the cumulative regret defined in Section 1.6. Indeed, a learner that can sample directly from f would not reject a single sample, and would hence achieve $L_n^* = 0$. So $L_n(A)$ is equal to the

difference between $L_n(A)$ and L_n^*. Hence $L_n(A)$ is the cost of not knowing how to sample directly from f.

We now define the minimax expected cumulative regret for adaptive rejection sampling for $0 < s$,

$$\inf_{A \in \mathcal{A}} \sup_{f} \mathbb{E}_f(L_n(A)),$$

where the sup is taken over all (s, H)-Hölder densities and $\mathbb{E}_f(L_n(A))$ is the expectation of the loss of A on the problem defined by f. It is taken over the randomness of the algorithm A.

1.6.5 A suboptimal upper bound for adaptive rejection sampling

We describe the method presented in Erraqabi et al. (2016), pliable rejection sampling (PRS), as well as their upper bound on the minimax regret. It allows sampling from multivariate densities satisfying mild regularity assumptions. Assume f is bounded, (s, H)-Hölder for some $0 < s \leq 2$. PRS is a two-step adaptive algorithm, based on the use of non-parametric kernel methods for estimating the target density. Take $K = \prod_{i=1}^{d} K_0$, where K_0 is the Gaussian kernel:

$$K_0 : x \mapsto \frac{\exp(-x^2/2)}{\sqrt{2\pi}}.$$

Assume that PRS is given a budget of n evaluations of the function f.

In order to describe the algorithm by Erraqabi et al. (2016), we present the construction of their envelope at every time step. Set $N = n^{(2s+d)/(3s+d)}$. For every $t \leq N$, g_t is a uniform density in $[0, 1]^d$ and $M_t = 1 + H$. So for a density f defined on a compact domain, PRS first evaluates f on a number $N < n$ of points uniformly drawn in the domain of f.

At time step N, it uses these evaluations to produce an estimate of the density f using Kernel regression.

$$\hat{f}(x) = \frac{1}{Nh^d} \sum_{k=1}^{N} f(X_i) K\left(\frac{X_i - x}{h}\right).$$

Then it builds a proposal density using a high probability confidence bound on the estimate of f. So for $N < t \leq n$, we define a proposal distribution

$$g_t = \frac{\hat{f} + r_N}{\frac{1}{N}\sum_{i=1}^{N} f(X_i) + r_N} \mathbb{1}[0, 1]^d,$$

where for some constant C'',

$$r_N = C''(\log(Nd/\delta)/N)^{s/2s+d}.$$

The associated rejection constant is then the renormalization constant,

$$M_t = \frac{\frac{1}{N}\sum_{i=1}^{N} f(X_i) + r_N}{\frac{1}{N}\sum_{i=1}^{N} f(X_i) - 5r_N}.$$

PRS then applies rejection sampling $n - N$ times using such an envelope.

The proposal density multiplied by the rejection constant is proven to be with high probability a correct envelope, i.e., an upper bound for f. So with high probability,

this method obtains a *perfect sampler*, i.e., a sampler which outputs *i.i.d. samples from the density* f.

Remark 24.
- *Their algorithm only relies on two different envelopes. It is akin to an explore-then-commit bandit algorithm described in Section 1.6.2. This hints at some untapped potential in using such a strategy in this sequential setting.*
- $\hat{f}$ *is a Kernel estimate related to the one presented for nonparametric regression in Section 1.3.2.* h *defines a bandwidth and they end up considering* $h = (\log(N/\delta)/N)^{1/(2s+1)}$, *which corresponds to the size of a bandwidth in classical density estimation with noisy observations presented in Corollary 1. However, in the usual sampling setting, which is the one they consider in Erraqabi et al. (2016),* f *is evaluated without noise. This leads to suboptimal results in the same way as noiseless regression improves on the rate of noisy regression.*

Using their algorithm they obtain the following upper bound on the minimax regret in the class of adaptive rejection sampling algorithms.

Theorem 19. *Let* $\delta \in (0,1)$, $0 < s \leq 2$ *and* $H \geq 0$. *Then the minimax regret is upper bounded by*

$$\inf_{A \in \mathcal{A}} \sup_{f \in \mathcal{F}} \mathbb{E}_f(L_n(A)) \leq C \log(nd)^{s/(3s+d)} n^{1-s/(3s+d)},$$

where $\mathcal{F}$ *is the set of* (s, H)*-Hölder densities, and* C *is a constant depending on* s *and* d.

This means that it asymptotically accepts almost all the samples. However, there is no guarantee that this rate might not be improved using another algorithm. Indeed, no lower bound on the rejection rate over all algorithms is presented. This is a motivation for us to complete the analysis of adaptive rejection sampling and improve on the results from the literature.

1.6.6 Motivation for Chapter 4

With rejection sampling, a fundamental Monte Carlo method, one can sample from distributions admitting a probability density function that can be evaluated exactly at any given point. However, if it is not properly tuned, this technique leads to a high rejection rate, that is, a lot of wasted samples. Based on the principle of adaptively estimating the density by a simpler function using the information of the previous samples, we formalized the problem of adaptive rejection sampling in Section 1.6.4. Now, most results from the literature either rely on strong assumptions or lack proper theoretical performance guarantees. We will study this problem under the lens of the minimax framework and as explained in Section 1.6.5, there already exists an upper bound for minimax adaptive rejection sampling from Erraqabi et al. (2016). But the authors use tools which might not be the most well adapted to this problem leading to an upper bound on the minimax rate that can be improved upon. So crossing ideas from bandit theory and minimax statistical inference, we provide in Chapter 4

- a new adaptive rejection sampling algorithm yielding the best existing upper bound on the minimax cumulative regret.
- a lower bound on the class of all adaptive rejection sampling algorithms and all s-Hölder densities.

1.7 Proofs for Chapter 1

1.7.1 Proof of Proposition 2

We first show that $X_j \sim \mathcal{P}(nq_j)$ for any j. Let $\mathbf{i}$ be such that $\mathbf{i}^2 = -1$. We write the characteristic function of X_j for any t:

$$\mathbb{E}(e^{\mathbf{i}tX_1}) = \mathbb{E}(\mathbb{E}(e^{\mathbf{i}tX_1}|\hat{n})) = \sum_{j\geq 0}((1+q_1(e^{\mathbf{i}t}-1))^j \frac{n^j}{j!}e^{-n} = \exp(nq_1(e^{\mathbf{i}t}-1)),$$

which corresponds to the characteristic function associated with $\mathcal{P}(nq_1)$.

Then let us prove that $(X_j)_{j\leq d}$ are independent. We have

$$\begin{aligned}
\mathbb{P}(X_1 = x_1, \ldots, X_d = x_d) &= \sum_{l\geq 0} \mathbb{P}(X_1 = x_1, \ldots, X_d = x_d|\hat{n} = l)\mathbb{P}(\hat{n} = l) \\
&= \sum_l \frac{l!}{x_1!\ldots x_d!} q_1^{x_1}\ldots q_d^{x_d} \mathbf{1}\{\sum_{i=1}^d x_i = l\}\frac{n^l}{l!}e^{-n} \\
&= \frac{(\sum_i x_i)!}{x_1!\ldots x_d!} q_1^{x_1}\ldots q_d^{x_d}\frac{n^{\sum_{i=1}^d x_i}}{(\sum_{i=1}^d x_i)!}e^{-n} \\
&= \prod_{i=1}^d \mathbb{P}(X_i = x_i).
\end{aligned}$$

1.7.2 Proof of Theorem 3

We have

$$\mathbb{E}(X_l) = nq_l.$$

and $\mathbb{V}[X_l] = nq_l$. So the estimator X_l/n is unbiased and

$$\mathbb{V}(X_l/n) = q_l/n.$$

So

$$\mathbb{E}(\|\hat{\theta}-\theta\|_2^2) = \sum_{l=1}^d q_l/n = n^{-1}.$$

This result corresponds to an upper bound on the minimax mean squared error and we build an ℓ_u minimax estimation risk. Now for any vector Δ, we have for $0 < u \leq 2$

$$\|\Delta\|_{\ell_u} \leq d^{(1/u-1/2)}\|\Delta\|_{\ell 2}.$$

So we conclude for $0 < u \leq 2$,

$$\mathbb{E}(\|\hat{\theta}-\theta\|_{\ell_u}^u) \leq d^{(1-u/2)}n^{-u/2}.$$

1.7.3 Proof of Theorem 7

This lemma will provide the first four moments of a Poisson distribution.

Lemma 5. *Let $X \sim \mathcal{P}(\lambda)$. Then we have the following moments.*

$$\mathbb{E}(X) = \lambda, \quad \mathbb{E}(X^2) = \lambda^2+\lambda, \quad \mathbb{E}(X^3) = \lambda^3+3\lambda^2+\lambda \text{ and } \mathbb{E}(X^4) = \lambda^4+6\lambda^3+7\lambda^2+\lambda.$$

Let $\Delta_i = q_i - p_i$ for any $i \leq d$. The expected value is

$$E(T) = \sum_{i=1}^{d} (\mathbb{V}(X_i) + (n\Delta_i)^2) - n = \sum_{i=1}^{d} (n\Delta_i)^2,$$

because $\mathbb{E}(X_i) = \mathbb{V}(X_i) = nq_i$. And

$$\mathbb{V}(T) = \mathbb{V}(\sum_{i=1}^{d} [X_i - np_i]^2 - X_i) = \sum_{i=1}^{d} \mathbb{V}((X_i)^2 + (np_i)^2 - (2np_i + 1)X_i).$$

So

$$\mathbb{V}(T) = \sum [\mathbb{V}(X_i^2) - 2(2np_i + 1)\mathrm{Cov}(X_i^2, X_i) + (2np_i + 1)^2 \mathbb{V}(X_i)].$$

Now, by Lemma 5, we have

$$\begin{aligned}\mathbb{V}(X_i^2) &= \mathbb{E}(X_i^4) - \mathbb{E}(X_i^2)^2 = nq_i(1 + 7nq_i + 6(nq_i)^2 + (nq_i)^3) - (nq_i + (nq_i)^2)^2 \\ &= nq_i(1 + 6nq_i + 4(nq_i)^2),\end{aligned}$$

and

$$\begin{aligned}\mathrm{Cov}(X_i^2, X_i) &= \mathbb{E}(X_i^3) - \mathbb{E}(X_i^2)\mathbb{E}(X_i) = nq_i(1 + 3nq_i + (nq_i)^2) - nq_i(nq_i + (nq_i)^2) \\ &= nq_i(1 + 2nq_i).\end{aligned}$$

So

$$\begin{aligned}\mathbb{V}(T) &= \sum [nq_i(1 + 6nq_i + 4(nq_i)^2) - 2(2np_i + 1)nq_i(1 + 2nq_i) + (2np_i + 1)^2 nq_i] \\ &= \sum [2n^2 p_i^2 + 2n^2\Delta_i^2 + 4n^2\Delta_i p_i + 4n^3\Delta_i^3 + 4n^3\Delta_i^2 p_i] \\ &= 2n^2/d + 2n^2 \sum \Delta_i^2 + 4n^3 \sum \Delta_i^3 + 4n^3 \sum \Delta_i^2/d.\end{aligned}$$

Under $\mathcal{H}_0$. We have

$$\mathbb{E}(T) = 0, \quad \mathbb{V}(T) = 2n^2 \sum p_i^2 = 2n^2/d, \quad \sqrt{\mathbb{V}(T)} = \sqrt{2}n\|p\|_2 = \sqrt{2}nd^{-1/2}.$$

So by Chebyshev's inequality, with probability larger than $1 - \delta$,

$$T \leq \sqrt{2}nd^{-1/2}/\delta.$$

Under $\mathcal{H}_1$. We have

$$\mathbb{E}(T) = n^2 \sum \Delta_i^2.$$

And

$$\begin{aligned}\mathbb{V}(T) &\leq \sum (2n^2 p_i^2 + 2n^2\Delta_i^2 + 2(n^2 p_i^2 + n^2\Delta_i^2) + 2(n^4\Delta_i^4/16 + 16n^2\Delta_i^2) + 4n^3\Delta_i^2 p_i) \\ &\leq 36n^2 \sum p_i^2 + 36n^2 \sum \Delta_i^2 + \sum n^4\Delta_i^4(2/16 + 2/16) \\ &\leq 36n^2 \sum p_i^2 + 36n^2 \sum \Delta_i^2 + 1/4 \sum n^4\Delta_i^4.\end{aligned}$$

So

$$\begin{aligned}\sqrt{\mathbb{V}(T)} &\leq 6n\|p\|_2 + 6\sqrt{n^2 \sum \Delta_i^2} + 1/2n^2 \sum \Delta_i^2. \\ &\leq 6n\|p\|_2 + 36 + 3/4n^2 \sum \Delta_i^2.\end{aligned}$$

So by Chebyshev's inequality, with probability larger than $1-\delta$,

$$T \geq \mathbb{E}(T) - [6n\|p\|_2 + 36 + 3/4\mathbb{E}(T)]/\delta.$$

That is, with probability larger than $1-\delta$,

$$T \geq [1 - 3/(4\delta)]n^2 \sum \Delta_i^2 - (6n\|p\|_2 + 36)/\delta.$$

To sum up, we obtain the following condition

$$(2n\|p\|_2 + 68)/(\delta - 3/4) \leq n^2 \sum \Delta_i^2.$$

That is, the sufficient condition is

$$(\sqrt{2}n^{-1/2}\|p\|_2^{1/2} + \sqrt{68}n^{-1})/\sqrt{(\delta - 3/4)} \leq \sqrt{\sum \Delta_i^2}.$$

Now, for $0 < u \leq 2$,

$$\|\Delta\|_{\ell_u} \leq d^{(1/u-1/2)}\|\Delta\|_2.$$

So we have the following condition, for $0 < u \leq 2$,

$$\|\Delta\|_{\ell_u} \geq n^{-1/2}d^{(1/u-1/2)}(\sqrt{2}\|p\|_2^{1/2} + \sqrt{68}n^{-1/2})/\sqrt{(\delta - 3/4)}.$$

So if p is a uniform vector, we have the condition

$$\|\Delta\|_{\ell_u} \geq n^{-1/2}d^{(1/u-1/2)}(\sqrt{2}d^{-1/4} + \sqrt{68}n^{-1/2})/\sqrt{(\delta - 3/4)}.$$

1.7.4 Proofs of Lemmas 2 and 3

Proof of Lemma 2. Finding a lower bound on $\rho^*_\gamma(H_0, H_1, \mathcal{T}; \text{dist})$ amounts to finding a real number ρ such that $R(H_0, H_1(\rho, \text{dist}), \hat{\theta}_n) > \gamma$ for any $\hat{\theta}_n \in \mathcal{T}$. We will denote $\mathbb{P}_{\nu_0}$ (respectively, $\mathbb{P}_{\nu_{1,\rho}}$) the probability measure according to which f has distribution ν_0 (respectively, $\nu_{1,\rho}$) and $\mathcal{X} \sim f$.

Then

$$\begin{aligned}
R(H_0, H_1(\rho, \text{dist}), \hat{\theta}_n) &= \sup_{(f_0,f)\in H_0} \mathbb{P}_{f_0,f}(\hat{\theta}_n = 1) + \sup_{(f_0,f)\in H_1(\rho,\text{dist})} \mathbb{P}_{f_0,f}(\hat{\theta}_n = 0) \\
&\geq \mathbb{P}_{\nu_0}(\hat{\theta}_n = 1,\ (f_0, f) \in H_0) + \mathbb{P}_{\nu_{1,\rho}}(\hat{\theta}_n = 0,\ (f_0, f) \in H_1(\rho, \text{dist})) \\
&\geq \mathbb{P}_{\nu_0}(\hat{\theta}_n = 1) - \delta_0 + \mathbb{P}_{\nu_{1,\rho}}(\hat{\theta}_n = 0) - \delta_1,
\end{aligned}$$

since $\nu_0(H_0) \geq 1 - \delta_0$ and $\nu_{1,\rho}(H_1(\rho, \text{dist})) \geq 1 - \delta_1$.

So by definition of the total variation distance, we have for any $\hat{\theta}_n \in \mathcal{T}$,

$$\begin{aligned}
R(H_0, H_1(\rho, \text{dist}), \hat{\theta}_n) &\geq 1 - \mathbb{P}_{\nu_0}\left(\hat{\theta}_n = 0\right) + \mathbb{P}_{\nu_{1,\rho}}\left(\hat{\theta}_n = 0\right) - \delta_0 - \delta_1 \\
&\geq 1 - d_{TV}(\mathbb{P}_{\nu_0}, \mathbb{P}_{\nu_1}) - \delta_0 - \delta_1.
\end{aligned}$$

□

Proof. Proof of Lemma 3.

$$d_{TV}(\nu_0,\nu_1) = \int \left|\frac{d\nu_1}{d\nu_0} - 1\right| d\nu_0 = \mathbb{E}_{\nu_0}\left[\left|\frac{d\nu_1}{d\nu_0} - 1\right|\right]$$
$$\leq \left(\mathbb{E}_{\nu_0}\left[\left(\frac{d\nu_1}{d\nu_0}\right)^2\right] - 1\right)^{1/2} = \sqrt{\chi^2(\nu_0,\nu_1)}.$$

□

1.7.5 Proof of Theorem 10

The following lemma that we prove at the end of this subsection will give bounds on the empirical threshold.

Lemma 6. *With probability larger than* $1-\delta$,

$$n\|p\|_2\sqrt{1-1/(2\delta)} \leq \hat{t} \leq n\|p\|_2\sqrt{1+1/(2\delta)} + 2\sqrt{50/\delta}.$$

We have

$$\mathbb{E}(T) = \sum_{i=1}^{d} n^2(q_i - p_i)^2,$$

and

$$\begin{aligned}\mathbb{V}(T) &= \sum_{i=1}^{d} \mathbb{E}((X_i^{(1)} - Y_i^{(1)})^2)\mathbb{E}((X_i^{(2)} - Y_i^{(2)})^2) - \mathbb{E}(X_i^{(1)} - Y_i^{(1)})^2\mathbb{E}(X_i^{(2)} - Y_i^{(2)})^2 \\ &= \sum_{i=1}^{d}[(\mathbb{V}(X_i) + \mathbb{V}(Y_i) + n^2(q_i - p_i)^2)^2 - n^4(q_i - p_i)^4] \\ &= \sum_{i=1}^{d}[n^2(4p_i^2 + \Delta_i^2 + 4\Delta_i p_i) + n^3(2\Delta_i^3 + 4p_i\Delta_i^2)].\end{aligned}$$

So under $\mathcal{H}_0$,

$$\mathbb{E}(T) = 0, \quad \mathbb{V}(T) = 4n^2\sum p_i^2, \quad \sqrt{\mathbb{V}(T)} = 2n\sqrt{\sum p_i^2}.$$

By Chebyshev's inequality, with probability larger than $1-\delta$,

$$T \leq 2n\sqrt{\sum p_i^2}/\delta.$$

Now, under $\mathcal{H}_1$,

$$\mathbb{E}(T) = \sum n^2\Delta_i^2,$$

and

$$\begin{aligned}\mathbb{V}(T) &\leq \sum_{i=1}^{d}[n^2(6p_i^2+3\Delta_i^2)+2n^3(\Delta_i^3+p_i\Delta_i^2)]\\ &\leq \sum_{i=1}^{d}[6n^2p_i^2+3n^2\Delta_i^2+(16n^2\Delta_i^2+n^4\Delta_i^4/16)+(16n^2p_i^2+n^4\Delta_i^4/16)]\\ &\leq \sum_{i=1}^{d}[22n^2p_i^2+19n^2\Delta_i^2+2n^4\Delta_i^4/16].\end{aligned}$$

So

$$\begin{aligned}\sqrt{\mathbb{V}(T)} &\leq 5n\|p\|_2+5\sqrt{\sum n^2\Delta_i^2}+\sqrt{2}/4\sum_{i=1}^{d}n^2\Delta_i^2\\ &\leq 5n\|p\|_2+25+\sum n^2\Delta_i^2/4+\sqrt{2}/4\sum_{i=1}^{d}n^2\Delta_i^2.\end{aligned}$$

So

$$\sqrt{\mathbb{V}(T)} \leq 5n\|p\|_2+25+\sum n^2\Delta_i^2(1+\sqrt{2})/4.$$

By Chebyshev's inequality, with probability larger than $1-\delta$,

$$T \geq \sum n^2\Delta_i^2-(5n\|p\|_2+25+\sum n^2\Delta_i^2(1+\sqrt{2})/4)/\delta.$$

Summing up the results from both hypotheses together with Lemma 6, we end up with the condition

$$\sum n^2\Delta_i^2 \geq C(\|p\|_2+1)$$

That is,

$$\sqrt{\sum \Delta_i^2} \geq C(n^{-1}+n^{-1/2}\|p\|_2^{1/2})$$

So we have the following condition, for $0<u\leq 2$,

$$\|\Delta\|_{\ell_u} \geq C(n^{-1}d^{1/u-1/2}+n^{-1/2}d^{1/u-1/2}\|p\|_2^{1/2}).$$

That is, if p is a uniform vector,

$$\|\Delta\|_{\ell_u} \geq C(n^{-1}d^{1/u-1/2}+n^{-1/2}d^{1/u-3/4}).$$

Proof of Lemma 6. We first consider

$$\sum_i Y_i^{(1)}Y_i^{(2)},$$

and we have

$$\mathbb{E}(\sum_i Y_i^{(1)}Y_i^{(2)}) = \sum_i (np_i)^2.$$

Besides,

$$\begin{aligned}\mathbb{V}(\sum_i Y_i^{(1)}Y_i^{(2)}) &= \sum_i [\mathbb{E}[(Y_i^{(1)})^2]^2 - \mathbb{E}[(Y_i^{(1)})]^4] \\ &= \sum_i [(np_i + n^2p_i^2)^2 - n^4p_i^4] == \sum_i (n^2p_i^2 + 2n^3p_i^3).\end{aligned}$$

So

$$\begin{aligned}\sqrt{\mathbb{V}(\sum_i Y_i^{(1)}Y_i^{(2)})} &\leq \sqrt{\sum_i n^2p_i^2} + \sqrt{2\sum n^3p_i^3} \\ &\leq \sqrt{\sum_i n^2p_i^2} + \sqrt{\sum (16n^2p_i^2 + n^4p_i^4/16)} \\ &\leq 50 + \sum n^2p_i^2/2.\end{aligned}$$

So by Chebyshev's inequality, with probability larger than $1-\delta$,

$$|\sum_i Y_i^{(1)}Y_i^{(2)} - \sum_i (np_i)^2| \leq (50 + \sum n^2p_i^2/2)/\delta.$$

That is

$$\sum n^2p_i^2[1 - 1/(2\delta)] - 50/\delta \leq \sum_i Y_i^{(1)}Y_i^{(2)} \leq \sum n^2p_i^2[1 + 1/(2\delta)] + 50/\delta.$$

Now, taking

$$\hat{t} = (\sqrt{\sum Y_i^{(1)}Y_i^{(2)}} + \sqrt{50/\delta})/\sqrt{1 - 1/(2\delta)}.$$

So

$$n\|p\|_2 \leq \hat{t} \leq (n\|p\|_2\sqrt{1 + 1/(2\delta)} + 2\sqrt{50/\delta})/\sqrt{1 - 1/(2\delta)}.$$

□

1.7.6 Proof of Proposition 7

We have

$$\begin{aligned}R_t &= \sum_{t\leq T}(\sup_{i\in I}\mu_i - \mathbb{E}(\sum_{j\in J} X_t\mathbb{1}\{A_t = j\})) \\ &= \sum_{t\leq T}(\sup_{i\in I}\mu_i - \mathbb{E}(\sum_{j\in J} X_t\mathbb{1}\{A_t = j\}|A_t = j)\mathbb{E}(\mathbb{1}\{A_t = j\})) \\ &= \sum_{t\leq T}(\sup_{i\in I}\mu_i - \sum_{j\in J}\mu_j\mathbb{E}(\mathbb{1}\{A_t = j\})) \\ &= \sum_{t\leq T}(\sup_{i\in I}\mu_i \sum_{j\in J}\mathbb{E}(\mathbb{1}\{A_t = j\}) - \sum_{j\in J}\mu_j\mathbb{E}(\mathbb{1}\{A_t = j\})).\end{aligned}$$

So

$$R_T = \sum_{j\in I}\Delta_j \sum_{t\leq T}\mathbb{E}(\mathbb{1}\{A_t = j\}).$$

1.7.7 Proof of Theorem 16

Proof. For determining the regret, we will look at the expected number of times that arm 2, the suboptimal arm, is picked: $\sum_{t\le T}\mathbb{E}(\mathbb{1}\{A_t=2\})$. If $\varepsilon=0$, we have

$$\sum_{t\le T}\mathbb{E}(\mathbb{1}\{A_t=2\})\ge(1/2-\Delta)(T-1)/2.$$

Indeed, with probability $(1/2-\Delta)/2$, $A_1=1$ and $X_1=0$, and $A_2=2$ and $X_2=1$. So the estimate of μ_1 is 0 and the estimate of μ_2 is 1. So the algorithm will keep pulling arm 2 and the estimate of μ_2 will stay positive at every step. So by Proposition 7

$$R_T\ge(1/2-\Delta)(T-1)\Delta/2.$$

If $0<\varepsilon\le 1$, $\sum_{t\le T}\mathbb{E}(\mathbb{1}\{A_t=2\})\ge\sum_{t\le T}\varepsilon/2$. So by Proposition 7

$$R_T\ge\varepsilon T\Delta/2.$$

□

1.7.8 Proofs of Theorem 17 and Corollary 4

We start with the definition of subgaussian distributions.

Definition 17. *A random variable X is σ-subgaussian if for all $\lambda\in\mathbb{R}$, it holds that $\mathbb{E}[\exp(\lambda X)]\le\exp(\lambda^2\sigma^2/2)$.*

This definition is accompanied with the following concentration results which can be found in Lattimore and Szepesvári (2020) and Orabona (2019).

Theorem 20. *If X is σ-subgaussian, then for any $\varepsilon\ge 0$,*

$$\mathbb{P}(X\ge\varepsilon)\le\exp(-\varepsilon^2/(2\sigma^2)).$$

Corollary 5. *Let $X_i-\mu$ be independent, σ-subgaussian random variables. Then for any $\varepsilon\ge 0$,*

$$\mathbb{P}\left(\sum_{i=1}^n\frac{X_i}{n}\ge\mu+\varepsilon\right)\le\exp(-n\varepsilon^2/(2\sigma^2)),$$

and

$$\mathbb{P}\left(\sum_{i=1}^n\frac{X_i}{n}\le\mu-\varepsilon\right)\le\exp(-n\varepsilon^2/(2\sigma^2)).$$

Proof of Theorem 17. Let $\alpha>2$.

Let t^* be the largest index such that $T_2(t^*-1)\le 8\alpha\log T/\Delta^2$. If $t^*=T$, then $T_2(T)\le 8\alpha\log T/\Delta^2+1$. Assume from now on, $t^*<T$. Let $c_{t,s}=\sqrt{2\alpha\log(t)/s}$. We define the following events

$$B_1(t)=\left\{\sum_{i=1}^{t-1}X_i\mathbb{1}\{A_i=1\}\le\mu_1-c_{t,T_1(t-1)}\right\},$$

and

$$B_2(t)=\left\{\sum_{i=1}^{t-1}X_i\mathbb{1}\{A_i=2\}\ge\mu_2+c_{t,T_2(t-1)}\right\}.$$

Let $t > t^*$ such that $A_t = 2$. Assume $B_l(t)$ do not hold for $l \in \{1, 2\}$. Then

$$\begin{aligned}\sum_{i=1}^{t-1} X_i \mathbb{1}\{A_i = 1\} + c_{t,T_1(t-1)} &> \mu_1 = \mu_2 + \Delta \\ &> \sum_{i=1}^{t-1} X_i \mathbb{1}\{A_i = 2\} - c_{t,T_2(t-1)} + \Delta \\ &> \sum_{i=1}^{t-1} X_i \mathbb{1}\{A_i = 2\} - c_{t,T_2(t-1)} + 2c_{T,T_2(t-1)} \\ &> \sum_{i=1}^{t-1} X_i \mathbb{1}\{A_i = 2\} + c_{t,T_2(t-1)},\end{aligned}$$

by definition of t^*. So $A_t = 1$, which is a contradiction. So for any $t > t^*$, if $A_t = 2$, then $B_1(t) \cup B_2(t)$ holds.

We have

$$\begin{aligned}\mathbb{E}(T_2(T)) &= \mathbb{E}(T_2(t^*)) + \sum_{t=t^*+1}^{T} \mathbb{E}(\mathbb{1}\{A_t = i\}) \\ &\leq 8\alpha \log T/\Delta^2 + 1 + \sum_{t=t^*+1}^{T} \mathbb{E}(\mathbb{1}\{B_1(t) \cup B_2(t)\}).\end{aligned}$$

We have

$$B_1(t) = \{\sum_{i=1}^{t-1} X_i \mathbb{1}\{A_i = 1\} \leq \mu_1 - c_{t,T_1(t-1)}\} = \bigcup_{s=1}^{t-1} \{\bar{X}_s^{(1)} \leq \mu_1 - c_{t,s}\}.$$

Now, by Hoeffding's lemma, $X^{(1)} - \mu_1$ is 1-subgaussian, so we have by Corollary 5,

$$\mathbb{P}(\bar{X}_s^{(1)} \leq \mu_1 - c_{t,s}) \leq \exp(-\alpha \log(t)) = t^{-\alpha}.$$

So

$$\mathbb{P}(B_l(t)) \leq (t-1)t^{-\alpha}.$$

So

$$\mathbb{E}(T_2(T)) \leq 8\alpha \log T/\Delta^2 + 1 + 2\sum_{t=t^*+1}^{T} (t-1)t^{-\alpha} \leq 8\alpha \log T/\Delta^2 + 1 + 2\sum_{t=2}^{\infty} t^{1-\alpha}.$$

Now

$$\sum_{t=2}^{\infty} t^{1-\alpha} \leq \int_1^{\infty} x^{1-\alpha} dx = 1/(\alpha - 2).$$

So

$$\mathbb{E}(T_2(T)) \leq 8\alpha \log T/\Delta^2 + \alpha/(\alpha - 2).$$

□

Proof of Corollary 4. By application of Proposition 7 on the result obtained in Theorem 17, we have the regret, for some fixed $h > 0$,

$$\begin{aligned} R_T &= \Delta \mathbb{E}(T_2(T))(\mathbb{1}\{\Delta < h\} + \mathbb{1}\{\Delta \geq h\}) \\ &\leq T\Delta \mathbb{1}\{\Delta < h\} + [8\alpha \log T/\Delta + \alpha\Delta/(\alpha - 2)]\mathbb{1}\{\Delta \geq h\} \\ &\leq Th + 8\alpha \log T/h + \alpha\Delta/(\alpha - 2). \end{aligned}$$

We take $h = 2\sqrt{2\alpha \log(T)/T}$. So

$$R_T \leq 4\sqrt{2\alpha T \log(T)} + \alpha\Delta/(\alpha - 2).$$

□

1.7.9 Proof of Theorem 18

Let us assume that $\forall t \leq n, \forall x \in [0,1]^d, f(x) \leq M_t g_t(x)$. If X_t has been drawn at time t, and E_t denotes the event in which X_t is accepted, and $\tilde{\chi}_j$ denotes the set of the proposal samples drawn at time $j \leq n$ and of their images by f, then $\forall \Omega \subset [0,1]^d$ such that Ω is Lebesgue measurable, it holds:

$$\begin{aligned} &\mathbb{P}_{X_t \sim g_t, U \sim \mathcal{U}_{[0,1]}}\Big(\{X_t \in \Omega\} \cap E_t \;\Big|\; \bigcup_{j<t} \chi_j\Big) \\ &\qquad = \mathbb{P}_{X_t \sim g_t, U \sim \mathcal{U}_{[0,1]}}\Big(X_t \in \Omega; \frac{f(X_t)}{M_t g_t(X_t)} \geq U_t \;\Big|\; \bigcup_{j<t} \chi_j\Big) \\ &\qquad = \int_\Omega \frac{f(x)}{M_t g_t(x)} g_t(x) dx \\ &\qquad = \int_\Omega \frac{f(x)}{M_t} dx, \end{aligned}$$

because U_t is independent from X_t conditionally to $\bigcup_{j<t} \chi_j$.
Hence, since $\mathbb{P}_{X_t \sim g_t, U \sim \mathcal{U}_{[0,1]}}(E_t) = I_f/M_t$, we have:

$$\begin{aligned} \mathbb{P}_{X_t \sim g_t, U \sim \mathcal{U}_{[0,1]}}\Big(X_t \in \Omega \;\Big|\; E_t; \bigcup_{j<t} \chi_j\Big) &= \int_\Omega \frac{f(x)}{M_t} \left(\frac{M_t}{I_f}\right) dx \\ &= \int_\Omega \frac{f(x)}{I_f} dx. \end{aligned}$$

Thus $X_t | E_t$ is distributed according to f/I_f and is independent from the samples accepted before step t, since $X_t | E_t$ is independent from $\bigcup_{j<t} \chi_j$.

We have proved that the algorithm provides independent samples drawn according to the density f/I_f.

Chapter 2

Local minimax closeness testing of discrete distributions

2.1 Introduction

This chapter focuses on closeness testing that we already introduced in Section 1.4.3. In the statistical testing problem we will consider, the null hypothesis is true when both distributions are the same, and in the alternative hypothesis, they are different and separated in ℓ_1-norm. For a fixed number of samples, the goal is to find out how close both distributions can get to one another and still be distinguishable, depending on the shape of one of the distributions. In the following, we provide a formal setting for this problem.

2.1.1 Setting

Let $d \in \mathbb{N}^*$. We define the set of vectors of size d that correspond to multinomial distributions over d categories as

$$\mathbf{P} = \left\{ \pi \in (\mathbb{R}^+)^d, \sum_{i \leq d} \pi_i = 1 \right\}.$$

Let $\pi \in \mathbf{P}$. Define for any $i \in \mathbb{Z}$

$$S_\pi(i) = \{j \in \{1, \dots, d\} : \pi_j \in [2^{-i}, 2^{-i+1})\}.$$

Define

$$\mathbf{P}_\pi = \left\{ q \in \mathbf{P} : \forall i \in \mathbb{Z}, \frac{|S_\pi(i)|}{2} \leq \sum_{j=i-1}^{i+1} |S_q(j)| \leq \frac{3}{2} \sum_{j=i-2}^{i+2} |S_\pi(j)| \right\},$$

which represents a class of probability vectors very similar to π. Indeed, for any $q \in \mathbf{P}_\pi$, the discrete level sets S_q and S_π are close in size.

Let $p \in \mathbf{P}$, $q \in \mathbf{P}_\pi$ and $n \in \mathbb{N}^*$. The independent sample sets $(\mathcal{X}, \mathcal{Y})$ are obtained from the following two multinomial distributions.

$$\mathcal{X} \sim \mathcal{M}(n, p), \quad \mathcal{Y} \sim \mathcal{M}(n, q), \tag{2.1}$$

where $\mathcal{M}$ is the multinomial distribution. That is, for $i \leq n$, we have independent $\mathcal{X}_i$ taking value $j \in \{1, \dots, d\}$ with probability p_j – respectively, $\mathcal{Y}_i = j$ with probability q_j. In what follows, we write $\mathbb{P}_{p,q}$ for the probability associated to $(\mathcal{X}, \mathcal{Y})$.

For any vector $x \in \mathbb{R}^d$, let $\|x\|_t = (\sum_{i=1}^d |x_i|^t)^{1/t}$ denote the ℓ_t-norm of x for any $0 < t < \infty$. Let $\Phi = \{\varphi : \mathbb{N}^{2n} \to \{0,1\}\}$. For a fixed $\rho > 0$, we formalize the closeness testing problem as the following hypothesis sets

$$H_{0,\pi}^{(\mathrm{Clo})} = \{(p,q) \in (\mathbf{P}_\pi)^2; p = q\}, \quad H_{1,\pi}^{(\mathrm{Clo})}(\rho) = \{(p,q) \in \mathbf{P} \times \mathbf{P}_\pi; \|p-q\|_1 \geq \rho\}. \quad (2.2)$$

with the following set of tests using the observations of $(\mathcal{X}, \mathcal{Y})$

$$\mathcal{T}_n^{(\mathrm{Clo})} = \{\hat{\theta}; \hat{\theta} = \varphi(\mathcal{X}, \mathcal{Y}), \varphi \in \Phi\}.$$

Here ρ represents a *separation distance* between p and q which amounts to assuming that the null and alternative hypotheses are different enough.

Remark 25. *To the best of our knowledge, closeness testing has never known any formal definition as a hypothesis testing problem allowing for its local minimax analysis. Our formalization satisfies a few important criteria for the purpose of instance-optimal closeness testing. Firstly, the null hypothesis is composite as well as the alternative hypothesis, in contrast to identity testing whose null hypothesis is simple. That is, under any one of both hypotheses from Equation (2.2), q is allowed to be quite different from π, since there is no relation in the ordering of their entries. So there does not exist any test exploiting the full knowledge of p or q, and our problem is inherently harder than identity testing, where $q = \pi$, presented in Valiant and Valiant (2017) and Balakrishnan and Wasserman (2017a). Secondly, $q \in \mathbf{P}_\pi$ is still related to π in the sense discussed above, so that our results can be instance-optimal and depend on π. The results can vary greatly depending on π and a worst-case study from Chan et al. (2014) does not guarantee an optimal test in all cases. Intuitively, if π is the uniform distribution for example, the testing problem is more difficult than if π just has a few entries with non-zero probability. We want to capture this dependence on the distribution, as Valiant and Valiant (2017) and Balakrishnan and Wasserman (2017a) do for one-sample testing. Finally, we provide in Section 2.4 a discussion on how this formalization could be generalized to other set-ups.*

It is clear from Equation (2.2) that the vectors that are too close to q are removed from the alternative hypothesis. With π fixed, we want to find the smallest ρ such that both hypotheses are still distinguishable. The notion of distinguishability of both hypotheses is formalized by the definition of error risk and separation distance. The error risk is the sum of type I and type II error probabilities. For any separation distance $\rho > 0$ and probability vector π, we can define some testing problem from a set couple $(H_{0,\pi}, H_{1,\pi}(\rho))$ – e.g., $H_{0,\pi}^{(\mathrm{Clo})}$ and $H_{1,\pi}^{(\mathrm{Clo})}(\rho)$ for Equation (2.2). Then the separation distance given a test $\varphi(\mathcal{X}, \mathcal{Y}) = \hat{\theta}_n$ is

$$R(H_{0,\pi}, H_{1,\pi}(\rho), \hat{\theta}_n) = \sup_{(p,q) \in H_0(\pi)} \mathbb{P}_{p,q}(\hat{\theta}_n = 1) + \sup_{(p,q) \in H_1(\pi,\rho)} \mathbb{P}_{p,q}(\hat{\theta}_n = 0),$$

where we remind that $\mathbb{P}_{p,q}$ is the probability measure associated with $(\mathcal{X}, \mathcal{Y})$. Then, fixing some $\gamma \in (0,1)$, we say that a testing problem can be solved with error smaller than γ, if we can construct a uniformly γ-consistent test, that is, if there exists φ such that:

$$R(H_{0,\pi}, H_{1,\pi}(\rho), \hat{\theta}_n) \leq \gamma.$$

Clearly, $\rho \mapsto R(H_{0,\pi}, H_{1,\pi}(\rho), \hat{\theta}_n)$ is non-increasing, and greater or equal to one when $\rho = 0$. Define the separation distance for some fixed $\gamma \in (0,1)$ as

$$\rho_\gamma(H_{0,\pi}, H_{1,\pi}, \hat{\theta}_n) = \inf\{\rho > 0 : R(H_{0,\pi}, H_{1,\pi}(\rho), \hat{\theta}_n) \leq \gamma\}.$$

A good test φ such that $\varphi(\mathcal{X}, \mathcal{Y}) = \hat{\theta}_n$ is characterized by a small separation distance. So we define the local minimax separation distance, also known as local critical radius, as

$$\rho^*_\gamma(H_{0,\pi}, H_{1,\pi}, \mathcal{T}_n) = \inf_{\hat{\theta}_n \in \mathcal{T}_n} \rho_\gamma(H_{0,\pi}, H_{1,\pi}, \hat{\theta}_n).$$

Besides, it is possible to consider the global minimax separation distance defined as

$$\sup_{\pi \in \mathbf{P}} \rho^*_\gamma(H_{0,\pi}, H_{1,\pi}, \mathcal{T}_n).$$

A worst-case analysis is sufficient for finding the global minimax separation distance, so it is a weaker result than finding the local minimax separation distance.

A lot of relevant results from the literature that we present in Section 2.1.2 come from the field of property testing in computer science. So although this thesis focuses on rates in separation distance, we will link this concept with that of sample complexity, favoured in computer science. Sample complexity corresponds to the number of samples that are necessary and sufficient in order to achieve a certain testing error for a fixed separation distance. Formally, for a fixed ρ, since $n \mapsto R(H_{0,\pi}, H_{1,\pi}(\rho), \hat{\theta}_n)$ is non-increasing, the sample complexity for some fixed $\gamma \in (0, 1)$ is defined as

$$n_\gamma(H_{0,\pi}, H_{1,\pi}(\rho), \hat{\theta}_{n_\gamma}) = \inf\{n \in \mathbb{N} : R(H_{0,\pi}, H_{1,\pi}(\rho), \hat{\theta}_n) \leq \gamma\}.$$

Then the minimax sample complexity is

$$n^*_\gamma(H_{0,\pi}, H_{1,\pi}(\rho), \mathcal{T}_{n^*_\gamma}) = \inf_{\hat{\theta}_{n^*_\gamma} \in \mathcal{T}_{n^*_\gamma}} n_\gamma(H_{0,\pi}, H_{1,\pi}, \hat{\theta}_{n^*_\gamma}; \pi, \rho).$$

So the local minimax sample complexity is written as $n^*_\gamma(H_{0,\pi}, H_{1,\pi}(\rho), \mathcal{T}_{n^*_\gamma})$. And the global minimax sample complexity is $\sup_\pi n^*_\gamma(H_{0,\pi}, H_{1,\pi}(\rho), \mathcal{T}_{n^*_\gamma})$. If ρ^*_γ or n^*_γ are invertible, then it is possible to obtain one from the other. Let us define the inverses $\rho \mapsto (\rho^*_\gamma)^{-1}(H_{0,\pi}, H_{1,\pi}(\rho), \mathcal{T}_{n^*_\gamma})$ and $n \mapsto (n^*_\gamma)^{-1}(H_{0,\pi}, H_{1,\pi}(\rho), \mathcal{T}_n)$. Then $(\rho^*_\gamma)^{-1}(H_{0,\pi}, H_{1,\pi}(\rho), \mathcal{T}_{n^*_\gamma}) = n^*_\gamma(H_{0,\pi}, H_{1,\pi}(\rho), \mathcal{T}_{n^*_\gamma})$ and reciprocally.

Additional notations. We introduce the following notations. For a vector $u \in \mathbb{R}^d$, let s be a permutation of $\{1, \dots, d\}$ such that $u_{s(1)} \geq u_{s(2)} \geq \dots \geq u_{s(d)}$. We write $u_{(.)} := u_{s(.)}$. Set also $J_u = \min_{j \leq d} \left\{j : u_{(j)} \leq \frac{1}{n}\right\}$.

2.1.2 Literature review

Hypothesis testing is a classical statistical problem and we refer the reader to Neyman and Pearson (1933) and Lehmann and Romano (2006) for a more global perspective on the problem. In parallel to the study of hypothesis testing, there exists a broad literature on the related problem of property testing tackled by the theoretical computer science community, with seminal papers like Rubinfeld and Sudan (1996), Goldreich, Goldwasser, and Ron (1998).

In earlier studies, tests were built based on good asymptotic properties like having asymptotically normal limits, but this criterion often fails to produce tests which are efficient in high-dimensional cases notably, as stated in Balakrishnan and Wasserman (2017b). An alternative and popular take on the study of hypothesis testing is minimax optimality, with the seminal work of Ingster and Suslina (2012) on identity testing. The problem of identity testing consists in distinguishing whether a sample set $\mathcal{X} \sim \mathcal{M}(n, p)$ is drawn from a specified distribution $\pi \in \mathbf{P}$, versus a composite alternative separated from the null in ℓ_1-distance. We recall the formalization from

Section 1.4.4, and emphasize that this chapter will focus on separation distances in ℓ_1. Let $\Phi = \{\varphi : \mathbb{N}^n \to \{0,1\}\}$. Let $\pi \in \mathbf{P}$. We define the following sets for the definition of the hypothesis testing problem.

$$H_{0,\pi}^{(\mathrm{Id})} = \{(p,q); p = \pi = q\}, \quad H_{1,\pi}^{(\mathrm{Id})}(\rho) = \{(p,q); p \in \mathbf{P}, q = \pi, \|p-q\|_1 \geq \rho\}, \quad (2.3)$$

$$\mathcal{T}_d^{(\mathrm{Id})} = \{\hat{\theta}; \hat{\theta} = \varphi(X_1, \dots, X_n), \varphi \in \Phi\},$$

where $X \sim \mathcal{M}(n,p)$ is an independent random variable with respect to probability measure $\mathbb{P}_{p,q}$ for any $i \leq d$. We will consider local and global minimax rates for both the separation distance and sample complexity. Indeed, if either the number of sample points or the ℓ_1-separation between the distributions is reduced, then the problem becomes more difficult. Thus, it is possible to parametrize the difficulty of the problem using either the number of sample points n or the ℓ_1-separation distance ρ. Tables 2.1 and 2.2 capture the existing results in the literature on the global and local minimax separation distance and sample complexity for identity testing.

TABLE 2.1: Global minimax separation distance and sample complexity obtained for identity testing defined in Equation (2.3): $\sup_\pi \rho_\gamma^*(H_{0,\pi}^{(\mathrm{Id})}, H_{1,\pi}^{(\mathrm{Id})}, \mathcal{T}_n^{(\mathrm{Id})})$ and $\sup_\pi n_\gamma^*(H_{0,\pi}^{(\mathrm{Id})}, H_{1,\pi}^{(\mathrm{Id})}(\rho), \mathcal{T}_{n_\gamma^*}^{(\mathrm{Id})})$. The rates are only worst-case considerations, presented up to some constant depending only on γ.

	Separation distance	Sample complexity
Paninski (2008)	$d^{1/4}/\sqrt{n}$	$\sqrt{d}/\rho^2$

TABLE 2.2: Local minimax separation distance and sample complexity obtained for identity testing defined in Equation (2.3): $\rho_\gamma^*(H_{0,\pi}^{(\mathrm{Id})}, H_{1,\pi}^{(\mathrm{Id})}, \mathcal{T}_n^{(\mathrm{Id})})$ and $n_\gamma^*(H_{0,\pi}^{(\mathrm{Id})}, H_{1,\pi}^{(\mathrm{Id})}(\rho), \mathcal{T}_{n_\gamma^*}^{(\mathrm{Id})})$. The rates depend on the distribution in the null hypothesis and are up to some constant depending only on γ. Here $(x)_+ = \max(0,x)$, and $\mathbf{1}\{A_i\}$ equals 1 if A_i is true and 0 otherwise. So $\|\pi_{(i)}^{2/3}\mathbf{1}\{2 \leq i < m\}\|_1 = \sum_{2\leq i<m} \pi_{(i)}^{2/3}$ and $\|(\pi_{(i)}\mathbf{1}\{i \geq m\})_i\|_1 = \sum_{i\geq m} \pi_{(i)}$.

	Valiant and Valiant (2017)
Separation distance	$\min_m \left[\frac{\|(\pi_{(i)}^{2/3}\mathbf{1}\{2\leq i<m\})_i\|_1^{3/4}}{\sqrt{n}} \vee \frac{1}{n} \vee \|(\pi_{(i)}\mathbf{1}\{i \geq m\})_i\|_1\right]$
Sample complexity	$\min_m \left[\frac{1}{\rho} \vee \frac{\|(\pi_{(i)}^{2/3}\mathbf{1}\{2\leq i<m\})_i\|_1^{3/2}}{(\rho - \|(\pi_{(i)}\mathbf{1}\{i\geq m\})_i\|_1)_+^2}\right]$

Similarly for two-sample testing, Tables 2.3 and 2.4 capture the existing results in the literature on the global and local minimax separation distance and sample complexity.

First, let us consider the results obtained for the classical problem of identity testing presented in Equation (2.3). An upper bound on the global minimax sample complexity is given in Paninski (2008) and tightened in Valiant and Valiant (2017) for the class of multinomial distributions over a support of size d. The meaning of the global minimax sample complexity listed in Table 2.1 is that an optimal algorithm will be able to test with fixed non-trivial probability, using $\sqrt{d}/\rho^2$ samples, up to some explicit constant. This sample complexity can be translated into the separation distance presented in the same table, as justified in Section 2.1.1. The global minimax sample complexity is a worst-case analysis, that is, it corresponds to the rate obtained

TABLE 2.3: Bounds on the global minimax separation distance and sample complexity obtained for closeness testing defined in Equation (2.2): $\sup_\pi \rho^*_\gamma(H^{(\mathrm{Clo})}_{0,\pi}, H^{(\mathrm{Clo})}_{1,\pi}, \mathcal{T}^{(\mathrm{Clo})}_n)$ and $\sup_\pi n^*_\gamma(H^{(\mathrm{Clo})}_{0,\pi}, H^{(\mathrm{Clo})}_{1,\pi}(\rho), \mathcal{T}^{(\mathrm{Clo})}_{n^*_\gamma})$, up to some constant depending only on γ. The result in Batu et al. (2000) is only an upper bound (UB). The result in Chan et al. (2014) provides matching upper and lower bounds, and is hence global minimax optimal.

	Separation distance	Sample complexity
Batu et al. (2000) (UB only)	$d^{1/6}\log(d)^{1/4}/n^{1/4}$	$d^{2/3}\log(d)/\rho^4$
Chan et al. (2014) (minimax)	$\frac{d^{1/2}}{n^{3/4}} \vee \frac{d^{1/4}}{n^{1/2}}$	$\frac{d^{2/3}}{\rho^{4/3}} \vee \frac{d^{1/2}}{\rho^2}$

TABLE 2.4: Upper bounds (UB) on the local minimax separation distance and sample complexity obtained for closeness testing defined in Equation (2.2): $\rho^*_\gamma(H^{(\mathrm{Clo})}_{0,\pi}, H^{(\mathrm{Clo})}_{1,\pi}, \mathcal{T}^{(\mathrm{Clo})}_n)$ and $n^*_\gamma(H^{(\mathrm{Clo})}_{0,\pi}, H^{(\mathrm{Clo})}_{1,\pi}(\rho), \mathcal{T}^{(\mathrm{Clo})}_{n^*_\gamma})$. The rates are problem-dependent, even though both distributions are unknown. It is presented up to some polylog(dn) for the separation distance and up to a polylog(d/ρ) for the sample complexity. We present here a corollary from their Proposition 2.14, applied to our closeness testing problem – which is not defined in Diakonikolas and Kane (2016).

	Diakonikolas and Kane (2016)
Separation distance (UB)	$\frac{\|\mathbf{1}\{\pi<1/n\}\|_1^{1/2}\|\pi^2\mathbf{1}\{\pi<1/n\}\|_1^{1/4}}{\sqrt{n}} \vee \frac{\|\pi^{2/3}\|_1^{3/4}}{\sqrt{n}}$
Sample complexity (UB)	$\min_m \left(m \vee \frac{\|\mathbf{1}\{\pi<1/m\})\|_1\|\pi^2\mathbf{1}\{\pi<1/m\}\|_1^{1/2}}{\rho^2} \vee \frac{\|\pi^{2/3}\|_1^{3/2}}{\rho^2} \right)$

for the hardest problem overall. In the case of identity testing, the uniform distribution is the hardest distribution to test against.

From the observation that the sample complexity might take values substantially different from that of the worst case, the concept of minimaxity has been refined in recent lines of research. One such refinement corresponds to local minimaxity, also known as instance-optimality, where the local minimax sample complexity depends on π. Local minimax sample complexity and separation distance for identity testing are presented in Table 2.2. Valiant and Valiant (2017) obtains the local minimax sample complexity for Problem (2.3). Balakrishnan and Wasserman (2017a) makes their test more practical and expresses the rate in terms of local minimax separation distance. The reformulation of their bounds, presented in Table 2.2, comes from our Proposition 12. Note that the dependences in d in the local minimax sample complexity and separation distance are contained in the vector norms. Finally, Balakrishnan and Wasserman (2017a) also obtains the local minimax sample complexity and separation distance for identity testing in the continuous case with Lipschitz densities, but we focus on the discrete case here.

Let us now consider the literature involving closeness testing, for which we provide a formalization in Equation (2.2). The global minimax sample complexity and separation distance are summarized in Table 2.3, and upper bounds on the local minimax sample complexity and separation distance are given in Table 2.4. In the case of closeness testing, Batu et al. (2000) proposes a test and obtains a loose upper bound on the global minimax sample complexity. The actual global minimax sample complexity has been identified in Chan et al. (2014), using the tools developed in Valiant (2011). A very interesting message from Chan et al. (2014) is that there exists a substantial difference

between identity testing and closeness testing, and that the latter is harder. It is interesting to note that while the uniform distribution is the most difficult distribution to test in identity testing, π can be chosen in a different appropriate way in order to worsen the sample complexity and separation distance in closeness testing.

Again, distribution-dependent minimax sample complexity and separation distance might differ greatly from global minimax sample complexity and separation distance, respectively. Attempts at obtaining finer results have been made for closeness testing of continuous distributions. Indeed, a large variety of classes of distributions can be defined in the continuous case and it makes sense to obtain minimax rates over rather small classes of distributions. In Diakonikolas, Kane, and Nikishkin (2017), the authors focus on closeness testing over the class of piecewise constant probability distributions (referred to as h-histograms) and obtain minimax near-optimal testers. In the same way, in Diakonikolas, Kane, and Nikishkin (2015), the authors display optimal closeness testers for various families of structured distributions, with an emphasis on continuous distributions.

Now, as explained in the review of Balakrishnan and Wasserman (2017b), the definition of local minimaxity in closeness testing is more involved than in identity testing, and it is in fact an interesting open problem that we focus on in this chapter. The difficulty arises from the fact that both distributions are unknown, although we would like the minimax sample complexity and separation distance to depend on them. Indeed, in contrast to Problem (2.3) whose null hypothesis is simple, Problem (2.2) is composite-composite. So there is the additional difficulty of having to adapt to the unknown vector q. Now, the existence and the size of a difference in the local minimax rates between both problems depending on π are open questions. We remind that Chan et al. (2014) sheds light on such a gap, but only in the worst case of π, whereas we look for instance-based minimax optimality.

Diakonikolas and Kane (2016) constructs a test for closeness testing with sample complexity adaptive to one of the distributions (p, q), when either $p = q$ or $\|p-q\|_1 \geq \rho$. Their Proposition 2.14 states that their test achieves a sample complexity of

$$\min_m \left(m + \frac{\|\mathbf{1}\{q < 1/m\})\|_1 \|q^2 \mathbf{1}\{q < 1/m\}\|_1^{1/2}}{\rho^2} + \frac{\|q^{2/3}\|_1^{3/2}}{\rho^2} \right).$$

As explained at the end of Section 2.1.1, their sample complexity can be translated into a separation distance, which is useful as a comparison with our own results. So in Table 2.4, we present a corollary of their Proposition 2.14 in order to obtain a separation distance corresponding to an upper bound on $\rho^*_\gamma(H^{(\mathrm{Clo})}_{0,\pi}, H^{(\mathrm{Clo})}_{1,\pi}, \mathcal{T}^{(\mathrm{Clo})}_n)$ in our setting. Our corollary relies on the definition of $\mathbf{P}_\pi$. Indeed, $q \in \mathbf{P}_\pi$ has level sets with similar sizes to those of π, and therefore q has similar (partial) norms to π up to a multiplicative constant. The sample complexity from Diakonikolas and Kane (2016) matches the global minimax sample complexity $\sup_\pi n^*_\gamma(H^{(\mathrm{Clo})}_{0,\pi}, H^{(\mathrm{Clo})}_{1,\pi}(\rho), \mathcal{T}^{(\mathrm{Clo})}_{n^*_\gamma})$ obtained in Chan et al. (2014) for some choice of π and m. However they do not introduce any lower bound dependent on π, so the only known local lower bound comes from Valiant and Valiant (2017), which is found for the problem of identity testing in Equation (2.3). But the lower bound from Valiant and Valiant (2017) does not match the upper bound in Diakonikolas and Kane (2016) .

We now mention recent alternative viewpoints on the study of identity and closeness testing. In Acharya et al. (2012), the authors compare their closeness tester against an oracle tester which is given the knowledge of the underlying distribution q. When an oracle tester needs n samples, their closeness tester needs $n^{3/2}$ samples. Otherwise,

some studies have been made in closeness testing when the number of sample points for both distributions is not constrained to be the same in Bhattacharya and Valiant (2015), Diakonikolas and Kane (2016) and Kim, Balakrishnan, and Wasserman (2018). What is more, Diakonikolas et al. (2017) works on identity testing in the high probability case, instead of a fixed probability as it is done usually. That is to say, the authors introduce a global minimax optimal identity tester which discriminates both hypotheses with probability converging to 1.

2.1.3 Contributions

The following are the major contributions of this work:

- We provide a lower bound on the local minimax separation distance for closeness testing presented in Equation (2.2) – see Equation (2.4) for $u = 2.001$.
- We propose a test providing an upper bound that nearly matches the obtained lower bound for $u = 1/2$. So it is local minimax near-optimal for closeness testing, but the test is also practical, since it does not take π as a parameter even though the upper bound optimally depends on π.
- We point out the similarities and differences in regimes with local minimax identity testing.

More precisely we prove in Theorems 21 and 22 that the local minimax separation distance $\rho^*_\gamma(H^{(\mathrm{Clo})}_{0,\pi}, H^{(\mathrm{Clo})}_{1,\pi}, \mathcal{T}^{(\mathrm{Clo})}_n)$ up to some polylog(dn) is

$$\min_{I \geq J_\pi} \left[\frac{\sqrt{I}}{n} \vee \left(\sqrt{\frac{I}{n}} \|\pi^2 \exp(-un\pi)\|_1^{1/4} \right) \vee \|(\pi_{(i)} \mathbf{1}\{i \geq I\})_i\|_1 \right] \vee \frac{\left\| (\pi^{2/3}_{(i)} \mathbf{1}\{i \leq J_\pi\})_i \right\|_1^{3/4}}{\sqrt{n}} \vee \sqrt{\frac{1}{n}}, \tag{2.4}$$

where J_π and $\pi_{(.)}$ are defined in Section 2.1.1, $u = 2.001$ for the lower bound and $u = 1/2$ for the upper bound. The exponential and the powers are applied element-wise. Let I^* denote an I where the minimum in Equation (2.4) is reached.

The local minimax separation distance $\rho^*_\gamma(H^{(Id)}_{0,\pi}, H^{(Id)}_{1,\pi}, \mathcal{T}^{(\mathrm{Id})}_n)$ obtained in Valiant and Valiant (2017) and Balakrishnan and Wasserman (2017a) is

$$\min_m \left[\frac{\|(\pi^{2/3}_{(i)} \mathbf{1}\{2 \leq i < m\})_i\|_1^{3/4}}{\sqrt{n}} \vee \frac{1}{n} \vee \|(\pi_{(i)} \mathbf{1}\{i \geq m\})_i\|_1 \right].$$

We compare it with Equation (2.4). Indeed, $\rho^*_\gamma(H^{(Id)}_{0,\pi}, H^{(Id)}_{1,\pi}, \mathcal{T}^{(\mathrm{Id})}_n)$ also represents a lower bound on $\rho^*_\gamma(H^{(\mathrm{Clo})}_{0,\pi}, H^{(\mathrm{Clo})}_{1,\pi}, \mathcal{T}^{(\mathrm{Clo})}_n)$, as explained in Proposition 12. Let m^* denote an m where the minimum is reached.

Table 2.5 references the local minimax optimal separation distance we obtain for the closeness testing problem defined in Equation (2.2) and compares it with the upper bound from Diakonikolas and Kane (2016) and the local minimax optimal separation distance for identity testing found in Valiant and Valiant (2017) and Balakrishnan and Wasserman (2017a). In order to build Table 2.5, we classify the coefficients of π depending on their size and the corresponding contribution to the separation distance. As illustrated by the table, our local minimax separation distance fleshes out three

main regimes. Looking at the coefficients of π, for the indices smaller than J_π, the part of the separation distance corresponding to them is

$$\frac{\left\|(\pi_{(i)}^{2/3}\mathbf{1}\{i \leq J_\pi\})_i\right\|_1^{3/4}}{\sqrt{n}}.$$

As for the indices greater than I^*, the part of the separation distance corresponding to them is $\|(\pi_{(i)}\mathbf{1}\{i \geq I\})_i\|_1$. And so regarding the contribution of the indices smaller than J_π or greater than m^* to the separation distance, the regimes are the same as in the local minimax separation distance $\rho_\gamma^*(H_{0,\pi}^{(Id)}, H_{1,\pi}^{(Id)}, \mathcal{T}_n^{(\mathrm{Id})})$ for identity testing from Valiant and Valiant (2017). However regarding the coefficients corresponding to the indices between J_π and I^* the local minimax separation distance for closeness testing is not of same order as for identity testing. The difference concerning local minimax separation distances between identity testing (Valiant and Valiant (2017) and Balakrishnan and Wasserman (2017a)) and closeness testing then lies in the indices between J_π and m^*. Chan et al. (2014) also notes a difference in the global minimax rates between both problems and we have refined this intuition to make it depend on π.

We now detail the comparison with the paper by Diakonikolas and Kane (2016) that also studies the problem of local closeness testing. Diakonikolas and Kane (2016) also obtains an upper bound on the local minimax separation distance for closeness testing. Although it is adaptive to π and matching the one from Chan et al. (2014) in the worst case, their upper bound is not local minimax optimal. In fact, they capture two of the three different phases we describe in Table 2.5. But their regime corresponding to the very small coefficients, with indices greater than I^*, can be made tighter, matching the local minimax separation distance in identity testing.

We further illustrate the difference between the local minimax separation distance that we present and the upper bound from Diakonikolas and Kane (2016) with the following example. Take $d = n^4 + 2$ and $0 < h < 1/2$. Let $\pi_1 = 1/2$, $\pi_2 = 1/2 - h$, for any $3 \leq i \leq d$, $\pi_i = h/n^4$. Then, up to multiplicative constants depending on γ, results proved in this chapter lead to the minimax separation distance $1/\sqrt{n} + h$, whereas Diakonikolas and Kane (2016) obtains $n^{1/2}$ which leads to the trivial upper bound of 1. This highlights a gap in their upper bound with respect to the local minimax rate in some specific regimes. However, our main contribution with respect to Diakonikolas and Kane (2016) and to the rest of the literature is our lower bound. It is the evidence from a local perspective that two sample testing is more difficult than identity testing.

This chapter is organized as follows. In Section 2.2, an upper bound on the local minimax separation distance for Problem (2.2) is presented. This will entail the construction of a test based on multiple subtests. In Section 2.3, a lower bound that matches the upper bound up to logarithmic factors is proposed. Finally, the proofs of all the results presented in this chapter are left for the later sections.

2.2 Upper bound

In this section, we build a test composed of several tests for Problem (2.2). One of them is related to the test introduced in the context of identity testing in Valiant and Valiant (2017) and Balakrishnan and Wasserman (2017a). The others complement this test, in particular regarding what happens for smaller masses. Here, as explained

TABLE 2.5: Comparison of the upper bounds on the local minimax separation distances depending on which term dominates. Note that $J_\pi \leq I^* \leq m^*$, by definition of these quantities. Each index i belongs to some index range U and $\pi_{(i)}$ contributes to the separation distance rate differently depending on the index range U. The notation $|U|$ refers to the number of elements in U. Current paper corresponds to the local minimax separation distance that we prove for closeness testing as defined in Equation (2.2) with $u = 2.001$ for the lower bound and $u = 1/2$ for the upper bound. Valiant and Valiant (2017) and Balakrishnan and Wasserman (2017a) ([VV17,BW17]) present the local minimax separation distance found for identity testing defined in Equation (2.3), which corresponds to a lower bound for Problem (2.2). Diakonikolas and Kane (2016) ([DK16]) presents a rate which corresponds to an upper bound for Problem (2.2). All the separation distances are presented up to log-factors. $\mathbf{1}_U$ is the indicator function of U applied elementwise.

Index range	$U = \{1, \dots, J_\pi\}$	$U = \{J_\pi, \dots, I^*\}$
Contribution of the terms		
Current paper	$\frac{\left\|\pi_{(.)}^{2/3}\mathbf{1}_U\right\|_1^{3/4}}{\sqrt{n}}$	$\sqrt{\frac{\lvert U\rvert}{n}}\left[\|\pi^2\exp(-un\pi)\|_1^{1/4} \vee \frac{1}{\sqrt{n}}\right]$
Lower bound [VV17,BW17]	$\frac{\left\|\pi_{(.)}^{2/3}\mathbf{1}_U\right\|_1^{3/4}}{\sqrt{n}}$	$\frac{\left\|\pi_{(.)}^{2/3}\mathbf{1}_U\right\|_1^{3/4}}{\sqrt{n}}$
Upper bound from [DK16]	$\frac{\left\|\pi_{(.)}^{2/3}\mathbf{1}_U\right\|_1^{3/4}}{\sqrt{n}}$	$\sqrt{\frac{\lvert U\rvert}{n}}\|\pi_{(.)}^2\mathbf{1}_U\|_1^{1/4}$

Index range	$U = \{I^*, \dots, m^*\}$	$U = \{m^*, \dots, d\}$
Contribution of the terms		
Current paper	$\left\|\pi_{(.)}\mathbf{1}_U\right\|_1$	$\left\|\pi_{(.)}\mathbf{1}_U\right\|_1$
Lower bound from [VV17,BW17]	$\frac{\left\|\pi_{(.)}^{2/3}\mathbf{1}_U\right\|_1^{3/4}}{\sqrt{n}}$	$\left\|\pi_{(.)}\mathbf{1}_U\right\|_1$
Upper bound from [DK16]	$\sqrt{\frac{\lvert U\rvert}{n}}\|\pi_{(.)}^2\mathbf{1}_U\|_1^{1/4}$	$\sqrt{\frac{\lvert U\rvert}{n}}\|\pi_{(.)}^2\mathbf{1}_U\|_1^{1/4}$

in the setting, we observe the following independent sample sets.

$$\mathcal{X} \sim \mathcal{M}(n, p), \quad \mathcal{Y} \sim \mathcal{M}(n, q).$$

Assume from now on that $n \geq 3$. These two sample sets can each be split into 3 independent sample sets. That is, for any $j \leq 3$ and $i \leq d$, we consider the independent sample sets

$$\mathcal{X}^{(j)} \sim \mathcal{M}(\bar{n}, p), \quad \mathcal{Y}^{(j)} \sim \mathcal{M}(\bar{n}, q),$$

where $\bar{n} = \lfloor n/3 \rfloor$.

We will then apply a Poissonization trick in order to consider independent Poisson random variables instead of independent multinomial random variables – see Section 2.6.1 in the Appendix for the precise derivations. Firstly, for $j \in \{1, 2, 3\}$ and $m \in \{1, 2\}$, let $\bar{n}_m^{(j)}$ follow $\mathcal{P}(2\bar{n}/3)$ independently. Note that by concentration of Poisson random variables, we have with probability larger than $1 - 6\exp(-\bar{n}/12)$, that $\bar{n}_m^{(j)} \leq \bar{n}$ for all $j \in \{1, 2, 3\}$ and $m \in \{1, 2\}$ at the same time. With this in mind, we

define the following $6d$ counts. For any $i \leq d$ and $j \in \{1,2,3\}$, let

$$X_i^{(j)} = \sum_{r \leq \bar{n}_1^{(j)} \wedge \bar{n}} \mathbf{1}\{\mathcal{X}_r^{(j)} = i\}, \quad Y_i^{(j)} = \sum_{r \leq \bar{n}_2^{(j)} \wedge \bar{n}} \mathbf{1}\{\mathcal{Y}_r^{(j)} = i\}.$$

By definition of Poisson random variables, on the large probability event such that $\bar{n}_m^{(j)} \leq \bar{n}$ for all $j \in \{1,2,3\}$ and $m \in \{1,2\}$ at the same time, we have that the $X_i^{(j)}$ coincide with independent $\mathcal{P}(2p_i\bar{n}/3)$ and that $Y_i^{(j)}$ coincide with independent $\mathcal{P}(2q_i\bar{n}/3))$, for $i \leq d$ and $j \leq 3$. Sample splitting and Poissonization allow for simpler derivations of guarantees for tests and they can be done without loss of generality in our setting. That is why we will construct our tests based on $(X^{(j)}, Y^{(j)})_{j\leq 3}$. All the probability statements in this section will be with respect to $(\bar{n}_m^{(j)})_{m\leq 2, j\leq 3}$ and $(X^{(j)}, Y^{(j)})_{j\leq 3}$.

Sections 2.2.1–2.2.4 introduce the individual tests as well as their guarantees. Section 2.2.5 combines all the tests into one and produces a problem-dependent upper bound for our setting. The proofs for the upper bound are compiled in Appendix 2.6.

The general strategy behind our construction is to readjust the test presented in Valiant and Valiant (2017) to make it fit to our setting. Indeed, both distributions are unknown in our case, making instance-based minimax optimality all the more complicated. Instead of knowing π directly, it is estimated up to some multiplicative constant when possible. This will induce a gap with the local minimax separation distances for identity testing presented in Valiant and Valiant (2017) and Balakrishnan and Wasserman (2017a). Using other tests, we offset this gap partially. However, as shown in the lower bound, the upper bound is local minimax optimal and the difference in separation distances highlighted in the upper bound is actually fundamental to the problem of closeness testing, making it harder than identity testing in some regimes.

2.2.1 Pre-test: Detection of divergences coordinate-wise

We first define a pre-test. It is an initial test designed to detect cases where some coordinates of p and q are very different from one another. It relies on the ℓ_∞-distance between the observations.

Let $c > 0$, $\hat{q} = (Y^{(3)} \vee 1)/\bar{n}$ and $\hat{p} = (X^{(3)} \vee 1)/\bar{n}$, where the maximum is taken element-wise. The pre-test is defined as

$$\varphi_\infty(X^{(3)}, Y^{(3)}, c, n, d) = \begin{cases} 1, \text{ if there exists } i : |\hat{p}_i - \hat{q}_i| \geq c\sqrt{\frac{\hat{q}_i \log(\hat{q}_i^{-1} \wedge n)}{n}} + c\frac{\log(n)}{n}. \\ 0, \text{ otherwise.} \end{cases}$$

In order to simplify the notations, we will just write $\varphi_\infty(c)$ in the future.

Proposition 8. *Let $\delta \in (0,1)$. Then there exist $c_{\delta,\infty} > 0, \tilde{c}_{\delta,\infty} > 0$ large enough depending only on δ such that the following holds.*

- *If $p = q$, then with probability larger than $1 - 2\delta - 7n^{-1} - 6\exp(-n/100)$,*

$$\varphi_\infty(c_{\delta,\infty}) = 0.$$

- *If there exists $i \leq d$ such that*

$$|p_i - q_i| \geq \tilde{c}_{\delta,\infty}\sqrt{q_i \frac{\log(q_i^{-1} \wedge n)}{n}} + \tilde{c}_{\delta,\infty}\frac{\log(n)}{n},$$

then with probability larger than $1 - 2\delta - 7n^{-1} - 6\exp(-n/100)$,

$$\varphi_\infty(c_{\delta,\infty}) = 1.$$

2.2.2 Definition of the 2/3-test on large coefficients

We now consider a test that is related to the one in Valiant and Valiant (2017) and Balakrishnan and Wasserman (2017a) based on a weighted ℓ_2-norm. But here, the weights are constructed empirically. Such an empirical twist on an existing test in order to obtain adaptive results was also explored in Diakonikolas and Kane (2016). The objective is to detect differences in the coefficients that are larger than $1/n$ in an efficient way.

Let $c > 0$. Set

$$T_{2/3} = \sum_{i \le d} \hat{q}_i^{-2/3}(X_i^{(1)} - Y_i^{(1)})(X_i^{(2)} - Y_i^{(2)}). \tag{2.5}$$

We also define $\hat{t}_{2/3} = \sqrt{n^{-2/3}\|(Y^{(1)})^{2/3}\|_1} + 1$, and

$$\varphi_{2/3}(c) := \varphi_{2/3}(X^{(1)}, Y^{(1)}, X^{(2)}, Y^{(2)}, X^{(3)}, \hat{q}, c, n, d) = \mathbf{1}\{T_{2/3} \ge c\hat{t}_{2/3}\}.$$

Proposition 9. *Let* $\delta > 0$. *Let* $c_{\delta,\infty}$ *defined as in Proposition* (8) *and* $n \ge 4\left(80e^4/\sqrt{\delta/2}\right)^3$. *Then there exist* $c_{\delta,2/3} > 0, \tilde{c}_{\delta,2/3} > 0$ *large enough depending only on* δ *such that the following holds.*

- *If* $p = q$, *then with probability larger than* $1 - 3\delta - 7n^{-1} - 6\exp(-n/100)$,

$$\varphi_{2/3}(c_{\delta,2/3}) = 0 \qquad \text{and} \qquad \varphi_\infty(c_{\delta,\infty}) = 0.$$

- *If*

$$\left\|(p-q)\mathbf{1}\{nq \ge 1\}\right\|_1^2 \ge \frac{\tilde{c}_{\delta,2/3}}{n}\left(\left\|q^2\frac{1}{(q \vee n^{-1})^{4/3}}\right\|_1^{3/2} \vee \left\|q^2\frac{1}{(q \vee n^{-1})^{4/3}}\right\|_1\right),$$

then with probability larger than $1 - 3\delta - 7n^{-1} - 6\exp(-n/100)$,

$$\varphi_{2/3}(c_{\delta,2/3}) = 1 \qquad \text{or} \qquad \varphi_\infty(c_{\delta,\infty}) = 1.$$

This proposition provides an upper bound on the local minimax separation distance. It is related to the upper bound on the local minimax sample complexity obtained in Proposition 2.14 in Diakonikolas and Kane (2016), and Table 2.5 makes a detailed comparison between the results obtained in Diakonikolas and Kane (2016) and ours. In Diakonikolas and Kane (2016), the authors partition the distribution into different empirical level sets. Then for each level set, they apply a standard ℓ_2-test to the pseudo-distributions restricted to that level set. In contrast, we apply only one test with appropriate weights. This is analog with comparing the max test to the 2/3 test both depicted in Balakrishnan and Wasserman (2017a). In the test statistic from Diakonikolas and Kane (2016), the partitioning of the distributions is empirical. Comparatively, we modified the 2/3-test statistic in order to make the weights empirical.

Remark 26. • *Let us compare $T_{2/3}$ defined in Equation (2.5) with the test statistic presented in Valiant and Valiant (2017):*

$$\sum_i \frac{(X_i - nq_i)^2 - X_i}{q_i^{2/3}}. \tag{2.6}$$

We start by explaining their construction. Equation (2.6) is a modified chi-squared statistic producing a local minimax optimal test for identity testing. Now, in closeness testing, q is unknown. That is the reason why we estimate q using $\hat{q}$ and ensure its value cannot be 0 in the denominator. This constraint leads to rates in separation distance which are different from those obtained with the statistic from Equation (2.6) for identity testing. Our other tests tackle the case corresponding to $Y^{(3)} = 0$ as well as possible, but the rates will remain worse than those in identity testing. Such a gap will prove to be intrinsic to closeness testing as we find a lower bound our upper bound.

- *The threshold $\hat{t}_{2/3}$ associated with the definition of $\varphi_{2/3}$ is stochastic. But if π is known, then the problem can be reduced to identity testing and the threshold can be made deterministic as in Valiant and Valiant (2017) and Balakrishnan and Wasserman (2017a), with value $\sqrt{\sum \pi_i^{2/3}} + 1$.*

2.2.3 Definition of the ℓ_2-test for intermediate coefficients

We now construct a test for intermediate coefficients, i.e., those that are too small to have weights computed in a meaningful way using the method in Section 2.2.2. For these coefficients, we simply suggest an ℓ_2-test that is related to the one carried out in Chan et al. (2014) and Diakonikolas and Kane (2016). And we apply this test only on coordinates that we empirically find as being small.

Set

$$T_2 = \sum_{i \le d} (X_i^{(1)} - Y_i^{(1)})(X_i^{(2)} - Y_i^{(2)}) \mathbf{1}\{Y_i^{(3)} = 0\}, \tag{2.7}$$

and

$$\hat{t}_2 = \sqrt{\|Y^{(1)} Y^{(2)} \mathbf{1}\{Y^{(3)} = 0\}\|_1} + \log(n)^2.$$

Write $\varphi_2(c) := \varphi_2(X^{(1)}, Y^{(1)}, X^{(2)}, Y^{(2)}, Y^{(3)}, c, n, d) = \mathbf{1}\{T_2 \ge c\hat{t}_2\}$.

Proposition 10. *Let $\delta \in (0,1)$. Let $c_{\delta,\infty}$ defined as in Proposition (8) and assume $\varphi_\infty(c_{\delta,\infty}) = 0$. We write $s(.)$ such that $q_{s(.)} = q_{(.)}$. There exist $c_{\delta,2} > 0, \tilde{c}_{\delta,2} > 0$ large enough depending only on δ such that the following holds.*

- *If $p = q$, then with probability larger than $1 - 3\delta - 7n^{-1} - 6\exp(-n/100)$,*

$$\varphi_2(c_{\delta,2}) = 0 \qquad \text{and} \qquad \varphi_\infty(c_{\delta,\infty}) = 0.$$

- *If there exists $I \ge J_q$ such that*

$$\left(\sum_{i=J_q}^{I} |p_{s(i)} - q_{s(i)}| \right)^2 \ge c_{\delta,2} \frac{I - J_q}{n} \left[\frac{\log^2(n)}{n} \vee \left(\sqrt{\|q^2 \exp(-nq)\|_1} \right) \right],$$

then with probability larger than $1 - 3\delta - 7n^{-1} - 6\exp(-n/100)$,

$$\varphi_2(c_{\delta,2}) = 1 \qquad \text{or} \qquad \varphi_\infty(c_{\delta,\infty}) = 1.$$

This test based on the ℓ_2-statistic tackles a particular regime where the coefficients of the distribution q are neither too small nor too large. Such an application of an ℓ_2-statistic to an ℓ_1-closeness testing problem is reminiscent of Chan et al. (2014) and Diakonikolas and Kane (2016). In particular, like in Diakonikolas and Kane (2016), we restrict the application of this test to a section of the distribution that is constructed empirically.

Remark 27.
- *Note that T_2 defined in Equation (2.7) is based on the ℓ_2-separation between both samples in the same way as $T_{2/3}$ defined in Equation (2.5). However, T_2 is not reweighted since it focuses on the case when $Y^{(3)} = 0$. This is comparable to the test statistic presented in Diakonikolas and Kane (2016). Indeed, their statistic is not rescaled using the values of q, but they partition it regrouping coefficients of q of the same order instead. Our statistic amounts to doing just that, except that we focus on smaller coefficients only and we partition q empirically.*
- *Once again, we define an empirical threshold $\hat{t}_2$. With the knowledge of π, we would obtain the following deterministic threshold instead:* $\sqrt{\sum_{\pi_i \le n} (n\pi_i)^2} + \log^2(n)$.

2.2.4 Definition of the ℓ_1-test for small coefficients

Finally we define another test to exclude situations where the ℓ_1-norm of the small coefficients in p and q are very different.

Set

$$T_1 = \sum_{i \le d} (X_i^{(1)} - Y_i^{(1)}) \mathbf{1}\{Y_i^{(3)} = 0\}.$$

Write $\varphi_1(c) := \varphi_1(X^{(1)}, Y^{(1)}, Y^{(3)}, c, n, d) = \mathbf{1}\{T_1 \ge c\sqrt{n}\}$.

Proposition 11. *Let $\delta \in (0,1)$. Let $c_{\delta,\infty}, \tilde{c}_{\delta,\infty}$ defined as in Proposition (8). Assume $n \ge 13\delta^{-1}(1 + 9\tilde{c}_{\delta,\infty} \log(n/3)/2)^2$. We write s such that $q_{s(.)} = q_{(.)}$. Then there exist $c_{\delta,1} > 0, \tilde{c}_{\delta,1} > 0$ large enough depending only on δ such that the following holds.*

- *If $p = q$, then with probability larger than $1 - 3\delta - 7n^{-1} - 6\exp(-n/100)$,*

$$\varphi_1(c_{\delta,1}) = 0 \qquad \text{and} \qquad \varphi_\infty(c_{\delta,\infty}) = 0.$$

- *If*

$$\left\| (p-q)\mathbf{1}\{nq \ge 1\} \right\|_1^2 \ge \frac{\tilde{c}_{\delta,1}}{n} \left(\left\| q^2 \frac{1}{(q \vee n^{-1})^{4/3}} \right\|_1^{3/2} \vee \left\| q^2 \frac{1}{(q \vee n^{-1})^{4/3}} \right\|_1 \right),$$

then with probability larger than $1 - 3\delta - 7n^{-1} - 6\exp(-n/100)$,

$$\varphi_1(c_{\delta,1}) = 1 \qquad \text{or} \qquad \varphi_\infty(c_{\delta,\infty}) = 1.$$

As stated in Proposition 11, this test captures the case of large ℓ_1-deviation at places where p and q have small coefficients. This is mainly interesting for cases where there are extremely many small coefficients, making a very crude test the most meaningful tool to use. The pathological cases addressed here contribute to the differences in separation distances with Diakonikolas and Kane (2016).

2.2.5 Combination of the four tests

To conclude, we combine all four tests by taking the maximum value that they output, effectively rejecting the null hypothesis whenever one of the tests is rejected.

Let $\varphi(c_\infty, c_{2/3}, c_2, c_1) = \varphi_\infty(c_\infty) \vee \varphi_{2/3}(c_{2/3}) \vee \varphi_2(c_2) \vee \varphi_1(c_1)$, where $c_\infty, c_{2/3}, c_2, c_1$ are positive constants.

Theorem 21. *Let $\delta < 1$. There exist $c_{\delta,\infty}, c_{\delta,2/3}, c_{\delta,2}, c_{\delta,1}, \tilde{c}_{\delta,\infty}, \tilde{c}_\delta > 0$ that depend only on δ such that the following holds. Let $n \geq \left[13\delta^{-1}(1 + 9\tilde{c}_{\delta,\infty}\log(n/3)/2)^2\right] \vee \left[3\left(80e^4/\sqrt{\delta/2}\right)^3\right]$.*

- *If $p = q$, then with probability larger than $1 - 5\delta - 7n^{-1} - 6\exp(-n/100)$,*

$$\varphi(c_{\delta,\infty}, c_{\delta,2/3}, c_{\delta,2}, c_{\delta,1}) = 0.$$

- *If*

$$\begin{aligned}
&\sum |p_i - q_i| \\
&\geq \tilde{c}_\delta \left\{ \min_{I \geq J_\pi} \left[\left(\sqrt{I}\frac{\log(n)}{n} \right) \vee \left(\frac{\sqrt{I}}{\sqrt{n}} \| q^2 \exp(-nq) \|_1^{1/4} \right) \vee \| (q_{(i)} \mathbf{1}\{i \geq I\})_i \|_1 \right] \right\} \\
&\qquad \vee \left[\frac{\left\| q^2 \frac{1}{(q \vee n^{-1})^{4/3}} \right\|_1^{3/4}}{\sqrt{n}} \right] \vee \left[\sqrt{\frac{\log(n)}{n}} \right].
\end{aligned}$$

  then with probability larger than $1 - 5\delta - 7n^{-1} - 6\exp(-n/100)$,

$$\varphi(c_{\delta,\infty}, c_{\delta,2/3}, c_{\delta,2}, c_{\delta,1}) = 1.$$

Then the theorem can be formulated as the following upper bound.

Corollary 6. *Let $\gamma > 0$. There exists a constant $c_\gamma > 0$ that depends only on γ such that*

$$\begin{aligned}
&\rho_\gamma^*(H_{0,\pi}^{(\mathrm{Clo})}, H_{1,\pi}^{(\mathrm{Clo})}, \mathcal{T}_n^{(Clo)}) \\
&\quad \leq c_\gamma \left\{ \min_{I \geq J_\pi} \left[\left(\sqrt{I}\frac{\log(n)}{n} \right) \vee \left(\frac{\sqrt{I}}{\sqrt{n}} \| \pi^2 \exp(-n\pi/2) \|_1^{1/4} \right) \vee \| (\pi_{(i)} \mathbf{1}\{i \geq I\})_i \|_1 \right] \right\} \\
&\quad \vee \left[\frac{\left\| (\pi_{(i)}^{2/3} \mathbf{1}\{i \leq J_\pi\})_i \right\|_1^{3/4}}{\sqrt{n}} \right] \vee \left[\sqrt{\frac{\log(n)}{n}} \right].
\end{aligned}$$

Thus, once we have aggregated all four tests, we end up with an upper bound on the local minimax separation distance for closeness testing defined in Equation (2.2). Most importantly, the knowledge of π is not exploited by the test. So our method reaches the displayed rate adaptively to π. That is, the separation distance does not just consider the worst π. Instead, it depends on π although it is not an input parameter in the test. In Table 2.5, the contributions of the different coefficients from π are summarized into different regimes, along with the regimes obtained in Valiant and Valiant (2017) and Diakonikolas and Kane (2016). Our upper bound improves upon that of Diakonikolas and Kane (2016) as emphasized in Section 2.1.3. We manage

to obtain separation distances comparable to those found in identity testing defined in Equation (2.3). In particular, the terms $\|(\pi_{(i)}\mathbf{1}\{i \geq I\})_i\|_1$ and $\frac{\left\|(\pi_{(i)}^{2/3}\mathbf{1}\{i\leq J_\pi\})_i\right\|_1^{3/4}}{\sqrt{n}}$ can also be found in identity testing. However, the differences that we point out in the upper bound turn out to be fundamental to closeness testing. Indeed, we present a matching lower bound in the following section, which represents our main contribution.

2.3 Lower bound

This section will focus on the presentation of a lower bound on the local minimax separation distance for closeness testing defined in Equation (2.2). Since the lower bound will match the upper bound previously presented, our test will turn out to be local minimax optimal.

Theorem 22. *Let $\pi \in \mathbf{P}$ and $\gamma, v > 0$. Assume $n \geq 2^8$. There exists a constant $c_{\gamma,v} > 0$ that depends only on γ, v such that the following holds.*

$$\rho_\gamma^*(H_{0,\pi}^{(\mathrm{Clo})}, H_{1,\pi}^{(\mathrm{Clo})}, \mathcal{T}_n^{(Clo)}) \geq c_{\gamma,v}\left\{\min_{I\geq J_\pi}\left[\frac{\sqrt{I}}{n} \vee \left(\sqrt{\frac{I}{n}}\|\pi^2\exp(-(2+v)n\pi)\|_1^{1/4}\right)\right.\right.$$
$$\left.\left. \vee \|(\pi_{(i)}\mathbf{1}\{i \geq I\})_i\|_1\right]\right\} \vee \frac{\left\|(\pi_{(i)}^{2/3}\mathbf{1}\{i \leq J_\pi\})_i\right\|_1^{3/4}}{\sqrt{n}} \vee \sqrt{\frac{1}{n}}.$$

The details of the proof can be found in Section 2.7 of the Appendix. But we provide the intuition through the following sketch of the proof.

Sketch of the proof of Theorem 22. The construction of the lower bound can be decomposed into three propositions. We first state Proposition 12, which is a corollary from Valiant and Valiant (2017) and Balakrishnan and Wasserman (2017a) and it will provide an initial lower bound on the local minimax separation distance. We will refine this lower bound using Propositions 13 and 14, which constitute our main contributions. The general strategy is the same for both propositions. At first, we reduce the testing problem to a smaller one that is difficult enough and which is not yet covered by Proposition 12. Afterwards, the idea is to hide the discrepancies between distributions in the smaller coefficients, which is justified by the thresholding effect already witnessed in the upper bound. Indeed coefficients corresponding to low probabilities have a great chance of generating 0's. So the information on the coefficients being small to different degrees is lost.

Proposition 12 relies on the fact that two-sample testing is at least as hard as its one-sample counterpart. It is also the most convenient formulation of the local minimax separation distance from Valiant and Valiant (2017) and Balakrishnan and Wasserman (2017a) in order to compare it with our results.

Proposition 12. *Let $\pi \in \mathbf{P}$ and $\gamma > 0$. There exists a constant $c_\gamma > 0$ that depends only on γ such that*

$$\rho_\gamma^*(H_{0,\pi}^{(\mathrm{Clo})}, H_{1,\pi}^{(\mathrm{Clo})}, \mathcal{T}_n^{(Clo)})$$
$$\geq c_\gamma \min_m\left[\frac{\|(\pi_{(i)}^{2/3}\mathbf{1}\{2 \leq i < m\})_i\|_1^{3/4}}{\sqrt{n}} \vee \frac{1}{n} \vee \|(\pi_{(i)}\mathbf{1}\{i \geq m\})_i\|_1\right].$$

The next proposition is a novel construction, which settles the case for small coefficients.

Proposition 13. *Consider some $\pi \in \mathbf{P}$ and $\gamma > 0$. Set for $v \geq 0$ and with the convention $\min_{j \leq d} \emptyset = d$,*

$$I_{v,\pi} = \min_{J_\pi \leq j \leq d} \Big\{ \{j : \pi_{(j)} \leq \sqrt{C_\pi / j}\} \cap \{j : \sum_{i \geq j} \exp(-2n\pi_{(i)}) \pi_{(i)}^2 \leq C_\pi\} \\ \cap \{j : \sum_{i \geq j} \pi_{(i)} \leq \sum_{J_\pi \leq i < j} \pi_{(i)}\} \Big\},$$

where $C_\pi = \frac{\sqrt{\sum_i \pi_i^2 \exp(-2(1+v)n\pi_i)}}{n}$. There exist constants $c_{\gamma,v}, c'_{\gamma,v}, c''_{\gamma,v} > 0$ that depend only on γ, v such that the following holds. Assume that $\|\pi^2 \exp(-2(1+v)n\pi)\|_2^2 \geq \frac{c''_{\gamma,v}}{n^2}$ and $n \geq 2^8$. Then

$$\rho^*_\gamma(H^{(\mathrm{Clo})}_{0,\pi}, H^{(\mathrm{Clo})}_{1,\pi}, \mathcal{T}^{(Clo)}_n) \\ \geq c_{\gamma,v} \Bigg[\Big[\|(\pi_{(i)} \mathbf{1}\{i \geq I_{v,\pi}\})_i\|_1 \vee \frac{I_{v,\pi} - J_\pi}{\sqrt{I_{v,\pi} n}} \|\pi^2 \exp(-2(1+v)\pi)\|_1^{1/4} \Big] \\ \wedge \|(\pi_{(i)} \mathbf{1}\{i \geq J_\pi\})_i\|_1 \Bigg] - \frac{c'_{\gamma,v}}{\sqrt{n}},$$

where J_π is defined at the end of Section 2.1.1.

The proof of this proposition and the following one is based on a classical Bayesian approach for minimax lower bounds. It heavily relies on explicit choices of prior distributions over the couples (p, q) either corresponding to hypothesis set $H^{(\mathrm{Clo})}_{0,\pi}$ or $H^{(\mathrm{Clo})}_{1,\pi}(\rho)$. The goal is then to show that the chosen priors are so close that the risk $R(H^{(\mathrm{Clo})}_{0,\pi}, H^{(\mathrm{Clo})}_{1,\pi}(\rho), \hat{\theta}_n)$ is at least as large as γ for a fixed n. Details on the general approach are provided in Appendix 2.7.2.

The brunt of our contribution relies on the definition of appropriate priors. The priors are enforced to have support in $\mathbf{P}^2$, as detailed in the proof with ideas related to the Poissonization trick. But this is only a technical difficulty which is not fundamental from an information theoretic perspective. A more crucial step regards constructing prior distributions on "non-normalised" versions of the vectors (p, q). We use the notation $(\tilde{p}, \tilde{q})$ for the "non-normalised" vectors associated with the prior distributions.

Let us now present the prior distributions on the parameters $(\tilde{p}, \tilde{q})$ defined for the proof of Proposition 13. π and n are fixed, and the priors critically revolve around π and perturbations thereof in order to obtain a local minimax optimal lower bound. We start by defining an index set $\mathcal{A}$ corresponding to a subset of elements of π containing a fixed proportion of each significant level set S_π. Then $\mathcal{A}$ is a set of indices such that $(\pi_i)_{i \in \mathcal{A}}$ is a vector with a similar shape to π and the elements from $\mathcal{A}^C$ can be used in order to define normalised (p, q).

Under both the null and the alternative hypotheses, the prior distributions are defined such that for any $i \in \mathcal{A}^C$, $\tilde{p}_i = \tilde{q}_i = \pi_i$. We now consider the definition of $(\tilde{p}, \tilde{q})$ on $\mathcal{A}$. Under any of both hypotheses, the elements of $\tilde{q}$ restricted to $\mathcal{A}$ are taken at random uniformly from the elements of π restricted to $\mathcal{A}$.

- Under the null hypothesis, $\tilde{p}$ is set equal to $\tilde{q}$.
- Under the alternative hypothesis, $\tilde{p}$ is a stochastic vector that differs from $\tilde{q}$ in the following way:

– All coordinates larger than $1/n$ are set equal to those of $\tilde{q}$.
– For the other coordinates, set $\tilde{p}_i = \tilde{q}_i(1+\xi_i)$, where ξ_i is uniform on $\{-\varepsilon_i^*, \varepsilon_i^*\}$, and ε_i^* is defined in an implicit way in Lemma 15.

The quantities ε_i^*'s are defined to satisfy the conditions in Lemma 15. The intuition associated with those conditions are the following.

- There is no deviation for the larger coefficients, i.e., $\tilde{p} = \tilde{q}$ for coefficients larger than $1/n$.
- The ℓ_2-separation and the ℓ_∞-distance between $\tilde{p}$ and $\tilde{q}$ are upper bounded with high probability, making the discrepancy hard to detect.
- The ℓ_1-distance between $\tilde{p}$ and $\tilde{q}$ is lower bounded with high probability by the local minimax separation distance to be proven.
- The way $\tilde{q}$ deviates from π creates some uncertainty. This makes it difficult to leverage any knowledge on $\tilde{q}$ for constructing the test besides the fact that (the normalised version of) $\tilde{q}$ is in $\mathbf{P}_\pi$.

Finally, the following proposition complements Proposition 13 in the case where the tail coefficients are very small.

Proposition 14. *Let $\pi \in \mathbf{P}$ and $\gamma, v > 0$. There exist constants $\tilde{c}_{\gamma,v}, c_{\gamma,v}, c'_{\gamma,v} > 0$ that depend only on γ, v such that the following holds. Assume that $\|\pi^2 \exp(-2(1+v)\pi)\|_1 \leq \tilde{c}_{\gamma,v}/n^2$. Then*

$$\rho_\gamma^*(H_{0,\pi}^{(\mathrm{Clo})}, H_{1,\pi}^{(\mathrm{Clo})}, \mathcal{T}_n^{(Clo)}) \geq c_{\gamma,v}\|(\pi_{(i)}\mathbf{1}\{i \geq J_\pi)_i\}\|_1 - c'_{\gamma,v}\sqrt{\|(\pi_{(i)}^2\mathbf{1}\{i \geq J_\pi\})_i\|_1},$$

where J_π is defined in Section 2.1.1.

This proposition refines Proposition 13 in the specific case where $\|\pi^2 \exp(-2(1+v)\pi)\|_1$ is small, and the construction of the priors is related, but simpler. Combining Propositions 12, 13 and 14 lead to the lower bound in Theorem 22.

Thus a lower bound is constructed for the local minimax separation distance, which characterizes the difficulty of closeness testing defined in Equation (2.2). In fact, the lower bound matches the upper bound up to log terms. Thus, we have a good envelope of the local minimax rate. We firstly conclude explicitly that there exist some π such that $\rho_\gamma^*(H_{0,\pi}^{(\mathrm{Clo})}, H_{1,\pi}^{(\mathrm{Clo})}, \mathcal{T}_n^{(\mathrm{Clo})}) > \rho_\gamma^*(H_{0,\pi}^{(\mathrm{Id})}, H_{1,\pi}^{(\mathrm{Id})}, \mathcal{T}_n^{(\mathrm{Id})})$, that is, two-sample testing is strictly harder than one-sample testing for some distributions π. Secondly, the result highlights the location of the gap in further detail than the worst-case study of Chan et al. (2014). We provide a detailed comparison between results in Section 2.1.3 using Table 2.5.

2.4 Conclusion & Discussion

In this chapter, we have established the local minimax near-optimal separation distance for the closeness testing problem defined in Equation (2.2). It represents the first near-tight lower bound for local minimax closeness testing, and the first test that matches it up to log terms. The minimax rate is adaptive to π in the following sense. The test we construct only takes samples from p and q, but its testing rate optimally depends on π, as evidenced by the lower bound. The construction of the lower bound heavily relies on our formalization of closeness testing from Equation (2.2). Such

a formalization is critically different from its identity testing counterpart, because of q remaining unfixed. So we end up considering a testing problem, where both hypotheses are composite. Comparing our local minimax separation distance with the one achievable in local minimax identity testing, a gap can be noted. Indeed, closeness testing turns out to be more difficult, especially when there are terms which are rather small without being negligible (corresponding to the indices between J_π and m^*). But it is also noteworthy that both rates match otherwise.

On the horizon, the corresponding local minimax sample complexity for closeness testing has yet to be found. Besides, the upper bound could be made tighter in order to bridge the gap caused by the log factors. Finally, our analysis focuses on discrete distributions but the formalization of the problem of closeness testing presented in this chapter generalizes well to other settings. Indeed, our formalization relies on q being restrained to the set $\mathbf{P}_\pi$. Now, an analog set to $\mathbf{P}_\pi$ can be defined with an additional regularity condition in a continuous setting. So the extension of our study to densities still remains a major direction to be explored and it would be interesting to extend the framework from this chapter as Balakrishnan and Wasserman (2017a) does for Valiant and Valiant (2017) in the context of identity testing.

2.5 Preliminary results on the Poisson distribution

The proofs to our theorems will be provided for Poisson distributions which can be translated into results for multinomial distributions. Similar considerations of independent Poisson samples in order to simplify the proofs are made in Chan et al. (2014) and Valiant and Valiant (2017).

We first provide an equivalence result between the samples from a multinomial distribution and samples from independent Poisson distributions.

We remind Theorem 2 presented in Section 1.3.1.

Theorem 23. *Let $n \in \mathbb{R}^+$, $p \in \mathbf{P}$. Let $\hat{n} \sim \mathcal{P}(n)$. Let the conditional distribution of ξ be $\mathcal{M}(\hat{n}, p)$, conditionally on $\hat{n}$. For any $i \leq d$, we have $X_i = \sum_{j=1}^{\hat{n}} \mathbf{1}\{\xi_j = i\}$. Then we have independent*

$$X_i \sim \mathcal{P}(np_i).$$

Now, the following lemma states that Poisson samples concentrate around their mean.

Lemma 7. *If $Z \sim \mathcal{P}(\lambda)$, where $\lambda > 0$,*

$$\mathbb{P}(|Z - \lambda| \geq \lambda/2) \leq 2 \exp\left(-\frac{\lambda}{12}\right).$$

Proof. If $Z \sim \mathcal{P}(\lambda)$, where $\lambda > 0$, we have, by concentration of the Poisson random variables, that for any $t \geq 0$,

$$\mathbb{P}(|Z - \lambda| \geq t) \leq 2 \exp\left(-\frac{t^2}{2(\lambda + t)}\right).$$

In particular,

$$\mathbb{P}(|Z - \lambda| \geq \lambda/2) \leq 2 \exp\left(-\frac{\lambda}{12}\right).$$

□

2.6 Proofs of the upper bounds: Propositions 8, 9, 10, 11 and Theorem 21

For any $i \in \{1,2,3\}$, we write $\mathbb{E}^{(i)}, \mathbb{V}^{(i)}$ for the expectation and variance with respect to $(X^{(i)}, Y^{(i)})$ and $\bar{n}_m^{(i)}$. $\mathbb{E}$ and $\mathbb{V}$ denote the expectation and variance with respect to all sample sets and all $\bar{n}_m^{(j)}$. We write for all $i \leq d$, $\Delta_i = p_i - q_i$. Assume without loss of generality that q is ordered such that $q_1 \geq q_2 \geq \ldots \geq q_d$. We remind the reader about the following notation: $\pi_{(1)} \geq \pi_{(2)} \geq \ldots \geq \pi_{(d)}$. Throughout Section 2.6, let $I \geq J_q$ and we write $J := J_q$.

2.6.1 From multinomial samples to independent Poisson samples

Let $\hat{n} \sim \mathcal{P}(n)$. We define the following independent random variables $\mathcal{Z}_1, \ldots, \mathcal{Z}_{\hat{n}}$ each taking value in $\{1, \ldots, d\}$ according to the probability vector p, and we set $m = \lfloor 3n/2 \rfloor \wedge \hat{n}$.

We define $Z_i = \sum_{j=1}^{m} \mathbf{1}\{\mathcal{Z}_j = i\}$ and $\tilde{Z}_i = \sum_{j=1}^{\hat{n}} \mathbf{1}\{\mathcal{Z}_j = i\}$ for any $i \leq d$. By Theorem 2, we have independent

$$\tilde{Z}_i \sim \mathcal{P}(np_i),$$

for any $i \leq d$. Note that $(\tilde{Z}_i)_i$ coincides with $(Z_i)_i$ on the event where $\hat{n} \leq \lfloor 3n/2 \rfloor$.

Also, we have by Lemma (7)

$$\mathbb{P}(\hat{n} \leq 3n/2) \geq 1 - \exp\left(-\frac{n}{12}\right).$$

And so on an event of probability larger than $1 - \exp\left(-\frac{n}{12}\right)$, the $(Z_i)_i$ coincides with the $(\tilde{Z}_i)_i$, i.e. with independent $\mathcal{P}(np_i)$ samples.

Applying this to each of our sample sets $\mathcal{X}^{(j)}, \mathcal{Y}^{(j)}$ respectively associated with $\bar{n}_m^{(j)}$ for $j \in \{1,2,3\}$ and $m \in \{1,2\}$, we finally obtain that on an event of probability larger than $1 - 6\exp(-\bar{n}/18)$, $(X_i^{(j)})_i$ coincides with independent $\mathcal{P}(\bar{n}p_i/6)$, and $(Y_i^{(j)})_i$ coincides with independent $\mathcal{P}(2\bar{n}q_i/3)$.

From this point on, we will therefore assume that

$$X^{(j)} \sim \mathcal{P}(2\bar{n}p/3), \quad Y^{(j)} \sim \mathcal{P}(2\bar{n}q/3),$$

and that they are independent accross j. In what follows we will only consider events intersected with that event of probability larger than $1 - 6\exp(-n/18)$ where $\bar{n}_m^{(j)} \leq \bar{n}$. In what follows, since we always reason up to multiplicative constants, we will write n instead of $2\bar{n}/3$ to simplify notations.

2.6.2 Proof of Proposition 8

In order to derive the guarantees on the pre-test stated in Proposition 8, we first provide the following lemma. The deviation from a Poisson random variable to its expected value will be bounded depending on the outcome of the random variable, and then depending on its expected value.

Lemma 8. *Let $\lambda \in (\mathbb{R}^+)^d$ such that $\sum_i \lambda_i = n$. Let independent $Z_i \sim \mathcal{P}(\lambda_i)$ for any $i \leq d$. Let $\bar{z} = Z/n$. Let $\delta \in (0,1)$ and $a := 16\frac{\log(2n/\delta)}{n}$. With probability larger than*

$1-\delta-n^{-1}$ *and for all* $i \le d$ *we have*

$$|\bar{z}_i - \lambda_i/n| \le 2\sqrt{\frac{12(\bar{z}_i \vee a)\log(96\log(2n/\delta)(\bar{z}_i^{-1} \wedge a^{-1})/\delta)}{n}} + 2\frac{\log(96\log(2n/\delta)(a^{-1})/\delta)}{n},$$

and

$$2\sqrt{\frac{12(\bar{z}_i \vee a)\log(96\log(2n/\delta)(\bar{z}_i^{-1} \wedge a^{-1})/\delta)}{n}} + 2\frac{\log(96\log(2n/\delta)(a^{-1}/\delta)}{n} \le 12\sqrt{\frac{((\lambda_i/n) \vee a)\log\left(384\log(2n/\delta)\left((\lambda_i/n)^{-1} \wedge a^{-1}\right)/\delta\right)}{n}} + 2\frac{\log\left(384\log(2n/\delta)a^{-1}/\delta\right)}{n}.$$

So an immediate corollary to this lemma is the following:

Corollary 7. *With probability larger than* $1-2\delta-2n^{-1}-6\exp(-n/18)$, $\varphi_\infty(c_{\delta,\infty}) = 1$ *if there exists* $i \le d$ *such that*

- *if* $q_i \ge 16\frac{\log(2n/\delta)}{n}$*:*

$$|\Delta_i| \ge 50\sqrt{\frac{q_i \log\left(384\log(2n/\delta)q_i^{-1}/\delta\right)}{n}} + 300\frac{\log\left(384\log(2n/\delta)a^{-1}/\delta\right)}{n}.$$

- *if* $q_i \le 16\frac{\log(2n/\delta)}{n}$*:*

$$|\Delta_i| \ge 50\sqrt{\frac{a\log\left(384\log(2n/\delta)a^{-1}/\delta\right)}{n}} + 300\frac{\log\left(384\log(2n/\delta)a^{-1}/\delta\right)}{n}.$$

If $\Delta = 0$*, then* $\varphi_\infty(c_{\delta,\infty}) = 0$ *with probability larger than* $1-2\delta-2n^{-1}-6\exp(-n/18)$.

This corollary implies that there exists a universal constant $c > 0$ such that the preliminary test rejects the null hypothesis on an event of probability larger than $1-2\delta-2n^{-1}-6\exp(-n/18)$, when Δ_i is such that

$$|\Delta_i| \ge c\sqrt{q_i\frac{\log((q_i^{-1} \wedge n)/\delta)}{n}} + c\frac{\log(n/\delta)}{n}.$$

This leads to the result stated in Proposition 8.

Proof of Lemma 8. **Analysis of the small λ_i's.** We consider every i such that $\lambda_i \le n^{-2}$. Then for any such i,

$$\mathbb{P}(\bar{z}_i > 1/n) = 1-(1+\lambda_i)e^{-\lambda_i} \le 1-(1+\lambda_i)(1-\lambda_i) = \lambda_i^2.$$

So

$$\mathbb{P}(\cup_{j:\lambda_j \le n^{-2}}\{\bar{z}_j > 1/n\}) \le \sum_j \lambda_j^2 \le \frac{1}{n^2}\sum_j \lambda_j = 1/n.$$

So with probability larger than $1-1/n$, we have for all $\lambda_i \le 1/n^2$ at the same time that

$$\bar{z}_i \le 1/n.$$

Let $a = 16\frac{\log(2n/\delta)}{n}$. Then with probability larger than $1-1/n$, for every $\lambda_i \le 1/n^2$ at the same time,

$$|\bar{z}_i - \lambda_i/n| \le 2\sqrt{\frac{12(\bar{z}_i \vee a)\log(96\log(2n/\delta)(\bar{z}_i^{-1}\wedge a^{-1})/\delta)}{n}} + 2\frac{\log(96\log(2n/\delta)(\bar{z}_i^{-1}\wedge a^{-1})/\delta)}{n},$$

and

$$2\sqrt{\frac{12(\bar{z}_i \vee a)\log(96\log(2n/\delta)(\bar{z}_i^{-1}\wedge a^{-1})/\delta)}{n}} + 2\frac{\log(96\log(2n/\delta)(\bar{z}_i^{-1}\wedge a^{-1})/\delta)}{n}$$
$$\le 2\sqrt{\frac{36((\lambda_i/n)\vee a)\log\left(384\log(2n/\delta)\left((\lambda_i/n)^{-1}\wedge a^{-1}\right)/\delta\right)}{n}} + 2\frac{\log\left(384\log(2n/\delta)\left((\lambda_i/n)^{-1}\wedge a^{-1}\right)/\delta\right)}{n}.$$

Analysis of the large λ_i's. We consider every i such that $\lambda_i > n^{-2}$.

If $Z_i \sim \mathcal{P}(\lambda_i)$, where $\lambda_i > 0$, we have, by concentration of the Poisson random variables, that for any $t \ge 0$,

$$\mathbb{P}(|Z_i - \lambda_i| \ge t) \le 2\exp\left(-\frac{t^2}{2(\lambda_i + t)}\right).$$

We set $\tilde{\delta}_i$ as $2\exp\left(-\frac{t^2}{2(\lambda_i+t)}\right)$, the inequality implies that with probability larger than $1-\tilde{\delta}_i$,

$$|\bar{z}_i - \lambda_i/n| \le 2\sqrt{\frac{(\lambda_i/n)\log(2/\tilde{\delta}_i)}{n}} + 2\frac{\log(2/\tilde{\delta}_i)}{n}. \tag{2.8}$$

So we write $\tilde{\delta}_i = \lambda_i\delta/n$. Then, since $\sum_i \lambda_i = n$, we have with probability larger than $1-\delta$, for every i such that $\lambda_i > n^{-2}$ at the same time

$$|\bar{z}_i - \lambda_i/n| \le 2\sqrt{\frac{(\lambda_i/n)\log(2n/(\lambda_i\delta))}{n}} + 2\frac{\log(2n/(\lambda_i\delta))}{n}$$
$$\le 2\sqrt{\frac{(\lambda_i/n)\log(2n/[(\lambda_i\vee n^{-2})\delta])}{n}} + 2\frac{\log(2n/[(\lambda_i\vee n^{-2})\delta])}{n}. \tag{2.9}$$

By considering two subcases, let us prove the following inequality on an event of probability larger than $1-\delta$, for all i

$$((\lambda_i/n)\vee a)/4 \le \bar{z}_i \vee a \le 3((\lambda_i/n)\vee a). \tag{2.10}$$

Subcase $\lambda_i/n \geq a$. By Equation (2.9), we have on an event of probability larger than $1-\delta$, that for all i such that $\lambda_i/n \geq 16\frac{\log(2n/\delta)}{n} = a$,

$$|\bar{z}_i - \lambda_i/n| \leq 2\sqrt{\frac{(\lambda_i/n)\log(2n/\delta)}{n}} + 2\frac{\log(2n/\delta)}{n} \leq 5\lambda_i/(8n).$$

So

$$\lambda_i/(4n) \leq \bar{z}_i \leq 3\lambda_i/n.$$

That is,

$$((\lambda_i/n) \vee a)/4 \leq \bar{z}_i \vee a \leq 3((\lambda_i/n) \vee a).$$

Subcase $n^{-3} < \lambda_i/n < a$. By Equation (2.9), we have on the same event of probability larger than $1-\delta$, that for all i such that $n^{-3} < \lambda_i/n < 16\frac{\log(2n/\delta)}{n} = a$,

$$|\bar{z}_i - \lambda_i/n| \leq 14\frac{\log(2n/\delta)}{n} + 6\frac{\log(2n/\delta)}{n} \leq 20\frac{\log(2n/\delta)}{n} \leq 2a.$$

So $\bar{z}_i \leq 3a$, and then,

$$((\lambda_i/n) \vee a)/4 = a/4 \leq \bar{z}_i \vee a \leq 3a = 3((\lambda_i/n) \vee a).$$

Conclusion for the large λ_i's.

Let us first reformulate Equation (2.9) using the definition of a. We have with probability larger than $1-\delta$ that for all i,

$$|\bar{z}_i - \lambda_i/n| \leq 2\sqrt{\frac{3(\lambda_i/n)\log(32\log(2n/\delta)/[((\lambda_i/n) \vee a)\delta])}{n}} + 6\frac{\log(32\log(2n/\delta)/[((\lambda_i/n) \vee a)\delta])}{n}.$$

So by application of Equation (2.10), we get that with probability larger than $1-\delta$ and for all i we have

$$|\bar{z}_i - \lambda_i/n| \leq 2\sqrt{\frac{12(\bar{z}_i \vee a)\log(96\log(2n/\delta)(\bar{z}_i^{-1} \wedge a^{-1})/\delta)}{n}} + 2\frac{\log(96\log(2n/\delta)(\bar{z}_i^{-1} \wedge a^{-1})/\delta)}{n},$$

and

$$2\sqrt{\frac{12(\bar{z}_i \vee a)\log(96\log(2n/\delta)(\bar{z}_i^{-1} \wedge a^{-1})/\delta)}{n}} + 2\frac{\log(96\log(2n/\delta)(\bar{z}_i^{-1} \wedge a^{-1})/\delta)}{n}$$
$$\leq 2\sqrt{\frac{36((\lambda_i/n) \vee a)\log\left(384\log(2n/\delta)\left((\lambda_i/n)^{-1} \wedge a^{-1}\right)/\delta\right)}{n}} + 2\frac{\log\left(384\log(2n/\delta)\left((\lambda_i/n)^{-1} \wedge a^{-1}\right)/\delta\right)}{n}.$$

□

2.6.3 Proof of Proposition 9

Proposition 9 provides guarantees on the test $\varphi_{2/3}$. In order to prove it, let us first consider the associated statistic $T_{2/3}$.

Expression of the test statistic.

We have

$$T_{2/3} = \sum_{i \le d} \hat{q}_i^{-2/3} (X_i^{(1)} - Y_i^{(1)})(X_i^{(2)} - Y_i^{(2)}).$$

So taking the expectation highlights different terms associated with different independent sub-samples.

$$\mathbb{E}T_{2/3} = \sum_{i \le d} \mathbb{E}^{(3)}(\hat{q}_i^{-2/3}) \mathbb{E}^{(1)}(X_i^{(1)} - Y_i^{(1)}) \mathbb{E}^{(2)}(X_i^{(2)} - Y_i^{(2)}),$$

and

$$\begin{aligned} \mathbb{V}T_{2/3} &= \sum_{i \le d} \mathbb{V}\Big[\hat{q}_i^{-2/3}(X_i^{(1)} - Y_i^{(1)})(X_i^{(2)} - Y_i^{(2)})\Big] \\ &\le \sum_{i \le d} \mathbb{E}^{(3)}(\hat{q}_i^{-4/3}) \mathbb{E}^{(1)}[(X_i^{(1)} - Y_i^{(1)})^2] \mathbb{E}^{(2)}[(X_i^{(2)} - Y_i^{(2)})^2]. \end{aligned}$$

Terms that depend on the first and second sub-samples. We have

$$\mathbb{E}^{(1)}(X_i^{(1)} - Y_i^{(1)}) \mathbb{E}^{(2)}(X_i^{(2)} - Y_i^{(2)}) = n^2 \Delta_i^2.$$

and

$$\begin{aligned} \mathbb{E}^{(1)}[(X_i^{(1)} - Y_i^{(1)})^2] \mathbb{E}^{(2)}[(X_i^{(2)} - Y_i^{(2)})^2] &= \Big[\mathbb{E}^{(1)}[(X_i^{(1)} - Y_i^{(1)})^2]\Big]^2 \\ &= \Big[\mathbb{E}^{(1)}[(X_i^{(1)} - Y_i^{(1)} - n\Delta_i)^2] + n^2\Delta_i^2\Big]^2 \\ &= [n(p_i + q_i) + n^2\Delta_i^2]^2. \end{aligned}$$

Terms that depend on the third sub-sample. Now, the following lemma will help us control the terms associated with $\hat{q}$.

Lemma 9. *Assume that $Z \sim \mathcal{P}(\lambda)$. Then for $r \in \{2/3, 4/3\}$*

$$\frac{1}{2}\Big(\frac{1}{(e^2\lambda) \vee 1}\Big)^r \le \mathbb{E}[(Z \vee 1)^{-r}] \le 6\Big(\frac{e^2}{\lambda \vee 1}\Big)^r.$$

The proof of the lemma is at the end of the section. By direct application of Lemma 9, we have with probability larger than $1 - 6\exp(-n/18)$ with respect to $(\bar{n}_m^{(j)})_{m \le 2, j \le 3}$,

$$\mathbb{E}^{(3)}\left[\hat{q}_i^{-2/3}\right] \ge \frac{n^{2/3}}{2e^2}\Big(\frac{1}{(nq_i) \vee 1}\Big)^{2/3},$$

and

$$\mathbb{E}^{(3)}\left[\hat{q}_i^{-4/3}\right] \le 6e^4 n^{4/3}\Big(\frac{1}{(nq_i) \vee 1}\Big)^{4/3}.$$

Bound on the expectation and variance for $T_{2/3}$. We obtain with probability larger than $1-6\exp(-n/18)$ with respect to $(\bar{n}_m^{(j)})_{m\le 2, j\le 3}$

$$\mathbb{E}T_{2/3} \ge \sum_{i\le d} \frac{n^{2/3}}{2e^2}\Big(\frac{1}{(nq_i)\vee 1}\Big)^{2/3}\Big[n^2\Delta_i^2\Big] = \frac{n^2}{2e^2}\Big\|\Delta^2\Big(\frac{1}{q\vee n^{-1}}\Big)^{2/3}\Big\|_1. \tag{2.11}$$

and

$$\begin{aligned}
\mathbb{V}(T_{2/3}) &\le \sum_{i\le d} 6e^4 n^{4/3}\Big(\frac{1}{(nq_i)\vee 1}\Big)^{4/3}[n(p_i+q_i)+n^2\Delta_i^2]^2 \\
&\le 12e^4 n^{4/3}\bigg[\Big\|\Big(\frac{1}{(nq)\vee 1}\Big)^{4/3} n^2(p+q)^2\Big\|_1 + n^4\Big\|\Big(\frac{1}{(nq)\vee 1}\Big)^{4/3}\Delta^4\Big\|_1\bigg] \\
&\le 100e^4 n^{4/3} n^2\bigg[\Big\|\Big(\frac{1}{(nq)\vee 1}\Big)^{4/3} q^2\Big\|_1 + \Big\|\Big(\frac{1}{(nq)\vee 1}\Big)^{4/3}\Delta^2\Big\|_1 \\
&\qquad + n^2\Big\|\Big(\frac{1}{(nq)\vee 1}\Big)^{4/3}\Delta^4\Big\|_1\bigg].
\end{aligned}$$

This implies with probability larger than $1-6\exp(-n/18)$ with respect to $(\bar{n}_m^{(j)})_{m\le 2, j\le 3}$,

$$\begin{aligned}
\sqrt{\mathbb{V}(T_{2/3})} \le 10e^2 n\bigg[&\sqrt{\Big\|\Big(\frac{1}{q\vee n^{-1}}\Big)^{4/3} q^2\Big\|_1} + \sqrt{\Big\|\Big(\frac{1}{q\vee n^{-1}}\Big)^{4/3}\Delta^2\Big\|_1} \\
&+ n\sqrt{\Big\|\Big(\frac{1}{q\vee n^{-1}}\Big)^{4/3}\Delta^4\Big\|_1}\bigg].
\end{aligned} \tag{2.12}$$

Analysis of $T_{2/3}$ under $H_{0,\pi}^{(\mathbf{Clo})}$ and $H_{1,\pi}^{(\mathbf{Clo})}(\rho)$. Let us inspect the behaviour of statistic $T_{2/3}$ under the null and the alternative hypotheses. We aim at showing that a test based on $T_{2/3}$ will have different outcomes under $H_{0,\pi}^{(\mathrm{Clo})}$ and $H_{1,\pi}^{(\mathrm{Clo})}$ with large probability.

Under $H_{0,\pi}^{(\mathrm{Clo})}$. We have with probability larger than $1-6\exp(-n/18)$ with respect to $(\bar{n}_m^{(j)})_{m\le 2, j\le 3}$,

$$\mathbb{E}T_{2/3} = 0, \quad \text{and} \quad \sqrt{\mathbb{V}(T_{2/3})} \le 20e^2 n\sqrt{\Big\|\Big(\frac{1}{q\vee n^{-1}}\Big)^{4/3} q^2\Big\|_1},$$

and so by Chebyshev's inequality with probability larger than $1-\alpha-6\exp(-n/18)$

$$T_{2/3} \le \alpha^{-1/2} 20e^2 n\sqrt{\Big\|\Big(\frac{1}{q\vee n^{-1}}\Big)^{4/3} q^2\Big\|_1}.$$

Under $H_{1,\pi}^{(\mathrm{Clo})}(\rho)$. We assume that for a large $C>0$

$$\begin{aligned}
\Big\|\Delta(\mathbf{1}\{i\le J\})_i\Big\|_1^2 &= \Big\|\Delta\mathbf{1}\{nq\ge 1\}\Big\|_1^2 \\
&\ge C\left(\frac{\Big\|\Big(\frac{1}{q\vee n^{-1}}\Big)^{4/3} q^2\Big\|_1^{3/2}}{n} \vee \frac{\Big\|\Big(\frac{1}{q\vee n^{-1}}\Big)^{4/3} q^2\Big\|_1}{n}\right),
\end{aligned}$$

which implies by Cauchy-Schwarz inequality,

$$\left\| \frac{\Delta^2}{q^{2/3}} \mathbf{1}\{nq \geq 1\} \right\|_1 \geq C \left(\frac{\sqrt{\left\| \left(\frac{1}{q \vee n^{-1}} \right)^{4/3} q^2 \right\|_1}}{n} \vee \frac{1}{n} \right),$$

and in particular that

$$\left\| \frac{\Delta^2}{(q \vee n^{-1})^{2/3}} \right\|_1 \geq C/n. \tag{2.13}$$

Moreover if the pre-test does not reject the null, we have with probability larger than $1 - 2\delta - 2n^{-1} - 6\exp(-n/18)$, that there exists $0 < \tilde{c} < +\infty$ universal constant and $0 < \tilde{c}_\delta < +\infty$ that depends only on δ such that for any i

$$\frac{\Delta_i^2}{(q_i \vee n^{-1})^{2/3}} \leq \frac{\tilde{c}^2}{n} \frac{(q_i \vee (\log(n/\delta)/n)) \log(n/\delta)}{(q_i \vee n^{-1})^{2/3}} \leq \frac{\tilde{c}_\delta^2}{n},$$

So with probability larger than $1 - 2\delta - 2n^{-1} - 6\exp(-n/18)$,

$$\left\| \frac{\Delta^4}{(q \vee 1/n)^{4/3}} \right\|_1 \leq \frac{\tilde{c}_\delta^2}{n} \left\| \frac{\Delta^2}{(q \vee 1/n)^{2/3}} \right\|_1,$$

i.e., by Equation (2.13),

$$\sqrt{\left\| \frac{\Delta^4}{(q \vee 1/n)^{4/3}} \right\|_1} \leq \sqrt{\frac{\tilde{c}_\delta^2}{n} \left\| \frac{\Delta^2}{(q \vee 1/n)^{2/3}} \right\|_1} \leq \sqrt{\frac{\tilde{c}_\delta^2}{C}} \left\| \frac{\Delta^2}{(q \vee 1/n)^{2/3}} \right\|_1. \tag{2.14}$$

We have from Equation (2.11):

$$2e^2 \mathbb{E} T_{2/3}/n^2 \geq \left\| \Delta^2 \left(\frac{1}{q \vee n^{-1}} \right)^{2/3} \right\|_1.$$

And from Equation (2.12),

$$\sqrt{\mathbb{V}(T_{2/3})}/n^2 \leq 10e^2 \left[\frac{\sqrt{\left\| \left(\frac{1}{q \vee n^{-1}} \right)^{4/3} q^2 \right\|_1}}{n} + \frac{\sqrt{\left\| \left(\frac{1}{q \vee n^{-1}} \right)^{4/3} \Delta^2 \right\|_1}}{n} + \sqrt{\left\| \left(\frac{1}{q \vee n^{-1}} \right)^{4/3} \Delta^4 \right\|_1} \right].$$

Let us compare the terms involved in the upper bound on $\sqrt{\mathbb{V}(T_{2/3})}/n^2$ with the lower bound on $\mathbb{E} T_{2/3}/n^2$.

For the first term, we have by Equation (2.13):

$$10e^2 \frac{\sqrt{\left\| \left(\frac{1}{q \vee n^{-1}} \right)^{4/3} q^2 \right\|_1}}{n} \leq \frac{20e^4}{C} \mathbb{E} T_{2/3}/n^2.$$

We have for the second term:

$$10e^2\frac{\sqrt{\left\|\left(\frac{1}{q\vee n^{-1}}\right)^{4/3}\Delta^2\right\|_1}}{n}\leq 20e^4n^{-2/3}\sqrt{\left\|\left(\frac{1}{q\vee n^{-1}}\right)^{2/3}\Delta^2\right\|_1}.$$

Since $a^2+b^2\geq 2ab$ for any a,b,

$$10e^2\frac{\sqrt{\left\|\left(\frac{1}{q\vee n^{-1}}\right)^{4/3}\Delta^2\right\|_1}}{n}\leq 10e^4(1/n+n^{-1/3}\mathbb{E}T_{2/3}/n^2).$$

which, from Equation (2.13), yields

$$10e^2\frac{\sqrt{\left\|\left(\frac{1}{q\vee n^{-1}}\right)^{4/3}\Delta^2\right\|_1}}{n}\leq 10e^4(1/C+n^{-1/3})\mathbb{E}T_{2/3}/n^2.$$

Then for the third term, we have shown in Equation (2.14) that with probability larger than $1-2\delta-2n^{-1}-6\exp(-n/18)$,

$$10e^2\sqrt{\left\|\left(\frac{1}{q\vee n^{-1}}\right)^{4/3}\Delta^4\right\|_1}\leq 20e^4\sqrt{\frac{\tilde{c}_\delta^2}{C}}\mathbb{E}T_{2/3}/n^2.$$

And so we have by Chebyshev's inequality, with probability larger than $1-2\delta-2n^{-1}-\alpha-6\exp(-n/18)$:

$$|T_{2/3}-\mathbb{E}T_{2/3}|\leq\frac{20e^4}{\sqrt{\alpha}}(1/(2C)+n^{-1/3}/2+\sqrt{\tilde{c}_\delta^2/C}+1/C)\mathbb{E}T_{2/3}.$$

Now, if $n\geq\left(\frac{80e^4}{\sqrt{\alpha}}\right)^3$ and $C\geq\frac{40e^4}{\sqrt{\alpha}}\left(\frac{20e^4\tilde{c}_\delta^2}{\sqrt{\alpha}}\vee 1\right)$:

$$|T_{2/3}-\mathbb{E}T_{2/3}|\leq\mathbb{E}T_{2/3}/2.$$

Finally, if $n\geq\left(\frac{80e^4}{\sqrt{\alpha}}\right)^3$ and $C\geq\frac{40e^4}{\sqrt{\alpha}}\left(\frac{20e^4\tilde{c}_\delta^2}{\sqrt{\alpha}}\vee 1\right)$, with probability greater than $1-2\delta-2n^{-1}-\alpha-6\exp(-n/18)$:

$$\begin{aligned}T_{2/3}&\geq\mathbb{E}T_{2/3}/2\\&\geq\frac{Cn}{2}\left(\sqrt{\left\|\left(\frac{1}{q\vee n^{-1}}\right)^{4/3}q^2\right\|_1}+1\right),\end{aligned}\tag{2.15}$$

where the last inequality comes from Equations (2.11) and (2.13).

Analysis of $\hat{t}_{2/3}$. Test $\varphi_{2/3}$ compares statistic $T_{2/3}$ with threshold $\hat{t}_{2/3}$, which is empirical. So let us study the variations of $\hat{t}_{2/3}$. Applying Corollary 8 below gives guarantees on the empirical threshold $\hat{t}_{2/3}$. These can be used in conjunction with the guarantees on the statistic $T_{2/3}$ in order to conclude the proof of Proposition 9.

Theorem 24. *Let* $C_{2/3}=\sqrt{2\delta^{-1}e^{8/3}+1}+\sqrt{(2^{1/3}+e)}$.

With probability greater than $1-\beta$*:*

$$(e^{-2/3}/2+1)\left\|\left(\frac{1}{q\vee n^{-1}}\right)^{4/3}q^2\right\|_1 \leq n^{-2/3}\|(Y^{(1)})^{2/3}\|_1 + C_{2/3}/\sqrt{\beta} \leq \left\|\left(\frac{1}{q\vee n^{-1}}\right)^{4/3}q^2\right\|_1 + 2C_{2/3}/\sqrt{\beta} + 2\delta^{-1}.$$

The proof of this theorem is in Section 2.6.7.1. And the following corollary is obtained immediately from the theorem.

Corollary 8. *We define*

$$\hat{t}_{2/3} = \alpha^{-1/2}20en\sqrt{n^{-2/3}\|(Y^{(1)})^{2/3}\|_1 + C_{2/3}/\sqrt{\alpha}}.$$

Then if $C \geq (8e^6/20e^{-1}\sqrt{\alpha}) \vee (8^2e^{10}c_\delta^2\alpha/100) \vee (\alpha^{-1/2}40e^7\sqrt{2\delta^{-1}(C_{2/3}+1)})$*, we have with probability greater than* $1-\alpha-6\exp(-n/18)$*:*

$$\alpha^{-1/2}20en\sqrt{\left\|\left(\frac{1}{q\vee n^{-1}}\right)^{4/3}q^2\right\|_1} \leq \hat{t}_{2/3} \leq e^{-6}\frac{C}{2}n\left(\sqrt{\left\|\left(\frac{1}{q\vee n^{-1}}\right)^{4/3}q^2\right\|_1}+1\right).$$

Let us now sum up the results leading to Proposition 9. Under $H_{0,\pi}^{(\mathrm{Clo})}$, with probability larger than $1-\delta/2-2\delta-2n^{-1}-6\exp(-n/18)$,

$$T_{2/3} \leq (\delta/2)^{-1/2}20en\sqrt{\left\|\left(\frac{1}{q\vee n^{-1}}\right)^{4/3}q^2\right\|_1}.$$

And so if $C \geq \left[\frac{40e^4}{\sqrt{\alpha}}\left(\frac{20e^4\tilde{c}_\delta^2}{\sqrt{\alpha}}\vee 1\right)\right] \vee \left(\alpha^{-1/2}40e^7\sqrt{2C_{2/3}(\delta^{-1/2}+\alpha^{-1/2})}\right)$, we have with probability greater than $1-\delta/2-6\exp(-n/18)$:

$$(\delta/2)^{-1/2}20en\sqrt{\left\|\left(\frac{1}{q\vee n^{-1}}\right)^{4/3}q^2\right\|_1} \leq \hat{t}_{2/3}.$$

So, under $H_{0,\pi}^{(\mathrm{Clo})}$, with probability larger than $1-3\delta-2n^{-1}-6\exp(-n/18)$,

$$T_{2/3} \leq \hat{t}_{2/3}.$$

Under $H_{1,\pi}^{(\mathrm{Clo})}(\rho)$, if $C \geq \left[\frac{40e^4}{\sqrt{\delta/2}}\left(\frac{20e^4\tilde{c}_\delta^2}{\sqrt{\delta/2}}\vee 1\right)\right] \vee \left((\delta/2)^{-1/2}80e^7\sqrt{2C_{2/3}\delta^{-1/2}}\right)$ and $n \geq \left(\frac{80e^4}{\sqrt{\delta/2}}\right)^3$, we have with probability larger than $1-\delta/2-2\delta-2n^{-1}-6\exp(-n/18)$,

$$e^{-6}\frac{Cn}{2}\left(\sqrt{\left\|\left(\frac{1}{q\vee n^{-1}}\right)^{4/3}q^2\right\|_1}+1\right) \leq T_{2/3}.$$

And if $C \geq \left[\frac{40e^4}{\sqrt{\delta/2}}\left(\frac{20e^4\tilde{c}_\delta^2}{\sqrt{\delta/2}} \vee 1\right)\right] \vee \left((\delta/2)^{-1/2}80e^7\sqrt{2C_{2/3}\delta^{-1/2}}\right)$ and $n \geq \left(\frac{80e^4}{\sqrt{\delta/2}}\right)^3$, we have with probability larger than $1 - \delta/2 - 6\exp(-n/18)$,

$$\hat{t}_{2/3} \leq e^{-6}\frac{Cn}{2}\left(\sqrt{\left\|\left(\frac{1}{q \vee n^{-1}}\right)^{4/3} q^2\right\|_{2/3}} + 1\right).$$

So, under $H_{1,\pi}^{(\mathrm{Clo})}(\rho)$, with probability larger than $1 - 3\delta - 2n^{-1} - 6\exp(-n/18)$,

$$\hat{t}_{2/3} \leq T_{2/3}.$$

Proof of Lemma 9. We have by definition of the Poisson distribution that

$$\begin{aligned}\mathbb{E}[(Z \vee 1)^{-r}] &= \exp(-\lambda) + \exp(-\lambda)\sum_{i\geq 1}\frac{\lambda^i}{i!}i^{-r} \\ &= \exp(-\lambda) + \exp(-\lambda)\sum_{1\leq i\leq \lambda/e^2}\frac{\lambda^i}{i!}i^{-r} + \exp(-\lambda)\sum_{1\vee(\lambda/e^2)<i}\frac{\lambda^i}{i!}i^{-r}.\end{aligned}$$

And so we have

$$\begin{aligned}\mathbb{E}[(Z \vee 1)^{-r}] &\leq \exp(-\lambda) + \exp(-\lambda)\sum_{1\leq i\leq \lambda/e^2}\frac{\lambda^i}{i!}i^{-r} + \left(\frac{e^2}{\lambda} \wedge 1\right)^r \\ &\leq \exp(-\lambda) + \exp(-\lambda)\sum_{1\leq i\leq \lambda/e^2}\frac{\lambda^i e^i}{i^i} + \left(\frac{e^2}{\lambda} \wedge 1\right)^r,\end{aligned}$$

since $i! \geq i^i/e^i$ and $i \geq 1$. Then

$$\begin{aligned}\mathbb{E}[(Z \vee 1)^{-r}] &\leq \exp(-\lambda) + \exp(-\lambda)\sum_{1\leq i\leq \lambda/e^2}\exp(i\log(\lambda) + i - i\log(i)) \\ &\quad + \left(\frac{1}{1 \vee (\lambda/e^2)}\right)^r \\ &\leq \exp(-\lambda) + \lambda\exp\left(-\lambda + \frac{\lambda}{e^2}\log(\lambda) + \frac{\lambda}{e^2} - \frac{\lambda}{e^2}\log\left(\frac{\lambda}{e^2}\right)\right) \\ &\quad + \left(\frac{1}{1 \vee (\lambda/e^2)}\right)^r \\ &\leq \exp(-\lambda) + \lambda\exp(-\lambda/2) + \left(\frac{1}{1 \vee (\lambda/e^2)}\right)^r \\ &\leq 5\exp(-\lambda/4) + \left(\frac{1}{1 \vee (\lambda/e^2)}\right)^r \leq 6\left(\frac{e^2}{1 \vee \lambda}\right)^r,\end{aligned}$$

since $r \in \{2/3, 4/3\}$. Now let us prove the other inequality.

$$\begin{aligned}\mathbb{E}[(Z \vee 1)^{-r}] &\geq \exp(-\lambda) + \exp(-\lambda)\sum_{i<e^2\lambda}\frac{\lambda^i}{i!}i^{-r} \\ &\geq \exp(-\lambda) + (e^2\lambda)^{-r}\exp(-\lambda)\sum_{i<e^2\lambda}\frac{\lambda^i}{i!}.\end{aligned}$$

So, since $i! \geq i^i/e^i$,

$$\begin{aligned}
\mathbb{E}[(Z \vee 1)^{-r}] &\geq \exp(-\lambda) + \Big(\frac{1}{(e^2\lambda) \vee 1}\Big)^r \Big[1 - \exp(-\lambda) \sum_{i \geq e^2\lambda} \frac{\lambda^i e^i}{i^i}\Big] \\
&= \exp(-\lambda) \\
&\quad + \Big(\frac{1}{(e^2\lambda) \vee 1}\Big)^r \Big[1 - \exp(-\lambda) \sum_{i \geq e^2\lambda} \exp(i\log(\lambda) + i - i\log(i)\Big] \\
&\geq \exp(-\lambda) \\
&\quad + \Big(\frac{1}{(e^2\lambda) \vee 1}\Big)^r \Big[1 - \exp(-\lambda) \sum_{i \geq e^2\lambda} \exp(i\log(\lambda) + i - i\log(e^2\lambda))\Big] \\
&= \exp(-\lambda) + \Big(\frac{1}{(e^2\lambda) \vee 1}\Big)^r \Big[1 - \exp(-\lambda) \sum_{i \geq e^2\lambda} \exp(-i)\Big] \\
&\geq \exp(-\lambda) + \Big(\frac{1}{(e^2\lambda) \vee 1}\Big)^r \Big[1 - 2\exp(-\lambda - e^2\lambda)\Big].
\end{aligned}$$

If $\lambda \geq \frac{\log(4)}{1+e^2}$, then

$$\mathbb{E}[(Z \vee 1)^{-r}] \geq \Big(\frac{1}{(e^2\lambda) \vee 1}\Big)^r \Big[1 - 2\exp(-\lambda - e^2\lambda)\Big] \geq \frac{1}{2}\Big(\frac{1}{(e^2\lambda) \vee 1}\Big)^r.$$

If $\lambda < \frac{\log(4)}{1+e^2}$, then

$$\mathbb{E}[(Z \vee 1)^{-r}] \geq \exp(-\lambda) \geq \exp\left(-\frac{\log(4)}{1+e^2}\right) \geq 1/2 \geq \frac{1}{2}\Big(\frac{1}{(e^2\lambda) \vee 1}\Big)^r.$$

So in any case,

$$\mathbb{E}[(Z \vee 1)^{-r}] \geq \frac{1}{2}\Big(\frac{1}{(e^2\lambda) \vee 1}\Big)^r.$$

□

2.6.4 Proof of Proposition 10

Proposition 10 provides guarantees on the test φ_2. The structure of its proof will be identical to that of Proposition 9. We first study T_2.

Expression of the test statistic.

We have

$$T_2 = \sum_{i \leq d} (X_i^{(1)} - Y_i^{(1)})(X_i^{(2)} - Y_i^{(2)})\mathbf{1}\{Y_i^{(3)} = 0\}.$$

And so

$$\mathbb{E}T_2 = \sum_{i \leq d} \mathbb{E}^{(1)}(X_i^{(1)} - Y_i^{(1)})\mathbb{E}^{(2)}(X_i^{(2)} - Y_i^{(2)})\mathbb{E}^{(3)}\mathbf{1}\{Y_i^{(3)} = 0\},$$

and

$$\begin{aligned}\mathbb{V}T_2 &\le \sum_{i\le d}\mathbb{V}\Big[(X_i^{(1)}-Y_i^{(1)})(X_i^{(2)}-Y_i^{(2)})\mathbf{1}\{Y_i^{(3)}=0\}\Big]\\ &\le \sum_{i\le d}\mathbb{E}^{(1)}[(X_i^{(1)}-Y_i^{(1)})^2]\mathbb{E}^{(2)}[(X_i^{(2)}-Y_i^{(2)})^2]\mathbb{E}^{(3)}\mathbf{1}\{Y_i^{(3)}=0\}.\end{aligned}$$

We will bound every term separately.

Terms that depend on the first and second sub-sample. We have with probability larger than $1-6\exp(-n/18)$ with respect to $(\bar{n}_m^{(j)})_{m\le 2,j\le 3}$,

$$\mathbb{E}^{(1)}(X_i^{(1)}-Y_i^{(1)})\mathbb{E}^{(2)}(X_i^{(2)}-Y_i^{(2)}) = n^2\Delta_i^2.$$

and

$$\begin{aligned}\mathbb{E}^{(1)}[(X_i^{(1)}-Y_i^{(1)})^2]\mathbb{E}^{(2)}[(X_i^{(2)}-Y_i^{(2)})^2] &= \Big[\mathbb{E}^{(1)}[(X_i^{(1)}-Y_i^{(1)})^2]\Big]^2\\ &= \Big[\mathbb{E}^{(1)}[(X_i^{(1)}-Y_i^{(1)}-n\Delta_i)^2]+n^2\Delta_i^2\Big]^2\\ &= [n(p_i+q_i)+n^2\Delta_i^2]^2.\end{aligned}$$

Terms that depend on the third sub-sample. We define

$$R_i := \mathbb{E}^{(3)}\mathbf{1}\{Y_i^{(3)}=0\}, \qquad \text{and so} \qquad R_i = \exp(-nq_i).$$

Bound on the expectation and variance for T_2. We have with probability larger than $1-6\exp(-n/18)$ with respect to $(\bar{n}_m^{(j)})_{m\le 2,j\le 3}$,

$$\mathbb{E}T_2 = \sum_{i\le d}\Big[n^2\Delta_i^2R_i\Big] = n^2\|\Delta^2R\|_1 = n^2\|\Delta^2\exp(-nq)\|_1. \tag{2.16}$$

And

$$\begin{aligned}\mathbb{V}T_2 &\le \sum_{i\le d}\Big[n(p_i+q_i)+n^2\Delta_i^2\Big]^2R_i\\ &\le 4\sum_{i\le d}\Big[n^2q_i^2+n^2\Delta_i^2+n^4\Delta_i^4\Big]R_i\\ &\le 4\Big[n^2\|q^2R\|_1+n^2\|\Delta^2R\|_1+n^4\|R\Delta^4\|_1\Big],\end{aligned}$$

and so with probability larger than $1-6\exp(-n/18)$ with respect to $(\bar{n}_m^{(j)})_{m\le 2,j\le 3}$,

$$\begin{aligned}\sqrt{\mathbb{V}T_2} &\le 2n\Big[\sqrt{\|q^2R\|_1}+\sqrt{\|\Delta^2R\|_1}+n\sqrt{\|R\Delta^4\|_1}\Big]\\ &\le 2\Big[\sqrt{n^2\|q^2\exp(-nq)\|_1}+\sqrt{n^2\|\Delta^2\exp(-nq)\|_1}+n^2\sqrt{\|\Delta^4\exp(-nq)\|_1}\Big].\end{aligned} \tag{2.17}$$

Analysis of T_2 under $H_{0,\pi}^{(\mathbf{Clo})}$ and $H_{1,\pi}^{(\mathbf{Clo})}(\rho)$. Let us inspect the behaviour of statistic T_2 under both hypotheses.

Under $H_{0,\pi}^{(Clo)}$. We have then with probability larger than $1-6\exp(-n/18)$ with respect to $(\bar{n}_m^{(j)})_{m\leq 2, j\leq 3}$,

$$\mathbb{E}T_2 = 0,$$

$$\sqrt{\mathbb{V}T_2} \leq 2n\sqrt{\|q^2\exp(-nq)\|_1}.$$

And so by Chebyshev's inequality, with probability larger than $1-\alpha-6\exp(-n/18)$

$$T_2 \leq 2\alpha^{-1/2}\sqrt{\|(nq)^2\exp(-nq)\|_1}.$$

Under $H_{1,\pi}^{(Clo)}(\rho)$. Assume that for $C>0$ large we have

$$\|\Delta(\mathbf{1}\{I\geq i\geq J\})_i\|_1^2 \geq Ce^2\frac{I-J}{n}\Big[\frac{\log^2(n)}{n}\vee\Big(\sqrt{\|q^2\exp(-nq)\|_1}\Big)\Big].$$

By Cauchy-Schwarz inequality and since for any $I\geq i\geq J$, $nq_i\leq 1$, this implies

$$n^2\|\Delta^2\exp(-nq)\|_1 \geq C\Big[\log^2(n)\vee\Big(n\sqrt{\|q^2\exp(-nq)\|_1}\Big)\Big]. \tag{2.18}$$

If the pre-test accepts the null, then with probability larger than $1-2\delta-2n^{-1}-6\exp(-n/18)$ on the third sub-sample only, there exists $\tilde{c}_\delta$ that depends only on δ such that for all $i\leq d$ we have $|\Delta_i|\leq \tilde{c}_\delta\Big[\sqrt{\frac{q_i\log(n)}{n}}\vee\frac{\log(n)}{n}\Big]$ and so

$$\begin{aligned}
&n^2\sqrt{\|\Delta^4\exp(-nq)\|_1}\\
&\leq \tilde{c}_\delta n^2\sqrt{\|\exp(-nq)\Big[\frac{q^2\log(n)^2}{n^2}\mathbf{1}\{nq\geq 2\log(n)\}+\frac{\Delta^2\log(n)^2}{n^2}\mathbf{1}\{nq\leq 2\log(n)\}\Big]\|_1}\\
&\leq \tilde{c}_\delta n^2\Big[\sqrt{\|\exp(-nq)\frac{q^2\log(n)^2}{n^2}\mathbf{1}\{nq\geq 2\log(n)\}\|_1}\\
&+\sqrt{\|\exp(-nq)\frac{\Delta^2\log(n)^2}{n^2}\mathbf{1}\{nq\leq 2\log(n)\}\|_1}\Big]\\
&\leq \tilde{c}_\delta n^2\Big[\sqrt{\|\exp(-nq)\frac{q^2\log(n)^2}{n^2}\mathbf{1}\{nq\geq 2\log(n)\}\|_1}+\frac{\log(n)}{n}\sqrt{\|\exp(-nq)\Delta^2\|_1}\Big]\\
&\leq \tilde{c}_\delta n^2\Big[\sqrt{n^{-4}\log(n)^2\|q^2\mathbf{1}\{nq\geq 2\log(n)\}\|_1}+\frac{\log(n)}{n}\sqrt{\|\exp(-nq)\Delta^2\|_1}\Big]\\
&\leq \tilde{c}_\delta\Big[\log(n)+n\log(n)\sqrt{\|\exp(-nq)\Delta^2\|_1}\Big]\\
&\leq \tilde{c}_\delta\log(n)\Big[1+n\sqrt{\|\exp(-nq)\Delta^2\|_1}\Big],
\end{aligned} \tag{2.19}$$

since $\sum_i q_i = 1$.
We have from Equation (2.16):

$$\mathbb{E}T_2/n^2 = \|\Delta^2\exp(-nq)\|_1.$$

And from Equation (2.17),

$$\sqrt{\mathbb{V}T_2}/n^2 \leq 2\Big[\sqrt{\|q^2\exp(-nq)\|_1}/n + \sqrt{\|\Delta^2\exp(-nq)\|_1}/n + \sqrt{\|\Delta^4\exp(-nq)\|_1}\Big].$$

Let us compare the terms of $\sqrt{\mathbb{V}T_2}/n^2$ with $\mathbb{E}T_2/n^2$.

For the first term, we use Equation (2.18), and we get:

$$\sqrt{\|q^2\exp(-nq)\|_1}/n \leq \|\Delta^2\exp(-nq)\|_1/C = \frac{1}{C}\mathbb{E}T_2/n^2.$$

For the second term, we have:

$$\sqrt{\|\Delta^2\exp(-nq)\|_1}/n \leq \sqrt{\|\Delta^2\exp(-nq)\|_1}\log(n)/n.$$

So using Equation (2.18), we have:

$$\sqrt{\|\Delta^2\exp(-nq)\|_1}/n \leq \frac{1}{\sqrt{C}}\mathbb{E}T_2/n^2.$$

For the third term, using Equation (2.19), we have with probability larger than $1-2\delta-2n^{-1}-6\exp(-n/18)$:

$$\sqrt{\|\Delta^4\exp(-nq)\|_1} \leq \tilde{c}_\delta\Big[n^{-2}\log(n) + \frac{\log(n)}{n}\sqrt{\|\exp(-nq)\Delta^2\|_1}\Big].$$

So by Equation (2.18), we have with probability larger than $1-2\delta-2n^{-1}-6\exp(-n/18)$:

$$\begin{aligned}\sqrt{\|\Delta^4\exp(-nq)\|_1} &\leq \tilde{c}_\delta\left[n^{-2}\log^2(n) + \frac{1}{\sqrt{C}}\mathbb{E}T_2/n^2\right]\\ &\leq \tilde{c}_\delta(1/C + 1/\sqrt{C})\mathbb{E}T_2/n^2.\end{aligned}$$

And so we have by Chebyshev's inequality, with probability larger than $1-\alpha-2\delta-2n^{-1}-6\exp(-n/18)$:

$$|T_2 - \mathbb{E}T_2| \leq 2/\sqrt{\alpha}(1/C + 1/\sqrt{C} + \tilde{c}_\delta(1/C + 1/\sqrt{C}))\mathbb{E}T_2.$$

So if $C \geq 1$, with probability larger than $1-\alpha-2\delta-2n^{-1}-6\exp(-n/18)$:

$$|T_2 - \mathbb{E}T_2| \leq \frac{4}{\sqrt{C\alpha}}(1+\tilde{c}_\delta)\mathbb{E}T_2.$$

So if $C \geq \big[8\alpha^{-1/2}(1+\tilde{c}_\delta)\big]^2$, we have with probability larger than $1-\alpha-2\delta-2n^{-1}-6\exp(-n/18)$:

$$|T_2 - \mathbb{E}T_2| \leq \mathbb{E}T_2/2.$$

Finally, if $C \geq \big[8\alpha^{-1/2}(1+\tilde{c}_\delta)\big]^2$, we have with probability larger than $1-\alpha-2\delta-2n^{-1}-6\exp(-n/18)$:

$$T_2 \geq \mathbb{E}T_2/2 \geq \frac{C}{2}\Big[\log^2(n) \vee \Big(n\sqrt{\|q^2\exp(-nq)\|_1}\Big)\Big].$$

Analysis of $\hat{t}_2$.

Test φ_2 compares statistic T_2 with empirical threshold $\hat{t_2}$. So let us study the variations of $\hat{t_2}$. Applying Corollary 9 below gives guarantees on the empirical threshold $\hat{t}_2$. These can be used in conjunction with the guarantees on the statistic T_2 in order to conclude the proof of Proposition 10.

Theorem 25. *We have with probability larger than* $1-\delta-6\exp(-n/18)$*:*

$$|\|Y^{(1)}Y^{(2)}\mathbf{1}\{Y^{(3)}=0\}\|_1 - \|(nq)^2e^{-nq}\|_1| \leq \frac{1}{\sqrt{\delta}}(\|(nq)^2e^{-nq}\|_1/2 + 1005\log(n)^4).$$

The proof of this theorem is in Section 2.6.7.2.

Corollary 9. *We define*

$$\hat{t}_2 = 2\alpha^{-1/2}(1-1/(2\sqrt{\delta}))^{-1/2}\sqrt{\|Y^{(1)}Y^{(2)}\mathbf{1}\{Y^{(3)}=0\}\|_1 + \frac{1005}{\sqrt{\delta}}\log(n)^4}.$$

If $C \geq \left[8\alpha^{-1/2}(1+\tilde{c}_\delta\sqrt{2\delta^{-1}})\right]^2 \vee \frac{8\cdot 45\alpha^{-1/2}}{(\sqrt{\delta}-1/2)^{1/2}}$*, we have with probability greater than* $1-\delta-6\exp(-n/18)$*:*

$$2\alpha^{-1/2}\sqrt{\|(nq)^2\exp(-nq)\|_1} \leq \hat{t}_2 \leq \frac{C}{2}\left[\log^2(n) \vee \left(n\sqrt{\|q^2\exp(-nq)\|_1}\right)\right].$$

Proof of Corollary 9. By application of Theorem 25, we have with probability greater than $1-\delta-6\exp(-n/18)$:

$$2\alpha^{-1/2}\sqrt{\|(nq)^2\exp(-nq)\|_1} \leq \hat{t}_2 \leq 2\alpha^{-1/2}\sqrt{\frac{2\sqrt{\delta}+1}{2\sqrt{\delta}-1}\|(nq)^2e^{-nq}\|_1 + 2010(\sqrt{\delta}-1/2)^{-1}\log(n)^4}.$$

So,

$$\hat{t}_2 \leq 2\alpha^{-1/2}\left(\sqrt{\frac{2\sqrt{\delta}+1}{2\sqrt{\delta}-1}\|(nq)^2e^{-nq}\|_1} + \sqrt{2010(\sqrt{\delta}-1/2)^{-1}\log(n)^4}\right).$$

Finally,

$$\hat{t}_2 \leq \frac{4\cdot 45\alpha^{-1/2}}{(\sqrt{\delta}-1/2)^{1/2}}\left(\sqrt{\|(nq)^2e^{-nq}\|_1} \vee \log(n)^2\right).$$

□

Let us now sum up the results leading to Proposition 10. Under $H_{0,\pi}^{(\mathrm{Clo})}$, with probability larger than $1-\delta/2-2\delta-2n^{-1}-6\exp(-n/18)$,

$$T_2 \leq 2(\delta/2)^{-1/2}\sqrt{\|(nq)^2\exp(-nq)\|_1}.$$

And if $C \geq \left[8(\delta/2)^{-1/2}(1+\tilde{c}_\delta\sqrt{2\delta^{-1}})\right]^2 \vee \frac{8\cdot 45\alpha^{-1/2}}{(\sqrt{\delta}-1/2)^{1/2}}$, we have with probability greater than $1-\delta/2-6\exp(-n/18)$:

$$2(\delta/2)^{-1/2}\sqrt{\|(nq)^2\exp(-nq)\|_1} \leq \hat{t}_2$$

So, under $H_{0,\pi}^{(\mathrm{Clo})}$, with probability larger than $1-3\delta-2n^{-1}-6\exp(-n/18)$,

$$T_2 \le \hat{t}_2.$$

Under $H_{1,\pi}^{(\mathrm{Clo})}(\rho)$, if $C \ge \left[8(\delta/2)^{-1/2}(1+\tilde{c}_\delta\sqrt{2\delta^{-1}})\right]^2$, we have with probability larger than $1-\delta/2-2\delta-2n^{-1}-6\exp(-n/18)$:

$$T_2 \ge C/2\Big[\log^2(n) \vee \Big(n\sqrt{\|q^2\exp(-nq)\|_1}\Big)\Big].$$

If $C \ge \left[8(\delta/2)^{-1/2}(1+\tilde{c}_\delta\sqrt{2\delta^{-1}})\right]^2 \vee \frac{8\cdot 45(\delta/2)^{-1/2}}{(\sqrt{\delta}-1/2)^{1/2}}$, we have with probability greater than $1-\delta/2-6\exp(-n/18)$:

$$\hat{t}_2 \le \frac{C}{2}\Big[\log^2(n) \vee \Big(n\sqrt{\|q^2\exp(-nq)\|_1}\Big)\Big].$$

So, under $H_{1,\pi}^{(\mathrm{Clo})}(\rho)$, with probability larger than $1-3\delta-2n^{-1}-6\exp(-n/18)$,

$$\hat{t}_2 \le T_2.$$

2.6.5 Proof of Proposition 11

Proposition 11 gives guarantees on test φ_1. This time, the proof will only focus on the variations of T_1 since the threshold is not empirical.

Analysis of the moments of T_1.

We have

$$T_1 = \sum_i (X^{(1)} - Y^{(1)})\mathbf{1}\{Y_i^{(3)} = 0\}.$$

So with probability larger than $1-6\exp(-n/18)$ with respect to $(\bar{n}_m^{(j)})_{m\le 2, j\le 3}$

$$\mathbb{E}T_1 = n\sum_i \Delta_i \exp(-nq_i). \tag{2.20}$$

And

$$\begin{aligned}
\mathbb{V}T_1 &\le \sum_i \mathbb{E}^{(1)}(X^{(1)} - Y^{(1)})^2 \mathbb{E}^{(3)}\mathbf{1}\{Y_i^{(3)} = 0\} \\
&\le \sum_i [n(p_i+q_i) + n^2\Delta_i^2]\exp(-nq_i) \\
&\le 2n\|q\exp(-nq)\|_1 + \Big|n\sum_i \Delta_i\exp(-nq_i)\Big| + n^2\|\exp(-nq)\Delta^2\|_1,
\end{aligned}$$

which implies

$$\sqrt{\mathbb{V}T_1} \le \sqrt{2n\|q\exp(-nq)\|_1} + \sqrt{n\left|\sum_i \Delta_i\exp(-nq_i)\right|} + n\sqrt{\|\Delta^2\exp(-nq)\|_1}. \tag{2.21}$$

Analysis of T_1 under $H_{0,\pi}^{(\mathbf{Clo})}$ and $H_{1,\pi}^{(\mathbf{Clo})}(\rho)$. Let us inspect the behaviour of statistic T_1 under both hypotheses.

Under $H_{0,\pi}^{(\mathrm{Clo})}$. We have with probability larger than $1-6\exp(-n/18)$ with respect to $(\bar{n}_m^{(j)})_{m\le 2, j\le 3}$, $\mathbb{E}T_1 = 0$ and $\sqrt{\mathbb{V}T_1} \le \sqrt{2n\|q\exp(-nq)\|_1}$. So by Chebyshev's inequality with probability larger than $1-\alpha-6\exp(-n/18)$

$$T_1 \le \sqrt{\frac{2n\|q\exp(-nq)\|_1}{\alpha}} \le \sqrt{2n/\alpha}.$$

Note that the result of Proposition 11 is based on the reunion of two conditions. That is the reason why we will divide the study of $H_{1,\pi}^{(\mathrm{Clo})}(\rho)$ into two.

Under $H_{1,\pi}^{(\mathrm{Clo})}(\rho)$*, analysis 1.* Assume first that

$$\|\Delta(\mathbf{1}\{i\ge I\})_i\|_1 \ge C\Big[\|q(\mathbf{1}\{i\ge I\})_i\|_1 \vee \sqrt{\frac{\log(n)}{n}}\Big], \tag{2.22}$$

and

$$\|\Delta(\mathbf{1}\{i\ge I\})_i\|_1 \ge 2\|\Delta(\mathbf{1}\{i< I\})_i\|_1. \tag{2.23}$$

We have

$$\sum_{i\ge I}(\Delta_i + 2q_i) = \sum_{i\ge I}(p_i+q_i) \ge \|\Delta(\mathbf{1}\{i\ge I\})_i\|_1.$$

So by Equation (2.22),

$$\sum_{i\ge I}\Delta_i \ge (C-2)\sum_{i\ge I} q_i, \tag{2.24}$$

and

$$\sum_{i\ge I}\Delta_i C/(C-2) \ge \|\Delta(\mathbf{1}\{i\ge I\})_i\|_1. \tag{2.25}$$

Then since for any $i\ge I$, $q_i \le 1/n$, Equation (2.24) yields:

$$\sum_i \Delta_i e^{-nq_i} \ge \sum_{i\ge I}\Delta_i e^{-nq_i} \ge \sum_{i\ge I}\Delta_i e^{-1} \ge (C-2)e^{-1}\sum_{i\ge I} q_i.$$

And again Equation (2.22) gives:

$$\sum_{i\ge I}\Delta_i + 2\sum_{i\ge I} q_i \ge C\sqrt{\frac{\log(n)}{n}}.$$

So

$$\sum_{i\ge I}\Delta_i C/(C-2) \ge C\sqrt{\frac{\log(n)}{n}}.$$

So for C large enough, we end up with:

$$\sum_i \Delta_i \exp(-nq_i) \ge \frac{C}{2}\Big[\|q(\mathbf{1}\{i\ge I\})_i\|_1 \vee \sqrt{\frac{\log(n)}{n}}\Big].$$

We then have by Equation (2.20):

$$\mathbb{E}T_1 = n\sum_i \Delta_i \exp(-nq_i) \ge \frac{C}{2}\Big[\big(n\|q(\mathbf{1}\{i\ge I\})_i\|_1\big) \vee \sqrt{n}\Big]. \tag{2.26}$$

Now considering Equations (2.23) and (2.25), we have for C large enough,

$$3\sum_{i\geq I}\Delta_i \geq 2\sum_{i<I}|\Delta_i|.$$

So

$$9\sum_{i\geq I}\Delta_i \geq 2\sum_{i}|\Delta_i|,$$

that is, by Equation (2.26),

$$\frac{9}{2}\mathbb{E}T_1/n \geq \|\exp(-nq)\Delta\|_1. \tag{2.27}$$

And if the pre-test did not reject the null, then with probability larger than $1-2\delta-2n^{-1}-6\exp(-n/18)$, there exists $+\infty > c_\delta > 0$ that only depends on δ and such that

$$|\Delta_i| < c_\delta\left(\sqrt{q_i\frac{\log(n)}{n}}\vee\frac{\log(n)}{n}\right).$$

If $q_i \geq \log(n)/n$, then $|\Delta_i| < c_\delta\sqrt{q_i\log(n)/n}$. So

$$n\sqrt{\|\exp(-nq)\Delta^2\|_1} \leq c_\delta\sqrt{\log(n)\|q\|_1} = c_\delta\sqrt{\log(n)}.$$

If $q_i < \log(n)/n$, then $|\Delta_i| < c_\delta\log(n)/n$. So

$$n\sqrt{\|\exp(-nq)\Delta^2\|_1} \leq \sqrt{c_\delta n\log(n)\|\exp(-nq)\Delta\|_1}.$$

Using Equation (2.21), we end up with probability larger than $1-2\delta-2n^{-1}-6\exp(-n/18)$:

$$\sqrt{\mathbb{V}T_1} \leq (\sqrt{2n\|q\exp(-nq)\|_1}+c_\delta\sqrt{\log n})+\sqrt{n|\sum_i\Delta_i\exp(-nq_i)|} + c_\delta\sqrt{n\log(n)}\sqrt{\|\exp(-nq)\Delta\|_1}.$$

Now let us compare the terms from the standard deviation $\sqrt{\mathbb{V}T_1}$ with $\mathbb{E}T_1$.

For the first term, we have

$$(\sqrt{2n\|q\exp(-nq)\|_1}+c_\delta\sqrt{\log n}) \leq (2+c_\delta)\sqrt{n} \leq \frac{2(2+c_\delta)}{C}\mathbb{E}T_1.$$

For the second term,

$$\sqrt{n|\sum_i\Delta_i\exp(-nq_i)|} = \sqrt{\mathbb{E}T_1}n^{-1/4}n^{1/4}.$$

So, since $2ab \leq a^2+b^2$ for any a, b, we have:

$$\sqrt{n|\sum_i\Delta_i\exp(-nq_i)|} \leq (n^{-1/2}\mathbb{E}T_1+\sqrt{n})/2 \leq (n^{-1/2}+2/C)\mathbb{E}T_1/2.$$

For the third term, in the same way,

$$c_\delta \sqrt{n \log(n)} \sqrt{\| \exp(-nq)\Delta\|_1} \leq c_\delta (\sqrt{n} \| \exp(-nq)\Delta\|_1 \log(n) + \sqrt{n})/2.$$

So we have by Equation (2.27):

$$c_\delta \sqrt{n \log(n)} \sqrt{\| \exp(-nq)\Delta\|_1} \leq c_\delta (9/2 \log(n)/\sqrt{n} + 2/C)\mathbb{E}T_1/2.$$

And so by Chebyshev's inequality, with probability larger than $1 - \alpha - 2\delta - 2n^{-1} - 6\exp(-n/18)$, we have

$$|T_1 - \mathbb{E}T_1| \leq \mathbb{E}T_1 \alpha^{-1/2}(2(2+c_\delta)\sqrt{2\delta^{-1}}/C + (n^{-1/2} + 2/C)/2 + c_\delta(\frac{9}{2}\frac{\log(n)}{\sqrt{n}} + 2/C)/2).$$

So if $C \geq 4\alpha^{-1/2}(1+4\sqrt{2\delta^{-1}} + c_\delta(1+2\sqrt{2\delta^{-1}}))$, and $2n^{-1/2}\alpha^{-1/2}(1+9c_\delta \log(n)/2) \leq 1$ (which is satisfied for n large enough), we have with probability larger than $1 - \alpha - 2\delta - 2n^{-1} - 6\exp(-n/18)$:

$$T_1 \geq \mathbb{E}T_1/2 \geq \frac{C}{4}\Big[\big(n\|q(\mathbf{1}\{i \geq I\})_i\|_1\big) \vee \sqrt{n}\Big].$$

So we have

$$T_1 \geq \frac{C}{4}\sqrt{n}.$$

Under $H^{(\mathrm{Clo})}_{1,\pi}(\rho)$, analysis 2. The analysis remains the same as analysis 1, with I replaced by J.

So the assumptions become:

$$\|\Delta(\mathbf{1}\{i \geq J\})_i\|_1 \geq C\Big[\|q(\mathbf{1}\{i \geq J\})_i\|_1 \vee \sqrt{\frac{\log(n)}{n}}\Big],$$

and

$$\|\Delta(\mathbf{1}\{i \geq J\})_i\|_1 \geq 2\|\Delta(\mathbf{1}\{i < J\})_i\|_1.$$

We then obtain, if $C \geq 4\alpha^{-1/2}(1 + 4\sqrt{2\delta^{-1}} + c_\delta(1 + 2\sqrt{2\delta^{-1}}))$, and $2n^{-1/2}\alpha^{-1/2}(1 + 9c_\delta \log(n)/2) \leq 1$, we have with probability larger than $1-\alpha-2\delta-2n^{-1}-6\exp(-n/18)$:

$$T_1 \geq \frac{C}{4}\Big[\big(n\|q(\mathbf{1}\{i \geq J\})_i\|_1\big) \vee \sqrt{n}\Big].$$

So we have

$$T_1 \geq \frac{C}{4}\sqrt{n}.$$

Finally, the guarantees on the statistic T_1 allow us to conclude the proof.

2.6.6 Proofs of Theorem 21 and Corollary 6

Let us prove Theorem 21 by combining all the guarantees on the ensemble of tests. From Propositions 9, 10, 11, we know that whenever $\Delta = 0$, all tests accept the null with probability larger than $1 - 5\delta - 2n^{-1} - 6\exp(-n/18)$. Besides, for $\tilde{c}_\delta$ large enough

depending only on δ, whenever there exists $I \geq J_q$ such that

$$\begin{aligned}&\|\Delta\|_1\\ &\geq \tilde{c}_\delta \Bigg\{ \left[\left(\sqrt{I - J_q} \frac{\log(n)}{n} \right) \vee \left(\frac{\sqrt{I - J_q}}{\sqrt{n}} \|q^2 \exp(-2nq)\|_1^{1/4} \right) \vee \|q(\mathbf{1}\{i \geq I\})_i\|_1 \right] \\ &\qquad \wedge \|q(\mathbf{1}\{i \geq J_q\})_i\|_1 \Bigg\} \vee \left[\frac{\left\| q^2 \frac{1}{(q \vee n^{-1})^{4/3}} \right\|_1^{3/4}}{\sqrt{n}} \right] \vee \left[\sqrt{\frac{\log(n)}{n}} \right],\end{aligned}$$

at least one test (and so the final test) rejects the null with probability larger than $1 - 5\delta - 2n^{-1} - 6\exp(-n/18)$.

2.6.7 Proofs for the thresholds: Theorems 24 and 25

2.6.7.1 Proof of Theorem 24 for threshold $\hat{t}_{2/3}$

Lemma 10. *Let $Z \sim \mathcal{P}(\lambda)$, where $\lambda \geq 0$. It holds that if $\lambda \geq 1$,*

$$e^{-2/3}\lambda^{2/3}/2 \leq \mathbb{E}(Z^{2/3}) \leq \lambda^{2/3},$$

and if $\lambda \leq 1$,

$$\lambda e^{-\lambda} \leq \mathbb{E}(Z^{2/3}) \leq \lambda.$$

Proof of Lemma 10.
Upper bound on the expectation. The function $t \to t^{2/3}$ is concave. So by application of Jensen's inequality, we have:

$$\mathbb{E}(Z^{2/3}) \leq \lambda^{2/3}.$$

Also we have by definition of the Poisson distribution

$$\begin{aligned}\mathbb{E}(Z^{2/3}) &= \sum_{i=1}^{\infty} \frac{\lambda^i}{(i-1)!} e^{-\lambda} i^{-1/3} \\ &= \lambda e^{-\lambda} \sum_{j \geq 0} \frac{\lambda^j}{j!} (j+1)^{-1/3} \leq \lambda.\end{aligned}$$

We conclude that if $\lambda \geq 1$,

$$\mathbb{E}(Z^{2/3}) \leq \lambda^{2/3},$$

and if $\lambda \leq 1$,

$$\mathbb{E}(Z^{2/3}) \leq \lambda.$$

Lower bound on the expectation in the case $\lambda \geq e^{-2}$. We have by definition of the Poisson distribution

$$\begin{aligned}\mathbb{E}(Z^{2/3}) &= \sum_{i=1}^{\infty} \frac{\lambda^i}{(i-1)!} e^{-\lambda} i^{-1/3} \\ &\geq \lambda e^{-\lambda} \sum_{0 \leq j \leq e^2\lambda - 1} \frac{\lambda^j}{j!} (j+1)^{-1/3} \\ &\geq e^{-2/3}\lambda^{2/3} e^{-\lambda} \sum_{0 \leq j \leq e^2\lambda - 1} \frac{\lambda^j}{j!},\end{aligned}$$

because $e^2\lambda - 1 \geq 0$ here. Then since $j! \geq \sqrt{2\pi} j^j e^{-j}$,

$$\begin{aligned}
\mathbb{E}(Z^{2/3}) &\geq e^{-2/3}\lambda^{2/3}\left(1 - \frac{e^{-\lambda}}{\sqrt{2\pi}} \sum_{j \geq \lfloor e^2\lambda \rfloor} \frac{\lambda^j e^j}{j^j}\right) \\
&\geq e^{-2/3}\lambda^{2/3}\left(1 - \frac{e^{-\lambda}}{\sqrt{2\pi}} \sum_{j \geq \lfloor e^2\lambda \rfloor} \frac{\lambda^j e^j}{(c_{\lfloor\rfloor} e^2\lambda)^j}\right) \\
&\geq e^{-2/3}\lambda^{2/3}\left(1 - \frac{e^{-\lambda}}{\sqrt{2\pi}} \sum_{j \geq \lfloor e^2\lambda \rfloor} (c_{\lfloor\rfloor} e)^{-j}\right),
\end{aligned}$$

where $1/2 \leq c_{\lfloor\rfloor} \leq 1$ such that $c_{\lfloor\rfloor} e^2\lambda = \lfloor e^2\lambda \rfloor$ because $e^2\lambda \geq 1$. Finally,

$$\mathbb{E}(Z^{2/3}) \geq e^{-2/3}\lambda^{2/3}\left(1 - \frac{e^{-\lambda}}{\sqrt{2\pi}} (c_{\lfloor\rfloor} e)^{-\lfloor e^2\lambda \rfloor} \frac{1}{1 - (c_{\lfloor\rfloor} e)^{-1}}\right).$$

In particular, if $\lambda \geq 1$,

$$\mathbb{E}(Z^{2/3}) \geq e^{-2/3}\lambda^{2/3}/2.$$

Lower bound in all cases. Without any assumption on λ it holds that $\mathbb{E}(Z^{2/3}) \geq \lambda e^{-\lambda}$.

Conclusion on the lower bound.

So, if $\lambda \geq 1$, $\mathbb{E}(Z^{2/3}) \geq e^{-2/3}\lambda^{2/3}/2$, and if $\lambda \leq 1$, $\mathbb{E}(Z^{2/3}) \geq \lambda e^{-\lambda}$

□

Lemma 11. *Let $Z \sim \mathcal{P}(\lambda)$, where $\lambda \geq 0$. It holds if $\lambda \geq e^{-2}$ that*

$$\mathbb{E}(Z^{4/3}) \leq \lambda^{4/3} e^{8/3} + e^{-\lambda} \frac{1}{1 - e^{-1/2}},$$

and if $\lambda < e^{-2}$ that

$$\mathbb{E}(Z^{4/3}) \leq e^{-\lambda}\lambda(2^{1/3} + e).$$

Proof of Lemma 11. Assume that $\lambda \geq e^{-2}$. We have by definition of the Poisson distribution

$$\begin{aligned}
\mathbb{E}(Z^{4/3}) &= \sum_{i \geq 1} \frac{\lambda^i}{i!} e^{-\lambda} i^{4/3} \\
&= \sum_{1 \leq i \leq e^2\lambda} \frac{\lambda^i}{i!} e^{-\lambda} i^{4/3} + \sum_{i > e^2\lambda} \frac{\lambda^i}{i!} e^{-\lambda} i^{4/3} \\
&\leq \lambda^{4/3} e^{8/3} + e^{-\lambda} \sum_{i \geq e^2\lambda} \frac{\lambda^i e^i}{i^i} i^{4/3},
\end{aligned}$$

using the inequality: $i! \geq i^i/e^i$. Then,

$$\begin{aligned}
\mathbb{E}(Z^{4/3}) &\leq \lambda^{4/3}e^{8/3} + e^{-\lambda}\sum_{i\geq e^2\lambda}\frac{\lambda^i e^i}{(e^2\lambda)^i}i^{4/3} \\
&= \lambda^{4/3}e^{8/3} + e^{-\lambda}\sum_{i\geq e^2\lambda} i^{4/3}e^{-i} \\
&\leq \lambda^{4/3}e^{8/3} + e^{-\lambda}\sum_{i\geq e^2\lambda} e^{-i/2} \\
&\leq \lambda^{4/3}e^{8/3} + e^{-\lambda}\frac{1}{1-e^{-1/2}}.
\end{aligned}$$

Now assume that $e^2\lambda < 1$. Then

$$\begin{aligned}
E(Z^{4/3}) &= \sum_{i\geq 1}\frac{\lambda^i}{i!}e^{-\lambda}i^{4/3} \\
&\leq e^{-\lambda}\lambda\left(1 + \sum_{j\geq 0}\frac{(j+2)^{1/3}}{j+1}\frac{1}{j!}\right) \\
&\leq e^{-\lambda}\lambda\left(1 + 2^{1/3} + \sum_{j\geq 1}\frac{1}{j!}\right) \\
&= e^{-\lambda}\lambda(2^{1/3} + e)
\end{aligned}$$

□

Proof of Theorem 24. By application of Lemma 10, we have the following bounds on the expectation of the empirical threshold with probability larger than $1-6\exp(-n/18)$ with respect to $(\bar{n}_m^{(j)})_{m\leq 2, j\leq 3}$,

$$\begin{aligned}
&\left\|\frac{e^{-2/3}q^{2/3}}{2}\mathbf{1}\{q \geq 1/n\}\right\|_1 \\
&\quad + \left\|n^{1/3}qe^{-nq}\mathbf{1}\{q \leq 1/n\}\right\|_1 \leq n^{-2/3}\mathbb{E}\|(Y^{(1)})^{2/3}\|_1 \leq \left\|q^{2/3}\mathbf{1}\{q \geq 1/n\}\right\|_1 \\
&\qquad\qquad + \|q\mathbf{1}\{q \leq 1/n\}\|_1 .
\end{aligned}$$

Now let us consider the standard deviation of the empirical threshold. We have by application of Lemma 11, with probability larger than $1 - 6\exp(-n/18)$ with respect

to $(\bar{n}_m^{(j)})_{m\le 2, j\le 3}$

$$
\begin{aligned}
& n^{-2/3}\sqrt{\mathbb{V}\|(Y^{(1)})^{2/3}\|_1} \\
& \le \sqrt{\|[q^{4/3}e^{8/3} + \frac{1}{1-e^{-1/2}} n^{-4/3} e^{-nq}]\mathbf{1}\{q \ge 1/n\}\|_1} \\
& \quad + \sqrt{\|n^{-1/3}(2^{1/3}+e)qe^{-nq}\mathbf{1}\{q \le 1/n\}\|_1} \\
& \le \sqrt{\|q^{4/3}e^{8/3}\mathbf{1}\{q \ge 1/n\}\|_1 + 1} + \sqrt{\|n^{-1/3}(2^{1/3}+e)qe^{-nq}\mathbf{1}\{q \le 1/n\}\|_1} \\
& \le \sqrt{\|q^{4/3}e^{8/3}\mathbf{1}\{q \ge 1/n\}\|_1 + 1} + \sqrt{(2^{1/3}+e)} \\
& \le \sqrt{e^{8/3}+1} + \sqrt{(2^{1/3}+e)} = C_1.
\end{aligned}
$$

Then by application of Chebyshev's inequality, we have with probability greater than $1 - \beta - 6\exp(-n/18)$,

$$
\begin{aligned}
& \left\| \frac{e^{-2/3}q^{2/3}}{2}\mathbf{1}\{q \ge 1/n\} \right\|_1 \\
& + \left\| n^{1/3}qe^{-nq}\mathbf{1}\{q \le 1/n\} \right\|_1 - \frac{C_1}{\sqrt{\beta}} \\
& \qquad \le n^{-2/3}\|(Y^{(1)})^{2/3}\|_1 \\
& \qquad \le \left\| q^{2/3}\mathbf{1}\{q \ge 1/n\} \right\|_1 \\
& \qquad + \|q\mathbf{1}\{q \le 1/n\}\|_1 + \frac{C_1}{\sqrt{\beta}}.
\end{aligned}
$$

Now, on the one hand, we have that

$$
n^{4/3}q^2\mathbf{1}\{q \le 1/n\} \le n^{1/3}q\mathbf{1}\{q \le 1/n\},
$$

so

$$
\left\| \Big(\frac{1}{q \vee n^{-1}}\Big)^{4/3} q^2\mathbf{1}\{q \le 1/n\} \right\|_1 \le \|n^{4/3}q^2\mathbf{1}\{q \le 1/n\}\|_1 \le \|n^{1/3}q\mathbf{1}\{q \le 1/n\}\|_1.
$$

And on the other hand,

$$
\left\| \Big(\frac{1}{q \vee n^{-1}}\Big)^{4/3} q^2\mathbf{1}\{q \ge 1/n\} \right\|_1 \le \|q^{2/3}\mathbf{1}\{q \ge 1/n\}\|_1.
$$

So with probability larger than $1 - \beta - 6\exp(-n/18)$ we have

$$
\begin{aligned}
& (e^{-2/3}/2 + 1)\left\| \Big(\frac{1}{q \vee n^{-1}}\Big)^{4/3} q^2 \right\|_1 \\
& \qquad \le n^{-2/3}\|(Y^{(1)})^{2/3}\|_1 + C_{2/3}/\sqrt{\beta} \\
& \qquad \le \left\| q^{2/3}\mathbf{1}\{q \ge 1/n\} \right\|_1 + 1 + 2C_{2/3}/\sqrt{\beta}.
\end{aligned}
$$

□

2.6.7.2 Proof of Theorem 25 for threshold $\hat{t}_2$

Lemma 12. *Consider three independent random vectors $Y^{(1)}$, $Y^{(2)}$ and $Y^{(3)}$ distributed according to $\mathcal{P}(nq)$. We obtain the following expectation:*

$$\mathbb{E}(\|Y^{(1)}Y^{(2)}\mathbf{1}\{Y^{(3)}=0\}\|_1) = \|(nq)^2e^{-nq}\|_1.$$

Proof. Firstly,

$$\mathbb{E}(\|Y^{(1)}Y^{(2)}\mathbf{1}\{Y^{(3)}=0\}\|_1) = \|\mathbb{E}(Y^{(1)}Y^{(2)}\mathbf{1}\{Y^{(3)}=0\})\|_1.$$

Now

$$\mathbb{E}(Y^{(1)}Y^{(2)}\mathbf{1}\{Y^{(3)}=0\}) = \mathbb{E}(Y^{(1)})\mathbb{E}(Y^{(2)})\mathbb{P}(Y^{(3)}=0) = (nq)^2e^{-nq}.$$

So

$$\mathbb{E}(\|Y^{(1)}Y^{(2)}\mathbf{1}\{Y^{(3)}=0\}\|_1) = \|(nq)^2e^{-nq}\|_1.$$

□

Lemma 13. *Consider three independent random vectors $Z^{(1)}$, $Z^{(2)}$ and $Z^{(3)}$ distributed according to $\mathcal{P}(nq)$ and whose elements are independent, too. We obtain the following variance:*

$$\mathbb{V}(\|Z^{(1)}Z^{(2)}\mathbf{1}\{Z^{(3)}=0\}\|_1) = \|((nq)^2+nq)^2e^{-nq} - (nq)^4e^{-2nq}\|_1.$$

Proof. Each sample $Z^{(i)}$ consists in a vector of independent elements. So

$$\mathbb{V}(\|Z^{(1)}Z^{(2)}\mathbf{1}\{Z^{(3)}=0\}\|_1) = \|\mathbb{V}(Z^{(1)}Z^{(2)}\mathbf{1}\{Z^{(3)}=0\})\|_1.$$

Now

$$\mathbb{V}(Z^{(1)}Z^{(2)}\mathbf{1}\{Z^{(3)}=0\}) = \left[\mathbb{E}((Z^{(1)})^2)\mathbb{E}((Z^{(2)})^2) - \mathbb{E}((Z^{(1)}))^2\mathbb{E}((Z^{(2)}))^2\right]\mathbb{P}(Z^{(3)}=0)$$

by independence between $Z^{(1)}$, $Z^{(2)}$ and $Z^{(3)}$.

And

$$\mathbb{E}((Z^{(1)})^2) = E((Z^{(2)})^2) = (nq)^2+nq.$$

So

$$\mathbb{V}(\|Z^{(1)}Z^{(2)}\mathbf{1}\{Z^{(3)}=0\}\|_1) = \|((nq)^2+nq)^2e^{-nq} - (nq)^4e^{-2nq}\|_1.$$

□

Proof of Theorem 25. By application of lemma 12, we have for the expectation of the empirical threshold with probability larger than $1-6\exp(-n/18)$ with respect to $(\bar{n}_m^{(j)})_{m\leq 2,j\leq 3}$:

$$\mathbb{E}(\|Y^{(1)}Y^{(2)}\mathbf{1}\{Y^{(3)}=0\}\|_1) = \|(nq)^2e^{-nq}\|_1.$$

Then by application of lemma 13, we have for the standard deviation of the empirical threshold with probability larger than $1-6\exp(-n/18)$ with respect to $(\bar{n}_m^{(j)})_{m\leq 2,j\leq 3}$:

$$\sqrt{\mathbb{V}(\|Y^{(1)}Y^{(2)}\mathbf{1}\{Y^{(3)}=0\}\|_1)} \leq \sqrt{2}(\sqrt{\|(nq)^4e^{-nq}\|_1} + \sqrt{\|(nq)^2e^{-nq}\|_1}).$$

In particular,

$$\mathbb{E}(\|Y^{(1)}Y^{(2)}\mathbf{1}\{Y^{(3)}=0\}\|_1/n^2) = \|q^2e^{-nq}\|_1,$$

and

$$\sqrt{\mathbb{V}(\|Y^{(1)}Y^{(2)}\mathbf{1}\{Y^{(3)}=0\}\|_1/n^2)} \leq \sqrt{2}(\sqrt{\|q^4e^{-nq}\|_1} + \sqrt{\|q^2e^{-nq}\|_1}/n).$$

Let us compare both terms of the standard deviation with the expectation.

Firstly,

$$\begin{aligned}\sqrt{\|q^2e^{-nq}\|_1}/n &\leq 1/2(\|q^2e^{-nq}\|_1\sqrt{\delta}/4 + 4/(\sqrt{\delta}n^2)) \\ &\leq 1/2(\|q^2e^{-nq}\|_1\sqrt{\delta}/4 + 4\log(n)^4/(\sqrt{\delta}n^2)).\end{aligned}$$

Secondly, for an upper bound on $\sqrt{\|q^4e^{-nq}\|_1}$, we consider two regimes.

Study of the large q_i's.

We consider $q_i \geq 5\log(n)/n$. Then we have the following upper bound on the number of such q_i's,

$$\#\{i|q_i \geq 5\log(n)/n\} \leq 1/(5\log(n)/n).$$

So

$$\sqrt{\|q^4e^{-nq}\mathbf{1}\{q \geq 5\log(n)/n\}\|_1} \leq n^{-2}/\sqrt{5\log n} \leq 3\log(n)^4/n^2.$$

Study of the small q_i's.

We consider $q_i < 5\log(n)/n$.

$$\begin{aligned}\sqrt{\|q^4e^{-nq}\mathbf{1}\{q < 5\log(n)/n\}\|_1} &\leq 5\log(n)/n\sqrt{\|q^2e^{-nq}\|_1} \\ &\leq 1/2(\|q^2e^{-nq}\|_1\sqrt{\delta}/4 + 100\log(n)^2/(\sqrt{\delta}n^2)) \\ &\leq 1/2(\|q^2e^{-nq}\|_1\sqrt{\delta}/4 + 2000\log(n)^4/(\sqrt{\delta}n^2)).\end{aligned}$$

Finally,

$$\begin{aligned}&\sqrt{\mathbb{V}(\|Y^{(1)}Y^{(2)}\mathbf{1}\{Y^{(3)}=0\}\|_1/n^2)} \\ &\leq \frac{\sqrt{\delta}}{2}\mathbb{E}(\|Y^{(1)}Y^{(2)}\mathbf{1}\{Y^{(3)}=0\}\|_1/n^2) + 1005\log(n)^4/(\sqrt{\delta}n^2).\end{aligned}$$

So by application of Chebyshev's inequality, we have with probability greater than $1-\delta-6\exp(-n/18)$:

$$|\|Y^{(1)}Y^{(2)}\mathbf{1}\{Y^{(3)}=0\}\|_1 - \|(nq)^2e^{-nq}\|_1| \leq 1/2\|(nq)^2e^{-nq}\|_1 + \frac{1005}{\delta}\log(n)^4.$$

□

2.7 Proofs of the lower bounds: Propositions 12, 13, 14 and Theorem 22

2.7.1 Proof of Propositions 12

The lower bound obtained in Valiant and Valiant (2017) and Balakrishnan and Wasserman (2017a) for identity testing will also be useful to us as a lower bound for closeness testing.

Proof of Proposition 12 and adaptation to Theorem 22. As a corollary from Theorem 1 in Balakrishnan and Wasserman (2017a), there exists a constant $c'_\gamma > 0$ that depends

only on γ such that for any $q \in \mathbf{P}_\pi$

$$\rho^*_\gamma(H^{(\mathrm{Clo})}_{0,\pi},\ H^{(\mathrm{Clo})}_{1,\pi}(\rho^*_\gamma), \mathcal{T}^{(\mathrm{Clo})}_n) \geq c'_\gamma \min_I \left[\frac{\|q^{2/3}_{(.)}(\mathbf{1}\{2 \leq i < I\})_i\|_1^{3/4}}{\sqrt{n}} \vee \frac{1}{n} + \|q_{(.)}(\mathbf{1}\{i \geq I\})_i\|_1 \right].$$

In particular, taking $q = \pi$, there exists a constant $c'_\gamma > 0$ that depends only on γ and such that

$$\rho^*_\gamma(H^{(\mathrm{Clo})}_{0,\pi},\ H^{(\mathrm{Clo})}_{1,\pi}(\rho^*_\gamma), \mathcal{T}^{(\mathrm{Id})}_n) \geq c'_\gamma \min_I \left[\frac{\|\pi^{2/3}_{(.)}(\mathbf{1}\{2 \leq i < I\})_i\|_1^{3/4}}{\sqrt{n}} \vee \frac{1}{n} + \|\pi_{(.)}(\mathbf{1}\{i \geq I\})_i\|_1 \right].$$

□

We then adapt Proposition 12 to the purpose of obtaining Theorem 22.

Proposition 15. *Let $\pi \in \mathbf{P}$ and $\gamma > 0$. There exists a constant $c_\gamma > 0$ that depends only on γ such that $\rho^*_\gamma(H^{(\mathrm{Clo})}_{0,\pi},\ H^{(\mathrm{Clo})}_{1,\pi}(\rho^*_\gamma), \mathcal{T}^{(Clo)}_n) \geq \frac{c_\gamma}{\sqrt{n}} \left[\left\| \pi^{2/3} \frac{1}{(\pi \vee n^{-1})^{4/3}} \right\|_1^{3/4} \vee 1 \right]$.*

Proof of Proposition 15. From Proposition 12, there exists a constant $c'_\gamma > 0$ that depends only on γ and such that

$$\rho^*_\gamma(H^{(\mathrm{Clo})}_{0,\pi},\ H^{(\mathrm{Clo})}_{1,\pi}(\rho^*_\gamma), \mathcal{T}^{(\mathrm{Clo})}_n) \geq c'_\gamma \min_I \left[\frac{\|\pi^{2/3}_{(.)}(\mathbf{1}\{2 \leq i < I\})_i\|_1^{3/4}}{\sqrt{n}} \vee \frac{1}{n} \vee \|\pi_{(.)}(\mathbf{1}\{i \geq I\})_i\|_1 \right].$$

Let I^* denote one of the I's where the minimum from the right-hand side of the previous inequality is attained.

Case 1: $\|\pi_{(.)}(\mathbf{1}\{i > I^*\})_i\|_1 > 1/2$. The result follows immediately.

Case 2: $\|\pi_{(.)}(\mathbf{1}\{i > I^*\})_i\|_1 \leq 1/2$. So $\|\pi^{2/3}_{(.)}(\mathbf{1}\{i \leq I^*\})_i\|_1 \geq 1/2$ since $\|\pi\|_1 = 1$, implying that $\rho^*_\gamma \geq c'_\gamma (1/2)^{3/4}/\sqrt{n}$.

Subcase 1: $I^* \geq J_\pi$. We have $\|\pi_{(.)}(\mathbf{1}\{i > I^*\})_i\|_1 \geq \|\pi^2(\mathbf{1}\{i > I^*\})_i\|_1 \sqrt{n}$.

Subcase 2: $I^* < J_\pi$. Having for all $J_\pi \geq i > I^*$, $\pi_{(i)} \geq 1/n$ implies that we have

$$\|\pi^{2/3}_{(.)}(\mathbf{1}\{J_\pi \geq i > I^*\})_i\|_1 \leq n^{1/3} \|\pi_{(.)}(\mathbf{1}\{J_\pi \geq i > I^*\})_i\|_1.$$

And so

$$\|\pi^{2/3}_{(.)}(\mathbf{1}\{J_\pi \geq i > I^*\})_i\|_1^{3/4} \leq n^{1/2} \|\pi_{(.)}(\mathbf{1}\{J_\pi \geq i > I^*\})_i\|_1$$

since $\|\pi_{(.)}(\mathbf{1}\{J_\pi \geq i > I^*\})_i\|_1 \leq 1$.

Finally

$$\|\pi^{2/3}_{(.)}(\mathbf{1}\{J_\pi \geq i > I^*\})_i\|_1^{3/4}/\sqrt{n} \leq \|\pi_{(.)}(\mathbf{1}\{J_\pi \geq i > I^*\})_i\|_1,$$

which implies that I^* must be larger than J_π. This concludes the proof in any case. □

This concludes the proof of Proposition 12 .

2.7.2 Classical method for proving lower bounds: the Bayesian approach

Let us fix some $\gamma \in (0,1)$. Finding a lower bound on $\rho^*_\gamma(H^{(\mathrm{Clo})}_{0,\pi}, H^{(\mathrm{Clo})}_{1,\pi}, \mathcal{T}^{(\mathrm{Clo})}_n))$ amounts to finding a real number ρ such that $R(H^{(\mathrm{Clo})}_{0,\pi}, H^{(\mathrm{Clo})}_{1,\pi}(\rho), \hat{\theta}_n) > \gamma$ for any test φ. We now apply a Bayesian approach. Let $(\alpha,\beta) \in [0,1)$ such that $\alpha+\beta<1$. Let ν_0 and $\nu_{1,\rho}$ be distributions such that for any $\rho>0$,

$$\nu_0(H^{(\mathrm{Clo})}_{0,\pi}) \geq 1-\alpha, \quad \nu_{1,\rho}(H^{(\mathrm{Clo})}_{1,\pi}(\rho)) \geq 1-\beta.$$

So for any $\hat{\theta}_n \in \mathcal{T}^{(\mathrm{Clo})}_n$, we in the same way as in Lemma 2 from Section 1.4.2.

$$R(H^{(\mathrm{Clo})}_{0,\pi}, H^{(\mathrm{Clo})}_{1,\pi}(\rho), \hat{\theta}_n) \geq 1 - d_{TV}(\mathbb{P}_{\nu_0}, \mathbb{P}_{\nu_1}) - \alpha - \beta, \tag{2.28}$$

where d_{TV} denotes the total variation distance. Thus, the lower bound that is obtained heavily relies on the choice of ν_0 and ν_1.

2.7.3 Proof of Proposition 13

Let us prove the lower bound stated in Proposition 13. It heavily relies on the Bayesian approach presented in Section 2.7.2. Let us recall the definition of $I_{v,\pi}$. Set for $v \geq 0$, with the convention $\min_{j\leq d} \emptyset = d$,

$$\begin{aligned} I_{v,\pi} = \min_{J_\pi \leq j \leq d} \Big\{ \{j : \pi_{(j)} \leq \sqrt{C_\pi/j}\} \cap \{j : \sum_{i\geq j} \exp(-2n\pi_{(i)})\pi^2_{(i)} \leq C_\pi\} \\ \cap \{j : \sum_{i\geq j} \pi_{(i)} \leq \sum_{J_\pi \leq i < j} \pi_{(i)}\} \Big\}, \end{aligned} \tag{2.29}$$

where

$$C_\pi = \frac{\sqrt{\sum_i \pi_i^2 \exp(-2(1+v)n\pi_i)}}{n}. \tag{2.30}$$

In what follows, we also consider for any $i \leq d$, the quantity $\varepsilon^*_i \in [0,1/2]$ that we will specify later.

Definition of some measures. We assume that $\pi \in \mathbf{P}$ is fixed such that for any $i \leq j \leq d$, we have $\pi_i \geq \pi_j$. Let $0<\delta\leq 1/8$ and $4[1 \vee (32\log(1/\delta))^2] \leq M \leq \sqrt{n}$. We define $\mathcal{A} \subset \{1,\dots,d\}$ such that for any $i \in \mathbb{Z}$ where $|S_\pi(\lfloor \log_2(n)\rfloor + i)| > a\sqrt{\log((|i|+1)/\gamma)}$, we have that $\mathcal{A} \cap S_\pi(\lfloor \log_2(n)\rfloor + i)$ are the $\lfloor |S_\pi(\lfloor \log_2(n)\rfloor + i)|/M \rfloor$ largest elements of $S_\pi(\lfloor \log_2(n)\rfloor + i)$. Note that $\sum_{\mathcal{A}} \pi_i \leq 2/M \leq 1/2$. Let $\mathcal{A}' = \mathcal{A} \cap \{J_\pi, \dots, d\}$.

Let $U_{\pi,\mathcal{A}}$ be the discrete distribution that is uniform over $\{\pi_i, i \in \mathcal{A}\}$, and we will write for any $i \in \mathcal{A}$, $U_{\pi,\mathcal{A}}(\{\pi_i\}) = 1/|\mathcal{A}|$. We will now work on the definition of appropriate measures corresponding to (q,p). Conditional to two vectors $q,p \in \mathbb{R}^{+d}$, we define

$$\Lambda_{q,p} = \prod_i \left(\mathcal{P}(nq_i) \otimes \mathcal{P}(np_i)\right).$$

Definition of Λ_0: First, for any $i \in \mathcal{A}$ we will consider independent $q_i \sim U_{\pi,\mathcal{A}}$, and otherwise set $q_i = \pi_i$ for any $i \notin \mathcal{A}$. We write Λ_0 for the distribution $\Lambda_{q,q}$ when q is defined as before:

$$\Lambda_0 = \mathbb{E}_q(\Lambda_{q,q}),$$

where $\mathbb{E}_q$ is the expectation according to the distribution of q.

Definition of Λ_1: We consider q defined as above. For any $i \in \mathcal{A}$, we know that there exists j_i such that $q_i = \pi_{j_i}$. Let us write ξ_i for independent random variables that are uniform in $\{\varepsilon^*_{j_i}, -\varepsilon^*_{j_i}\}$ if $i \in \mathcal{A}'$, and 0 otherwise. Then set for any $i \in \mathcal{A}$: define $p_i = q_i(1+\xi_i)$ and for any $i \notin \mathcal{A}$: define $p_i = \pi_i$. We write Λ_1 for the distribution $\Lambda_{q,p}$ averaged over q, p. So

$$\Lambda_1 = \mathbb{E}_{q,p}(\Lambda_{q,p}),$$

where $\mathbb{E}_{q,p}$ is the expectation according to the distribution of q, p – i.e. according to q, ξ.

Definition of $\tilde{\Lambda}_0$: Now, for any $i \in \mathcal{A}$ we will consider independent $\tilde{q}_i = q_i \sim U_{\pi,\mathcal{A}}$. Then for any $i \notin \mathcal{A}$: set $\tilde{q}_i = \pi_i \left(1 - \frac{\sum_{l\in\mathcal{A}}(\tilde{q}_l - \pi_l)}{\sum_{j\notin\mathcal{A}}\pi_j}\right)$. We write $\tilde{\Lambda}_0$ for the distribution $\Lambda_{\tilde{q},\tilde{q}}$ when $\tilde{q}$ is defined as before:

$$\tilde{\Lambda}_0 = \mathbb{E}_{\tilde{q}}(\Lambda_{\tilde{q},\tilde{q}}),$$

where $\mathbb{E}_{\tilde{q}}$ is the expectation according to the distribution of $\tilde{q}$ – i.e. according to q.

Definition of $\tilde{\Lambda}_1$: We consider $q, \xi, \tilde{q}$ defined as above. Then for any $i \in \mathcal{A}$, set $\tilde{p}_i = p_i = q_i(1+\xi_i)$. For any $i \notin \mathcal{A}$, set $\tilde{p}_i = \pi_i \left(1 - \frac{\sum_{l\in\mathcal{A}}(\tilde{p}_l - \pi_l)}{\sum_{j\notin\mathcal{A}}\pi_j}\right)$. We write Λ_1 for the distribution $\Lambda_{\tilde{q},\tilde{p}}$ averaged over $\tilde{q}, \tilde{p}$. So

$$\Lambda_1 = \mathbb{E}_{\tilde{q},\tilde{p}}(\Lambda_{\tilde{q},\tilde{p}}),$$

where $\mathbb{E}_{\tilde{q},\tilde{p}}$ is the expectation according to the distribution of $\tilde{q}, \tilde{p}$ – i.e. according to q, ξ. Note that $\sum_i \tilde{q}_i = 1 = \sum_i \tilde{p}_i$.

Properties of $\tilde{\Lambda}_0$ and $\tilde{\Lambda}_1$, and bound on their total variation distance. We first prove the following lemma, which implies that $\tilde{\Lambda}_0$ and $\tilde{\Lambda}_1$ take values in $\mathbf{P}_\pi$ with high probability.

Lemma 14. *Assume that $M \geq 4(16\log 2/\delta)^2$, and that $a > 2$. There exists a universal constant $c > 0$ such that we have with probability larger than $1 - \delta - c\gamma$ with respect to $\tilde{q}$ that $\tilde{q} \in \mathbf{P}_\pi$, and with probability larger than $1 - \delta - c\gamma$ with respect to $\tilde{p}$ that $\tilde{p} \in \mathbf{P}_\pi$.*

We now turn to $d_{TV}(\Lambda_0, \Lambda_1)$ and state the following lemmas which will help us conclude on a bound on $d_{TV}(\Lambda_0, \Lambda_1)$.

Lemma 15. *Let $\pi \in (\mathbb{R}^+)^d$ such that $\sum_i \pi_i \leq 1$ and such that it is ordered in decreasing order, i.e. $\forall i \leq j \leq d$, $\pi_i \geq \pi_j$. We remind the reader that $J := J_\pi$. Let $1 > u > 0, v \geq 0$. Then there exists $\varepsilon^* \in \mathbb{R}^d$ such that for any $i \leq d$*

- *$\varepsilon^*_i \in [0, 1/2]$ and $\varepsilon^*_i = 0$ for any i such that $\pi_i \geq 1/n$.*
- *$\sum_i \pi_i^2 \varepsilon_i^{*2} \exp(-2n\pi_i) \leq u\frac{\sqrt{\sum_i \pi_i^2 \exp(-2(1+v)n\pi_i)}}{n} := uC_\pi$.*
- *we have $\pi_i\varepsilon^*_i \leq \sqrt{u}\Big[(1/n) \wedge \sqrt{C_\pi/(2I_{v,\pi})} \wedge \pi_i/2\Big]$.*
- *and we have*

$$\sum_i \pi_i\varepsilon^*_i \geq \left[\Big[\sum_{i\geq I_{v,\pi}} \frac{\sqrt{u}\pi_i}{\sqrt{2}}\Big] \vee \frac{\sqrt{uC_\pi}(I_{v,\pi} - J)}{\sqrt{2I_{v,\pi}}}\right] \wedge \Big[\sqrt{\frac{u}{8}}\sum_{i\geq J}\pi_i\Big].$$

We show the following lemma which will help us conclude on a bound on $d_{TV}(\Lambda_0, \Lambda_1)$.

Lemma 16. *Let π satisfying the hypotheses of Lemma 15. We take ε^* associated with π as in Lemma 15 for some $u > 0, v > 0$. Write $\lambda = n\pi$. There exists a constant $\tilde{c}_v > 0$ that depends only on v such that the following holds. Let $\xi = \xi_{|\theta}$ be a random variable that depends on θ and takes a random value uniformly in $\{\varepsilon_j^*, -\varepsilon_j^*\}$ when $\theta = \lambda_j = n\pi_j$. Write $\mathcal{U}_\lambda$ for the uniform distribution over $\{\lambda_i = n\pi_i, i \leq d\}$. We set for $\varepsilon^* \in [0,1]^d$, conditionally on θ:*

$$\nu_{0|\theta} = \mathcal{P}(\theta)^{\otimes 2}, \qquad \nu_0 = \mathbb{E}_{\theta\sim\mathcal{U}_\lambda}(\nu_{0|\theta}),$$

and

$$\nu_{1|\theta,\xi} = \mathcal{P}(\theta) \otimes \frac{\left[\mathcal{P}(\theta(1+\xi)) + \mathcal{P}(\theta(1-\xi))\right]}{2}, \quad \nu_1 = \mathbb{E}_{\theta\sim\mathcal{U}_\lambda,\xi}(\nu_{1|\theta,\xi}).$$

We have

$$d_{TV}(\nu_0^{\otimes d}, \nu_1^{\otimes d}) \leq \sqrt{\tilde{c}_v u},$$

for $u \leq \tilde{c}_v^{-1}$, i.e. for u smaller than a constant that depends only on v.

Let $u > 0$ and $v > 0$. We first apply Lemma 15 to π sorted in decreasing order, which leads to the definition of a vector denoted as $\bar{\varepsilon}^*$. Then we apply Lemma 15 to π restricted to $\mathcal{A}$ and sorted in decreasing order, which defines a vector denoted as $\tilde{\varepsilon}^*$.

Since

$$\sum_{i:|S_\pi(\lfloor\log_2(n)\rfloor+i)|\leq a\sqrt{\log((|i|+1)/\gamma)}} \sum_{j\in S_\pi(\lfloor\log_2(n)\rfloor+i)} \pi_j^2 \exp(-2(1+v)n\pi_j) \leq 4a\frac{\sqrt{\log(1/\gamma)}}{n^2},$$

we have by definition of ε^* in Lemma 15 that if $\|\pi^2 \exp(-2(1+v)n\pi\|_2^2 \geq 8aM\frac{\sqrt{\log(1/\gamma)}}{n^2}$ then $\bar{\varepsilon}^*/(8M) \leq \tilde{\varepsilon}_i^*$, where we assume that $M \geq 4(16\log 2/\delta)^2$, and that $a > 2$.

From now on we take $\varepsilon^* = \bar{\varepsilon}^*/(8M)$. Since $\bar{\varepsilon}^*/(8M) \leq \tilde{\varepsilon}_i^*$, we can apply Lemma 16 to the restriction of π and ε^* to $\mathcal{A}$, so we have that there exists $\tilde{c}_v > 0$ such that for $u \leq \tilde{c}_v^{-1}$

$$d_{TV}(\Lambda_0, \Lambda_1) \leq \sqrt{\tilde{c}_v u}.$$

Total variation distance between Poisson distributions, and Multinomial distributions. We define the following distribution: $M_{\tilde{q},\tilde{p}|n_1,n_2} = \mathcal{M}(n_1, \tilde{q}) \otimes \mathcal{M}(n_2, \tilde{p})$. Let $\mathcal{D} = \mathcal{P}(n\sum_i q_i)$ and $\mathcal{D}' = \mathcal{P}(n\sum_i p_i)$.

By definition of the total variation distance, we have

$$\begin{aligned}
&d_{TV}\left[\mathbb{E}_{\tilde{q},(\hat{n}_1,\hat{n}_2)\sim\mathcal{D}^{\otimes 2}}(M_{\tilde{q},\tilde{q}|\hat{n}_1,\hat{n}_2}), \mathbb{E}_{\tilde{q},\tilde{p},(\hat{n}_1,\hat{n}_2)\sim\mathcal{D}\otimes\mathcal{D}'}(M_{\tilde{q},\tilde{p}|\hat{n}_1,\hat{n}_2})\right] \\
&\quad d_{TV}\Bigg[\mathbb{E}_{\tilde{q},(\hat{n}_1,\hat{n}_2)\sim\mathcal{D}^{\otimes 2}|(\hat{n}_1,\hat{n}_2)\in[n/2,3n/2]^2}(M_{\tilde{q},\tilde{q}|\hat{n}_1,\hat{n}_2}), \\
&\qquad\qquad \mathbb{E}_{\tilde{q},\tilde{p},(\hat{n}_1,\hat{n}_2)\sim\mathcal{D}\otimes\mathcal{D}'|(\hat{n}_1,\hat{n}_2)\in[n/2,3n/2]^2}(M_{\tilde{q},\tilde{p}|\hat{n}_1,\hat{n}_2})\Bigg] \\
&\leq \frac{}{\min_{\mathcal{D}''\in\{\mathcal{D}^{\otimes 2},\mathcal{D}\otimes\mathcal{D}'\}} \mathbb{P}_{\tilde{q},\tilde{p},(\hat{n}_1,\hat{n}_2)\sim\mathcal{D}''}(\hat{n}\in[n/2,3n/2])} \\
&\quad + \max_{\mathcal{D}''\in\{\mathcal{D}^{\otimes 2},\mathcal{D}\otimes\mathcal{D}'\}} \mathbb{P}_{\tilde{q},\tilde{p},(\hat{n}_1,\hat{n}_2)\sim\mathcal{D}''}((\hat{n}_1,\hat{n}_2)\notin[n/2,3n/2]^2).
\end{aligned}$$

This implies for $M \geq 4(32\log(1/\delta))^2$

$$
\begin{aligned}
&d_{TV}\left[\mathbb{E}_{\tilde{q},(\hat{n}_1,\hat{n}_2)\sim\mathcal{D}^{\otimes 2}}(M_{\tilde{q},\tilde{q}|\hat{n}_1,\hat{n}_2}), \mathbb{E}_{\tilde{q},\tilde{p},(\hat{n}_1,\hat{n}_2)\sim\mathcal{D}\otimes\mathcal{D}'}(M_{\tilde{q},\tilde{p}|\hat{n}_1,\hat{n}_2})\right] \\
&\quad d_{TV}\Bigg[\mathbb{E}_{\tilde{q},(\hat{n}_1,\hat{n}_2)\sim\mathcal{D}^{\otimes 2}|(\hat{n}_1,\hat{n}_2)\in[n/2,3n/2]^2}(M_{\tilde{q},\tilde{q}|\hat{n}_1,\hat{n}_2}), \\
&\qquad\qquad \mathbb{E}_{\tilde{q},\tilde{p},(\hat{n}_1,\hat{n}_2)\sim\mathcal{D}\otimes\mathcal{D}'|(\hat{n}_1,\hat{n}_2)\in[n/2,3n/2]^2}(M_{\tilde{q},\tilde{p}|\hat{n}_1,\hat{n}_2})\Bigg] \\
&\leq \frac{}{1-2\exp(-n/48)-\delta} \\
&\quad + 2\exp(-n/48)+\delta \\
&\leq d_{TV}\Bigg[\mathbb{E}_{\tilde{q},(\hat{n}_1,\hat{n}_2)\sim\mathcal{D}^{\otimes 2}|(\hat{n}_1,\hat{n}_2)\in[n/2,3n/2]^2}(M_{\tilde{q},\tilde{q}|\hat{n}_1,\hat{n}_2}), \\
&\qquad\qquad \mathbb{E}_{\tilde{q},\tilde{p},(\hat{n}_1,\hat{n}_2)\sim\mathcal{D}\otimes\mathcal{D}'|(\hat{n}_1,\hat{n}_2)\in[n/2,3n/2]^2}(M_{\tilde{q},\tilde{p}|\hat{n}_1,\hat{n}_2})\Bigg] + 8\exp(-n/48)+4\delta,
\end{aligned}
$$

for $n \geq 2^8$ and $\delta \leq 1/8$. Now we have by Theorem 2,

$$
d_{TV}\left[\mathbb{E}_{\tilde{q},(\hat{n}_1,\hat{n}_2)\sim\mathcal{D}^{\otimes 2}}(M_{\tilde{q},\tilde{q}|\hat{n}_1,\hat{n}_2}), \mathbb{E}_{\tilde{q},\tilde{p},(\hat{n}_1,\hat{n}_2)\sim\mathcal{D}\otimes\mathcal{D}'}(M_{\tilde{q},\tilde{p}|\hat{n}_1,\hat{n}_2})\right] = d_{TV}(\Lambda_0,\Lambda_1).
$$

And so when combined with the previously displayed equation

$$
\begin{aligned}
d_{TV}(\Lambda_0,\Lambda_1) \geq\ & -8\exp(-n/48)-4\delta \\
&+ d_{TV}\Bigg[\mathbb{E}_{\tilde{q},(\hat{n}_1,\hat{n}_2)\sim\mathcal{D}^{\otimes 2}|(\hat{n}_1\hat{n}_2)\in[n/2,3n/2]^2}(M_{\tilde{q},\tilde{q}|\hat{n}_1,\hat{n}_2}), \\
&\qquad\qquad \mathbb{E}_{\tilde{q},\tilde{p},(\hat{n}_1,\hat{n}_2)\sim\mathcal{D}\otimes\mathcal{D}'|(\hat{n}_1,\hat{n}_2)\in[n/2,3n/2]^2}(M_{\tilde{q},\tilde{p}|\hat{n}_1,\hat{n}_2})\Bigg].
\end{aligned} \tag{2.31}
$$

Conclusion. Consider an event of probability larger than $1-\delta$ with respect to q,ξ such that for some $\bar{\rho}>0$ we have

$$
\sum_i |\tilde{q}_i - \tilde{p}_i| = \sum_{i\in\mathcal{A}} q_i|\xi_i| = \sum_{i\in\mathcal{A}'} q_i|\xi_i| \geq \bar{\rho}.
$$

So we have by Equation (2.31) and Lemma 14 that under the condition that $n \geq 2^8 \vee M^2$, $\delta \leq 1/8$, $M \geq 4[1 \vee (32\log(1/\delta))^2]$, and $\|\pi^2\exp(-2(1+v)n\pi\|_2^2 \geq 8aM\frac{\sqrt{\log(1/\gamma)}}{n^2}$, then for any test $\hat{\theta}_n \in \mathcal{T}_n^{(\mathrm{Clo})}$

$$
R(H_{0,\pi}^{(\mathrm{Clo})}, H_{1,\pi}^{(\mathrm{Clo})}(\bar{\rho}), \hat{\theta}_{\lfloor n/2\rfloor}) \geq 1 - 8\exp\left(-\frac{n}{48}\right) - 7\delta - c\gamma - \sqrt{\tilde{c}_v u}.
$$

We now present the following lemma.

Lemma 17. *It holds with probability larger than* $1-\delta$ *that*

$$\sum_{i\in\mathcal{A}}|\xi_i|q_i$$
$$\geq \frac{1}{8M^2}\left[\Big[\sum_{i\geq I_{v,\pi}}\frac{\sqrt{u}\pi_i}{\sqrt{2}}\Big]\vee\frac{\sqrt{u}(I_{v,\pi}-J)}{\sqrt{2I_{v,\pi}}}\sqrt{\frac{\sqrt{\sum_i\pi_i^2\exp(-2(1+v)n\pi_i)}}{n}}\right]\wedge\sqrt{\frac{u}{8}}\right]$$
$$-\frac{1}{\sqrt{nM\delta}}-8\frac{a(1+\log(1/\gamma))}{n}.$$

This implies that we can take $\bar{\rho}$ as in the lemma, which concludes the proof.

Proof of Lemma 14. Study of $|S_{\tilde{q}}(\lfloor\log_2(n)\rfloor+i)\cap\mathcal{A}|$ and $|S_{\tilde{p}}(\lfloor\log_2(n)\rfloor+i)\cap\mathcal{A}|$. For any i such that $|S_\pi(\lfloor\log_2(n)\rfloor+i)|>a\sqrt{\log((|i|+1)/\gamma)}$, we have $|S_{\tilde{q}}(\lfloor\log_2(n)\rfloor+i)\cap\mathcal{A}|\sim\mathrm{Bin}(|\mathcal{A}|,\lfloor|S_\pi(\lfloor\log_2(n)\rfloor+i)|/2\rfloor/|\mathcal{A}|)$, by definition of the distribution of $\tilde{q}$ on $\mathcal{A}$. By Hoeffding's inequality, we have for any $\varepsilon>0$, that with probability larger than $1-2\exp(-2\varepsilon^2 n)$ with respect to the distribution of $\tilde{q}$

$$|S_{\tilde{q}}(\lfloor\log_2(n)\rfloor+i)\cap\mathcal{A}|-\lfloor|S_\pi(\lfloor\log_2(n)\rfloor+i)\cap\mathcal{A}|/2\rfloor|\leq\varepsilon n.$$

So with probability larger than $1-2\exp(-2|S_\pi(\lfloor\log_2(n)\rfloor+i)|^2/16)$ according to the distribution of $\tilde{q}$

$$||S_{\tilde{q}}(\lfloor\log_2(n)\rfloor+i)\cap\mathcal{A}|-\lfloor|S_\pi(\lfloor\log_2(n)\rfloor+i)\cap\mathcal{A}|/2\rfloor|\leq|S_\pi(\lfloor\log_2(n)\rfloor+i)\cap\mathcal{A}|/4.$$

So for any i such that $|S_\pi(\lfloor\log_2(n)\rfloor+i)|>a\sqrt{\log((|i|+1)/\gamma)}$, we have with probability larger than $1-2\exp(-a^2\log((|i|+1)/\gamma))$ with respect to the distribution of $\tilde{q}$ that

$$||S_{\tilde{q}}(\lfloor\log_2(n)\rfloor+i)\cap\mathcal{A}|-\lfloor|S_\pi(\lfloor\log_2(n)\rfloor+i)\cap\mathcal{A}|/2\rfloor|\leq|S_\pi(\lfloor\log_2(n)\rfloor+i)\cap\mathcal{A}|/4.$$

So whenever $a>1$, there exists a constant $c_a>0$ that depends only on a and such that we have with probability larger than $1-2\sum_{i\in\mathbb{Z}}\exp(-a^2\log((|i|+1)/\gamma))\geq 1-c_a\gamma$ with respect to the distribution of $\tilde{p}$ that for all i such that $|S_\pi(\lfloor\log_2(n)\rfloor+i)|>a\sqrt{\log((|i|+1)/\gamma)}$ at the same time,

$$\frac{3}{4}|S_\pi(\lfloor\log_2(n)\rfloor+i)\cap\mathcal{A}|\leq|S_{\tilde{q}}(\lfloor\log_2(n)\rfloor+i)\cap\mathcal{A}|\leq\frac{5}{4}|S_\pi(\lfloor\log_2(n)\rfloor+i)\cap\mathcal{A}|. \quad (2.32)$$

Now since $\varepsilon_i^*\in[0,1/2]$ we know that for any $i\in\mathcal{A}$ we have

$$\frac{1}{2}\tilde{q}_i\leq\tilde{p}_i\leq\frac{3}{2}\tilde{q}_i.$$

And so we also know that with probability larger than $1-c_a\gamma$ with respect to the distribution of $\tilde{p}$ for all $i\in\mathbb{Z}$ such that $|S_\pi(\lfloor\log_2(n)\rfloor+i)|>a\sqrt{\log((|i|+1)/\gamma)}$ at the same time,

$$\frac{3}{4}|S_\pi(\lfloor\log_2(n)\rfloor+i)\cap\mathcal{A}|\leq\sum_{j=-1}^{j+1}|S_{\tilde{q}}(\lfloor\log_2(n)\rfloor+j)\cap\mathcal{A}|\leq\frac{5}{4}\sum_{j=-2}^{j+2}|S_\pi(\lfloor\log_2(n)\rfloor+i)\cap\mathcal{A}|. \quad (2.33)$$

Study of the rescaled coefficients outside $\mathcal{A}$. We define the following events

$$H_p = \{|\sum_{l\in\mathcal{A}}(p_l-\pi_l)| \leq 16M^{-1}\log(2/\delta)\}, \qquad H_q = \{|\sum_{l\in\mathcal{A}}(q_l-\pi_l)| \leq 16M^{-1}\log(2/\delta)\}. \tag{2.34}$$

We remind that for $i \in \mathcal{A}$ we have $\tilde{p}_i = p_i$ and $\tilde{q}_i = q_i$. So the events H_p and H_q are very informative with respect to $\tilde{p}, \tilde{q}$. Now, $\max_{j\in\mathcal{A}}\pi_j \leq 1$ and since $|S_{\tilde{q}}(i)\cap\mathcal{A}|/|S_{\tilde{q}}(i)| \leq 1/M$ for any $i \in \mathbb{Z}$ and $\sqrt{n} \geq M$. So by Bernstein's inequality, we have with probability larger than $1-\delta$ with respect to p and $\tilde{p}$ that H_p holds and with probability larger than $1-\delta$ with respect to q and $\tilde{q}$ that H_q holds.

So if $\sqrt{n} \geq M \geq 4(16\log(2/\delta))^2$, we have with probability larger than $1-\delta$ with respect to $\tilde{q}$ that for any $j \notin \mathcal{A}$,

$$\frac{1}{2}\pi_j \leq \tilde{q}_j \leq \frac{3}{2}\pi_j,$$

and also with probability larger than $1-\delta$ with respect to $\tilde{p}$ that for any $j \notin \mathcal{A}$,

$$\frac{1}{2}\pi_j \leq \tilde{p}_j \leq \frac{3}{2}\pi_j.$$

And so finally we have with probability larger than $1-\delta$ with respect to $\tilde{q}$ that for any i,

$$|S_\pi(i)\cap\mathcal{A}^C| \leq \sum_{j=i-1}^{i+1}|S_{\tilde{q}}(j)\cap\mathcal{A}^C| \leq \sum_{j=i-2}^{i+2}|S_\pi(i)\cap\mathcal{A}^C|.$$

Similarly we have with probability larger than $1-\delta$ with respect to $\tilde{p}$ that for any i,

$$|S_\pi(i)\cap\mathcal{A}^C| \leq \sum_{j=i-1}^{i+1}|S_{\tilde{p}}(j)\cap\mathcal{A}^C| \leq \sum_{j=i-2}^{i+2}|S_\pi(i)\cap\mathcal{A}^C|.$$

Conclusion. Combining both studies on $\mathcal{A}$ and $\mathcal{A}^C$, we get that if $M \geq 4(16\log 2/\delta)^2$, we have with probability larger than $1-\delta-c_a\gamma$ with respect to $\tilde{q}$ that $\tilde{q} \in \mathbf{P}_\pi$, and with probability larger than $1-\delta-c_a\gamma$ with respect to $\tilde{p}$ that $\tilde{p} \in \mathbf{P}_\pi$.

□

Proof of Lemma 16. Define the discrete uniform distribution $\mathcal{U}_\lambda$ such that $\mathcal{U}_\lambda(\{\lambda_i\}) = 1/d$. We will now work on the definition of appropriate measures for (p,q). Let $\theta \sim \mathcal{U}_\lambda$ and ξ taking value ε_i^* when θ takes value λ_i. We reparametrize ξ by λ, and we set

$$\xi_\theta = \frac{1}{\Big|\{i : n\pi_i = \theta\}\Big|}\sum_{\{i:n\pi_i=\theta\}}\varepsilon_i^*,$$

with the convention $0/0 = 0$. Note that by definition of ε^*, we have from Lemma 15

- $\xi_\theta \in [0,1]$ and $\xi_\theta = 0$ for any $\theta \geq 1$.
- $\xi_\theta\theta \leq \sqrt{u}\Big[n\sqrt{C/I_{v,\pi}} \wedge 1\Big]$.

- By definition of $\mathcal{U}_\lambda$ and Lemma 15

$$\int \theta^2 \xi_\theta^2 e^{-2\theta} d\mathcal{U}_\lambda(\theta) = \frac{n^2}{d} \sum_i \pi_i^2 \varepsilon_i^{*2} e^{-2n\pi_i} \leq \frac{n^2}{d} u \frac{\sqrt{\sum_i \pi_i^2 e^{-2(1+v)n\pi_i}}}{n}$$
$$= u \sqrt{\frac{\int \theta^2 e^{-2(1+v)\theta} d\mathcal{U}_\lambda(\theta)}{d}}. \quad (2.35)$$

Bound on the total variation. Let us dominate the total variation distance with the chi-squared distance χ^2.

For two distributions $\tilde{\nu}_1, \tilde{\nu}_0$ such that $\tilde{\nu}_1$ is absolutely continuous with respect to $\tilde{\nu}_0$, then

$$d_{TV}(\tilde{\nu}_0, \tilde{\nu}_1) = \int \left| \frac{d\tilde{\nu}_1}{d\tilde{\nu}_0} - 1 \right| d\tilde{\nu}_0 = \mathbb{E}_{\tilde{\nu}_0} \left[\left| \frac{d\tilde{\nu}_1}{d\tilde{\nu}_0} - 1 \right| \right]$$
$$\leq \left(\mathbb{E}_{\tilde{\nu}_0} \left[\left(\frac{d\tilde{\nu}_1}{d\tilde{\nu}_0} \right)^2 \right] - 1 \right)^{1/2} = \sqrt{\chi^2(\tilde{\nu}_0, \tilde{\nu}_1)}.$$

By the tensorization property of the chi-squared distance and by application of the inequality above to $\nu_0^{\otimes d}, \nu_1^{\otimes d}$, we have

$$d_{TV}(\nu_0^{\otimes d}, \nu_1^{\otimes d}) \leq \sqrt{\chi^2(\nu_0^{\otimes d}, \nu_1^{\otimes d})} = \sqrt{(1 + \chi^2(\nu_0, \nu_1))^d - 1}. \quad (2.36)$$

Now, we have by the law of total probability for any $m, m' \geq 0$

$$\nu_0(m, m') = \int \frac{e^{-2\theta} \theta^{m+m'}}{m! m'!} d\mathcal{U}_\lambda(\theta),$$

and

$$\nu_1(m, m') = \int \frac{1}{2} \frac{e^{-2\theta} \theta^{m+m'}}{m! m'!} (e^{\xi_\theta \theta} (1 - \xi_\theta)^{m'} + e^{-\xi_\theta \theta} (1 + \xi_\theta)^{m'}) d\mathcal{U}_\lambda(\theta).$$

So

$$\chi^2(\nu_0, \nu_1)$$
$$= \sum_{m,m'} \frac{(\int \theta^{m+m'} e^{-2\theta} [-e^{\xi_\theta \theta} (1 - \xi_\theta)^{m'}/2 - e^{-\xi_\theta \theta} (1 + \xi_\theta)^{m'}/2 + 1] d\mathcal{U}_\lambda(\theta))^2}{m! m'! \int \theta^{m+m'} e^{-2\theta} d\mathcal{U}_\lambda(\theta)}$$
$$= \sum_{m,m'} \frac{\int \int (\theta\theta')^{m+m'} e^{-2(\theta+\theta')} D_\theta(m) D_{\theta'}(m) d\mathcal{U}_\lambda(\theta) d\mathcal{U}_\lambda(\theta')}{m! m'! \int \theta^{m+m'} e^{-2\theta} d\mathcal{U}_\lambda(\theta)}, \quad (2.37)$$

where

$$D_\theta(m) = -\frac{e^{\xi_\theta \theta} (1 - \xi_\theta)^m}{2} - \frac{e^{-\xi_\theta \theta} (1 + \xi_\theta)^m}{2} + 1.$$

We will analyse the terms of this sum depending on the value of $m + m'$.

Analysis of the terms in Equation (2.37). Term for $m + m' = 0$. We have

$$D_\theta(0) = -\cosh(\xi_\theta \theta) + 1 \qquad \text{and} \qquad D_\theta(0) D_{\theta'}(0) \leq (\theta\theta' \xi_\theta \xi_{\theta'})^2,$$

since $\xi_\theta\theta \leq 1$.

And so

$$\frac{\int\int e^{-2(\theta+\theta')}D_\theta(0)D_{\theta'}(0)d\mathcal{U}_\lambda(\theta)d\mathcal{U}_\lambda(\theta')}{\int e^{-2\theta}d\mathcal{U}_\lambda(\theta)} \leq \frac{\left(\int e^{-2\theta}(\theta\xi_\theta)^2 d\mathcal{U}_\lambda(\theta)\right)^2}{\int e^{-2\theta}d\mathcal{U}_\lambda(\theta)} \leq u\frac{\int \theta^2 e^{-2(1+v)\theta}d\mathcal{U}_\lambda(\theta)}{d\int e^{-2\theta}d\mathcal{U}_\lambda(\theta)} \leq \frac{c_v u}{d},$$

where we obtained the second inequality by Equation (2.35) and for $c_v < +\infty$ that depends only on $v > 0$ and such that

$$c_v = \sup_{\theta>0}\left[e^{-2v\theta}(1 \vee \theta^2)\right]. \tag{2.38}$$

Term for $m + m' = 1$. We have then

$$D_\theta(1) = -\cosh(\xi_\theta\theta) + 1 + \xi_\theta \sinh(\theta\xi_\theta)$$

and so since $\xi_\theta \in [0, 1]$ and $\theta\xi_\theta \in [0, \xi_\theta]$, we have

$$D_\theta(1)D_{\theta'}(1) \leq \theta\theta'(\xi_\theta\xi_{\theta'})^2.$$

So the term for $m + m' = 1$ can be bounded as

$$\begin{aligned}\frac{\int\int \theta\theta' e^{-2(\theta+\theta')}(D_\theta(0)D_{\theta'}(0) + D_\theta(1)D_{\theta'}(1))d\mathcal{U}_\lambda(\theta)d\mathcal{U}_\lambda(\theta')}{\int \theta e^{-2\theta}d\mathcal{U}_\lambda(\theta)} &\leq \frac{1}{\int \theta e^{-2\theta}d\mathcal{U}_\lambda(\theta)}\Big(\int e^{-2\theta}2(\theta\xi_\theta)^2 d\mathcal{U}_\lambda(\theta)\Big)^2 \\ &\leq 4u\frac{\int \theta^2 e^{-2(1+v)\theta}d\mathcal{U}_\lambda(\theta)}{d\int \theta e^{-2\theta}d\mathcal{U}_\lambda(\theta)} \leq 4c_v\frac{u}{d},\end{aligned}$$

where we obtained the second inequality by Equation (2.35) and the last by Definition of c_v in Equation (2.38).

Term for $m + m' = 2$. We have then

$$D_\theta(2) = -\cosh(\xi_\theta\theta) + 1 + 2\xi_\theta \sinh(\theta\xi_\theta) - \xi_\theta^2 \cosh(\theta\xi_\theta)$$

and again since $\xi_\theta \in [0, 1]$ and $\theta\xi_\theta \in [0, \xi_\theta]$, we have

$$D_\theta(2)D_{\theta'}(2) = 4(\xi_\theta\xi_{\theta'})^2.$$

So the term for $m + m' = 2$ can be bounded as

$$\begin{aligned}\frac{\int\int (\theta\theta')^2 e^{-2(\theta+\theta')}(D_\theta(0)D_{\theta'}(0)/2 + D_\theta(1)D_{\theta'}(1) + D_\theta(2)D_{\theta'}(2)/2)d\mathcal{U}_\lambda(\theta)d\mathcal{U}_\lambda(\theta')}{\int \theta^2 e^{-2\theta}d\mathcal{U}_\lambda(\theta)} &\leq \frac{1}{\int \theta^2 e^{-2\theta}d\mathcal{U}_\lambda(\theta)}\Big(\int e^{-2\theta}4(\theta\xi_\theta)^2 d\mathcal{U}_\lambda(\theta)\Big)^2 \\ &\leq \frac{16u}{d}c_v,\end{aligned}$$

where we obtain the second inequality by Equation (2.35).

Term for $m+m' \geq 3$. We have

$$D_\theta(m) \leq 2^{m+2}\xi_\theta^2. \tag{2.39}$$

Subcase 1: $m+m'=3$. We have by Equation (2.39)

$$\begin{aligned}\frac{\int\int(\theta\theta')^3 e^{-2\theta} D_\theta(m) D_{\theta'}(m) d\mathcal{U}_\lambda(\theta) d\mathcal{U}_\lambda(\theta')}{m!m'! \int \theta^3 e^{-2\theta} d\mathcal{U}_\lambda(\theta)} &\leq \frac{\left(\int e^{-2\theta}\theta^3 (2^{m+2}\xi_\theta^2) d\mathcal{U}_\lambda(\theta)\right)^2}{m!m'! \int \theta^3 e^{-2\theta} d\mathcal{U}_\lambda(\theta)} \\ &\leq 2^{2m+4} \frac{\left(\int e^{-2\theta}\theta(\theta\xi_\theta)^2 d\mathcal{U}_\lambda(\theta)\right)^2}{\int \theta^3 e^{-2\theta} d\mathcal{U}_\lambda(\theta)} \\ &= 2^{2m+4} \frac{n^3\left[\sum_{i\geq J} \pi_i (\pi_i \varepsilon_i^*)^2\right]^2}{d\sum_i \pi_i^3 e^{-2\pi_i n}}.\end{aligned}$$

Finally, by definition of ξ_θ and ε_i^* and since in any case $\varepsilon_i^*\pi_i \leq \sqrt{uC_\pi/I_{v,\pi}}$ (see Lemma 15), we have

$$\frac{\int\int(\theta\theta')^3 e^{-2\theta} D_\theta(m) D_{\theta'}(m) d\mathcal{U}_\lambda(\theta) d\mathcal{U}_\lambda(\theta')}{m!m'! \int \theta^3 e^{-2\theta} d\mathcal{U}_\lambda(\theta)} \leq 2^{2m+4} \frac{n^3\left[\sum_{i\geq J} \pi_i \frac{uC_\pi}{2I_{v,\pi}}\right]^2}{d\sum_i \pi_i^3 e^{-2\pi_i n}}.$$

This implies, since $\sum_{J\leq i<I_{v,\pi}} \pi_i \geq \sum_{I_{v,\pi}\leq i} \pi_i$ in the definition of $I_{v,\pi}$ (see Lemma 15),

$$\begin{aligned}\frac{\int\int(\theta\theta')^3 e^{-2\theta} D_\theta(m) D_{\theta'}(m) d\mathcal{U}_\lambda(\theta) d\mathcal{U}_\lambda(\theta')}{m!m'! \int \theta^3 e^{-2\theta} d\mathcal{U}_\lambda(\theta)} &\leq 2^{2m+2} \frac{u^2 C_\pi^2}{I_{v,\pi}^2} \frac{n^3\left[\sum_{J\leq i<I_{v,\pi}} \pi_i\right]^2}{n\sum_i \pi_i^3 e^{-2\pi_i n}} \\ &\leq 2^{2m+2} \frac{u^2 C_\pi^2}{I_{v,\pi}} \frac{n^3\left[\sum_{J\leq i<I_{v,\pi}} \pi_i^2\right]}{n\sum_i \pi_i^3 e^{-2\pi_i n}},\end{aligned}$$

by Cauchy-Schwarz inequality. Then

$$\begin{aligned}&\frac{\int\int(\theta\theta')^3 e^{-2\theta} D_\theta(m) D_{\theta'}(m) d\mathcal{U}_\lambda(\theta) d\mathcal{U}_\lambda(\theta')}{m!m'! \int \theta^3 e^{-2\theta} d\mathcal{U}_\lambda(\theta)} \\ &\qquad \leq 2^{2m+2} \frac{u^2}{I_{v,\pi}} \frac{n\left[\sum_i \pi_i^2 \exp(-2(1+v)n\pi_i)\right]\left[\sum_{J\leq i<I_{v,\pi}} \pi_i^2\right]}{n\sum_i \pi_i^3 e^{-2\pi_i n}},\end{aligned}$$

by Definition of C_π in Equation (2.30). In particular,

$$\begin{aligned}\sum_i \pi_i^2 \exp(-2(1+v)n\pi_i) &= \sum_{i<I_{v,\pi}} \pi_i^2 \exp(-2(1+v)n\pi_i) + \sum_{I_{v,\pi}\leq i} \pi_i^2 \exp(-2(1+v)n\pi_i) \\ &\leq 2e^{2(1+v)} \sum_{i<I_{v,\pi}} \pi_i^2 \exp(-2(1+v)n\pi_i),\end{aligned}$$

since $\sum_{I_{v,\pi}\le i}\pi_i \le \sum_{J\le i<I_{v,\pi}}\pi_i$ and for all $i \ge J$ we have $n\pi_i \le 1$. So, once we plug the last inequality in, we obtain:

$$\begin{aligned}
&\frac{\int\int(\theta\theta')^3 e^{-2\theta} D_\theta(m) D_{\theta'}(m) d\mathcal{U}_\lambda(\theta) d\mathcal{U}_\lambda(\theta')}{m!m'!\int\theta^3 e^{-2\theta} d\mathcal{U}_\lambda(\theta)} \\
&\qquad \le 2^{2m+3} e^{2(1+v)} \frac{u^2}{I_{v,\pi}} \frac{n\Big[\sum_{i<I_{v,\pi}} \pi_i^2 \exp(-2(1+v)n\pi_i)\Big]\Big[\sum_{J\le i<I_{v,\pi}}\pi_i^2\Big]}{d\sum_i \pi_i^3 e^{-2\pi_i n}} \\
&\qquad \le 2^{2m+3} e^{2(1+v)} \frac{u^2}{I_{v,\pi}} \frac{n\Big[\sum_{i<I_{v,\pi}} \pi_i^2 \exp(-2(1+v)n\pi_i)\Big]\Big[\sum_{J\le i<I_{v,\pi}}\pi_i^2\Big]}{d\sum_{i\le I_{v,\pi}} \pi_i^3 e^{-2\pi_i n}} \\
&\qquad \le 2^{2m+3} e^{2(1+v)} u^2 \frac{n\Big[\sum_{i\le I_{v,\pi}} \pi_i^4 \exp(-2(1+v)n\pi_i)\Big]}{d\sum_i \pi_i^3 e^{-2\pi_i n}},
\end{aligned}$$

because for any $a_1 \ge \ldots \ge a_{I_{v,\pi}} \ge 0$, $b_1 \ge \ldots \ge b_{I_{v,\pi}} \ge 0$, we have $\sum a_i \sum b_j \le I_{v,\pi}\sum a_i b_i$. Then $n\pi_i e^{-2vn\pi_i} \le c_v$ by Equation (2.38) and for any i, this implies

$$\frac{\int\int(\theta\theta')^3 e^{-2\theta} D_\theta(m) D_{\theta'}(m) d\mathcal{U}_\lambda(\theta) d\mathcal{U}_\lambda(\theta')}{m!m'!\int\theta^3 e^{-2\theta} d\mathcal{U}_\lambda(\theta)} \le 2^{2m+3} e^{2(1+v)} c_v \frac{u^2}{d} \le 2^9 e^{2(1+v)} c_v \frac{u^2}{d}.$$

Subcase 2: $m + m' \ge 4$. We have by Equation (2.39)

$$\begin{aligned}
&\frac{\int\int(\theta\theta')^{m+m'} e^{-2(\theta+\theta')} D_\theta(m) D_{\theta'}(m) d\mathcal{U}_\lambda(\theta) d\mathcal{U}_\lambda(\theta')}{m!m'!\int\theta^{m+m'} e^{-2\theta} d\mathcal{U}_\lambda(\theta)} \\
&\qquad \le \frac{\Big(\int e^{-2\theta}\theta^{m+m'}(2^{m+2}\xi_\theta^2) d\mathcal{U}_\lambda(\theta)\Big)^2}{m!m'!\int\theta^{m+m'} e^{-2\theta} d\mathcal{U}_\lambda(\theta)} \\
&\qquad = \frac{2^{2m+4}}{m!m'!} \frac{\Big(\int e^{-2\theta}\theta^{m+m'}\xi_\theta^2 d\mathcal{U}_\lambda(\theta)\Big)^2}{\int\theta^{m+m'} e^{-2\theta} d\mathcal{U}_\lambda(\theta)} \\
&\qquad \le \frac{2^{2m+4}}{m!m'!} \int e^{-2\theta}\theta^{m+m'}\xi_\theta^4 d\mathcal{U}_\lambda(\theta),
\end{aligned}$$

where the last inequality comes by application of Cauchy-Schwarz inequality. And so since $\varepsilon_\theta = 0$ for any $\theta \ge 1$, we have

$$\begin{aligned}
&\frac{\int\int(\theta\theta')^{m+m'} e^{-2(\theta+\theta')} D_\theta(m) D_{\theta'}(m) d\mathcal{U}_\lambda(\theta) d\mathcal{U}_\lambda(\theta')}{m!m'!\int\theta^{m+m'} e^{-2\theta} d\mathcal{U}_\lambda(\theta)} \\
&\qquad \le \frac{2^{2m+4}}{m!m'!} \int e^{-2\theta}\theta^{m+m'-4}(\theta\xi_\theta)^4 \mathbf{1}\{\theta \le 1\} d\mathcal{U}_\lambda(\theta).
\end{aligned}$$

Then, since $m + m' \geq 4$,

$$\begin{aligned}\frac{\int\int(\theta\theta')^{m+m'}e^{-2(\theta+\theta')}D_\theta(m)D_{\theta'}(m)d\mathcal{U}_\lambda(\theta)d\mathcal{U}_\lambda(\theta')}{m!m'!\int\theta^{m+m'}e^{-2\theta}d\mathcal{U}_\lambda(\theta)} &\leq \frac{2^{2m+4}}{m!m'!}\int(\theta\xi_\theta)^4\mathbf{1}\{\theta\leq 1\}d\mathcal{U}_\lambda(\theta)\\ &= \frac{2^{2m+4}}{m!m'!}\frac{n^4}{d}\Big[\sum_i(\pi_i\varepsilon_i^*)^4\Big]\\ &\leq \frac{2^{2m+4}e^2}{m!m'!}\frac{n^4}{d}\frac{u^2C_\pi^2}{I_{v,\pi}},\end{aligned}$$

since, by definition of ε_i^* in Lemma 15, $\pi_i\varepsilon_i^* \leq \sqrt{uC_\pi/I_{v,\pi}}$ and $\sum_i(\pi_i\varepsilon_i^*)^2 \leq uC_\pi e^2$ using the fact that $\varepsilon_i^* = 0$ for $\pi_i \geq 1/n$. By Equation (2.40), this implies

$$\frac{\int\int(\theta\theta')^{m+m'}e^{-2(\theta+\theta')}D_\theta(m)D_{\theta'}(m)d\mathcal{U}_\lambda(\theta)d\mathcal{U}_\lambda(\theta')}{m!m'!\int\theta^{m+m'}e^{-2\theta}d\mathcal{U}_\lambda(\theta)} \leq \frac{2^{2m+4}e^2}{m!m'!}\frac{2u^2}{d}.$$

Conclusion on the distance between the two distributions. Now, we plug the bounds we found for each term back in Equation (2.37) and we obtain

$$\begin{aligned}\chi^2(\nu_0,\nu_1) &\leq \sum_{m,m'}\frac{\int\int(\theta\theta')^{m+m'}e^{-2(\theta+\theta')}D_\theta(m)D_{\theta'}(m)d\mathcal{U}_\lambda(\theta)d\mathcal{U}_\lambda(\theta')}{m!m'!\int\theta^{m+m'}e^{-2\theta}d\mathcal{U}_\lambda(\theta)}\\ &\leq \frac{c_v u}{d} + 4\frac{c_v u}{d} + \frac{16u}{d}c_v + 2^9e^{2(1+v)}\frac{u}{d}c_v + \sum_{m,m':m+m'\geq 4}\frac{2^{2m+4}e^2}{m!m'!}\frac{2u^2}{d}\\ &\leq (5+16+2^9e^{2(1+v)})\frac{u}{d}c_v + \sum_{m,m'}\frac{2^{2m+4}e^2}{m!m'!}\frac{2u}{d}\\ &= (5+16+2^9e^{2(1+v)})\frac{u}{d}c_v + \frac{2^5e^7u}{d}\\ &\leq (5c_v + 16c_v + 2^9e^{2(1+v)}c_v + 2^5e^7)\frac{u}{d} \leq \tilde{c}_v\frac{u}{d},\end{aligned}$$

for $u \leq 1$ and where $\tilde{c}_v$ is a constant that depends only on v and $\tilde{c}_v$ is bounded away from 0 for $v > 0$. And so by Equation (2.36) we have

$$d_{TV}(\nu_0^{\otimes d},\nu_1^{\otimes d}) \leq \sqrt{\Big(1+\tilde{c}_v\frac{u}{d}\Big)^d - 1} \leq \sqrt{\exp(\tilde{c}_v u) - 1} \leq \sqrt{\tilde{c}_v u},$$

for $u \leq \tilde{c}_v^{-1}$, i.e. for u smaller than a constant that depends only on v. □

Proof of Lemma 17. Note that

$$\mathbb{E}_{\tilde{q},\tilde{p}}\sum_i|\tilde{p}_i - \tilde{q}_i| = \mathbb{E}_{q,\xi}\Big[\sum_{i\in\mathcal{A}}|\xi_i|q_i\Big] = \sum_{i\in\mathcal{A}}\varepsilon_i^*\pi_i,$$

and

$$\mathbb{V}_q\Big[\sum_{i\in\mathcal{A}}|\xi_i|q_i\Big] = \sum_{i\in\mathcal{A}}\mathbb{V}_{q_1\sim U_{\pi,\mathcal{A}}}(|\xi_1|q_1) \leq \sum_{i\in\mathcal{A}}(\varepsilon_i^*\pi_i)^2.$$

Let $\alpha > 0$. By Chebyshev's inequality, we know that with probability larger than $1-\alpha$ with respect to (q, ξ),

$$\sum_{i\in\mathcal{A}} |\xi_i| q_i \geq \sum_{i\in\mathcal{A}} \varepsilon_i^* \pi_i - \sqrt{\frac{\sum_{i\in\mathcal{A}} \varepsilon_i^{*2} \pi_i^2}{\alpha}}.$$

Now, by definition of $\mathcal{A}$ and $(\varepsilon_i^*)_i$ we have

$$\sum_{i\in\mathcal{A}} \varepsilon_i^* \pi_i \geq \frac{1}{8M^2} \sum_i \bar{\varepsilon}_i^* \pi_i - \sum_{i\in\mathbb{N}:|S_\pi(\lfloor \log_2(n)\rfloor + i)| \leq a\sqrt{\log(i/\gamma)}} \sum_{j\in S_\pi(\lfloor \log_2(n)\rfloor + i)} \pi_j.$$

First note that

$$\sum_{i\in\mathbb{N}:|S_\pi(\lfloor \log_2(n)\rfloor + i)| \leq a\sqrt{\log(i/\gamma)}} \sum_{j\in S_\pi(\lfloor \log_2(n)\rfloor + i)} \pi_j \leq \sum_{i\in\mathbb{N}} a\sqrt{\log(i/\gamma)} \frac{2^{-i}}{n} \leq 8\frac{a(1+\log(1/\gamma))}{n}.$$

Also by application of Lemma 15, where we associate $\bar{\varepsilon}_i^*$ to the ordered version of π, we have

$$\sum_i \bar{\varepsilon}_i^* \pi_i \geq \left[\left[\sum_{i\geq I_{v,\pi}} \frac{\sqrt{u}\pi_i}{\sqrt{2}} \right] \vee \frac{\sqrt{u}(I_{v,\pi} - J)}{\sqrt{2I_{v,\pi}}} \sqrt{\frac{\sqrt{\sum_i \pi_i^2 \exp(-2(1+v)n\pi_i)}}{n}} \right] \wedge \sqrt{\frac{u}{8}} \right].$$

So we conclude that

$$\begin{aligned}
&\sum_{i\in\mathcal{A}} \varepsilon_i^* \pi_i \\
&\geq \frac{1}{8M^2} \left[\left[\sum_{i\geq I_{v,\pi}} \frac{\sqrt{u}\pi_i}{\sqrt{2}} \right] \vee \frac{\sqrt{u}(I_{v,\pi} - J)}{\sqrt{2I_{v,\pi}}} \sqrt{\frac{\sqrt{\sum_i \pi_i^2 \exp(-2(1+v)n\pi_i)}}{n}} \right] \wedge \sqrt{\frac{u}{8}} \right] \\
&- 8\frac{a(1+\log(1/\gamma))}{n}.
\end{aligned}$$

So we have with probability larger than $1-\delta$,

$$\begin{aligned}
&\sum_{i\in\mathcal{A}} |\xi_i| q_i \\
&\geq \frac{1}{8M^2} \left[\left[\sum_{i\geq I_{v,\pi}} \frac{\sqrt{u}\pi_i}{\sqrt{2}} \right] \vee \frac{\sqrt{u}(I_{v,\pi} - J)}{\sqrt{2I_{v,\pi}}} \sqrt{\frac{\sqrt{\sum_i \pi_i^2 \exp(-2(1+v)n\pi_i)}}{n}} \right] \wedge \sqrt{\frac{u}{8}} \right] \\
&- \frac{1}{\sqrt{nM\delta}} - 8\frac{a(1+\log(1/\gamma))}{n}.
\end{aligned}$$

□

Proof of Lemma 15. We prove this lemma by defining suitable ε_i^*'s.

Step 1: Proof that $\sqrt{C_\pi/I_{v,\pi}} \leq \frac{\sqrt{2}}{n}$. We have

$$\begin{aligned} C_\pi^2 n^2 &\leq \sum_i \pi_i^2 \exp(-2(1+v)n\pi_i) \\ &= \sum_{i \leq I_{v,\pi}} \pi_i^2 \exp(-2(1+v)n\pi_i) + \sum_{i \geq I_{v,\pi}} \pi_i^2 \exp(-2(1+v)n\pi_i) \\ &\leq \frac{I_{v,\pi}}{n^2} + uC_\pi, \end{aligned}$$

as $\pi_i^2 \exp(-2n\pi_i) \leq \frac{1}{n^2}$. So we have that $C_\pi \leq \frac{2u}{n^2}$, or $C_\pi \leq \frac{\sqrt{2I_{v,\pi}}}{n^2}$, and so in any case

$$C_\pi \leq \frac{\sqrt{2I_{v,\pi}}}{n^2} \vee \frac{2u}{n^2},$$

which implies

$$\frac{\sqrt{C_\pi}}{I_{v,\pi}^{1/4}} \leq \frac{2^{1/4}}{n} \vee \frac{\sqrt{2u}}{nI_{v,\pi}^{1/4}} \leq \frac{\sqrt{2}}{n}, \tag{2.40}$$

since $0 < u < 1$.

Step 2: Definition of ε_i^* for $i \geq I_{v,\pi}$ or $i < J$. Take for all $i < J$ that $\varepsilon_i^* = 0$. Take for all other $i \geq I_{v,\pi}$

$$\varepsilon_i^* = \sqrt{u/2}.$$

We have for any $i \geq I_{v,\pi}$

- $\varepsilon_i^* \in [0,1]$, and $\varepsilon_i^* \pi_i \leq \sqrt{u}\Big[(1/n) \wedge \sqrt{C_\pi/(2I_{v,\pi})}\Big]$, since by definition of $I_{v,\pi}$ we know that $\pi_i \leq (1/n) \wedge \sqrt{C_\pi/I_{v,\pi}}$ if $i \geq I_{v,\pi}$
- by definition of $I_{v,\pi}$ we have

$$\sum_{i \geq I_{v,\pi}} \pi_i^2 \varepsilon_i^{*2} \exp(-2n\pi_i) \leq \frac{uC_\pi}{2}$$

- and also

$$\sum_{i \geq I_{v,\pi}} \varepsilon_i^* \pi_i = \sqrt{\frac{u}{2}} \sum_{i \geq I_{v,\pi}} \pi_i. \tag{2.41}$$

Step 3: Definition of ε_i^* for $i < I_{v,\pi}$ in three different cases. If $I_{v,\pi} \leq J$, the ε_i^* are already defined for all $i \geq J$, and by definition of ε_i^*,

$$\sum_i \varepsilon_i^* \pi_i \geq \sum_{i \geq J} \frac{\sqrt{u}\pi_i}{\sqrt{2}} = \Big[\sum_{i \geq J} \frac{\sqrt{u}\pi_i}{\sqrt{2}}\Big] \vee \Big[\frac{(I_{v,\pi} - J)\sqrt{uC_\pi}}{\sqrt{2I_{v,\pi}}}\Big]$$

This concludes the proof in that case. We assume from now on that $I_{v,\pi} > J$, then by definition of $I_{v,\pi}$, at least one of the constraints in Equation (2.29) must be saturated.

Case 1: third constraint saturated but not the first one: $\sum_{i \geq I_{v,\pi}-1} \pi_i > \sum_{J \leq i < I_{v,\pi}-1} \pi_i$ and $\pi_{I_{v,\pi}-1} \leq \sqrt{C_\pi/I_{v,\pi}} \wedge (1/n)$. We set $\varepsilon_{I_{v,\pi}-1}^* = \sqrt{\frac{u}{2}}$ and for any $i < I_{v,\pi} - 1$, we set $\varepsilon_i^* = 0$. Note that $\varepsilon_{I_{v,\pi}-1}^* \leq 1$ and $\varepsilon_{I_{v,\pi}-1}^* \pi_{I_{v,\pi}-1} \leq \sqrt{u}\Big[(1/n) \wedge$

$\sqrt{C_\pi/I_{v,\pi}}\Big]$. We also have by definition of ε_i^* for $i \geq I_{v,\pi}$ and by Equation (2.29)

$$\sum_{i \geq I_{v,\pi}} \pi_i^2 \varepsilon_i^{*2} \exp(-2n\pi_i) \leq \frac{uC_\pi}{2},$$

and so

$$\sum_i \pi_i^2 \varepsilon_i^{*2} \exp(-2n\pi_i) = \sum_{i \geq I_{v,\pi}-1} \pi_i^2 \varepsilon_i^{*2} \exp(-2n\pi_i) \leq uC_\pi.$$

Moreover by saturation of the third constraint

$$\sum_{i \geq I_{v,\pi}-1} \pi_i \varepsilon_i^* = \sqrt{\frac{u}{2}} \sum_{i \geq I_{v,\pi}-1} \pi_i \geq \sqrt{\frac{u}{2}} \sum_{J \leq i < I_{v,\pi}-1} \pi_i,$$

and so

$$\sum_i \pi_i \varepsilon_i^* = \sum_{i \geq I_{v,\pi}-1} \pi_i \varepsilon_i^* \geq \sqrt{\frac{u}{8}} \sum_{J \leq i} \pi_i.$$

This concludes the proof in this case.

Case 2: second constraint saturated but not the first one: $\sum_{i \geq I_{v,\pi}-1} \pi_i^2 \exp(-2n\pi_i) > C_\pi$ **and** $\pi_{I_{v,\pi}-1} \leq \sqrt{C_\pi/I_{v,\pi}}$**.** We have

$$\sum_{i \geq I_{v,\pi}-1} \pi_i^2 \geq \sum_{i \geq I_{v,\pi}-1} \pi_i^2 \exp(-2n\pi_i) \geq C_\pi. \tag{2.42}$$

Moreover by definition of ε_i^* for $i \geq I_{v,\pi}$ and by Equation (2.29) we have

$$\sum_{i \geq I_{v,\pi}} \varepsilon_i^{*2} \pi_i^2 \exp(-2n\pi_i) \leq \frac{uC_\pi}{2}.$$

Set $\varepsilon^*_{I_{v,\pi}-1} = \sqrt{u/2}$ and for all $i < I_{v,\pi} - 1$, we set $\varepsilon_i^* = 0$. Note that $\varepsilon^*_{I_{v,\pi}-1} \leq 1$ and $\varepsilon^*_{I_{v,\pi}-1} \pi_{I_{v,\pi}-1} \leq \sqrt{u}\Big[\sqrt{C_\pi/(2I_{v,\pi})} \wedge (1/n)\Big]$. So from the last displayed equation and the definition of ε_i^*

$$\sum_{i \geq I_{v,\pi}-1} \varepsilon_i^{*2} \pi_i^2 \exp(-2n\pi_i) \leq uC_\pi,$$

and by Equation (2.42)

$$\sum_i \varepsilon_i^{*2} \pi_i^2 = \sum_{i \geq I_{v,\pi}-1} \varepsilon_i^{*2} \pi_i^2 = \frac{u}{2} \sum_{i \geq I_{v,\pi}-1} \pi_i^2 \geq \frac{uC_\pi}{2}.$$

Since for all $i \geq I_{v,\pi} - 1$ we have $\pi_i \leq \pi_{I_{v,\pi}-1} \leq \sqrt{C_\pi/I_{v,\pi}}$ and $\varepsilon_i^* \leq 1$, we have thus

$$\sum_i \varepsilon_i^* \pi_i = \sum_{i \geq I_{v,\pi}-1} \varepsilon_i^* \pi_i \geq \sqrt{\frac{u}{2}} \frac{C_\pi}{\pi_{I_{v,\pi}-1}} \geq \sqrt{\frac{uC_\pi I_{v,\pi}}{2}} \geq \sqrt{\frac{uC_\pi}{2}} \frac{I_{v,\pi} - J}{\sqrt{I_{v,\pi}}}.$$

This concludes the proof in this case with Equation (2.41).

Case 3: first constraint saturated, i.e. $\pi_{I_{v,\pi}-1} > \sqrt{C_\pi/I_{v,\pi}}$**.** We set for any $i < J$, $\varepsilon_i^* = 0$ and for any $J \le i < I_{v,\pi}$,

$$\varepsilon_i^* = \frac{\sqrt{uC_\pi}}{\sqrt{2I_{v,\pi}}\pi_i}.$$

Note that for any i

$$\varepsilon_i^* \in [0,1], \quad \text{and,} \quad \varepsilon_i^*\pi_i \le \frac{\sqrt{uC_\pi}}{\sqrt{2I_{v,\pi}}} \le \sqrt{u}\Big[\sqrt{C_\pi/(2I_{v,\pi})} \wedge (1/n)\Big],$$

by Equation (2.40). Moreover we have

$$\sum_{J\le i<I_{v,\pi}} \varepsilon_i^{*2}\pi_i^2 \exp(-2n\pi_i) \le \frac{uC_\pi}{2},$$

and so by definition of $I_{v,\pi}$ in Equation (2.29) and of the ε_i^* we have

$$\sum_i \varepsilon_i^{*2}\pi_i^2 \exp(-2n\pi_i) \le uC_\pi.$$

Moreover

$$\sum_{J\le i<I_{v,\pi}} \varepsilon_i^*\pi_i \ge \sqrt{\frac{uC_\pi}{2}}\frac{I_{v,\pi}-J}{\sqrt{I_{v,\pi}}}.$$

This concludes the proof in this case with Equation (2.41). □

2.7.4 Proof of Proposition 14

The proof of this proposition is similar to the proof of Proposition 13, except that the measures Λ_1 and $\tilde{\Lambda}_1$ change, and that we need to adapt Lemma 16. We therefore take the same notations as in the proof of Proposition 13, but redefine $\xi, p, \tilde{p}, \Lambda_1, \tilde{\Lambda}_1$.

Definition of Λ_1: We consider $q, \mathcal{A}, \mathcal{A}'$ defined as in the proof of Proposition 13. Write

$$\bar{m} = n\int \bar{p}\mathbf{1}\{\bar{p} \le 1/n\}d\mathcal{U}_{\pi,\mathcal{A}}(\bar{p}) \le 1.$$

For any $i \in \mathcal{A}$, let ξ_i be a random variable that is uniform in $\{0, 2\bar{m}\}$ if $i \in \mathcal{A}'$, and equal to q_i otherwise. For any $i \notin \mathcal{A}$: define $p_i = q_i = \pi_i$. We write Λ_1 for the distribution $\Lambda_{q,p}$ averaged over q, p as

$$\Lambda_1 = \mathbb{E}_{q,p}(\Lambda_{q,p}),$$

where $\mathbb{E}_{q,p}$ is the expectation according to the distribution of (q,p), i.e. with respect to q, ξ.
$\tilde{\Lambda}_1$ and $\tilde{p}$ are then redefined as in Proposition 13 as a renormalised version of Λ_1, p using $\mathcal{A}^C$ such that $\sum_i \tilde{p}_i = 1$.

We prove the following lemma in order to conclude on a bound on $d_{TV}(\Lambda_0, \Lambda_1)$.

Lemma 18. *Let $\pi \in \mathbb{R}^{+d}$ be such that $\sum_i \pi_i \le 1$ be a vector ordered in decreasing order, and let $v > 0$. There exists a universal constant $h > 0$ such that if $\|\pi^2 \exp(-2(1+v)n\pi)\|_1 \le h/n^2$, then the following holds.*

Write $\lambda = n\pi$. Let $\mathcal{U}_\lambda$ be the uniform distribution over the values of the vector $\lambda = n\pi$. Define the probability distribution

$$V = \frac{1}{2}[\delta_{2\bar{m}} + \delta_0],$$

i.e. $2\bar{m}$ times the outcome of a Bernoulli distribution of parameter $1/2$. We now consider

$$\nu_0' = \nu_0 = \int \mathcal{P}(\theta)^{\otimes 2} d\mathcal{U}_\lambda(\theta),$$

and

$$\nu_1' = \int\int \mathcal{P}(\theta) \otimes \mathcal{P}(\theta') dV(\theta') d\mathcal{U}_\lambda(\theta).$$

Then we have

$$d_{TV}(\nu_0'^{\otimes d}, \nu_1'^{\otimes d}) \leq 69 e^{2(1+v)} h.$$

We use this lemma instead of Lemma 16 in the proof of Proposition 13. By application of Lemma 18 on π restricted to $\mathcal{A}$, we have $d_{TV}(\Lambda_0, \Lambda_1) \leq 69 e^{2(1+v)} h$. We can then proceed as in the proof of Proposition 13 to prove the proposition, together with the use of the following lemma instead of Lemma 17, to conclude the proof.

Lemma 19. *It holds with probability larger than $1 - \alpha$ that*

$$\sum_i |\tilde{p}_i - \tilde{q}_i| \geq \frac{1}{4M} \|\pi \mathbf{1}\{i \geq J\}\|_1 - 2\sqrt{\frac{1}{\alpha n}} - 8a\frac{(1 + \log(1/\gamma))}{n} := \tilde{\rho}.$$

Proof of Lemma 19. We remind that $\tilde{q} \sim \mathcal{U}_\pi^{\otimes d}$, and $n\tilde{p} \sim V_{(nq)}^{\otimes d}$ and are independent. Note that as in the proof of Lemma 17

$$\mathbb{E}_{(q,p)} \sum_i |\tilde{q}_i - \tilde{p}_i| = \sum_{i \in \mathcal{A}'} \frac{1}{2}\Big[\pi_i + |\pi_i - 2\bar{m}|\Big] \geq \frac{1}{4M} \|\pi \mathbf{1}\{\pi \leq 1/n\}\|_1 - 8a\frac{(1 + \log(1/\gamma))}{n},$$

and

$$\mathbb{V}_{(q,p)} \sum_{i \leq d} |q_i - p_i| = d \cdot \mathbb{V}_{(q_1,p_1)} |q_1 - p_1| \leq 4 \|\pi^2 \mathbf{1}\{\pi \leq 1/n\}\|_1 \leq 4/n.$$

We set for $\alpha > 0$

$$\Theta = \Big\{ \sum_{i \leq d} |q_i - p_i| \geq \frac{1}{4M} \|\pi \mathbf{1}\{\pi \leq 1/n\}\|_1 - 2\sqrt{\frac{\|\pi^2 \mathbf{1}\{\pi \leq 1/n\}\|_1}{\alpha}} \Big\}.$$

So by Chebyshev's inequality, we know that with probability larger than $1 - \alpha$,

$$\sum_{i \leq d} |q_i - p_i| \geq \frac{1}{4M} \|\pi \mathbf{1}\{i \geq J\}\|_1 - 2\sqrt{\frac{1}{\alpha n}} - 8a\frac{(1 + \log(1/\gamma))}{n} := \tilde{\rho}.$$

□

Proof of Lemma 18. By assumption, we have that

$$\|\pi^2 \mathbf{1}\{\pi n \leq 1\}\|_1 \leq e^{2(1+v)} h / n^2.$$

We also define

$$\kappa = \int \theta^2 \mathbf{1}\{\theta \le 1\} d\mathcal{U}_\lambda(\theta) \le e^{2(1+v)} h/n. \tag{2.43}$$

Definition of two measures for (p, q). Now, we have by definition for any $m, m' \ge 0$

$$\nu_0'(m, m') = \int \frac{e^{-2\theta}\theta^{m+m'}}{m!m'!} d\mathcal{U}_\lambda(\theta),$$

and for any θ in the support of $\mathcal{U}_\lambda$, we have $\theta \le 1$. So

$$\nu_1'(m, m') = \int \frac{e^{-\theta}\theta^m}{m!} \mathbf{1}\{\theta \le 1\} d\mathcal{U}_\lambda(\theta) \cdot \frac{1}{2}\Big[\frac{e^{-2\bar{m}}(2\bar{m})^{m'}}{m'!} + \mathbf{1}\{m' = 0\}\Big].$$

Bound on the total variation. We have

$$\begin{aligned}
&d_{TV}(\nu_0^{'\otimes d}, \nu_1^{'\otimes d}) \le d \cdot d_{TV}(\nu_0', \nu_1') \\
&\le d \sum_{n,n'} \Big| \Big[\int \mathbf{1}\{\theta \le 1\} \frac{e^{-2\theta}\theta^{m+m'}}{m!m'!} d\mathcal{U}_\lambda(\theta) \Big] \\
&\qquad - \Big[\int \frac{e^{-\theta}\theta^m}{m!} \mathbf{1}\{\theta \le 1\} d\mathcal{U}_\lambda(\theta) \cdot \frac{1}{2}\Big(\frac{e^{-2\bar{M}}(2\bar{m})^{m'}}{m'!} + \mathbf{1}\{m' = 0\}\Big)\Big] \Big| \\
&= d \sum_{m,m'} \Big| \int \mathbf{1}\{\theta \le 1\} \frac{e^{-2\theta}\theta^{m+m'}}{m!m'!} d\mathcal{U}_\lambda(\theta) \\
&\qquad - \int \frac{e^{-\theta}\theta^m}{m!} \mathbf{1}\{\theta \le 1\} d\mathcal{U}_\lambda(\theta) \cdot \frac{1}{2}\Big(\frac{e^{-2\bar{m}}(2\bar{m})^{m'}}{m'!} + \mathbf{1}\{m' = 0\}\Big) \Big|.
\end{aligned}$$

And so

$$\begin{aligned}
&d_{TV}(\nu_0^{'\otimes d}, \nu_1^{'\otimes d}) \\
&\le d \Bigg[\Big| \int \mathbf{1}\{\theta \le 1\} e^{-2\theta} d\mathcal{U}_\lambda(\theta) - \int e^{-\theta} \mathbf{1}\{\theta \le 1\} d\mathcal{U}_\lambda(\theta) \cdot \frac{1}{2}\big(e^{-2\bar{m}} + 1\big) \Big| \\
&\quad + \Big| \int \mathbf{1}\{\theta \le 1\} e^{-2\theta}\theta d\mathcal{U}_\lambda(\theta) - \int e^{-\theta}\theta \mathbf{1}\{\theta \le 1\} d\mathcal{U}_\lambda(\theta) \cdot \frac{1}{2}\big(e^{-2\bar{m}} + 1\big) \Big| \\
&\quad + \Big| \int \mathbf{1}\{\theta \le 1\} e^{-2\theta}\theta d\mathcal{U}_\lambda(\theta) - \int e^{-\theta} \mathbf{1}\{\theta \le 1\} d\mathcal{U}_\lambda(\theta) \cdot \frac{1}{2} e^{-2\bar{m}}(2\bar{m}) \Big| \\
&\quad + \sum_{m,m': m+m' \ge 2} \Big| \int \mathbf{1}\{\theta \le 1\} \frac{e^{-2\theta}\theta^{m+m'}}{m!m'!} d\mathcal{U}_\lambda(\theta) \\
&\qquad\qquad - \int \frac{e^{-\theta}\theta^m}{m!} \mathbf{1}\{\theta \le 1\} d\mathcal{U}_\lambda(\theta) \cdot \frac{1}{2}\Big(\frac{e^{-2\bar{m}}(2\bar{m})^{m'}}{m'!} + \mathbf{1}\{m' = 0\}\Big) \Big| \Bigg].
\end{aligned}$$

Since for any $0 \le x \le 2$ we have $|e^{-x} - 1 + x| \le x^2/2$ and $|e^{-x} - 1| \le x$, we have

$$\begin{aligned}
&d_{TV}(\nu_0'^{\otimes d}, \nu_1'^{\otimes d}) \\
&\le d\Bigg[\Big|\int \mathbf{1}\{\theta \le 1\}(1-2\theta)d\mathcal{U}_\lambda(\theta) - \int (1-\theta)\mathbf{1}\{\theta \le 1\}d\mathcal{U}_\lambda(\theta)\cdot\frac{1}{2}\big((1-2\bar m)+1\big)\Big| \\
&\qquad + 2\bar m^2\zeta + 3\kappa \\
&+ \Big|\int \mathbf{1}\{\theta \le 1\}\theta d\mathcal{U}_\lambda(\theta) - \int \theta\mathbf{1}\{\theta \le 1\}d\mathcal{U}_\lambda(\theta)\cdot\frac{1}{2}\big((1-2\bar m)+1\big)\Big| + 3\kappa + \bar m^2\zeta \\
&+ \Big|\int \mathbf{1}\{\theta \le 1\}\theta d\mathcal{U}_\lambda(\theta) - \int (1-\theta)\mathbf{1}\{\theta \le 1\}d\mathcal{U}_\lambda(\theta)\cdot\bar m\Big| + 2\bar m^2\zeta + 4\kappa \\
&+ \sum_{m,m':m+m'\ge 2}\frac{1}{m!m'!}\Big|\int \mathbf{1}\{\theta \le 1\}\theta^{m+m'}d\mathcal{U}_\lambda(\theta) \\
&\qquad\qquad + \int \theta^m\mathbf{1}\{\theta \le 1\}d\mathcal{U}_\lambda(\theta)\cdot\frac{1}{2}\big((2\bar m)^{m'} + \mathbf{1}\{m'=0\}\big)\Big|\Bigg].
\end{aligned}$$

Since by Cauchy-Schwarz inequality we have

$$\begin{aligned}
(\bar m\zeta)^2 &= \Big[\int \theta\mathbf{1}\{\theta \le 1\}d\mathcal{U}_\lambda(\theta)\Big]^2 \\
&\le \Big[\int \theta^2\mathbf{1}\{\theta \le 1\}d\mathcal{U}_\lambda(\theta)\Big]\Big[\int \mathbf{1}\{\theta \le 1\}d\mathcal{U}_\lambda(\theta)\Big] = \zeta\kappa,
\end{aligned}$$

then we have

$$\begin{aligned}
&d_{TV}(\nu_0'^{\otimes d}, \nu_1'^{\otimes d}) \\
&\le d\Bigg[18\kappa + \sum_{m,m':m+m'\ge 2}\frac{1}{m!m'!}\Big(\int \mathbf{1}\{\theta \le 1\}\theta^{m+m'}d\mathcal{U}_\lambda(\theta) \\
&\qquad\qquad + \int \theta^m\mathbf{1}\{\theta \le 1\}d\mathcal{U}_\lambda(\theta)\cdot\frac{1}{2}\big((2\bar m)^{m'} + \mathbf{1}\{m'=0\}\big)\Big)\Bigg].
\end{aligned}$$

Then, considering the cases $(m = 0, m' \ge 2)$, $(m = 1, m' \ge 2)$ and $(m \ge 2, m' = 0)$, we have

$$\begin{aligned}
&d_{TV}(\nu_0'^{\otimes d}, \nu_1'^{\otimes d}) \\
&\le d\Bigg[18\kappa + \sum_{m,m':m+m'\ge 2}\frac{2^{m'}}{m!m'!}\Big(\int \mathbf{1}\{\theta \le 1\}\theta^2 d\mathcal{U}_\lambda(\theta) + \int \theta^2\mathbf{1}\{\theta \le 1\}d\mathcal{U}_\lambda(\theta)\Big) \\
&\qquad + e^2\zeta\bar m^2 + e^1\kappa\Bigg] \\
&\le d\big[18\kappa + 2e^3\kappa + e^2\kappa + e^1\kappa\big] \le 69d\kappa.
\end{aligned}$$

And so finally

$$d_{TV}(\nu_0'^{\otimes d}, \nu_1'^{\otimes d}) \le 69e^{2(1+v)}h,$$

by Equation (2.43). □

2.7.5 Proof of Theorem 22

Combining Propositions 15, 13, and 14, we obtain that no test exists for the testing problem (2.2) with type I plus type II error smaller than $1-\alpha-4\tilde{c}_v u - 34e^{2(1+v)}\varepsilon$ whenever

$$\rho \le c''\Bigg\{\Bigg[\|\pi(\mathbf{1}\{i \ge I_{v,\pi}\})\|_1 \vee \Bigg(\frac{I_{v,\pi}-J}{\sqrt{I_{v,\pi}}}\frac{\left(\|\pi^2\exp(-2(1+v)n\pi)\|_1 \vee n^{-2}\right)^{1/4}}{\sqrt{n}}\Bigg)\Bigg] \wedge \|\pi(\mathbf{1}\{i \ge J\})\|_1\Bigg\} \vee \frac{\|\pi^2\frac{1}{(\pi\vee n^{-1})^{4/3}}\|_1^{3/4}}{\sqrt{n}} \vee \frac{1}{\sqrt{n}},$$

where $c'' > 0$ is some small enough constant that depends only on $u, \alpha, v, \varepsilon$.

And so there exists constants $c_{\gamma,v} > 0$ that depend only on γ, v such that there is no test φ which is uniformly γ-consistent, for the problem (2.2) with

$$\rho \le c_{\gamma,v}\Bigg\{\min_{I\ge J_\pi}\Bigg[\frac{\sqrt{I}}{n} \vee \Bigg(\sqrt{\frac{I}{n}}\|\pi^2\exp(-2n\pi)\|_1^{1/4}\Bigg) \vee \|\pi_{(.)}(\mathbf{1}\{i \ge I)_i\|_1\Bigg] \wedge \|\pi_{(.)}(\mathbf{1}\{i \ge J_\pi)_i\|_1\Bigg\} \vee \frac{\left\|\pi^2\frac{1}{(\pi\vee n^{-1})^{4/3}}\right\|_1^{3/4}}{\sqrt{n}} \vee \sqrt{\frac{1}{n}},$$

since for any $I \ge J_\pi$ we have

$$\frac{J_\pi}{\sqrt{I}}\frac{1}{n} \le n^{-1/2},$$

and

$$\frac{J_\pi}{\sqrt{In}}\|\pi^2\exp(-2n\pi)\|_1^{1/4} \le \Bigg[\frac{\left\|\pi^2\frac{1}{(\pi\vee n^{-1})^{4/3}}\right\|_1^{3/4}}{\sqrt{n}}\Bigg] \vee \sqrt{\frac{1}{n}}.$$

The final result follows if we take I^* as an I where the minimum is attained as in the theorem, since

$$\Bigg[\Bigg(\sqrt{I^*-J_\pi}\frac{\log(n)}{n}\Bigg) \vee \Bigg(\frac{\sqrt{I^*-J_\pi}}{\sqrt{n}}\|\pi^2\exp(-n\pi)\|_1^{1/4}\Bigg) \vee \|\pi_{(.)}(\mathbf{1}\{i \ge I^*\})_i\|_1\Bigg] \le \|\pi_{(.)}(\mathbf{1}\{i \ge J_\pi\})_i\|_1,$$

and since

$$\left\|\pi^2\frac{1}{(\pi\vee n^{-1})^{4/3}}\right\|_1 \ge \|\pi_{(.)}^{2/3}(\mathbf{1}\{i \ge J_\pi\})_i\|_1.$$

Chapter 3

Minimax identity testing under local differential privacy

3.1 Introduction

Ensuring user privacy is at the core of the development of Artificial Intelligence. Indeed datasets can contain extremely sensitive information, and someone with access to a privatized training set or the outcome of an algorithm should not be able to retrieve the original dataset. However, classical anonymization and cryptographic approaches fail to prevent the disclosure of sensitive information in the context of learning. Indeed, with the example of a hospital's database, removing names and social security numbers from databases does not prevent the identification of patients using a combination of other attributes like gender, age or illnesses. Dinur and Nissim (2003) cites Cystic Fibrosis as an example which exist with a frequency of around 1/3000. Hence differential privacy mechanisms were developed to cope with such issues. Such considerations can be traced back to Warner (1965), Duncan and Lambert (1986), Duncan and Lambert (1989), and Fienberg and Steele (1998). As early as in 1965, Warner (1965) presented the first privacy mechanism which is now a baseline method for binary data: Randomized response. Another important result is presented in the works of Duncan and Lambert (1986), Duncan and Lambert (1989), and Fienberg and Steele (1998), where they expose a trade-off between statistical utility, or in other terms performance, and privacy in a limited-disclosure setting.

We will tackle differential privacy discussed in Section 1.5, and in particular the stronger local differential privacy condition. Let $X_1, \dots, X_n$ unobserved random variables taking values in $[0, 1]$, which are i.i.d. with density f with respect to the Lebesgue measure. We observe $Z_1, \dots, Z_n$ which are α-local differentially private views of $X_1, \dots, X_n$. This notion has been extensively studied through the concept of local algorithms, especially in the context of privacy-preserving data mining Warner (1965), Agrawal and Srikant (2000), Agrawal and Aggarwal (2001), van den Hout and van der Heijden (2002), Evfimievski, Gehrke, and Srikant (2003), Agrawal and Haritsa (2005), Mishra and Sandler (2006), Jank, Shmueli, and Wang (2008), and Kasiviswanathan et al. (2011). Now note that local differential privacy can account for possible dependencies between Z_i's, corresponding to the interactive case. The role of interactivity has been further studied in Joseph et al. (2019), Butucea, Rohde, and Steinberger (2020), and Berrett and Butucea (2020). Recent results detailed in Duchi, Jordan, and Wainwright (2013b), Duchi, Wainwright, and Jordan (2013), and Duchi, Jordan, and Wainwright (2013c) give information processing inequalities depending on the local privacy constraint via the parameter α. Those can be used to obtain Fano or Le Cam-type inequalities in order to obtain a minimax lower bound for estimation or testing problems. Our proof also relies on Le Cam's inequality, albeit in a more refined way in order to obtain minimax optimal results.

In continuity with the first two chapters, we will consider a testing problem, where we want to design our tests so that they reject the null hypothesis $\mathcal{H}_0 : f = f_0$ if the data is not actually generated from the given model with a given confidence level. Assuming that f and f_0 belong to $\mathbb{L}_2([0,1]) = \left\{f : [0,1] \to \mathbb{R}, \|f\|_2^2 = \int_0^1 f^2(x)dx < \infty\right\}$, it is natural to propose a test based on an estimation of the squared $\mathbb{L}_2$-distance $\|f - f_0\|_2^2$ between f and f_0. In order to test whether $f = f_0$ from the observation of an i.i.d sample set $(X_1, \dots, X_n)$ with common density f, Neyman (1937) introduces an orthonormal basis $\{f_0, \phi_l, l \geq 0\}$ of $\mathbb{L}_2([0,1])$. The identity hypothesis is rejected if the estimator $\sum_{l=0}^{D-1}(\sum_{i=1}^n \phi_l(X_i)/n)^2$ exceeds some threshold, where D is a given integer depending on n. Data-driven versions of this test, where the parameter D is chosen to minimize some penalized criterion, have been introduced by Bickel and Ritov (1992), Ledwina (1994), Kallenberg W. (1995), and Inglot T. (1996).

Additionally, we want to find the limitations of a test by determining how close both hypotheses can get while remaining separated by the testing procedure. We will do this in a minimax testing framework, in a similar way to Section 1.4.

We will restrict our study to two cases, firstly with multinomial distributions covering the discrete case. We will also work in the continuous case with Besov balls for which non-private results already exist, making them meaningful for comparisons. Other motivations for the use of Besov balls are discussed in Section 1.3.3.

We recall a few non-private results from the literature. For Hölder classes with smoothness parameter $s > 0$, Ingster (1993) establishes the asymptotic minimax rate of testing $n^{-2s/(4s+1)}$. The test proposed in their paper is not adaptive since it makes use of a known smoothness parameter s. Minimax optimal adaptive identity tests over Hölder or Besov classes of alternatives are provided in Ingster (2000b) and Fromont and Laurent (2006b). These tests achieve the separation rate $(n/\sqrt{\log\log(n)})^{-2s/(4s+1)}$ over a wide range of regularity classes (Hölder or Besov balls) with smoothness parameter $s > 0$. The $\log\log(n)$ term is the optimal price to pay for adaptation to the unknown parameter $s > 0$.

In the discrete case, the goal is to distinguish between d-dimensional probability vectors p and q using samples from the multinomial distribution with parameters p and n. In Section 1.4, we review a few results that we recall here. Paninski (2008) obtain that the minimax optimal rate with respect to the ℓ_1-distance, $\sum_{i=1}^d |q_i - p_i|$, is $d^{1/4}/\sqrt{n}$. An extension is the study of local minimax rates as in Valiant and Valiant (2017), where the rate is made minimax optimal for any p instead of just in the worst choice of p. Finally, Balakrishnan and Wasserman (2017a) presents local minimax rates of testing both in the discrete and continuous cases.

A few problems have already been tackled in order to obtain minimax rates under local privacy constraint. The main question is whether the minimax rates are affected by the local privacy constraint and to quantify the degradation of the rate in that case. We define a sample degradation of $C(\alpha)$ in the following way. If n is the necessary and sufficient sample size in order to solve the classical non-private version of a problem, the α-local differential private problem is solved with $nC(\alpha)$ samples. For a few problems, a degradation of the effective sample size by a multiplicative constant is found. In Duchi, Wainwright, and Jordan (2013), they obtain minimax estimation rates for multinomial distributions in dimension d and find a sample degradation of α^2/d. In Duchi, Jordan, and Wainwright (2018), they also find a multiplicative sample degradation of α^2/d for generalized linear models, and α^2 for median estimation. However, in other problems, a polynomial degradation is noted. For one-dimensional mean estimation, the usual minimax rate is $n^{-(1\wedge(2-2/k))}$, whereas the private rate from Duchi, Jordan, and Wainwright (2018) is $(n\alpha^2)^{-(0\wedge(1-1/k))}$ for original observations

X satisfying $\mathbb{E}(X) \in [-1, 1]$ and $\mathbb{E}(|X|^k) < \infty$. As for the problem of nonparametric density estimation presented in Duchi, Jordan, and Wainwright (2018), the rate goes from $n^{-2s/(2s+1)}$ to $(n\alpha^2)^{-2s/(2s+2)}$ over an elliptical Sobolev space with smoothness s. This result was extended in Butucea et al. (2020) over Besov ellipsoids. The classical minimax mean squared errors were presented in Yu (1997), Yang and Barron (1999), and Tsybakov (2004).

Goodness-of-fit testing has been studied extensively under a global differential privacy constraint in Gaboardi et al. (2016), Cai, Daskalakis, and Kamath (2017), Aliakbarpour, Diakonikolas, and Rubinfeld (2018), Acharya, Sun, and Zhang (2018) and Canonne et al. (2019). Further steps into covering other testing problems under global differential privacy have been taken already with works like Aliakbarpour et al. (2019).

Our contributions can be summarized in the following way. Under non-interactive local differential privacy, we provide optimal separation rates for identity testing over Besov balls in the continuous case. To the best of our knowledge, we are the first to provide quantitative guarantees in such a continuous setting. We also provide minimax separation rates for multinomial distributions. In particular, we establish a lower bound that is completely novel in the definition of the prior distributions leading to optimal rates, and in the way we tackle privacy. Indeed, naive applications of previous information processing inequalities under local privacy lead to suboptimal lower bounds. Finally, we provide an adaptive version of our test, which is independent of the smoothness parameter s and rate-optimal up to a logarithmic factor. So in shorter terms:

- We provide the first minimax lower bound for the problem of identity testing under local privacy constraint over Besov balls.
- We present the first minimax optimal test with the associated local differentially private channel in this continuous setting.
- The test is made adaptive to the smoothness parameter of the unknown density up to a logarithmic term.
- A minimax optimal test under local privacy can be derived for multinomial distributions as well.

We start with citing results pertaining to the study of identity testing in the discrete case under local privacy. Gaboardi and Rogers (2017) take another point of view from ours and provide asymptotic distributions for a chi-squared statistic applied to noisy observations satisfying the local differential privacy condition. Sheffet (2018) takes a closer approach to ours and determines a sufficient number of samples for testing between $p = q$ and fixed $\sum |q_i - p_i|$, which has been improved upon by Acharya et al. (2018). Finally, in parallel with the writing of Lam-Weil, Laurent, and Loubes (2020), Berrett and Butucea (2020) have provided minimax optimal rates of testing for discrete distributions under local privacy, in both ℓ_1 and ℓ_2 norms. In particular, they tackle both interactive and non-interactive privacy channels and point out a discrepancy in the rates between both cases.

Now, the following papers tackle the continuous case. Butucea et al. (2020) provides minimax optimal rates for density estimation over Besov ellipsoids under local differential privacy. Following this paper, we apply Laplace noise to the projection of the observations onto a wavelet basis, although we tackle the different problem of density testing. The difference between density estimation and testing is fundamental and leads in our case to faster rates. A problem closer to density testing is the

estimation of the quadratic functional presented in Butucea, Rohde, and Steinberger (2020), where they find minimax rates over Besov ellipsoids under local differential privacy. They rely on the proof of the lower bound in the non-interactive case given in a preliminary version of Lam-Weil, Laurent, and Loubes (2020). It was refined in order to improve on the rate in α, reaching an optimal rate for low values of α.

The results presented in this chapter iterates on the first version of Lam-Weil, Laurent, and Loubes (2020), which only focused on the continuous case. We extend its scope and construct a unified setting to tackle both Besov classes and multinomial distributions, leading to minimax optimal results in both settings.

The rest of the chapter is articulated as follows. In Section 3.2, we detail our setting and sum up our results. A lower bound on the minimax separation distance for identity testing is introduced in Section 3.3. Then we introduce a test and a privacy mechanism in Section 3.4. This leads to an upper bound which matches the lower bound. However, in the continuous case, the proposed test depends on a smoothness parameter which is unknown in general. That is the reason why we present a version of the test in Section 3.5 that is adaptive to s. Afterwards, we conclude the chapter with a final discussion in Section 3.6. Finally, the proofs of all the results presented in this chapter are contained in Section 3.7 and discussions on possible alternatives for the proof of the lower bound in Section 3.8.

All along the chapter, C will denote some absolute constant, $c(a, b, \ldots), C(a, b, \ldots)$ will be constants depending only on their arguments. The constants may vary from line to line.

3.2 Setting

3.2.1 Non-interactive differential privacy

We now recall the definition of non-interactive differential privacy. Let n be some positive integer and $\alpha > 0$. Let f_0, f be densities in $\mathbb{L}_2([0,1])$ with respect to the Lebesgue measure. Let $X_1, \ldots, X_n$ be i.i.d. random variables with density f. We define $Z_1, \ldots, Z_n$ satisfying for all $1 \leq i \leq n$

$$\sup_{S \in \mathcal{Z}_i, (x,x') \in (\mathbb{R}^d)^2} \frac{Q_i(Z_i \in S | X_i = x)}{Q_i(Z_i \in S | X_i = x')} \leq \exp(\alpha). \tag{3.1}$$

Let $\mathcal{Q}_\alpha$ be the set of joint distributions $Q = \prod Q_i$ such that Q_i satisfies the condition in Equation (3.1) for all $1 \leq i \leq n$.

3.2.2 A unified setting for discrete and continuous distributions

We present a unified setting and end up dealing with densities in $\mathbb{L}_2([0,1])$ in both the continuous and discrete cases. In the discrete case, $\widetilde{X_1}, \ldots, \widetilde{X_n}$ are i.i.d. random variables taking their values in d classes denoted by $\{1, 2 \ldots, d\}$ according to the probability vector $q = (q_1, q_2, \ldots, q_d)$. For a given probability vector $p = (p_1, p_2, \ldots, p_d)$, we want to test the null hypothesis $\mathcal{H}_0 : p = q$ against the alternative $\mathcal{H}_1 : p \neq q$. In order to have a unified setting, we transform these discrete observations into continuous observations $X_1 \ldots, X_n$ with values in $[0,1]$ by the following process. For all $k \in \{1, \ldots, d\}$, if we observe $\widetilde{X_i} = k$, we generate X_i by a uniform distribution on the interval $[(k-1)/d, k/d)$. Note that the variables $X_1 \ldots, X_n$ are i.i.d. with common

density f defined for all $x \in [0,1]$ by

$$f(x) = \sum_{k=1}^{d} dq_k \mathbb{I}_{[\frac{k-1}{d}, \frac{k}{d})}(x).$$

Similarly, for the probability vector p, we define the corresponding density f_0 for $x \in [0,1]$ by

$$f_0(x) = \sum_{k=1}^{d} dp_k \mathbb{I}_{[\frac{k-1}{d}, \frac{k}{d})}(x).$$

So we have the equivalence $p = q \iff f = f_0$. The following equation highlights the connection between the separation rates for densities and for probability vectors. We have

$$\|f - f_0\|_2^2 = d \sum_{k=1}^{d} (q_k - p_k)^2. \tag{3.2}$$

3.2.3 Minimax identity testing under local differential privacy

We now define a privacy mechanism and a testing procedure based on the private views $Z_1, \ldots, Z_n$. We want to test

$$\mathcal{H}_0 : f = f_0, \qquad \text{versus} \qquad \mathcal{H}_1 : f \neq f_0, \tag{3.3}$$

from α-local differentially private views of $X_1, \ldots, X_n$.

The twist on classical identity testing is in the fact that the samples $(X_1, \ldots, X_n)$ from f are unobserved, we only observe their private views. We now aim at formalizing this hypothesis testing problem for multinomials and Besov densities. Let $X_i \sim f$ be an independent random variable for any $i \leq n$ with respect to probability measure $\mathbb{P}_f$. Let $\mathcal{Z}^{(\alpha)}(X_i)$ be the set of random variables Z_i ranging in $\mathcal{Z}_i$ such that Z_i is an α-private transformation of X_i with respect to Markov kernel Q_i. Let $\Phi(Z_i) = \{\varphi : \mathcal{Z}_i \to \{0,1\}\}$. We then define in both cases,

$$H_0(f_0) = \{(f_0, f); f_0 = f\}$$

and

$$\mathcal{T}_n^{(\alpha)} = \{\hat{\theta}; \hat{\theta} = \varphi(Z_1, \ldots, Z_n), \varphi \in \Phi(Z_i), Z_i \in \mathcal{Z}^{(\alpha)}(X_i), i \leq n\}.$$

There remains to define the alternative hypothesis set.

1. In the discrete case, we define

$$\mathcal{D} = \left\{ f \in \mathbb{L}_2([0,1]); \exists q = (q_1, \ldots, q_d) \in \mathbb{R}^d, f = \sum_{j=1}^{d} q_j \mathbb{I}_{[(j-1)/d, j/d)} \right\}, \tag{3.4}$$

   which is associated with the class of densities for multinomial distributions over d classes. Let f_0 be a fixed density in $\mathcal{D}$. Then we define for any $\rho > 0$,

$$H_1(f_0, \rho) = \{(f_0, f); f \in \mathcal{D}, f \geq 0, \int f = 1, \|f_0 - f\|_2 \geq \rho\}.$$

2. In the continuous case, we take a density $f \in \mathbb{L}_2([0,1])$ and we will be considering Besov balls. To define these classes, we consider a pair of compactly supported

and bounded wavelets (ϕ, ψ) such that for all J in $\mathbb{N}$,

$$\left\{2^{J/2}\phi(2^J(\cdot)-k), k \in \Lambda(J)\right\} \cup \left\{2^{j/2}\psi(2^j(\cdot)-k), j \geq J, k \in \Lambda(j)\right\}$$

is an orthonormal basis of $\mathbb{L}_2([0,1])$. For the sake of simplicity, we consider the Haar basis where $\phi = \mathbb{1}_{[0,1)}$ and $\psi = \mathbb{1}_{[0,1/2)} - \mathbb{1}_{[1/2,1)}$. In this case, for all $j \in \mathbb{N}$, $\Lambda(j) = \left\{0, 1, \dots 2^j - 1\right\}$.

We denote for all $j \geq 0, k \in \Lambda(j)$, $\alpha_{j,k}(f) = \int 2^{j/2} f\phi(2^j(\cdot)-k)$ and $\beta_{j,k}(f) = \int 2^{j/2} f\psi(2^j(\cdot)-k)$. For $R > 0$ and $s > 0$, the Besov ball $\widetilde{B}_{s,2,\infty}(R)$ with radius R associated with the Haar basis is defined as

$$\widetilde{B}_{s,2,\infty}(R) = \left\{ f \in \mathbb{L}_2([0,1]), \forall j \geq 0, \sum_{k \in \Lambda(j)} \beta_{jk}^2(f) \leq R^2 2^{-2js} \right\}.$$

Now note that, if $s \leq 1$, then there is an equivalence between the definition of $\widetilde{B}_{s,2,\infty}(R)$ and the definition of the corresponding Besov space using moduli of smoothness – see e.g. Theorem 4.3.2 in Giné and Nickl (2016). And for larger s, Besov spaces defined with Daubechies wavelets satisfy this equivalence property.

We introduce the following class of alternatives : for any $s > 0$ and $R > 0$, we define the set $\mathcal{B}_{s,2,\infty}(R)$ as follows

$$\mathcal{B}_{s,2,\infty}(R) = \left\{ f \in \mathbb{L}_2([0,1]), f - f_0 \in \widetilde{B}_{s,2,\infty}(R) \right\}. \tag{3.5}$$

Note that the class $\mathcal{B}_{s,2,\infty}(R)$ depends on f_0 since only the regularity for the difference $f - f_0$ is required to establish the separation rates. Nevertheless, for the sake of simplicity we omit f_0 in the notation of this set. Then we end up considering for any $\rho > 0$,

$$H_{1,s}(f_0, \rho) = \{(f_0, f); f \in \mathcal{B}_{s,2,\infty}(R), f \geq 0, \int f = 1, \|f_0 - f\|_2 \geq \rho\},$$

For $\alpha > 0$ and $\gamma \in (0,1)$, let $\mathcal{T}_{\gamma,n}^{(\alpha)}(f_0) = \{\hat{\theta}_n \in \mathcal{T}_n^{(\alpha)}; \mathbb{P}_{f_0}(\hat{\theta}_n = 1) \leq \gamma\}$ and $\Delta_{Q,\gamma,n} \in \mathcal{T}_{\gamma,n}^{(\alpha)}(f_0)$. The index Q in the notation of $\Delta_{Q,\gamma,n}$ is a reminder that the test relies on observations transformed by some α-private Markov kernel Q. That is, there exists an α-local differentially private channel $Q \in \mathcal{Q}_\alpha$ and a test function φ such that $\Delta_{Q,\gamma,n} = \varphi(Z_1, \dots, Z_n)$. Then

$$\mathbb{P}_{Q_{f_0}^n}(\varphi(Z_1, \dots, Z_n) = 1) \leq \gamma,$$

where

$$\mathbb{P}_{Q_{f_0}^n}((Z_1, \dots, Z_n) \in \prod_{i=1}^n S_i) = \int \prod_i Q_i(Z_i \in S_i | X_i = x_i) f_0(x_i) dx_i,$$

and Q_i is the i-th marginal channel of Q.

Then the uniform separation rate $\rho_\beta(f_0, H_1, \Delta_{Q,\gamma,n})$ of the test $\Delta_{Q,\gamma,n}$ with respect to H_1, is defined for all β in $(0,1)$ as

$$\rho_\beta(f_0, H_1, \Delta_{Q,\gamma,n}) = \inf\{\rho > 0; \sup_{f \in H_1(f_0,\rho)} \mathbb{P}_f(\Delta_{Q,\gamma,n} = 0) \leq \beta\}. \tag{3.6}$$

where $(X_1, \dots, X_n)$ are i.i.d. samples with common density f according to $\mathbb{P}_f$.

The uniform separation rate is then the smallest value in the sense of the $\mathbb{L}_2$-norm of $(f - f_0)$ for which the second kind error of the test is uniformly controlled by β over H_1. The combination of a privacy channel and a test with level γ having optimal performances should then have the smallest possible uniform separation rate (up to a multiplicative constant) over $\mathcal{C}$. To quantify this, we introduce the minimax separation distance of testing defined by

$$\rho^*_\beta(f_0, H_1, \mathcal{T}^{(\alpha)}_{\gamma,n}) = \inf_{\Delta_{Q,\gamma,n} \in \mathcal{T}^{(\alpha)}_{\gamma,n}(f_0)} \rho_\beta(H_1, \Delta_{Q,\gamma,n}). \tag{3.7}$$

3.2.4 Overview of the results

For any $\alpha > 0$, we define $z_\alpha = e^{2\alpha} - e^{-2\alpha} = 2\sinh(2\alpha)$.

Continuous case. The results presented in Theorems 27 and 30 can be condensed into the following conclusion that holds if $nz_\alpha^2 \geq (\log n)^{1+3/(4s)}$, $s > 0$, $R > 0$, $\alpha \geq 1/\sqrt{n}$, $(\gamma, \beta) \in (0,1)^2$ such that $2\gamma + \beta < 1$,

$$\begin{aligned} c\,(s, R, \gamma, \beta)\,[(nz_\alpha^2)^{-2s/(4s+3)} \vee n^{-2s/(4s+1)}] \\ \leq \rho^*_\beta(\mathbb{1}_{[0,1]}, H_{1,s}, \mathcal{T}^{(\alpha)}_{\gamma,n}) \\ \leq C(s, R, \gamma, \beta)\left[(n\alpha^2)^{-2s/(4s+3)} \vee n^{-2s/(4s+1)}\right]. \end{aligned} \tag{3.8}$$

Remark 28.

- *Having $nz_\alpha^2 \geq (\log n)^{1+3/(4s)}$ and $\alpha \geq 1/\sqrt{n}$ reduces to wanting a sample set large enough, which is a classical non-restrictive assumption.*
- *The upper bound holds for any density $f_0 \in \mathbb{L}_2([0,1])$ and matches the lower bound when $f_0 = \mathbb{1}_{[0,1]}$, as shown in Equation (3.8). So we can deduce the minimax separation rate for identity testing under a local privacy constraint. It can be decomposed into two different regimes, where the rates of our upper and lower bounds match in n as well as in α, when α tends to 0. When α is larger than $n^{1/(4s+1)}$, then the minimax rate is of order $n^{-2s/(4s+1)}$, which coincides with the rate obtained in the non-private case in Ingster (1987). The other regime corresponds to α being smaller than $n^{1/(4s+1)}$. The minimax rate is then of order $(n\alpha^2)^{-2s/(4s+3)}$ and so we show a polynomial degradation in the rate due to the privacy constraints. Such a degradation has also been discovered in the problem of second moment estimation and mean estimation, as well as for the density estimation in Butucea et al. (2020). Very similar results have been found for the estimation of the quadratic functional under non-interactive privacy in Butucea, Rohde, and Steinberger (2020). Now, Butucea, Rohde, and Steinberger (2020) tackle local differential privacy with dependencies between the Z_i's as well, hinting at the possibility of obtaining better rates when the channels are allowed to be interactive.*
- *Due to having z_α instead of α, our upper and lower bounds do not match in α when α is larger than a constant but smaller than $n^{1/(4s+1)}$. This is not an issue in practice, since α will be taken small in order to guarantee privacy.*

Discrete case. The results presented in Theorems 26 and 29 can be condensed into the following conclusion that holds if $n \geq (z_\alpha^{-2} d^{3/2} \log d) \vee ((\alpha^2 d^{-1/2}) \wedge d^{1/2})$, $\alpha > 0$,

$(\gamma, \beta) \in (0,1)^2$ such that $2\gamma + \beta < 1$,

$$c(\gamma,\beta)\left[((nz_\alpha^2)^{-1/2}d^{1/4}) \vee (n^{-1/2}d^{-1/4})\right] \leq \rho_\beta^*(\mathbb{1}_{[0,1]}, H_1, \mathcal{T}_{\gamma,n}^{(\alpha)})/d^{1/2} \leq C(\gamma,\beta)\left[\left((n\alpha^2)^{-1/2}d^{1/4}\right) \vee \left(n^{-1/2}d^{-1/4}\right)\right]. \tag{3.9}$$

Remark 29.

- *Assuming that $nz_\alpha^2 \geq d^{3/2}\log d$ means that the problem gets harder with the dimension, which aligns with the interpretation of the private rate.*
- *We present matching bounds on $\rho_\beta^*(\mathbb{1}_{[0,1]}, H_1, \mathcal{T}_{\gamma,n}^{(\alpha)})/d^{1/2}$ since it is the usual rate of interest as justified by the combination of Definition 3.6 and Equation (3.2). Here again, we find two regimes corresponding to the classical rate taking over if α is larger than $\sqrt{d}$. So we can see that the local privacy condition leaves the rate in n unchanged, but the rate in d changes drastically for the testing problem with respect to the $\mathbb{L}_2$-norm. Indeed, the classical testing problem with $\mathbb{L}_2$-separation becomes easier as the number of dimensions grows, whereas the private rate exhibits the opposite behaviour.*
- *Simultaneously and independently of our work, Berrett and Butucea (2020) find similar results in the non-interactive case and they also show that the minimax rates are improved when considering interactive privacy channels.*

3.3 Lower bound

This section will focus on the presentation of a lower bound on the minimax separation rate defined in Equation (3.7) for the problem of identity testing under a non-interactive privacy constraint. The result is presented both in the discrete and the continuous cases, when f_0 is the uniform density over $[0,1]$. Butucea, Rohde, and Steinberger (2020) and Berrett and Butucea (2020) provide results also accounting for sequentially interactive private channels, defined in Equation 1.9. Combined with their upper bounds, our lower bound is critical in proving that, for identity testing and the related problem of estimation of the squared functional, there is an intrinsic gap between minimax optimality considering only non-interactive privacy channels and minimax optimality with sequentially interactive channels.

The outline of our lower bound proof relies on a classical scheme, which is recalled below. Nevertheless, the implementation of this scheme in the context of local differential privacy is far from being classical, and we do it in a novel way which leads to a tight lower bound. At the end of the section, a more naive approach will be presented and shown to lead to suboptimal results.

We apply a Bayesian approach related to the one presented in Section 1.4.2, where we will define a prior distribution which corresponds to a mixture of densities such that $\|f - f_0\|_2$ is large enough. Such a starting point has been largely employed for lower bounds in minimax testing, as described in Baraud (2002b). Its application is mainly due to Ingster (1993) and inequalities on the total variation distance from Le Cam (1986). The result of this approach is summarized in Lemma 2.

The idea is to establish the connection between the second kind error and the total variation distance between arbitrary distributions with respective supports in $H_0(f_0)$ and $H_1(f_0,\rho)$. It turns out that the closer the distributions from H_0 and $H_1(f_0,\rho)$ are allowed to be, the higher the potential second kind error. So if we are able to provide distributions from H_0 and $H_1(f_0,\rho)$ which are close from one another, we can

guarantee that the second kind error of any test will be high. The main difficulty lies in finding the right prior distribution $\nu_{1,\rho}$ over $H_1(f_0, \rho)$ appearing in Lemma 2.

In the discrete case, we obtain the following lower bound.

Theorem 26. *Let $(\gamma, \beta) \in (0,1)^2$ such that $2\gamma + \beta < 1$. Let $\alpha > 0$.*

We obtain the following lower bound for the α-private minimax separation rate defined by Equation (3.7) for non-interactive channels in $\mathcal{Q}_\alpha$ over the class of alternatives $\mathcal{D}$ in Equation (3.4)

$$\frac{\rho^*_\beta(\mathbb{1}_{[0,1]}, H_1, \mathcal{T}^{(\alpha)}_{\gamma,n})}{d^{1/2}} \geq c\,(\gamma, \beta)\left([(nz_\alpha^2)^{-1/2} d^{1/4} \wedge d^{-1/2}(\log d)^{-1/2}] \vee (n^{-1/2} d^{-1/4})\right).$$

Remark 30. *In parallel to our work, Berrett and Butucea (2020) focus on the case when $\alpha \leq 1$ and find similar results displayed in their Theorem 6.*

In the continuous case, we obtain the following theorem for Besov balls.

Theorem 27. *Let $(\gamma, \beta) \in (0,1)^2$ such that $2\gamma + \beta < 1$. Let $\alpha > 0, R > 0, s > 0$.*

We obtain the following lower bound for the α-private minimax separation rate defined by Equation (3.7) for non-interactive channels in $\mathcal{Q}_\alpha$ over the class of alternatives $\mathcal{B}_{s,2,\infty}(R)$ defined in Equation (3.5)

$$\rho^*_\beta(\mathbb{1}_{[0,1]}, H_{1,s}, \mathcal{T}^{(\alpha)}_{\gamma,n}) \geq c\,(\gamma, \beta, R)\,[[(nz_\alpha^2)^{-2s/(4s+3)} \wedge (\log n)^{-1/2}] \vee n^{-2s/(4s+1)}].$$

Remark 31. *These theorems represent a major part of our contributions and lead to the construction of the inequalities presented in Section 3.2.4. Note that $(nz_\alpha^2)^{-2s/(4s+3)} \wedge (\log n)^{-1/2}$ reduces to $(nz_\alpha^2)^{-2s/(4s+3)}$ for n large enough and we can reduce the formulation of Theorem 26 in the same way with a condition on n being large enough.*

Sketch of proof. We want to find the largest $\mathbb{L}_2$-distance between the initial density f_0 under the null hypothesis and the density in the alternative hypothesis such that their transformed counterparts by an α-private channel Q cannot be discriminated by a test. We will rely on the singular vectors of Q in order to define densities and their private counterparts with ease. Employing bounds on the singular values of Q, we define a mixture of densities such that they have a bounded $\mathbb{L}_2$-distance to $f_0 = \mathbb{1}_{[0,1]}$. We obtain a sufficient condition for the total variation distance between the densities in the private space to be small enough for both hypotheses to be indistinguishable. Then we ensure that the functions that we have defined are indeed densities, and in the continuous case belong to the regularity class $\mathcal{B}_{s,2,\infty}(R)$. Collecting all these elements, the conclusion relies on Lemma 2.

Remark 32. *The total variation distance is a good criterion in order to determine whether two distributions are distinguishable. Another natural idea to prove Theorem 27 is to bound the total variation distance between two private densities by the total variation distance between the densities of the original samples, up to some constants depending on the privacy constraints. Following this intuitive approach, we can provide a lower bound using Theorem 1 in Duchi, Jordan, and Wainwright (2013c) combined with Pinsker's inequality. This approach has been used with success in density estimation in Butucea et al. (2020). However, the resulting lower bound does not match the upper bound for the separation rates of identity testing presented in our Section 3.4. Details on the application of this approach to our setting are provided in Section 3.8.*

3.4 Definition of a test and privacy mechanism

We will firstly define a testing procedure coupled with a privacy mechanism. Their application provides an upper bound on the minimax separation rate for any density f_0. The bounds obtained are presented in the right-hand side of Equations (3.8) and (3.9) for the continuous and the discrete cases respectively. The test and privacy mechanism will turn out to be minimax optimal since the upper bounds will match the lower bounds obtained in Section 3.3.

Let us first propose a transformation of the data, satisfying the differential privacy constraints.

3.4.1 Privacy mechanism

We consider the privacy mechanism introduced in Butucea et al. (2020). It relies on Laplace noise, which is classical as a privacy mechanism. However, applying it to the correct basis with the corresponding scaling is critical in finding optimal results. We will be focusing on the Haar basis presented in Example 2. We denote by ϕ the indicator function on $[0,1)$

$$\forall x, \phi(x) = \mathbb{1}_{[0,1)}(x),$$

and for all integer $L \geq 1$, we set, for all $k \in \{0, \dots, L-1\}$, for all $x \in [0,1)$,

$$\phi_{L,k}(x) = \sqrt{L}\phi(Lx - k).$$

The integer L will be taken as $L = 2^J$ for some $J \geq 0$ in the continuous case, and we choose $L = d$ in the discrete case. We define, for all $i \in \{1, \dots, n\}$, the vector $Z_{i,L} = (Z_{i,L,k})_{k \in \{0,\dots,L-1\}}$, by

$$\forall k \in \{0, \dots, L-1\}\,, \; Z_{i,L,k} = \phi_{L,k}(X_i) + \sigma_L W_{i,L,k}, \tag{3.10}$$

where $(W_{i,L,k})_{1 \leq i \leq n, k \in \{0,\dots,L-1\}}$ are i.i.d. Laplace distributed random variables with variance 1 and

$$\sigma_L = 2\sqrt{2}\frac{\sqrt{L}}{\alpha}.$$

So the privacy mechanism relies on adding Laplace noise to the coefficients of the original observations in the Haar basis.

We now provide the following result, showing that we have indeed defined an α-private channel. The following result connects the definition of privacy with probability measures to privacy with probability density functions.

Lemma 20. *For any i, denote $q_{i,L}(\cdot|x)$ the density of the random vector $Z_{i,L}$ with respect to the probability measure μ_i conditionally to $X_i = x$. Then*

$$\sup_{S \in \mathcal{Z}_{i,L}, (x,x') \in [0,1]^2} \frac{Q_i(Z_{i,L} \in S | X_i = x)}{Q_i(Z_{i,L} \in S | X_i = x')} \leq e^\alpha$$

if and only if there exists $\Omega \in \mathcal{Z}_{i,L}$ with $\mu_i(Z_{i,L} \in \Omega) = 1$ such that

$$\frac{q_{i,L}(z|x)}{q_{i,L}(z|x')} \leq e^\alpha$$

for any $z \in \Omega$ and any $(x, x') \in [0,1]^2$.

Proof. Assume there exists Ω with $\mu_i(Z_{i,L} \in \Omega) = 1$ such that $\frac{q_{i,L}(z|x)}{q_{i,L}(z|x')} \leq e^{\alpha}$ for any $z \in \Omega$. Let $\tilde{S} \in \mathcal{Z}_{i,L}$ and $S = \tilde{S} \cap \Omega$.

$$\frac{Q_i(Z_{i,L} \in \tilde{S}|X_i = x)}{Q_i(Z_{i,L} \in \tilde{S}|X_i = x')} = \frac{Q_i(Z_{i,L} \in S|X_i = x)}{Q_i(Z_{i,L} \in S|X_i = x')}.$$

Then

$$\begin{aligned}\frac{Q_i(Z_{i,L} \in S|X_i = x)}{Q_i(Z_{i,L} \in S|X_i = x')} &= \frac{\int_S q_{i,L}(z|x)d\mu_i(z)}{\int_S q_{i,L}(z|x')d\mu_i(z)} \\ &\leq \frac{\int_S q_{i,L}(z|x')e^{\alpha}d\mu_i(z)}{\int_S q_{i,L}(z|x)e^{-\alpha}d\mu_i(z)} = \frac{Q_i(Z_{i,L} \in S|X_i = x')}{Q_i(Z_{i,L} \in S|X_i = x)}e^{2\alpha}.\end{aligned}$$

So

$$\frac{Q_i(Z_{i,L} \in \tilde{S}|X_i = x)}{Q_i(Z_{i,L} \in \tilde{S}|X_i = x')} \leq e^{\alpha}.$$

Assume that $Q \in \mathcal{Q}_{\alpha}$. Then for any $S \in \mathcal{Z}_{i,L}$, we have $Q_i(Z_{i,L} \in S|X_i = x) \leq e^{\alpha}Q_i(Z_{i,L} \in S|X_i = x')$. That is, for any $S \in \mathcal{Z}_{i,L}$,

$$\int_S (e^{\alpha}q_{i,L}(z|x') - q_{i,L}(z|x))d\mu_i(z) \geq 0.$$

So there exists Ω with $\mu_i(Z_{i,L} \in \Omega) = 1$ such that $\frac{q_{i,L}(z|x)}{q_{i,L}(z|x')} \leq e^{\alpha}$ for any $z \in \Omega$.

□

The lemma here justifies the use of the privacy mechanism presented in this section.

Lemma 21. *To each random variable X_i of the sample set $(X_1, \ldots, X_n)$, we associate the vector $Z_{i,L} = (Z_{i,L,k})_{k\in\{0,\ldots,L-1\}}$. The random vectors $(Z_{1,L}, \ldots, Z_{n,L})$ are non-interactive α-local differentially private views of the samples $(X_1, \ldots, X_n)$. Namely, they satisfy the condition in Equation (3.1).*

The proof can also be found in Butucea et al. (2020) (see Proposition 3.1). We recall here the main arguments for the sake of completeness.

Proof. The random vectors $(Z_{i,L})_{1\leq i\leq n}$ are i.i.d. by definition. For any x_i, x_i' in $[0,1]$, for any $z_i \in \mathbb{R}^L$,

$$\begin{aligned}\frac{q_{i,L}(z_i|x_i)}{q_{i,L}(z_i|x_i')} &= \prod_{k=0}^{L-1} \exp\left[\sqrt{2}\frac{|z_{i,k} - \varphi_{L,k}(x_i')| - |z_{i,k} - \varphi_{L,k}(x_i)|}{\sigma_L}\right] \\ &\leq \exp\left[\sum_{k=0}^{L-1} \frac{\sqrt{2}}{\sigma_L}\left(|\varphi_{L,k}(x_i')| + |\varphi_{L,k}(x_i)|\right)\right].\end{aligned}$$

Since $\varphi_{L,k}(x_i) \neq 0$ for a single value of $k \in \{0, \ldots, L-1\}$, we get

$$\frac{q_{i,L}(z_i|x_i)}{q_{i,L}(z_i|x_i')} \leq \exp\left[\frac{2\sqrt{2}\|\varphi_{L,k}\|_{\infty}}{\sigma_L}\right] \leq e^{\alpha},$$

by definition of σ_L, which concludes the proof by application of Lemma 20. □

3.4.2 Definition of the test

Let f_0 be some fixed density in $\mathbb{L}_2([0,1])$. Our aim is now to define a testing procedure for identity testing from the observation of the vectors $(Z_1, \dots, Z_n)$. Our test statistic $\hat{T}_L$ is defined as

$$\hat{T}_L = \frac{1}{n(n-1)} \sum_{i \neq l=1}^{n} \sum_{k=0}^{L-1} \left(Z_{i,L,k} - \alpha^0_{L,k}\right) \left(Z_{l,L,k} - \alpha^0_{L,k}\right), \tag{3.11}$$

where $\alpha^0_{L,k} = \int_0^1 \phi_{L,k}(x) f_0(x) dx$.

We consider the test function

$$\varphi_{L,\gamma,Q} : (Z_1, \dots, Z_n) \mapsto \mathbb{1}\{\hat{T}_L > t^0_L(1-\gamma)\}, \tag{3.12}$$

where $t^0_L(1-\gamma)$ denotes the $(1-\gamma)$-quantile of $\hat{T}_L$ under $\mathcal{H}_0$. Note that this quantile can be estimated by simulations, under the hypothesis $f = f_0$. We can indeed simulate the vector $(Z_1, \dots, Z_n)$ if the density of $(X_1, \dots, X_n)$ is assumed to be f_0. Hence the test rejects the null hypothesis $\mathcal{H}_0$ if

$$\hat{T}_L > t^0_L(1-\gamma).$$

The test is of level γ by definition of the threshold.

Remark 33.

- *In a similar way as in Fromont and Laurent (2006b), the test is based on an estimation of the quantity $\|f - f_0\|_2^2$. Note indeed that $\hat{T}_L$ is an unbiased estimator of $\|\Pi_{S_L}(f - f_0)\|_2^2$, where Π_{S_L} denotes the orthogonal projection in $\mathbb{L}_2([0,1])$ onto the space generated by the functions $(\phi_{L,k}, k \in \{0, \dots, L-1\})$. In the discrete case, f and f_0 belong to S_L and $\Pi_{S_L}(f - f_0) = f - f_0$. In this case,*

$$\|\Pi_{S_L}(f - f_0)\|_2^2 = \|f - f_0\|_2^2 = d \sum_{k=1}^{d} (q_k - p_k)^2.$$

- *Note that, in the discrete case, we obtain the following expression for the test statistic*

$$\hat{T}_d = \frac{d}{n(n-1)} \sum_{k=1}^{d} \sum_{i \neq l=1}^{n} \left(\mathbb{1}\{\widetilde{X_i} = k\} - p_k\right) \left(\mathbb{1}\{\widetilde{X_l} = k\} - p_k\right). \tag{3.13}$$

It is interesting to compare this expression with the χ^2 statistics, which can be written as

$$\sum_{k=1}^{d} \sum_{i,l=1}^{n} \frac{\left(\mathbb{1}\{\widetilde{X_i} = k\} - p_k\right) \left(\mathbb{1}\{\widetilde{X_l} = k\} - p_k\right)}{np_k}.$$

Hence, besides the normalization of each term in the sum by p_k in the χ^2 test, the main difference lies in the fact that we remove the diagonal terms (corresponding to $i = l$) in our test statistics.

In the next section, we provide non-asymptotic theoretical results for the power of this test.

3.4.3 Upper bound for the second kind error of the test

We first provide an upper bound for the second kind error of our test and privacy channel in a general setting.

Theorem 28. *Let $(X_1, \ldots, X_n)$ be i.i.d. with common density f on $[0,1]$. Let f_0 be some given density on $[0,1]$. We assume that f and f_0 belong to $\mathbb{L}_2([0,1])$. From the observation of the random vectors $(Z_1, \ldots, Z_n)$ defined by Equation (3.10), for a given $\alpha > 0$, we test the hypotheses*

$$\mathcal{H}_0 : f = f_0, \qquad \textit{versus} \qquad \mathcal{H}_1 : f \neq f_0.$$

We consider the test $\varphi_{L,\gamma,Q}$ defined by Equation (3.12) with $\hat{T}_L$ defined in Equation (3.11). The test is obviously of level γ by definition of the threshold $t_L^0(1-\gamma)$, namely we have

$$\mathbb{P}_{Q_{f_0}^n}\left(\hat{T}_L \geq t_L^0(1-\gamma)\right) \leq \gamma.$$

Under the assumption that

$$\|\Pi_{S_L}(f - f_0)\|_2^2 \geq \sqrt{\mathbb{V}_{Q_{f_0}^n}(\hat{T}_L)/\gamma} + \sqrt{\mathbb{V}_{Q_f^n}\left(\hat{T}_L\right)/\beta}, \tag{3.14}$$

the second kind error of the test is controlled by β, namely we have

$$\mathbb{P}_{Q_f^n}\left(\hat{T}_L \leq t_L^0(1-\gamma)\right) \leq \beta. \tag{3.15}$$

Moreover, we have

$$\mathbb{V}_{Q_f^n}\left(\hat{T}_L\right) \leq C\left[\frac{(\sqrt{L}\|f\|_2 + \sigma_L^2)}{n}\|\Pi_{S_L}(f - f_0)\|_2^2 + \frac{(\|f\|_2^2 + \sigma_L^4)L}{n^2}\right]. \tag{3.16}$$

We give here a sketch of proof of Theorem 28. The complete proof of this result is given in Section 3.7.2. Note that it is not fundamentally different from non-private proofs given in Fromont and Laurent (2006b).

Sketch of proof. We want to establish a condition on $f - f_0$, under which the second kind error of the test is controlled by β. Denoting by $t_L(\beta)$ the β-quantile of $\hat{T}_L$ under $\mathbb{P}_{Q_f^n}$, the condition in Equation (3.15) holds as soon as $t_L^0(1-\gamma) \leq t_L(\beta)$. Hence, we provide an upper bound for $t_L^0(1-\gamma)$ and a lower bound for $t_L(\beta)$. By Chebyshev's inequality, we obtain that on the one hand,

$$t_L^0(1-\gamma) \leq \sqrt{\mathbb{V}_{Q_{f_0}^n}(\hat{T}_L)/\gamma}, \tag{3.17}$$

and on the other hand,

$$\|\Pi_{S_L}(f - f_0)\|_2^2 - \sqrt{\mathbb{V}_{Q_f^n}\left(\hat{T}_L\right)/\beta} \leq t_L(\beta). \tag{3.18}$$

We deduce from the inequalities in Equations (3.17) and (3.18) that Equation (3.15) holds as soon as

$$\|\Pi_{S_L}(f - f_0)\|_2^2 \geq \sqrt{\mathbb{V}_{Q_{f_0}^n}(\hat{T}_L)/\gamma} + \sqrt{\mathbb{V}_{Q_f^n}\left(\hat{T}_L\right)/\beta}.$$

The main ingredient to control the variance terms is a control of the variance for U-statistics of order two which relies on Hoeffding's decomposition – see e.g. Serfling (2009) Lemma A p. 183. The proof is given in Section 3.7.2.

We obtain the following corollary of Theorem 28. It states a result that will be used in order to obtain an upper bound on the minimax rate both in the discrete and the continuous cases.

Corollary 10. *Under the same assumptions as in Theorem 28, we obtain that Equation (3.15) holds, that is, the second kind error of the test is controlled by β provided that*

$$\|\Pi_{S_L}(f - f_0)\|_2^2 \geq C(\gamma, \beta)\frac{(\|f\|_2 + \|f_0\|_2 + \sigma_L^2)\sqrt{L}}{n}. \tag{3.19}$$

In the next sections, we derive from this result upper bounds for the minimax separation rate over Besov balls in the continuous case, and conditions on the ℓ_2-distance between p and q to obtain a prescribed power for the test in the discrete case.

3.4.4 Upper bound for the separation distance in the discrete case

The following theorem provides a sufficient condition on the separation distance between the probability vectors p and q for both error kinds of the test to be controlled by γ and β, respectively. This sufficient condition corresponds to an upper bound on the minimax rate $\rho^*_\beta(f_0, H_1, T^{(\alpha)}_{\gamma,n})/d^{1/2}$ in the discrete case.

Theorem 29. *Let $p = (p_1, p_2, \ldots, p_d)$ be some given probability vector. We also set $(X_1, \ldots, X_n)$ as i.i.d. random variables with values in the finite set $\{1, 2 \ldots, d\}$ and with common distribution defined by the probability vector $q = (q_1, q_2, \ldots, q_d)$.*

From the observation of the random vectors $(Z_1, \ldots, Z_n)$ defined by Equation (3.10) for a given $\alpha > 0$ with $L = d$, we want to test the hypotheses

$$\mathcal{H}_0 : p = q, \qquad \textit{versus} \qquad \mathcal{H}_1 : p \neq q.$$

We consider the test $\varphi_{d,\gamma,Q}$ defined by Equation (3.12), which has a first kind error of γ. The second kind error of the test is controlled by β, provided that

$$\sqrt{\sum_{i=1}^{d}(q_i - p_i)^2} \geq C(\gamma, \beta)\frac{d^{-1/4}}{n^{1/2}}\left(d^{1/4}\left[\left(\sum_{k=1}^{d} p_k^2\right)^{1/4} + n^{-1/2}\right] + d^{1/2}\alpha^{-1}\right).$$

Since $\sum_{k=1}^{d} p_k^2 \leq 1$, the second kind error of the test is controlled by β, provided that

$$\sqrt{\sum_{i=1}^{d}(q_i - p_i)^2} \geq C(\gamma, \beta)n^{-1/2}(1 \vee [d^{1/4}\alpha^{-1}]). \tag{3.20}$$

Remark 34. *Equation (3.20) displays a rate that is optimal in d, n, when α is smaller than $d^{1/4}$. Besides, the rate in α matches the lower bound asymptotically when α converges to 0. The upper bound presented in Theorem 1 from Berrett and Butucea (2020) tackles the case when α is smaller than 1 and they find the same rate as ours in their Corollary 2. They present an additional test statistic in order to refine their rates when p is not a uniform vector.*

Corollary 11. *We assume that there exists an absolute constant κ such that*

$$d\sum_{k=1}^{d} p_k^2 \leq \kappa. \tag{3.21}$$

Then the second kind error of the test is controlled by β, provided that

$$\sqrt{\sum_i (q_i - p_i)^2} \geq C(\gamma, \beta, \kappa)\left[\left(d^{-1/4}n^{-1/2}\right) \vee \left(d^{1/4}n^{-1/2}\alpha^{-1}\right) \vee n^{-1}\right]. \tag{3.22}$$

Remark 35. *If we also assume the bound on $d\sum_{k=1}^{d} p_k^2$ as expressed in Equation (3.21), we find optimal rates in d, n if $n \geq (\alpha^2 d^{-1/2}) \wedge d^{1/2}$. The assumption in Equation (3.21) in Lemma 11 is equivalent to assuming that the function f_0 defined in Section 3.2.2 belongs to $\mathbb{L}^2([0,1])$. It restricts p to vectors that are close to being uniform. This coincides with the lower bound on the rate found when f_0 is a uniform density.*

3.4.5 Upper bound for the minimax separation rate over Besov balls

We provide an upper bound on the uniform separation rate for our test and privacy channel over Besov balls in Theorem 30.

Theorem 30. *Let $(X_1, \ldots, X_n)$ be i.i.d. with common density f on $[0,1]$. Let f_0 be some given density on $[0,1]$. We assume that f and f_0 belong to $\mathbb{L}_2([0,1])$.*

We observe the random vectors $(Z_1, \ldots, Z_n)$ defined by Equation (3.10) for a given $\alpha > 0$ with the following value for L : we assume that $L = L^*$, where $L^* = 2^{J^*}$, and J^* is the smallest integer J such that $2^J \geq (n\alpha^2)^{2/(4s+3)} \wedge n^{2/(4s+1)}$.

We want to test the hypotheses

$$\mathcal{H}_0 : f = f_0, \qquad \textit{versus} \qquad \mathcal{H}_1 : f \neq f_0.$$

We consider the test $\varphi_{L^*,\gamma,Q}$ defined by Equation (3.12). The uniform separation rate, defined by Equation (3.6), of the test $\varphi_{L^*,\gamma,Q}$ over $\mathcal{B}_{s,2,\infty}(R)$ defined by Equation (3.5) satisfies for all $n \in \mathbb{N}^*$, $R > 0$, $\alpha \geq 1/\sqrt{n}$, $(\gamma, \beta) \in (0,1)^2$ such that $\gamma + \beta < 1$

$$\begin{aligned}&\rho_\beta(f_0, H_{1,s}, \varphi_{L^*,\gamma,Q}(Z_1, \ldots, Z_n))\\ &\leq C(s, R, \|f_0\|_2, \gamma, \beta)\left[(n\alpha^2)^{-2s/(4s+3)} \vee n^{-2s/(4s+1)}\right].\end{aligned}$$

The proof of this result is in Section 3.7.2.

Remark 36.

- *When the sample set $(X_1, \ldots, X_n)$ is directly observed, Fromont and Laurent (2006b) propose a testing procedure with uniform separation rate over the set $\mathcal{B}_{s,2,\infty}(R)$ controlled by*

$$C(s, R, \gamma, \beta)n^{-2s/(4s+1)},$$

which is an optimal result, as proved in Ingster (1993). Hence we obtain here a loss in the uniform separation rate, due to the fact that we only observe α-differentially private views of the original sample. This loss occurs when $\alpha \leq n^{1/(4s+1)}$. Otherwise, we get the same rate as when the original sample is observed. Comparing this result with the lower bound from Section 3.3, we conclude that the rate is optimal.

- *Finally, having $\alpha < 1/\sqrt{n}$ represents an extreme case, where the sample size is really low in conjunction with a very strict privacy condition. In such a range of α, J^* is taken equal to 0, but this does not lead to optimal rates.*

The test proposed in Theorem 30 depends on the smoothness parameter s of the Besov ball $\mathcal{B}_{s,2,\infty}(R)$ via the parameter J^*. In a second step, we will propose a test which is adaptive to the smoothness parameter s. Namely, in Section 3.5, we construct an aggregated testing procedure, which is independent of the smoothness parameter and achieves the minimax separation rates established in Equation (3.8) over a wide range of Besov balls simultaneously, up to a logarithmic term.

3.5 Adaptive tests

In Section 3.4, we have defined in the continuous case a testing procedure which depends on the parameter J. The performances of the test depend on this parameter. We have optimized the choice of J to obtain the smallest possible upper bound for the separation rate over the set $\mathcal{B}_{s,2,\infty}(R)$. Nevertheless, the test is not adaptive since this optimal choice of J depends on the smoothness parameter s.

In order to obtain adaptive procedure, we propose, as in Fromont and Laurent (2006b) to aggregate a collection of tests. For this, we introduce the set

$$\mathcal{J} = \left\{J \in \mathbb{N}, 2^J \leq n^2\right\}$$

and the aggregated procedure will be based on the collection of test statistics $(\hat{T}_{2^J}, J \in \mathcal{J})$ defined by (3.11).

In Theorem 30, the testing procedure is based on the observation of the random vectors $(Z_1, \ldots, Z_n)$ defined by Equation (3.10) with $L = 2^{J^*}$ for the optimized value of J^*. Hence, the private views of the original sample depend on the unknown parameter s. In order to build the aggregated procedure, we can no more use the optimized value J^* of J and we need to observe the random vectors $(Z_1, \ldots, Z_n)$ for all $J \in \mathcal{J}$. In order to guarantee the α-local differential privacy, we have to increase slightly the variance of the Laplace perturbation. The privacy mechanism is specified in the following lemma.

Lemma 22. *We consider the set $\mathcal{J} = \left\{J \in \mathbb{N}, 2^J \leq n^2\right\}$. We define, for all $i \in \{1, \ldots, n\}$, for all $J \in \mathcal{J}$, the vector $\tilde{Z}_{i,2^J} = (\tilde{Z}_{i,2^J,k})_{k\in\{0,\ldots,2^J-1\}}$, by*

$$\forall k \in \left\{0, \ldots, 2^J - 1\right\}, \ \tilde{Z}_{i,2^J,k} = \phi_{2^J,k}(X_i) + \tilde{\sigma}_{2^J} W_{i,2^J,k}, \tag{3.23}$$

where $(W_{i,2^J,k})_{1\leq i\leq n, k\in\{0,\ldots,2^J-1\}}$ are i.i.d. Laplace distributed random variables with variance 1 and

$$\tilde{\sigma}_{2^J} = 2\sqrt{2}|\mathcal{J}|\frac{2^{J/2}}{\alpha}.$$

For all $1 \leq i \leq n$, we define the random vector $\tilde{Z}_i = (\tilde{Z}_{i,2^J}, J \in \mathcal{J})$. The random vectors $(\tilde{Z}_i, 1 \leq i \leq n)$ are non-interactive α-local differentially private views of the samples $(X_1, \ldots, X_n)$. Namely, they satisfy the condition in Equation (3.1).

Proof. The random vectors $(\tilde{Z}_i)_{1\leq i\leq n}$ are i.i.d. by definition. Let us denote by $\tilde{q}_i(\cdot|x_i)$ the density of the vector $\tilde{Z}_i$, conditionally to $X_i = x_i$. For any x_i, x_i' in $[0,1]$, for any

$z_i \in \mathbb{R}^{\sum_{J\in\mathcal{J}} 2^J}$,

$$\frac{\tilde{q}_i(z_i|x_i)}{\tilde{q}_i(z_i|x_i')} = \prod_{J\in\mathcal{J}}\prod_{k=0}^{2^J-1} \exp\left[\sqrt{2}\frac{\left|z_{i,k}-\phi_{2^J,k}(x_i')\right| - \left|z_{i,k}-\phi_{2^J,k}(x_i)\right|}{\tilde{\sigma}_{2^J}}\right]$$
$$\leq \exp\left[\sum_{J\in\mathcal{J}}\sum_{k=0}^{2^J-1}\frac{\sqrt{2}}{\tilde{\sigma}_{2^J}}\left(\left|\phi_{2^J,k}(x_i')\right| + \left|\phi_{2^J,k}(x_i)\right|\right)\right].$$

Since $\phi_{2^J,k}(x_i) \neq 0$ for a single value of $k \in \{0,\dots,2^J-1\}$, we get

$$\frac{\tilde{q}_i(z_i|x_i)}{\tilde{q}_i(z_i|x_i')} \leq \exp\left[2\sqrt{2}\sum_{J\in\mathcal{J}}\frac{\|\phi_{2^J,k}\|_\infty}{\tilde{\sigma}_{2^J}}\right] \leq e^\alpha,$$

by definition of $\tilde{\sigma}_2^J$, which concludes the proof by application of Lemma 20. □

Note that $|\mathcal{J}| \leq 1 + 2\log_2(n)$, hence we will have a logarithmic loss for the separation rates due to the privacy condition for the aggregated procedure.

Let us now define the adaptive test. We set, for all $J \in \mathcal{J}$,

$$\widetilde{T}_J = \frac{1}{n(n-1)}\sum_{i\neq l=1}^{n}\sum_{k=0}^{2^J-1}\left(\tilde{Z}_{i,2^J,k} - \alpha^0_{2^J,k}\right)\left(\tilde{Z}_{l,2^J,k} - \alpha^0_{2^J,k}\right). \tag{3.24}$$

For a given level $\gamma \in (0,1)$, the aggregated testing procedure rejects the hypothesis $\mathcal{H}_0 : f = f_0$ if

$$\exists J \in \mathcal{J},\ \tilde{T}_J > \tilde{t}^0_J(1-u_\gamma),$$

where u_γ is defined by

$$u_\gamma = \sup\left\{u \in (0,1), \mathbb{P}_{Q^n_{f_0}}\left(\sup_{J\in\mathcal{J}}\left(\tilde{T}_J - \tilde{t}^0_J(1-u_\gamma)\right) > 0\right) \leq \gamma\right\} \tag{3.25}$$

and $\tilde{t}^0_J(1-u_\gamma)$ denotes the $1-u_\gamma$ quantile of $\tilde{T}_J$ under $\mathcal{H}_0$. Hence u_γ is the least conservative choice leading to a γ-level test. We easily notice that $u_\gamma \geq \gamma/|\mathcal{J}|$. Indeed,

$$\mathbb{P}_{Q^n_{f_0}}\left(\sup_{J\in\mathcal{J}}\left(\widetilde{T}_J - \tilde{t}^0_J(1-\gamma/|\mathcal{J}|)\right) > 0\right) \leq \sum_{J\in\mathcal{J}}\mathbb{P}_{Q^n_{f_0}}\left(\widetilde{T}_J > \tilde{t}^0_J(1-\gamma/|\mathcal{J}|)\right)$$
$$\leq \sum_{J\in\mathcal{J}}\gamma/|\mathcal{J}| \leq \gamma.$$

Let us now consider the second kind error for the aggregated test, which is the probability to accept the null hypothesis $\mathcal{H}_0$, although the alternative hypothesis $\mathcal{H}_1$ holds. This quantity is upper bounded by the smallest second kind error of the tests of the collection, at the price that γ has been replaced by u_γ. Indeed,

$$\mathbb{P}_{Q^n_f}\left(\sup_{J\in\mathcal{J}}\left(\tilde{T}_J - \tilde{t}^0_J(1-u_\gamma)\right) \leq 0\right) = \mathbb{P}_{Q^n_f}\left(\cap_{J\in\mathcal{J}}\left(\tilde{T}_J \leq \tilde{t}^0_J(1-u_\gamma)\right)\right)$$
$$\leq \inf_{J\in\mathcal{J}}\mathbb{P}_{Q^n_f}\left(\tilde{T}_J \leq \tilde{t}^0_J(1-u_\gamma)\right). \tag{3.26}$$

We obtain the following theorem for the aggregated procedure.

Theorem 31. *Let $(X_1, \dots, X_n)$ be i.i.d. with common density f in $\mathbb{L}_2([0,1])$. Let f_0 be some given density in $\mathbb{L}_2([0,1])$. From the observation of the random vectors $(\tilde{Z}_i, 1 \le i \le n)$ defined in Lemma 22 for a given $\alpha > 0$, we want to test the hypotheses*

$$\mathcal{H}_0 : f = f_0, \qquad \textit{versus} \qquad \mathcal{H}_1 : f \neq f_0.$$

We assume that $n\alpha^2/\log^{5/2}(n) \ge 1$. We consider the set $\mathcal{J} = \left\{J \in \mathbb{N}, 2^J \le n^2\right\}$ and the aggregated test

$$\varphi^{\mathcal{J}}_{\gamma,Q} : (Z_1, \dots, Z_n) \mapsto \mathbb{1}\{\sup_{J \in \mathcal{J}} \left(\widetilde{T}_J - \tilde{t}^0_J(1-u_\gamma)\right) > 0\}$$

where $\widetilde{T}_J$ is defined by Equation (3.24) and u_γ by Equation (3.25).

The uniform separation rate, defined by Equation (3.6), of the test $\varphi^{\mathcal{J}}_{\gamma,Q}$ over the set $\mathcal{B}_{s,2,\infty}(R)$ defined by Equation (3.5) satisfies for all $n \in \mathbb{N}^*$, $s > 0$, $R > 0$, $\alpha > 0$, $(\gamma, \beta) \in (0,1)^2$ such that $\gamma + \beta < 1$,

$$\rho_\beta(f_0, H_{1,s}, \varphi^{\mathcal{J}}_{\gamma,Q}(Z_1, \dots, Z_n)) \le C(\|f_0\|_2, R, \gamma, \beta)\left[(n\alpha^2/\log^{5/2}(n))^{-2s/(4s+3)} \vee (n/\sqrt{\log(n)})^{-2s/(4s+1)}\right],$$

The proof of this result is in Section 3.7.3. We compare this result with the rates obtained in Theorem 30, which has been proved to be optimal. Here, we incur a logarithmic loss due to the adaptation. We recall that in the non-private setting, the separation rates obtained by Ingster (2000b) and Fromont and Laurent (2006b) for adaptive procedures over Besov balls was $\left(n/\sqrt{\log\log(n)}\right)^{-2s/(4s+1)}$. This result was proved to be optimal for adaptive tests in Ingster (2000b). In their paper, the log-log term is obtained from exponential inequalities for U-statistics involved in the testing procedure under the null hypothesis. In our setting, obtaining exponential inequalities is not trivial due to the Laplace noise. That is why our logarithmic loss originates from a simple upper bound on the variance of our test statistic under the null. The optimality of the adaptive rates presented in Theorem 31 remains an open question.

3.6 Discussion

Our study of minimax testing rates is in line with Ingster's work and we focus on separation rates in $\mathbb{L}_2$-norm for identity testing under local differential privacy. We construct a unified setting in order to tackle both discrete and continuous distributions. In the continuous case, we provide the first minimax optimal test and local differentially private channel for the problem of identity testing over Besov balls. This result also holds for multinomial distributions. Besides, in the continuous case, the test and channel remain optimal up to a log factor even if the smoothness parameter is unknown. Among our technical contributions, it is to note that we use a proof technique in the lower bound that does not involve Theorem 1 from Duchi, Jordan, and Wainwright (2013c). The minimax separation rate turns out to suffer from a polynomial degradation in the private case. However, we point out an elbow effect, where the rate is the same as the usual case up to some constant factor if α is large enough. Simultaneously and independently, Berrett and Butucea (2020) present minimax testing rates for the ℓ_1 and ℓ_2 norms in the discrete case. We define Besov balls using Haar wavelets, which are equivalent to Besov balls defined using moduli of smoothness when $s \le 1$. In

order for the equivalence to hold for any s, it is possible to define Besov balls using Daubechies wavelets instead. In the proof of our lower bound, we use the disjoint support property of the Haar wavelets, but this can be circumvented by taking fewer wavelets in the definition of the prior distributions. A more critical assumption is that $\phi_{L,k}^2 = \sqrt{L}\phi_{L,k}$. Future possible works could extend our results to larger Besov classes and study the optimality of the adaptive procedure. Finally, our bounds match when f_0 is a uniform density, and matching bounds for any f_0 remain to be proved under local differential privacy.

3.7 Proof of the results

3.7.1 Lower bound: proof of Theorem 27

An initial version of this proof has been presented in a preprint of Lam-Weil, Laurent, and Loubes (2020), which was then improved upon by Butucea, Rohde, and Steinberger (2020) in order to find the matching rate in α and to account for different channels Q_i for each initial observation X_i. The proof remains fundamentally the same, however. In line with the rest of the chapter, both the discrete and the continuous cases are treated in one unified setting.

In this section, $f_0 = \mathbb{1}_{[0,1]}$.

3.7.1.1 Preliminary results

The following lemma sheds light on the equivalence between the local differential privacy condition and a similar condition on the density of the channel.

Lemma 23. *Let $Q \in \mathcal{Q}_\alpha$ be an α-private channel and $i \leq n$. Let X_i be a random variable with density $f \in \mathbb{L}_2([0,1])$ with respect to the Lebesgue measure. Then there exists a probability measure with respect to which $Q_i(\cdot|x)$ is absolutely continuous for any $x \in [0,1]$.*

Proof. Let $\mu_i = \int_{[0,1]} Q_i(\cdot|x)f(x)dx$. Let $S \in \mathcal{Z}_i$ such that $\mu_i(S) = 0$. Then since $Q_i(S|x) \geq 0$ for any x, there exists x such that $Q_i(S|x) = 0$. Now by α-local differential privacy, $Q_i(S|x) = 0$ for any x. □

For the sake of completeness, we prove the following classical inequality between the total variation distance and the chi-squared distance. It will be used in order to reduce the study of the distance between the distributions to that of an expected squared likelihood ratio.

Lemma 24.

$$d_{TV}(\mathbb{P}_{Q^n_{\nu_\rho}}, \mathbb{P}_{Q^n_{f_0}}) \leq \frac{1}{2}\left(\mathbb{E}_{Q^n_{f_0}}\left[L^2_{Q^n_{\nu_\rho}}(Z_1,\dots,Z_n) - 1\right]\right)^{1/2},$$

where $L_{Q^n_{\nu_\rho}}(Z_1,\dots,Z_n)$ is the likelihood ratio between $Q^n_{\nu_\rho}$ and $Q^n_{f_0}$.

Proof. We have

$$\begin{aligned} d_{TV}(\mathbb{P}_{Q^n_{\nu_\rho}}, \mathbb{P}_{Q^n_{f_0}}) = \frac{1}{2}\int \left|L_{Q^n_{\nu_\rho}} - 1\right| d\mathbb{P}_{Q^n_{f_0}} &= \frac{1}{2}\mathbb{E}_{Q^n_{f_0}}\left[\left|L_{Q^n_{\nu_\rho}}(Z_1,\dots,Z_n) - 1\right|\right] \\ &\leq \frac{1}{2}\left(\mathbb{E}_{Q^n_{f_0}}\left[L^2_{Q^n_{\nu_\rho}}(Z_1,\dots,Z_n) - 1\right]\right)^{1/2}, \end{aligned}$$

by Cauchy-Schwarz inequality and since $\mathbb{E}_{Q^n_{f_0}}\left(L_{Q^n_{\nu_\rho}}(Z_1,\dots,Z_n)\right) = 1$. □

The following two lemmas can be interpreted as data processing inequalities. Lemma 25 describes the contraction of the total variation distance by a stochastic channel.

Lemma 25. *Let $\mathbb{P}_f, \mathbb{P}_g$ be probability measures over the sample space $[0,1]$ with respective densities f and g with respect to the Lebesgue measure. Let Q be a stochastic channel. Then*

$$d_{TV}(\mathbb{P}_f, \mathbb{P}_g) \geq d_{TV}(\mathbb{P}_{Q_f}, \mathbb{P}_{Q_g}).$$

Proof. For any measurable set S,

$$\begin{aligned}\int_{[0,1]} Q(S|x)(f(x)-g(x))dx &= \int_{[0,1]} Q(S|x)(f(x)-g(x))\mathbb{1}\{f-g\geq 0\}(x)dx \\ &\quad + \int_{[0,1]} Q(S|x)(f(x)-g(x))\mathbb{1}\{f-g< 0\}(x)dx.\end{aligned}$$

Now, since $0 \leq Q(S|x) \leq 1$ for any measurable set S and $x \in [0,1]$,

$$\begin{aligned}0 &\leq \int_{[0,1]} Q(S|x)\ (f(x)-g(x))\mathbb{1}\{f-g\geq 0\}(x)dx \\ &\leq \int_{[0,1]} (f(x)-g(x))\mathbb{1}\{f-g> 0\}(x)dx.\end{aligned}$$

and

$$\begin{aligned}0 &\geq \int_{[0,1]} Q(S|x)\ (f(x)-g(x))\mathbb{1}\{f-g< 0\}(x)dx \\ &\geq \int_{[0,1]} (f(x)-g(x))\mathbb{1}\{f-g< 0\}(x)dx.\end{aligned}$$

So for any measurable set S,

$$\begin{aligned}&\int_{[0,1]} (f(x)-g(x))\mathbb{1}\{f-g< 0\}(x)dx \\ &\qquad \leq \int_{[0,1]} Q(S|x)(f(x)-g(x))dx \\ &\qquad\qquad \leq \int_{[0,1]} (f(x)-g(x))\mathbb{1}\{f-g> 0\}(x)dx.\end{aligned}$$

That is, for any measurable set S

$$\begin{aligned}\left|\int_{[0,1]} Q(S|x)(f(x)-g(x))dx\right| &\leq \left|\int_{[0,1]} (f(x)-g(x))\mathbb{1}\{f-g> 0\}(x)dx\right| \\ &\quad \vee \left|\int_{[0,1]} (f(x)-g(x))\mathbb{1}\{f-g< 0\}(x)dx\right| \\ &= \sup_A \left|\int_A (f(x)-g(x))dx\right| = d_{TV}(\mathbb{P}_f, \mathbb{P}_g).\end{aligned}$$

□

3.7.1.2 Definition of prior distributions

By Lemma 23, let $Q \in \mathcal{Q}_\alpha$ be a non-interactive α-private channel with marginal conditional densities $q_i(z_i|x_i)$ with respect to probability measure μ_i over the respective sample space $\tilde{\Omega}_i$ for any $1 \le i \le n$. In the discrete case, we assume that p is a uniform probability vector. By Equation (3.2), we can consider the associated uniform density on $[0,1]$. So in both the continuous and the discrete cases, we end up considering a uniform density f_0 over $[0,1]$. Let $\tilde{f}_{0,i}(z_i) = \int_0^1 q_i(z_i|x) f_0(x) dx = \int_0^1 q_i(z_i|x) dx$ with the convention $0/0 = 0$. Let $\mu_i = \int_{[0,1]} Q_i(\cdot|x) f_0(x) dx$ and $K_i : \mathbb{L}_2([0,1]) \to \mathbb{L}_2(\tilde{\Omega}_i, d\mu_i)$ such that

$$K_i f = \int_0^1 q_i(\cdot|x) f(x) \frac{dx}{\sqrt{\tilde{f}_{0,i}(\cdot)}}.$$

Let K_i^* denote the adjoint of K_i. Then $K_i^* K_i$ is a symmetric integral operator with kernel

$$F_i(x,y) = \int \frac{q_i(z_i|x) q_i(z_i|y)}{\tilde{f}_{0,i}(z_i)} d\mu_i(z_i). \tag{3.27}$$

And by Fubini's theorem, for any $f \in \mathbb{L}_2([0,1])$:

$$K_i^* K_i f(\cdot) = \int_0^1 F_i(\cdot, y) f(y) dy.$$

Note that f_0 is an eigenfunction of $K_i^* K_i$ associated to the eigenvalue $\lambda_{0,i} = 1$ for all $1 \le i \le n$. Let

$$K = \sum_{i=1}^n K_i^* K_i / n,$$

which is symmetric and positive semidefinite, and $\lambda_0 = 1$ is an eigenvalue associated with f_0. It is an integral operator with kernel

$$F(x,y) = \sum_i F_i(x,y)/n.$$

We denote by ψ the difference of indicator functions: $\psi = \mathbb{1}_{[0,1/2)} - \mathbb{1}_{[1/2,1)}$ and for all integer $L \ge 1$, we set, for all $k \in \{0, \dots, L-1\}$, for all $x \in [0,1)$,

$$\psi_k(x) = \sqrt{L} \psi(Lx - k).$$

The integer L will be taken as $L = 2^J$ for some $J \ge 0$ in the continuous case, and we choose $L = d/2$ in the discrete case (we assume that d is even). We denote by V the linear subspace of $\mathbb{L}^2([0,1])$ generated by the functions $(f_0, \psi_k, k \in \{0, 1, \dots L-1\})$. Then we complete (f_0) into an orthogonal basis $(f_0, u_i)_{1 \le i \le L}$ of V with eigenfunctions of K such that $\int u_i(x) dx = 0$ by orthogonality with f_0 and $\|u_i\|_2 = 1$. We write the corresponding eigenvalues λ_i.

Let $z_\alpha = e^{2\alpha} - e^{-2\alpha} \le 2$ for any $\alpha \in (0,1]$. Let $\tilde{\lambda}_k = (\lambda_k / z_\alpha^2) \vee L^{-1} \ge L^{-1}$. Let

$$f_\eta(x) = f_0(x) + \varepsilon \sum_{j=1}^L \eta_j \tilde{\lambda}_j^{-1/2} u_j(x),$$

where $\eta \in \{-1,1\}^L$. For all $i \in \{1,\dots,L\}$, $u_i \in \mathrm{Span}(\psi_k, k \in \{0,1,\dots,L-1\})$, hence we write

$$u_i = \sum_{k=0}^{L-1} a_{i,k}\psi_k.$$

Then

$$f_\eta(x) = f_0(x) + \varepsilon \sum_{j=1}^{L}\sum_{k=0}^{L-1} \eta_j a_{j,k} \tilde{\lambda}_j^{-1/2} \psi_k(x).$$

We define ν_ρ as the uniform probability measure over $\{f_\eta : \eta \in \{-1,1\}^L\}$. Now, we can identify the distance between f_η and f_0. Let $l = \sum_{i=1}^{L} \mathbb{1}\{z_\alpha^{-2}\lambda_i > L^{-1}\}$. By definition and orthonormality of $(u_i)_{1\le i\le L}$, for any $\eta \in \{-1,1\}^L$

$$\begin{aligned}
\|f_\eta - f_0\|_2 &= \varepsilon \sqrt{\sum_{i=1}^{L} \tilde{\lambda}_i^{-1} \|u_i\|_2^2} = \varepsilon \sqrt{\sum_{i=1}^{L} \tilde{\lambda}_i^{-1}} \\
&= \varepsilon \sqrt{\sum z_\alpha^2 \lambda_i^{-1} \mathbb{1}\{z_\alpha^{-2}\lambda_i > L^{-1}\} + L \sum \mathbb{1}\{z_\alpha^{-2}\lambda_i \le L^{-1}\}} \\
&\ge \varepsilon \sqrt{z_\alpha^2 l^2 (\sum_i \lambda_i \mathbb{1}\{z_\alpha^{-2}\lambda_i > L^{-1}\})^{-1} + L(L-l)}, \qquad (3.28)
\end{aligned}$$

by Cauchy-Schwarz inequality.

So let us provide guarantees on the singular values in order to determine sufficient conditions for ε to lead to a lower bound on $\rho_n^*(\varphi_{\gamma,Q}, \mathcal{C}, \beta, f_0)$, depending on $\mathcal{C}$.

3.7.1.3 Obtaining the inequalities on the eigenvalues

Lemma 26. *Let K be defined as in Section 3.7.1.2 and $(\lambda_i^2)_{0\le i\le L}$ its eigenvalues associated with the orthonormal basis $(f_0, u_i)_{1\le i\le L}$. Then the following inequality holds.*

$$\sum_{k=1}^{L} \lambda_k \le z_\alpha^2.$$

Proof. We have

$$\begin{aligned}
\sum_{k=1}^{L} \lambda_k &= \sum_{k=1}^{L} \int_0^1 \int_0^1 \frac{u_k(x)u_k(y)}{n} \sum_{i=1}^{n} F_i(x,y)dxdy \\
&= \frac{1}{n}\sum_{i=1}^{n} \int_{\tilde{\Omega}_i} \sum_{k=1}^{L} \left(\int_0^1 \frac{q_i(z_i|x)}{\tilde{f}_{0,i}(z_i)} u_k(x)dx \right)^2 \tilde{f}_{0,i}(z_i)d\mu_i(z_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \int_{\tilde{\Omega}_i} \sum_{k=1}^{L} \left(\int_0^1 \left(\frac{q_i(z_i|x)}{\tilde{f}_{0,i}(z_i)} - e^{-2\alpha} \right) u_k(x)dx \right)^2 \tilde{f}_{0,i}(z_i)d\mu_i(z_i),
\end{aligned}$$

since $\int u_k(x)dx = 0$. Now we define $f_{z,i}(x) = \frac{q_i(z_i|x)}{\tilde{f}_{0,i}(z_i)} - e^{-2\alpha}$ and by Lemma 20,

$$0 \le e^{-\alpha} - e^{-2\alpha} \le f_{z,i}(x) = \left(\int_0^1 \frac{q_i(z_i|s)}{q_i(z_i|x)} ds \right)^{-1} - e^{-2\alpha} \le e^{\alpha} - e^{-2\alpha} \le e^{2\alpha} - e^{-2\alpha}.$$

So $\|f_{z,i}\|_2 \leq e^{2\alpha} - e^{-2\alpha}$. Then, by orthonormality of the u_k's, we apply Parseval's inequality:

$$\sum_{k=1}^{L} \left(\int_0^1 \left(\frac{q_i(z_i|x)}{\tilde{f}_{0,i}(z_i)} - e^{-2\alpha} \right) u_k(x) dx \right)^2 = \sum_{k=1}^{L} \langle f_{z,i}, u_k \rangle^2 = \| \sum_{k=1}^{L} \langle f_{z,i}, u_k \rangle u_k \|_2^2 \leq \|f_{z,i}\|_2^2 \leq z_\alpha^2.$$

Finally, $\int_{\tilde{\Omega}_i} \tilde{f}_{0,i}(z_i) d\mu_i(z_i) = 1$ leads to $\sum \lambda_k \leq z_\alpha^2$.

□

Then from Equation (3.28) and by application of Lemma 26,

$$\|f_\eta - f_0\|_2 \geq L\varepsilon \sqrt{(L^{-1}l)^2 + 1 - (L^{-1}l)} \geq L\varepsilon\sqrt{3/4}. \tag{3.29}$$

So for the discrete case, by Equation (3.2),

$$\sqrt{\sum_{i=1}^{d} (q_i - p_i)^2} \geq \sqrt{d}\varepsilon\sqrt{3/4}. \tag{3.30}$$

3.7.1.4 Information bound

Let $\varepsilon > 0$, for all $\eta \in \{-1, 1\}^L$, we define

$$\tilde{f}_{\eta,i}(z) = \tilde{f}_{0,i}(z) + \varepsilon \sum_{j=1}^{L} \eta_j \tilde{\lambda}_j^{-1/2} \int_0^1 q_i(z_i|x) u_j(x) dx.$$

We consider the expected squared likelihood ratio:

$$\begin{aligned}
&\mathbb{E}_{Q_{f_0}^n} \left[L^2_{Q^n_{\nu_\rho}}(Z_1, \ldots, Z_n) \right] \\
&= \mathbb{E}_{Q_{f_0}^n} \mathbb{E}_{\eta,\eta'} \prod_{i=1}^{n} \left(1 + \varepsilon \frac{\sum_{j=1}^{L} \tilde{\lambda}_j^{-1/2} \eta_j \int_0^1 q_i(Z_i|x) u_j(x) dx}{\tilde{f}_0(Z_i)} \right) \\
&\qquad \left(1 + \varepsilon \frac{\sum_{j=1}^{L} \tilde{\lambda}_j^{-1/2} \eta'_j \int_0^1 q_i(Z_i|x) u_j(x) dx}{\tilde{f}_0(Z_i)} \right) \\
&= \mathbb{E}_{Q_{f_0}^n} \mathbb{E}_{\eta,\eta'} \prod_{i=1}^{n} \left(1 + \frac{\varepsilon \sum_{j=1}^{L} \tilde{\lambda}_j^{-1/2} \eta_j \int_0^1 q_i(Z_i|x) u_j(x) dx}{\tilde{f}_0(Z_i)} \right. \\
&\qquad + \frac{\varepsilon \sum_{j=1}^{L} \tilde{\lambda}_j^{-1/2} \eta'_j \int_0^1 q_i(Z_i|x) u_j(x) dx}{\tilde{f}_0(Z_i)} \\
&\quad \left. + \frac{\varepsilon^2 \sum_{j,l=1}^{L} \tilde{\lambda}_j^{-1/2} \tilde{\lambda}_l^{-1/2} \eta_j \eta'_l \int_0^1 q_i(Z_i|x) u_j(x) dx \int_0^1 q_i(Z_i|y) u_l(y) dy}{\tilde{f}_0(Z_i)^2} \right).
\end{aligned}$$

Now, for any j,

$$\mathbb{E}_{Q_{f_0}} \left[\frac{\int_0^1 q_i(Z_i|x) u_j(x) dx}{\tilde{f}_0(Z_i)} \right] = \int_0^1 \int_{\tilde{\Omega}_i} q_i(z|x) d\mu_i(z) u_j(x) dx = \int_0^1 u_j(x) dx = 0,$$

by orthogonality with uniform vector f_0.

And, by Equation (3.27),

$$\mathbb{E}_{Q_{f_0}}\left[\frac{\int_0^1 q_i(Z_i|x)u_j(x)dx\int_0^1 q_i(Z_i|y)u_l(y)dy}{\tilde{f}_0(Z_i)^2}\right] = \int_0^1\int_0^1 F_i(x,y)u_j(x)u_l(y)dxdy.$$

So since $1+u \le \exp u$ for any u,

$$\mathbb{E}_{Q^n_{f_0}}\left[L^2_{Q^n_{\nu_\rho}}(Z_1,\dots,Z_n)\right] \le \mathbb{E}_{\eta,\eta'}\exp\left(\varepsilon^2\sum_{j,l=1}^{L}\tilde{\lambda}_j^{-1/2}\tilde{\lambda}_l^{-1/2}\eta_j\eta'_l n\int_0^1\int_0^1 F(x,y)u_j(x)u_l(y)dxdy\right).$$

Now

$$\int_0^1\int_0^1 F(x,y)u_j(x)u_l(y)dxdy = \lambda_j\int_0^1 u_j(x)u_l(x)dx = \lambda_j\mathbb{1}\{j=l\},$$

since u_j is an eigenfunction of K and by orthonormality. So

$$\mathbb{E}_{Q^n_{f_0}}\left[L^2_{Q^n_{\nu_\rho}}(Z_1,\dots,Z_n)\right] \le \mathbb{E}_{\eta,\eta'}\exp\left(n\varepsilon^2\sum_{j=1}^{L}\tilde{\lambda}_j^{-1}\eta_j\eta'_j\lambda_j\right) \le \mathbb{E}_{\eta,\eta'}\exp\left(n\varepsilon^2\sum_{j=1}^{L}\eta_j\eta'_j z_\alpha^2\right).$$

Then

$$\mathbb{E}_{Q^n_{f_0}}\left[L^2_{Q^n_{\nu_\rho}}(Z_1,\dots,Z_n)\right] \le \prod_{j=1}^{L}\cosh(n\varepsilon^2 z_\alpha^2) \le \prod_{j=1}^{L}\exp(n^2\varepsilon^4 z_\alpha^4) \le \exp(n^2\varepsilon^4 z_\alpha^4 L).$$

Then, in order to apply Lemma 2 combined with 24, let us find a sufficient condition for

$$\mathbb{E}_{Q^n_{f_0}}\left[L^2_{Q^n_{\nu_\rho}}(Z_1,\dots,Z_n)\right] < 1+4(1-\gamma-\beta-\gamma)^2.$$

So let us choose ε and J in order to ensure that

$$\exp(n^2\varepsilon^4 z_\alpha^4 L) < 1+4(1-2\gamma-\beta)^2,$$

i.e.

$$L\varepsilon^4 \le (nz_\alpha^2)^{-2}\log\left[1+4(1-2\gamma-\beta)^2\right],$$

i.e.

$$\varepsilon \le (nz_\alpha^2)^{-1/2}\left(\frac{\log\left[1+4(1-2\gamma-\beta)^2\right]}{L}\right)^{1/4}. \tag{3.31}$$

3.7.1.5 Sufficient condition for f_η to be non-negative

Lemma 27. *If*

$$\varepsilon \le \frac{L^{-1}}{\sqrt{2\log(2L/\gamma)}},$$

then there exists $A_\gamma \subset \{-1,1\}^L$ such that $\mathbb{P}_{\nu_\rho}(\eta \in A_\gamma) \ge 1-\gamma$ and for any $\eta \in A_\gamma$, f_η is a density.

Proof. Let

$$A_\gamma = \{\eta : |\sum_j \eta_j a_{j,k} \tilde{\lambda}_j^{-1/2}| \leq \sqrt{2L\log(2L/\gamma)}\}.$$

Since u_i is orthogonal to f_0, uniform density on $[0,1]$ for all i, we have $\int_0^1 f_\eta(x)dx = 1$ and we just have to prove that f_η is nonnegative. We remind the reader that

$$u_i = \sum_{k=0}^{L-1} a_{i,k}\psi_k.$$

The bases $(u_1, \ldots, u_L)$ and $(\psi_k, k \in \{0, 1, \ldots, L-1\})$ are orthonormal. This implies that the matrix $A = (a_{i,k})_{1\leq i\leq L, k\in\{0,1,\ldots,L-1\}}$ is orthogonal. So

$$\forall i, \sum_{k=0}^{L-1} a_{i,k}^2 = 1, \ \forall k, \sum_{i=1}^{L} a_{i,k}^2 = 1.$$

Hence we have for all $x \in [0,1]$,

$$(f_\eta - f_0)(x) = \sum_{k=0}^{L-1}\sum_{i=1}^{L} \eta_i \varepsilon \tilde{\lambda}_i^{-1/2} a_{i,k}\psi_k(x). \tag{3.32}$$

The functions $(\psi_k, k \in \{0, 1, \ldots, L-1\})$ have disjoint supports and $\sup_{x\in[0,1]} |\psi_k(x)| = L^{1/2}$. Hence f_η is nonnegative if and only if for any $k \in \Lambda(J)$

$$L^{1/2}\left|\sum_{i=1}^{L} \eta_i \varepsilon \tilde{\lambda}_i^{-1/2} a_{i,k}\right| \leq 1. \tag{3.33}$$

By definition of ν_ρ, we have that f_η is a density with probability larger than $1-\gamma$ under the prior ν_ρ as soon as Equation (3.33) holds with probability larger than $1-\gamma$. That is,

$$\mathbb{P}_{\nu_\rho}\left(\forall k \in \{0, 1, \ldots, L-1\}, L^{1/2}\left|\sum_{i=1}^{L} \eta_i \varepsilon \tilde{\lambda}_i^{-1/2} a_{i,k}\right| \leq 1\right) \geq 1-\gamma,$$

where $(\eta_1, \ldots, \eta_L)$ are i.i.d. Rademacher random variables. Using Hoeffding's inequality, we get for all $x > 0$, for all $k \in \{0, 1, \ldots, L-1\}$,

$$\mathbb{P}_{\nu_\rho}\left(\left|\sum_{i=1}^{L} \eta_i \varepsilon \tilde{\lambda}_i^{-1/2} a_{i,k}\right| > x\right) \leq 2\exp\left(\frac{-2x^2}{\sum_{i=1}^{L}(2\varepsilon\tilde{\lambda}_i^{-1/2}a_{ik})^2}\right).$$

Hence

$$\mathbb{P}_{\nu_\rho}\left(\exists k \in \{0, 1, \ldots, L-1\}, \left|\sum_{i=1}^{L} \eta_i \varepsilon \tilde{\lambda}_i^{-1/2} a_{i,k}\right| > x\right)$$
$$\leq 2L\exp\left(\frac{-x^2}{2\sum_{i=1}^{L}(\varepsilon\tilde{\lambda}_i^{-1/2}a_{ik})^2}\right).$$

So the probability of having the existence of some $k \in \{0, 1, \dots, L-1\}$ such that

$$\left| \sum_{i=1}^{L} \eta_i \varepsilon \tilde{\lambda}_i^{-1/2} a_{i,k} \right| > \sqrt{2 \sum_{i=1}^{L} (\varepsilon \tilde{\lambda}_i^{-1/2} a_{ik})^2 \log(2L/\gamma)}$$

is smaller than γ. Hence, f_η is a density with probability larger than $1-\gamma$ under the prior ν_ρ as soon as for any $k \in \Lambda(J)$,

$$2L \sum_{i=1}^{L} (\varepsilon \tilde{\lambda}_i^{-1/2} a_{ik})^2 \log(2L/\gamma) \leq 1.$$

Now by definition, $\tilde{\lambda}_i^{-1/2} \leq L^{1/2}$. So we have the sufficient condition

$$2L\varepsilon^2 L \log(2L/\gamma) \sum_{i=1}^{L} a_{ik}^2 \leq 1.$$

And $\sum_{i=1}^{L} a_{ik}^2 = 1$ leads to the following sufficient condition,

$$\varepsilon \leq \frac{L^{-1}}{\sqrt{2 \log(2L/\gamma)}}.$$

□

3.7.1.6 Sufficient conditions for $f_\eta \in \mathcal{F}_\rho(\mathcal{B}_{s,2,\infty}(R))$, only in the continuous case

We first prove the following points.

Lemma 28. *If*

$$\varepsilon \leq \frac{L^{-1}(1 \wedge RL^{-s})}{\sqrt{2 \log(2L/\gamma)}},$$

then there exists $A_\gamma \subset \{-1, 1\}^d$ *such that* $\mathbb{P}_{\nu_\rho}(\eta \in A_\gamma) \geq 1 - \gamma$ *and for any* $\eta \in A_\gamma$,

a) f_η *is a density.*

b) $f_\eta \in \mathcal{B}_{s,2,\infty}(R)$.

Proof. We consider the same event as in the proof of Lemma 27:

$$A_\gamma = \{\eta : |\sum_j \eta_j a_{j,k} \tilde{\lambda}_j^{-1/2}| \leq \sqrt{2L \log(2L/\gamma)}\}.$$

a) In the same way as in Lemma 27, f_η is a density.

b) For all $k \in \{0, 1, \dots, L-1\}$,

$$\langle f_\eta - f_0, \psi_k \rangle = \varepsilon \sum_{i=1}^{L} \eta_i \tilde{\lambda}_i^{-1/2} a_{i,k}.$$

Hence $f_\eta \in \mathcal{B}_{s,2,\infty}(R)$ if and only if

$$\sum_{k=0}^{L-1} \varepsilon^2 \left(\sum_{i=1}^{L} \eta_i \tilde{\lambda}_i^{-1/2} a_{i,k} \right)^2 \leq R^2 L^{-2s}.$$

But we also have that for any $\eta \in A_\gamma$,

$$\sum_{k=0}^{L-1} \varepsilon^2 \left(\sum_{i=1}^{L} \eta_i \tilde{\lambda}_i^{-1/2} a_{i,k} \right)^2 \leq \varepsilon^2 L^2 2 \log(2L/\gamma).$$

So $f_\eta \in \mathcal{B}_{s,2,\infty}(R)$ if

$$\varepsilon \leq RL^{-(s+1)} / \sqrt{2 \log(2L/\gamma)}.$$

3.7.1.7 Conclusion

1. Discrete case.

   So combining Equation (3.31) and Lemma 27, we obtain the following sufficient condition in order to apply Lemma 2:

   $$\varepsilon \leq \left[(nz_\alpha^2)^{-1/2} \left(\frac{\log\left[1 + 4(1 - 2\gamma - \beta)^2\right]}{L} \right)^{1/4} \right] \wedge \frac{L^{-1}}{\sqrt{2 \log(2L/\gamma)}}.$$

   So, by Equation (3.30), if

   $$\sqrt{\sum_{k=1}^{d} (q_k - p_k)^2} \leq \sqrt{3/4} \left(\left[(nz_\alpha^2)^{-1/2} d^{1/4} \left(\log\left[1 + 4(1 - 2\gamma - \beta)^2\right] \right)^{1/4} \right] \wedge \frac{d^{-1/2}}{\sqrt{2 \log(2d/\gamma)}} \right),$$

   then we can define densities f_η such that the errors are larger than γ and β. So

   $$\begin{aligned} &\rho_\beta^*(f_0, H_1, \mathcal{T}_{\gamma,n}^{(\alpha)}) / d^{1/2} \\ &\geq c\,(\gamma, \beta)\, [((nz_\alpha^2)^{-1/2} d^{1/4}) \wedge (d \log d)^{-1/2}]. \end{aligned}$$

   Now, we also have

   $$\rho_\beta^*(f_0, H_1, \mathcal{T}_{\gamma,n}^{(\alpha)}) \geq \rho_\beta^*(f_0, H_1, \mathcal{T}_{\gamma,n}^{(+\infty)}),$$

   where $\rho_\beta^*(f_0, H_1, \mathcal{T}_{\gamma,n}^{(+\infty)})$ corresponds to the case where there is no local differential privacy condition on Q. In particular, taking Q such that $Z = X$ with probability 1 reduces the private problem to the classical testing problem. Now, the data processing inequality in Lemma 25 justifies that such a Q is optimal by contraction of the total variation distance. And the classical result leads to having $\rho_\beta^*(f_0, H_1, \mathcal{T}_{\gamma,n}^{(+\infty)}) = c\,(\gamma, \beta)\, n^{-1/2} d^{-1/4}$.

   So, we have

   $$\rho_\beta^*(f_0, H_1, \mathcal{T}_{\gamma,n}^{(\alpha)}) / d^{1/2} \geq c\,(\gamma, \beta)\, [((nz_\alpha^2)^{-1/2} d^{1/4}) \wedge (d \log d)^{-1/2}] \vee (n^{-1/2} d^{-1/4}).$$

2. Continuous case.

So combining Equation (3.31) and Lemma 28, we obtain the following sufficient condition in order to apply Lemma 2:

$$\varepsilon \leq \left[(nz_\alpha^2)^{-1/2} \left(\frac{\log\left[1 + 4(1-2\gamma-\beta)^2\right]}{L} \right)^{1/4} \right] \wedge \frac{L^{-1}(1 \wedge RL^{-s})}{\sqrt{2\log(2L/\gamma)}}.$$

So, by Equation (3.29), if

$$\begin{aligned} &\|f - f_0\|_2 \\ &\leq \sqrt{3/4} \Bigg(\left[(nz_\alpha^2)^{-1/2} L^{3/4} \log^{1/4}\left(1 + 4(1-2\gamma-\beta)^2\right) \right] \wedge \frac{(1 \wedge RL^{-s})}{\sqrt{2\log(2L/\gamma)}} \Bigg), \end{aligned}$$

then, taking J as the largest integer such that $2^J \leq c\,(\gamma, \beta, R)\,(nz_\alpha^2)^{2/(4s+3)}$, we obtain:

$$\rho_\beta^*(f_0, H_{1,s}, \mathcal{T}_{\gamma,n}^{(\alpha)}) \geq c\,(\gamma, \beta, R)\,(nz_\alpha^2)^{-2s/(4s+3)} (\log n)^{-1/2}.$$

Now, we also have

$$\rho_\beta^*(f_0, H_{1,s}, \mathcal{T}_{\gamma,n}^{(\alpha)}) \geq \rho_\beta^*(f_0, H_{1,s}, \mathcal{T}_{\gamma,n}^{(+\infty)}),$$

where $\rho_\beta^*(f_0, H_{1,s}, \mathcal{T}_{\gamma,n}^{(+\infty)})$ corresponds to the case where there is no local differential privacy condition on Q. In particular, taking Q such that $Z = X$ with probability 1 reduces the private problem to the classical testing problem. Now, the data processing inequality in Lemma 25 justifies that such a Q is optimal by contraction of the total variation distance. And the classical result leads to having $\rho_\beta^*(f_0, H_{1,s}, \mathcal{T}_{\gamma,n}^{(+\infty)}) = c\,(\gamma, \beta, R)\, n^{-2s/(4s+1)}$.

So, we have

$$\rho_\beta^*(f_0, H_{1,s}, \mathcal{T}_{\gamma,n}^{(\alpha)}) \geq c\,(\gamma, \beta, R)\,[(nz_\alpha^2)^{-2s/(4s+3)} (\log n)^{-1/2} \vee n^{-2s/(4s+1)}].$$

□

3.7.2 Proof of the upper bound

In this section, f_0 is some fixed density in $\mathbb{L}_2([0,1])$.

3.7.2.1 Proof of Theorem 28

We prove the bound on the variance term $\mathbb{V}_{Q_f^n}\left(\hat{T}_L\right)$ given in Equation (3.16). Let us define

$$\hat{U}_L = \frac{1}{n(n-1)} \sum_{i \neq l=1}^{n} \sum_{k=0}^{L-1} (Z_{i,L,k} - \alpha_{L,k})\,(Z_{l,L,k} - \alpha_{L,k}),$$

$$\hat{V}_L = 2 \sum_{k=0}^{L-1} (\alpha_{L,k} - \alpha_{L,k}^0)\, \frac{1}{n} \sum_{i=1}^{n} (Z_{i,L,k} - \alpha_{L,k}),$$

where $\alpha_{L,k} = \int_0^1 \phi_{L,k}(x) f(x) dx$ and $\alpha^0_{L,k} = \int_0^1 \phi_{L,k}(x) f_0(x) dx$. Then we obtain the Hoeffding's decomposition of the U-statistic $\hat{T}_L$, namely

$$\hat{T}_L = \hat{U}_L + \hat{V}_L + \|\Pi_{S_L}(f - f_0)\|_2^2.$$

We first control the variance of the degenerate U-statistic $\hat{U}_L$ which can be written as

$$\hat{U}_L = \frac{1}{n(n-1)} \sum_{i \neq l = 1}^{n} h_L(Z_{i,L}, Z_{l,L}),$$

where

$$h_L(Z_{i,L}, Z_{l,L}) = \sum_{k=0}^{L-1} \left(Z_{i,L,k} - \alpha_{L,k}\right) \left(Z_{l,L,k} - \alpha_{L,k}\right).$$

In order to provide an upper bound for the variance $\mathbb{V}_{Q_f^n}(\hat{U}_L)$, let us first state a lemma controlling the variance of a U-statistic of order 2. This result is a particular case of Lemma 8 in Meynaoui et al. (2019).

Lemma 29. *Let h be a symmetric function with 2 inputs, $Z_1, \ldots, Z_n$ be i.i.d. random vectors and U_n be the U-statistic of order 2 defined by*

$$U_n = \frac{1}{n(n-1)} \sum_{i \neq l = 1}^{n} h(Z_i, Z_l).$$

The following inequality gives an upper bound on the variance of U_n,

$$\mathbb{V}(U_n) \leq C \left(\frac{\sigma^2}{n} + \frac{s^2}{n^2} \right),$$

where $\sigma^2 = \mathbb{V}\left(\mathbb{E}[h(Z_1, Z_2) \mid Z_1]\right)$ and $s^2 = \mathbb{V}\left(h(Z_1, Z_2)\right)$.

We have that $\mathbb{E}_{Q_f^n}[h_L(Z_1, Z_2) \mid Z_1] = 0$, hence the first term in the upper bound of the variance vanishes. In order to bound the term s^2, we write

$$\begin{aligned} h_L(Z_1, Z_2) = &\sum_{k=0}^{L-1} \left(\phi_{L,k}(X_1) - \alpha_{L,k}\right) \left(\phi_{L,k}(X_2) - \alpha_{L,k}\right) + \sigma_L^2 \sum_{k=0}^{L-1} W_{1,L,k} W_{2,L,k} \\ &+ \sigma_L \sum_{k=0}^{L-1} W_{1,L,k} \left(\phi_{L,k}(X_2) - \alpha_{L,k}\right) + \sigma_L \sum_{k=0}^{L-1} W_{2,L,k} \left(\phi_{L,k}(X_1) - \alpha_{L,k}\right). \end{aligned}$$

So, since $\mathbb{E}_{Q_f^n}\left(\phi_{L,k}(X_i) - \alpha_{L,k}\right) = 0$ and $\mathbb{E}(W_{i,L,k}) = 0$ for any i. Using independence properties, we therefore have

$$\begin{aligned} &\mathbb{V}_{Q_f^n}\left(h_L(Z_1, Z_2)\right) \\ &= \mathbb{V}_{Q_f^n}\left[\sum_{k=0}^{L-1} \left(\phi_{L,k}(X_1) - \alpha_{L,k}\right) \left(\phi_{L,k}(X_2) - \alpha_{L,k}\right) \right] + \mathbb{V}_{Q_f^n}\left[\sigma_L^2 \sum_{k=0}^{L-1} W_{1,L,k} W_{2,L,k} \right] \\ &\quad + 2\mathbb{V}_{Q_f^n}\left[\sigma_L \sum_{k=0}^{L-1} W_{1,L,k} \left(\phi_{L,k}(X_2) - \alpha_{L,k}\right) \right]. \end{aligned}$$

Now, by independence of X_1 and X_2,

$$\begin{aligned}
&\mathbb{V}_{Q_f^n}\left[\sum_{k=0}^{L-1}\left(\phi_{L,k}(X_1)-\alpha_{L,k}\right)\left(\phi_{L,k}(X_2)-\alpha_{L,k}\right)\right] \\
&= \sum_{k,k'=0}^{L-1}\mathbb{E}\left[\left(\phi_{L,k}(X_1)-\alpha_{L,k}\right)\left(\phi_{L,k'}(X_1)-\alpha_{L,k'}\right)\right] \\
&\qquad\qquad \mathbb{E}\left[\left(\phi_{L,k}(X_2)-\alpha_{L,k}\right)\left(\phi_{L,k'}(X_2)-\alpha_{L,k'}\right)\right] \\
&= \sum_{k,k'=0}^{L-1}\left[\int \phi_{L,k}\phi_{L,k'}f-\alpha_{L,k}\alpha_{L,k'}\right]^2.
\end{aligned}$$

So

$$\begin{aligned}
&\mathbb{V}_{Q_f^n}\left[\sum_{k=0}^{L-1}\left(\phi_{L,k}(X_1)-\alpha_{L,k}\right)\left(\phi_{L,k}(X_2)-\alpha_{L,k}\right)\right] \\
&= \int\int\left(\sum_{k=0}^{L-1}\phi_{L,k}(x)\phi_{L,k}(y)\right)^2 f(x)f(y)dxdy - 2\int\left(\sum_{k=0}^{L-1}\alpha_{L,k}\phi_{L,k}(x)\right)^2 f(x)dx \\
&\qquad + \left(\sum_{k=0}^{L-1}\alpha_{L,k}^2\right)^2.
\end{aligned}$$

In order to control the first term of the variance, note that by definition of the functions $\phi_{L,k}$, we have that, for all $x\in[0,1]$, $\phi_{L,k}(x)\phi_{L,k'}(x)=0$ if $k\neq k'$, and that $\phi_{L,k}^2=\sqrt{L}\phi_{L,k}$. Hence,

$$\begin{aligned}
\int\int\left(\sum_{k=0}^{L-1}\phi_{L,k}(x)\phi_{L,k}(y)\right)^2 f(x)f(y)dxdy &= L\sum_{k=0}^{L-1}\alpha_{L,k}^2 \\
&\leq L\|f\|_2^2.
\end{aligned}$$

Since the second term of the variance is nonpositive, and the third term is controlled by $\|f\|_2^4$,we obtain

$$\mathbb{V}_{Q_f^n}\left[\sum_{k=0}^{L-1}\left(\phi_{L,k}(X_1)-\alpha_{L,k}\right)\left(\phi_{L,k}(X_2)-\alpha_{L,k}\right)\right] \leq L\|f\|_2^2+\|f\|_2^4 \leq 2L\|f\|_2^2.$$

By independence of the variables $(W_{i,L,k})$,

$$\mathbb{V}\left(\sigma_L^2\sum_{k=0}^{L-1}W_{1,L,k}W_{2,L,k}\right) = \sigma_L^4\sum_{k=0}^{L-1}\mathbb{V}(W_{1,L,k}W_{2,L,k}) = L\sigma_L^4.$$

Finally, using again the independence of the variables $(W_{1,L,k})_{k\in\{0,\dots,L-1\}}$, and their independence with X_2,

$$\begin{aligned}
&\mathbb{V}_{Q_f^n}\left[\sigma_L \sum_{k=0}^{L-1} W_{1,L,k}\left(\phi_{L,k}(X_2)-\alpha_{L,k}\right)\right] \\
&= \sigma_L^2 \mathbb{E}_{Q_f^n}\left[\sum_{k,k'=0}^{L-1} W_{1,L,k}W_{1,L,k'}\left(\phi_{L,k}(X_2)-\alpha_{L,k}\right)\left(\phi_{L,k'}(X_2)-\alpha_{L,k'}\right)\right] \\
&= \sigma_L^2 \sum_{k=0}^{L-1} \mathbb{E}(W_{1,L,k}^2)\mathbb{E}_{Q_f^n}\left[\left(\phi_{L,k}(X_2)-\alpha_{L,k}\right)^2\right] \\
&\leq \sigma_L^2 \sum_{k=0}^{L-1} \int \phi_{L,k}^2 f \leq \sigma_L^2 L \sum_{k=0}^{L-1} \int_{k/L}^{(k+1)/L} f \leq \sigma_L^2 L
\end{aligned}$$

since $\int_0^1 f = 1$. This leads to the following upper bound for $\mathbb{V}_{Q_f^n}\left(h_L(Z_1,Z_2)\right)$,

$$\mathbb{V}_{Q_f^n}\left(h_L(Z_1,Z_2)\right) \leq (2\|f\|_2^2+\sigma_L^2+\sigma_L^4)L,$$

from which, by application of Lemma 29, we deduce that

$$\mathbb{V}_{Q_f^n}\left(\hat{U}_L\right) \leq 2\frac{(\|f\|_2^2+\sigma_L^4)L}{n^2}.$$

Let us now compute $\mathbb{V}_{Q_f^n}(\hat{V}_L)$. Since $\hat{V}_L$ is centered,

$$\begin{aligned}
\mathbb{V}_{Q_f^n}\left(\hat{V}_L\right) &= \mathbb{E}_{Q_f^n}(\hat{V}_L^2) \\
&= \frac{4}{n^2}\sum_{k,k'=0}^{L-1}\left(\alpha_{L,k}-\alpha_{L,k}^0\right)\left(\alpha_{L,k'}-\alpha_{L,k'}^0\right) \\
&\quad \sum_{i,l=1}^{n}\mathbb{E}_{Q_f^n}\left[(Z_{i,L,k}-\alpha_{L,k})(Z_{l,L,k'}-\alpha_{L,k'})\right].
\end{aligned}$$

Note that, if $i \neq l$,

$$\mathbb{E}_{Q_f^n}\left[(Z_{i,L,k}-\alpha_{L,k})(Z_{l,L,k'}-\alpha_{L,k'})\right] = 0.$$

Moreover,

$$\begin{aligned}
&\mathbb{E}_{Q_f^n}\left[(Z_{i,L,k}-\alpha_{L,k})(Z_{i,L,k'}-\alpha_{L,k'})\right] \\
&\qquad = \mathbb{E}\big[(\phi_{L,k}(X_i)-\alpha_{L,k})(\phi_{L,k'}(X_i)-\alpha_{L,k'}) + \sigma_L^2\mathbb{E}_{Q_f^n}(W_{i,L,k}W_{i,L,k'})\big] \\
&\qquad = \int \phi_{L,k}\phi_{L,k'}f - \alpha_{L,k}\alpha_{L,k'} + 2\sigma_L^2\mathbb{1}\{k=k'\}.
\end{aligned}$$

Hence,

$$
\begin{aligned}
&\mathbb{V}_{Q_f^n}\left(\hat{V}_L\right) \\
&= \frac{4}{n}\sum_{k,k'=0}^{L-1}\left(\alpha_{L,k}-\alpha_{L,k}^0\right)\left(\alpha_{L,k'}-\alpha_{L,k'}^0\right)\left(\int \phi_{L,k}\phi_{L,k'}f - \alpha_{L,k}\alpha_{L,k'} + 2\sigma_L^2\mathbb{1}\{k=k'\}\right) \\
&= \frac{4}{n}\int\left(\sum_{k=0}^{L-1}(\alpha_{L,k}-\alpha_{L,k}^0)\phi_{L,k}\right)^2 f - \frac{4}{n}\left(\sum_{k=0}^{L-1}\alpha_{L,k}(\alpha_{L,k}-\alpha_{L,k}^0)\right)^2 \\
&\qquad + \frac{8}{n}\sigma_L^2\sum_{k=0}^{L-1}(\alpha_{L,k}-\alpha_{L,k}^0)^2 \\
&\leq \frac{4}{n}\sum_{k=0}^{L-1}(\alpha_{L,k}-\alpha_{L,k}^0)^2\int\phi_{L,k}^2 f + \frac{8}{n}\sigma_L^2\sum_{k=0}^{L-1}(\alpha_{L,k}-\alpha_{L,k}^0)^2 \\
&\leq \frac{1}{n}\left(4\sqrt{L}\|f\|_2 + 8\sigma_L^2\right)\sum_{k=0}^{L-1}(\alpha_{L,k}-\alpha_{L,k}^0)^2
\end{aligned}
$$

since by Cauchy Schwarz inequality,

$$
0 \leq \int \phi_{L,k}^2 f = \sqrt{L}\int\phi_{L,k}f \leq \sqrt{L}\|\phi_{L,k}\|_2\|f\|_2 = \sqrt{L}\|f\|_2.
$$

We finally obtain,

$$
\mathbb{V}_{Q_f^n}\left(\hat{V}_L\right) \leq C\frac{(\sqrt{L}\|f\|_2+\sigma_L^2)}{n}\|\Pi_{S_L}(f-f_0)\|_2^2.
$$

Collecting the upper bounds for $\mathbb{V}_{Q_f^n}\left(\hat{U}_L\right)$ and for $\mathbb{V}_{Q_f^n}\left(\hat{V}_L\right)$, we obtain the inequality from Equation (3.16), that we remind here:

$$
\mathbb{V}_{Q_f^n}\left(\hat{T}_L\right) \leq C\left[\frac{(\sqrt{L}\|f\|_2+\sigma_L^2)}{n}\|\Pi_{S_L}(f-f_0)\|_2^2 + \frac{(\|f\|_2^2+\sigma_L^4)L}{n^2}\right].
$$

3.7.2.2 Proof of Corollary 10

From Equation (3.16) and taking $f = f_0$, we obtain

$$
\sqrt{\mathbb{V}_{Q_{f_0}^n}\left(\hat{T}_L\right)/\gamma} \leq C(\gamma)\frac{(\|f_0\|_2+\sigma_L^2)\sqrt{L}}{n}.
$$

Moreover, we deduce from (3.16) that

$$
\sqrt{\mathbb{V}_{Q_f^n}\left(\hat{T}_L\right)/\beta} \leq C(\beta)\left[\frac{(L^{1/4}\|f\|_2^{1/2}+\sigma_L)}{\sqrt{n}}\|\Pi_{S_L}(f-f_0)\|_2 + \frac{(\|f\|_2+\sigma_L^2)\sqrt{L}}{n}\right].
$$

Using the inequality between geometric and harmonic means, we get

$$
\sqrt{\mathbb{V}_{Q_f^n}\left(\hat{T}_L\right)/\beta} \leq \frac{1}{2}\|\Pi_{S_L}(f-f_0)\|_2^2 + C(\beta)\frac{(\|f\|_2+\sigma_L^2)\sqrt{L}}{n}.
$$

We conclude the proof by using the condition in Equation (3.14).

3.7.2.3 Proof of Theorem 29

We recall that we have defined

$$f = d\sum_{k=1}^{d} q_k \mathbb{I}_{[k/d,(k+1)/d)}, \;\; f_0 = d\sum_{k=1}^{d} p_k \mathbb{I}_{[k/d,(k+1)/d)}.$$

We obtain from Corollary 10 that the second kind error of the test is controlled by β if

$$\frac{3}{2}\|\Pi_{S_L}(f-f_0)\|_2^2 \geq C(\gamma,\beta)\frac{(\|f\|_2+\|f_0\|_2+\sigma_L^2)\sqrt{L}}{n}.$$

In the discrete case, by definition, f and f_0 belong to S_L, hence $\|\Pi_{S_L}(f-f_0)\|_2 = \|f-f_0\|_2$ and $\|f\|_2 \leq \|f_0\|_2 + \|f-f_0\|_2$. So we have the following sufficient condition.

$$\|f-f_0\|_2^2 \geq C(\gamma,\beta)\frac{(\|f_0\|_2+\sigma_L^2+\sqrt{L}/n)\sqrt{L}}{n}$$

Moreover, we have

$$\|f-f_0\|_2^2 = d\sum_{k=1}^{d}(q_k-p_k)^2.$$

We recall that $L=d$ and $\sigma_L = 2\sqrt{2d}/\alpha$. That is, the sufficient condition turns out to be

$$d\sum_{k=1}^{d}(q_k-p_k)^2 \geq C(\gamma,\beta)\frac{d^{1/2}}{n}\left(\|f_0\|_2 + d\alpha^{-2} + d^{1/2}n^{-1}\right).$$

By definition of f_0, we have that

$$\|f\|_2^2 = d\sum_{k=1}^{d} p_k^2.$$

Finally, we obtain the following condition

$$\sqrt{\sum_{i=1}^{d}(q_i-p_i)^2} \geq C(\gamma,\beta)\frac{d^{-1/4}}{n^{1/2}}\left(\left[d\sum_{k=1}^{d}p_k^2\right]^{1/4} + d^{1/2}\alpha^{-1} + d^{1/4}n^{-1/2}\right).$$

3.7.2.4 Proof of Theorem 30

We obtain from Corollary 10 that the second kind error of the test is controlled by β if

$$\|f-f_0\|_2^2 \geq \|f-f_0-\Pi_{S_L}(f-f_0)\|_2^2 + C(\|f_0\|_2,\|f\|_2,\gamma,\beta)\frac{(\sigma_L^2+1)\sqrt{L}}{n}.$$

Since $f-f_0 \in \mathcal{B}_{s,2,\infty}(R)$, setting $L=2^J$, we have, on one hand

$$\|f-f_0-\Pi_{S_L}(f-f_0)\|_2^2 \leq R^2 2^{-2Js},$$

and on the other hand, $\|f\|_2 \leq C(s,R,\|f_0\|_2)$. This leads to the sufficient condition

$$\|f-f_0\|_2^2 \geq R^2 2^{-2Js} + C(s,R,\|f_0\|_2,\gamma,\beta)\frac{(\sigma_L^2+1)2^{J/2}}{n}.$$

We recall that $\sigma_L = 2\sqrt{2L}/\alpha$. That is, the sufficient condition turns out to be:

$$\|f - f_0\|_2^2 \geq C(s, R, \|f_0\|_2, \gamma, \beta) \left(2^{-2Js} + \frac{2^{3J/2}}{\alpha^2 n} + \frac{2^{J/2}}{n} \right). \tag{3.34}$$

J^* being set as the smallest integer J such that $2^J \geq (n\alpha^2)^{2/(4s+3)} \wedge n^{2/(4s+1)}$, we consider two cases.

- If $1/\sqrt{n} \leq \alpha \leq n^{1/(4s+1)}$, then $(n\alpha^2)^{2/(4s+3)} \leq n^{2/(4s+1)}$ and the right-hand side of the inequality in Equation (3.34) for $J = J^*$ is upper bounded by

$$C(s, R, \|f_0\|_2, \gamma, \beta)(n\alpha^2)^{-4s/(4s+3)}.$$

- If $\alpha > n^{1/(4s+1)}$, then $(n\alpha^2)^{2/(4s+3)} > n^{2/(4s+1)}$ and the right-hand side of the inequality in Equation (3.34) for $J = J^*$ is upper bounded by

$$C(s, R, \|f_0\|_2, \gamma, \beta)n^{-4s/(4s+1)}.$$

Hence, the separation rate of our test over the set $\mathcal{B}_{s,2,\infty}(R)$ is controlled by

$$C(s, R, \|f_0\|_2, \gamma, \beta) \left[(n\alpha^2)^{-2s/(4s+3)} \vee n^{-2s/(4s+1)} \right],$$

which concludes the proof of Theorem 30.

3.7.3 Adaptivity: proof of Theorem 31

In this section, f_0 is some fixed density in $\mathbb{L}_2([0,1])$.

Using the inequality from Equation (3.26), and the fact that $u_\gamma \geq \gamma/|\mathcal{J}|$, we obtain that

$$\mathbb{P}_{Q_f^n} \left(\varphi_{\gamma,Q}^{\mathcal{J}} = 0 \right) \leq \beta \tag{3.35}$$

as soon as

$$\exists J \in \mathcal{J}, \mathbb{P}_{Q_f^n} \left(\tilde{T}_J \leq \tilde{t}_J^0(1 - u_\gamma) \right) \leq \beta.$$

We use the result of Corollary 10, for $L = 2^J$ for some $J \in \mathcal{J}$, where σ_L is replaced by $\tilde{\sigma}_{2^J}$ and γ is replaced by $\gamma/|\mathcal{J}|$.

Using the fact that $|\mathcal{J}| \leq C\log(n)$, we get that Equation (3.35) holds as soon as there exists $J \in \mathcal{J}$ such that

$$\|\Pi_{S_{2^J}}(f - f_0)\|^2 \geq C(\|f_0\|_2, \|f\|_2, \beta) \left(\frac{(\tilde{\sigma}_{2^J}^2 + 1)2^{J/2}}{n\sqrt{\gamma/|\mathcal{J}|}} \right),$$

or equivalently

$$\|f - f_0\|^2$$
$$\geq \inf_{J \in \mathcal{J}} \left[\|f - f_0 - \Pi_{S_{2^J}}(f - f_0)\|^2 + C(\|f_0\|_2, \|f\|_2, \gamma, \beta) \frac{(\tilde{\sigma}_J^2 + 1)2^{J/2}\sqrt{\log(n)}}{n} \right].$$

Assuming that $f \in \mathcal{B}_{s,2,\infty}(R)$, for some $s > 0$ and $R > 0$, we get that Equation (3.35) holds if

$$\|f - f_0\|^2 \geq \inf_{J \in \mathcal{J}} \left[R^2 2^{-2Js} + C(\|f_0\|_2, R, \gamma, \beta) \left(2^{J/2} + \frac{2^{3J/2}\log^2(n)}{\alpha^2} \right) \frac{\sqrt{\log(n)}}{n} \right].$$

Choosing $J \in \mathcal{J}$ as the smallest integer in $\mathcal{J}$ such that $2^J \geq (n^2\alpha^4/\log^5(n))^{1/(4s+3)} \wedge (n^2/\log(n))^{1/(4s+1)}$, we obtain the sufficient condition

$$\|f - f_0\|^2 \geq C(\|f_0\|_2, R, \gamma, \beta)\left[(n\alpha^2/\log^{5/2}(n))^{-4s/(4s+3)} \vee (n/\sqrt{\log(n)})^{-4s/(4s+1)}\right].$$

Hence, for all $s > 0$, $R > 0$, the separation rate of the aggregated test over the set $\mathcal{B}_{s,2,\infty}(R)$ is controlled by

$$C(\|f_0\|_2, R, \gamma, \beta)\Big[(n\alpha^2/\log^{5/2}(n))^{-2s/(4s+3)} \vee (n/\sqrt{\log(n)})^{-2s/(4s+1)}\Big],$$

which concludes the proof of Theorem 31.

3.8 Naive lower bound

As promised in Remark 32, we provide a lower bound using the main result of Duchi, Jordan, and Wainwright (2013c), but the resulting rate turns out to be suboptimal.

Theorem 32. *Let $(\gamma, \beta) \in (0,1)$ such that $\gamma + \beta < 1$, let $\alpha > 0, R > 0, s > 0$. We obtain the following lower bound for the α-private minimax separation rate defined by Equation (3.7) for non-interactive channels in $\mathcal{Q}_\alpha$ over the class of alternatives $H_{1,s}$*

$$\rho^*_\beta(\mathbb{1}_{[0,1]}, H_{1,s}, \mathcal{T}^{(\alpha)}_{\gamma,n}) \geq c\,(\gamma, \beta, R)\left(2^{-Js} \wedge \frac{1}{(e^\alpha - 1)\sqrt{n}}\right).$$

The proof will remain concise since some arguments are also presented in the proofs of our main results.

Proof. Let $f_0 = \mathbb{1}_{[0,1]}$. Let us first define the setup similarly to Section 3.7.1.2. Let $Q \in \mathcal{Q}_\alpha$ be a non-interactive α-private channel. We assume that f_0 is the uniform density on $[0,1]$. We define the function $\psi \in \mathbb{L}^2([0,1])$ by $\psi(x) = \mathbb{1}_{[0,\frac{1}{2})} - \mathbb{1}_{[\frac{1}{2},1)}$, and for some given $J \in \mathbb{N}$, that will be specified later, we define, for all $k \in \Lambda(J) = \{0, 1, \ldots, 2^J - 1\}$, $\psi_{J,k}(x) = 2^{J/2}\psi(2^J x - k)$. We denote by V the linear subspace of $\mathbb{L}^2([0,1])$ generated by the functions $(f_0, \psi_{J,k}, k \in \Lambda(J))$.

Let

$$f_\eta = f_0 + \rho 2^{-J/2}\sum_{i=1}^{2^J} \eta_i \psi_{J,i},$$

where $\psi_{J,i}$ for every i have disjoint supports, $\int \psi_{J,i} = 0$, $\int \psi^2_{J,i} = 1$ and $\|\psi_{J,i}\|_\infty = 2^{J/2}$.

It is possible to show that f_η is a density if $\rho \leq 1$ and it is in the Besov set $B_{s,2,\infty}(R, 2)$ if $\rho \leq R2^{-Js}$.

Note that by orthonormality,

$$\|f_\eta - f_0\|_2^2 = \rho^2.$$

Denote D_{KL} the Kullback-Leibler divergence. Consider Theorem 1 in Duchi, Jordan, and Wainwright (2013c), for any densities f, g and $Q \in \mathcal{Q}_\alpha$:

$$D_{KL}(\mathbb{P}_{Q_f}, \mathbb{P}_{Q_g}) + D_{KL}(\mathbb{P}_{Q_g}, \mathbb{P}_{Q_f}) \leq 4(e^\alpha - 1)^2 d_{TV}(\mathbb{P}_f, \mathbb{P}_g)^2. \tag{3.36}$$

We have

$$D_{KL}(\mathbb{P}_{Q^n_{f_0}}, \mathbb{P}_{Q^n_{\nu_\rho}}) \leq \frac{1}{2^{K-1}} \sum_{\eta \in \{-1,1\}^{K-1}} D_{KL}(\mathbb{P}_{Q^n_{f_0}}, \mathbb{P}_{Q^n_{f_\eta}}). \tag{3.37}$$

And by application of the Kullback-Leibler divergence over products of distributions,

$$D_{KL}(\mathbb{P}_{Q^n_{f_0}}, \mathbb{P}_{Q^n_{f_\eta}}) = nD_{KL}(\mathbb{P}_{Q_{f_0}}, \mathbb{P}_{Q_{f_\eta}}).$$

So by application of Pinsker's inequality on one side of the inequality in Equation (3.37) and using Equation (3.36) on the other side, this implies

$$2d_{TV}(\mathbb{P}_{Q^n_{f_0}}, \mathbb{P}_{Q^n_{\nu_\rho}})^2 \leq \frac{4(e^\alpha - 1)^2}{2^L} \sum_{\eta \in \{-1,1\}^L} nd_{TV}(\mathbb{P}_{f_0} - \mathbb{P}_{f_\eta})^2.$$

So

$$d_{TV}(\mathbb{P}_{Q^n_{f_0}}, \mathbb{P}_{Q^n_{\nu_\rho}}) \leq \sqrt{2}(e^\alpha - 1)\sqrt{n} d_{TV}(\mathbb{P}_{f_0} - \mathbb{P}_{f_\eta}).$$

And by application of Lemma 24,

$$d_{TV}(\mathbb{P}_{f_0}, \mathbb{P}_{f_\eta}) \leq \frac{1}{2} \left(\mathbb{E}_{f_0} \left[L^2_{f_\eta}(X_1) - 1 \right] \right)^{1/2}.$$

Now

$$\mathbb{E}_{f_0} \left[L^2_{f_\eta}(X_1) \right] = 1 + 2\rho 2^{-J/2} \sum_{i=1}^{2^J} \eta_i \mathbb{E}_{f_0}(\psi_{J,i}) + \rho^2 2^{-J} \sum_{i=1}^{2^J} \eta_i^2 \mathbb{E}_{f_0}(\psi^2_{J,i}),$$

since $\psi_{J,i}$ have disjoint supports.

So

$$\mathbb{E}_{f_0} \left[L^2_{f_\eta}(X_1) \right] = 1 + \rho^2.$$

Finally,

$$d_{TV}(\mathbb{P}_{Q^n_{f_0}}, \mathbb{P}_{Q^n_{\nu_\rho}}) \leq \sqrt{2}/2(e^\alpha - 1)\sqrt{n}\rho.$$

□

Remark 37. *Focusing on the following term from the naive lower bound on the minimax rate*

$$1/\sqrt{n},$$

we notice a gap with what we obtain using our proof:

$$2^{3J/4}/\sqrt{n}.$$

The source of the gap is in the inequality presented in Equation (3.37). Indeed, on the left-hand side, there is a distance describing testing with an alternative hypothesis composed of 2^J elements. Whereas on the right-hand side, we have the average distance corresponding to testing with only a simple alternative hypothesis. This inequality is nonetheless applied in order to obtain univariate distributions over which Theorem 1 from Duchi, Jordan, and Wainwright (2013c) is applicable.

Chapter 4

Minimax adaptive rejection sampling

4.1 Introduction

The breadth of applications requiring independent sampling from a probability distribution is sizable. Numerous classical statistical results, and in particular those involved in machine learning, rely on the independence assumption. For some densities, direct sampling may not be tractable, and the evaluation of the density at a given point may be costly. Rejection sampling (RS) is a well-known Monte-Carlo method for sampling from a density f on $\mathbb{R}^d$ when direct sampling is not tractable (see Von Neumann, 1951, Devroye, 1986). It assumes access to a density g, called the proposal density, and a positive constant M, called the rejection constant, such that f is upper-bounded by Mg, which is called the *envelope*. Sampling from g is assumed to be easy. At every step, the algorithm draws a proposal sample X from the density g and a point U from the uniform distribution on $[0, 1]$, and accepts X if U is smaller than the ratio of $f(X)$ and $Mg(X)$, otherwise it rejects X. The algorithm outputs all accepted samples, which can be proven to be i.i.d. samples from the density f. This is to be contrasted with Markov Chain Monte Carlo (MCMC) methods which produce a sequence of non dependent samples and therefore fulfill a different objective. Besides, the application of rejection sampling includes variational inference: Naesseth et al. (2016) and Naesseth et al. (2017) generalize the reparametrization trick to distributions which can be generated by rejection sampling.

Adaptive rejection sampling is a variant of rejection sampling motivated by the high number of rejected samples with standard rejection sampling. Given n, a *budget* of evaluations of f, the goal is to maximize $\hat{n}$, the number of output samples which have to be drawn independently from f. In other words, the ratio $\frac{n-\hat{n}}{n}$, also called *rejection rate*, is to be made as small as possible, like in standard rejection sampling. To achieve this maximal number of output samples, adaptive rejection sampling methods gradually improve the proposal function and the rejection constant by using the information given by the evaluations of f at the previous proposal samples. These samples are used to estimate and tightly bound f from above.

4.1.1 Literature review

Closely related works. Erraqabi et al. (2016) provides an adaptive rejection sampling algorithm together with theoretical guarantees, making this work very relevant in comparison with ours. We detail their approach in Section 1.6.5 and we will prove in this chapter that their results are suboptimal. Another related sampling method is A* sampling (Maddison, Tarlow, and Minka, 2014). It is close to the OS* algorithm from Dymetman, Bouchard, and Carter (2012) and relies on an extension of

the Gumbel-max trick. The trick enables the sampling from a categorical distribution over classes $i \in [1, \dots, n]$ with probability proportional to $\exp(\phi(i))$, where ϕ is an unnormalized mass. It uses the following property of the Gumbel distribution. Adding Gumbel noise to each of the $\phi(i)$'s and taking the argmax of the resulting variables returns i with a probability proportional to $\exp(\phi(i))$. Then, the authors generalize the notion of Gumbel-max trick to a continuous distribution. This method shows good empirical efficiency in the number of evaluations of the target density. However, the assumption that the density can be decomposed into a bounded function and a function that is easy to integrate and sample from, is rarely true in practice.

Other related works. Gilks and Wild (1992) introduced ARS: a technique of adaptive rejection sampling for one-dimensional log-concave and differentiable densities whose derivative can be evaluated. ARS sequentially builds a tight envelope of the density by exploiting the concavity of $\log(f)$ in order to bound it from above. At each step, it samples a point from a proposal density. It evaluates f at this point, and updates the current envelope to a new one which is closer to f. The proposal density and the envelope thus converge towards f, while the rejection constant converges towards 1. The rejection rate is thereby improved. Gilks (1992) also developed an alternative to this ARS algorithm for the case where the density is not differentiable or the derivative can not be evaluated. The main difference with the former method is that the computation of the new proposal does not require any evaluation of the derivative. For this algorithm, as for the one presented in Gilks, Best, and Tan (1995), the assumption that the density is log-concave represents a substantial constraint in practice. In particular, it restrains the use of ARS to unimodal densities.
An extension from Hörmann (1995) of ARS adapts it to T-concave densities, with T being a monotonically increasing transformation. However, this method still cannot be used with multimodal densities. In 1998, Evans and Swarz proposed a method applicable to multimodal densities presented in Evans and Swartz (1998) which extends the former one. It deals with T-transformed densities and spots the intervals where the transformed density is concave or convex. Then it applies an ARS-like method separately on each of these intervals. However it needs access to the inflection points, which is a strong requirement. A more general method in Görür and Teh (2011) consists of decomposing the log of the target density into a sum of a concave and convex functions. It deals with these two components separately. An obvious drawback of this technique is the necessity of the decomposition itself, which may be a difficult task. Similarly, Martino and Míguez (2011) deal with cases where the log-density can be expressed as a sum of composition of convex functions and of functions that are either convex or concave. This represents a relatively broad class of functions; other variants focusing on the computational cost of ARS have been explored in Martino (2017) and Martino and Louzada (2017).
For all the methods previously introduced, no theoretical efficiency guarantees are available.

A further attempt at improving simple rejection sampling resulted in Adaptive Rejection Metropolis Sampling (ARMS) (Gilks, Best, and Tan, 1995). ARMS extends ARS to cases where densities are no longer assumed to be log-concave. It builds a proposal function whose formula is close to the one in Gilks (1992). This time however, the proposal might not be an envelope, which would normally lead to oversampling in the regions where the proposal is smaller than the density. In ARMS, this is compensated with a Metropolis-Hastings control-step. One drawback of this method is that it outputs a Markov Chain, in which the samples are correlated. Moreover, the chain may be trapped in a single mode. Improved adaptive rejection Metropolis

(Martino, Read, and Luengo, 2012) modifies ARMS in order to ensure that the proposal density tends to the target density. In Meyer, Cai, and Perron (2008) an alternative is presented that uses polynomial interpolations as proposal functions. However, this method still yields correlated samples.

Markov Chain Monte Carlo (MCMC) methods (Metropolis and Ulam, 1949; Andrieu et al., 2003) represent a very popular set of generic approaches in order to sample from a distribution. Although they scale with dimension better than rejection sampling, they are not perfect samplers, as they do not produce i.i.d. samples, and can therefore not be applied to achieve our goals. Variants producing independent samples were proposed in Fill (1997) and Propp and Wilson (1998). However, to the best of our knowledge, no theoretical studies on the rejection rate of these variants is available in the literature.

Importance sampling is a problem close to rejection sampling, and adaptive importance sampling algorithms are also available (see e.g., Oh and Berger, 1992; Cappé et al., 2008; Ryu and Boyd, 2014). Among them, the algorithm in Zhang (1996) sequentially estimates the target function, whose integral has to be computed using kernel regression, similarly to the approach of Erraqabi et al. (2016). A recent notable method regarding discrete importance sampling was introduced in Canévet, Jose, and Fleuret (2016). In Delyon and Portier (2018), adaptive importance sampling is shown to be efficient in terms of asymptotic variance.

4.1.2 Our contributions

The above mentioned sampling methods either do not provide i.i.d samples, or do not come with theoretical efficiency guarantees, apart from Erraqabi et al. (2016) or Zhang (1996) and Delyon and Portier (2018) in importance sampling. In the present work, we propose the Nearest Neighbour Adaptive Rejection Sampling algorithm (NNARS), an adaptive rejection sampling technique which requires f to have s-Hölder regularity (see Assumption 2). Our contributions are threefold, since NNARS:

- is a *perfect sampler* for sampling from the density f.
- offers an *average rejection rate of order* $\log(n)^2 n^{s/d}$, if $s \leq 1$. This significantly improves the state of the art average rejection rate from Erraqabi et al. (2016) over s-Hölder densities, which is of order $(\log(nd)/n)^{\frac{s}{3s+d}}$.
- matches a *lower bound for the rejection rate* on the class of all adaptive rejection sampling algorithms and all s-Hölder densities. It gives an answer to the theoretical problem of quantifying the difficulty of adaptive rejection sampling in the minimax sense. So NNARS offers a near-optimal average rejection rate, in the minimax sense over the class of Hölder densities.

NNARS follows a common approach to that of most adaptive rejection sampling methods. It relies on non-parametric estimation of f. It improves this estimation iteratively, and as the latter gets closer to f, the envelope also approaches f. Our improvements consist of designing an optimal envelope, and updating the envelope as we get more information at carefully chosen times. This leads to an average rejection rate for NNARS which is minimax near-optimal (up to a logarithmic term) over the class of Hölder densities. No adaptive rejection algorithm can perform significantly better in this class. The proof of the minimax lower bound is also new to the best of our knowledge.

The optimal envelope we construct is a very simple one. For every known point of the target density f, we use the regularity assumptions on f in order to construct an

envelope which is piecewise constant. It stays constant in the neighborhood of every known point of f. Figure 4.1 depicts NNARS' first steps on a mixture of Gaussians in dimension 1.

In the second section of this chapter, we set the problem formally and discuss the assumptions that we make. In the third section, we introduce the NNARS algorithm and provide a theoretical upper bound on its rejection rate. In the fourth section, we present a minimax lower bound for the problem of adaptive rejection sampling. In the fifth section, we discuss our method and detail the open questions regarding NNARS. In the sixth section, we present experimental results on both simulated and real data that compare our strategy with state of the art algorithms for adaptive rejection sampling. The implementation of the code of NNARS can be found on the following webpage: `https://github.com/jlamweil/NNARS`. Finally, the proofs of all the results presented in this chapter are left for the later sections.

4.2 Setting

Let f be a bounded density defined on $[0,1]^d$. The objective is to provide an algorithm which outputs as many i.i.d. samples drawn according to f as possible, with a fixed number n of evaluations of f. We call n the budget.

4.2.1 Description of the problem

The framework that we consider is *sequential and adaptive rejection sampling* for which we already give a very detailed description in Section 1.6.4. So we will simply recall a few notions here.

The rejection sampling step is defined in Algorithm 2.

Algorithm 2 Rejection Sampling Step with (f, g, M) : $\mathrm{RSS}(f, g, M)$

Input: target density f, proposal density g, rejection constant M.
Output: Either a sample X from f, or nothing.
Sample $X \sim g$ and $U \sim \mathcal{U}_{[0,1]}$.
if $U \leq \frac{f(X)}{Mg(X)}$ **then**
  output X.
else
  output $\emptyset$.
end if

Then one defines the class of adaptive rejection sampling algorithms.

Definition 18. (Class of Adaptive Rejection Sampling (ARS) Algorithms) *An algorithm A is an ARS algorithm if, given f and n, at each step $t \in \{1 \dots n\}$:*

- *A chooses a density g_t, and a positive constant M_t, depending on $\Big\{(X_1, f(X_1)), \dots, (X_{t-1}, f(X_{t-1}))\Big\}$.*
- *A performs a Rejection Sampling Step with (f, g_t, M_t).*

The objective of an ARS algorithm is to sample as many i.i.d. points according to f as possible.

We also recall the following result from Section 1.6.4, giving a sufficient condition under which an adaptive rejection sampling algorithm outputs i.i.d. samples from f.

Theorem 33. *Given access to a positive, bounded density f defined on $[0,1]^d$, any Adaptive Rejection Sampling algorithm (as described above) satisfies:*
if $\forall t \leq n,\ \forall x \in [0,1]^d,\ f(x) \leq M_t g_t(x)$, the output $\mathcal{S}$ contains i.i.d. samples drawn according to f.

Definition of the loss. If the learner is a perfect sampler at every step, we define the loss as $L_n = n - \#\mathcal{S}$, which corresponds to the number of rejected samples. Otherwise, we just set $L_n = n$. Finally, we note that the rejection rate is L_n/n.

4.2.2 Assumptions

We make the following assumptions on f. They will be used by the algorithm and for the theoretical results.

Assumption 2.

- *The function f is (s,H)-Hölder with respect to $\mathbb{L}_\infty$ for some $0 < s \leq 1$ and $H \geq 0$,*
  i.e., $\forall x, y \in [0,1]^d,\ |f(x) - f(y)| \leq H\|x-y\|_\infty^s$, where $\|u\|_\infty = \max_i |u_i|$;
- *There exists $0 < c_f \leq 1$ such that $\forall x \in [0,1]^d,\ c_f < f(x)$.*

Let $\mathcal{F}_0 := \mathcal{F}_0(s, H, c_f, d)$ denote the set of functions satisfying Assumption 2 for given $0 < s \leq 1$, $H \geq 0$ and $0 < c_f \leq 1$.

Remarks. Here the domain of f is assumed to be $[0,1]^d$, but it could without loss of generality be relaxed to any hyperrectangle of $\mathbb{R}^d$. Besides, for any distribution with sub-Gaussian tails, this assumption is almost true. In practice, the diameter of the support is bounded by $O(\sqrt{\log n})$, where n is the number of evaluations, because of the vanishing tail property. The assumption of Hölder regularity is a usual regularity assumption in order to control for rates of convergence. It is also a mild one, considering that s can be chosen arbitrarily close to 0. Note however that we assume the knowledge of s and H for the NNARS algorithm. Since f is a Hölder regular density defined on the unit cube, we can obtain the following upper bound: $f(x) \leq 1 + H\ \forall x \in [0,1]^d$. As for the assumption involving the constant c_f, it is widespread in non-parametric density estimation. Besides, the algorithm will still produce exact independent samples from the target distribution without the latter assumption. It is important to note that f is chosen as a probability density for clarity, but it is not a required hypothesis. In the proofs, we study the general case when f is not assumed to be a probability density.

4.3 The NNARS Algorithm

The NNARS algorithm proceeds by constructing successive proposal functions g_t and rejection constants M_t that gradually approach f and 1, respectively. In turn, an increased number of evaluations of f should result in a more accurate estimate and thus in a better upper bound.

4.3.1 Description of the algorithm

The algorithm outlined in Algorithm 3 takes as inputs the budget n, and c_f, H, s as defined in Assumption 2. Let $\mathcal{S}$ denote its output. At each round k, the algorithm performs a number of RSS steps with specifically chosen g_k and M_k. We call χ_k the set of points generated at round k and of their images by f, whether they get accepted or rejected.

Initialization. The sets $\mathcal{S}$ and $\chi_k, k \in \mathbb{N}$ are initialized to $\emptyset$. g_1 is a uniform proposal on $[0,1]^d$. $M_1 = 1 + H$ is an upper bound on f. $N = N_1 = \lceil 2(10H)^{d/s} \log(n) c_f^{-1-d/s} \rceil$. For any function h defined on $[0,1]^d$, we set $I_h = \int_{[0,1]^d} h(x)dx$.

Loop. The algorithm proceeds in $K = \left\lceil \log_2(\frac{n}{N}) \right\rceil$ rounds, where $\lceil\ \rceil$ is the ceiling function, and $\log_2$ is the logarithm in base 2.
Each round $k \in \{1, \dots, K\}$ consists of the following steps.

1. Perform a Rejection Sampling Step $\mathrm{RSS}(f, g_k, M_k)$ N_k times. Add the accepted samples to $\mathcal{S}$. All proposal samples as well as their images by f produced in the Rejection Sampling Step are stored in χ_k, whether they are rejected or not.

2. Build an estimate $\hat{f}_{\cup_{i \le k} \chi_i}$ of f based on the evaluations of f at all points stored in $\cup_{i \le k} \chi_i$, thanks to the Approximate Nearest Neighbor Estimator, referred to in Definition 19, applied to the set χ_k.

3. Compute the proposal with the formula:

$$g_{k+1}(x) = \frac{\hat{f}_{\cup_{i \le k} \chi_i}(x) + \hat{r}_{\cup_{i \le k} \chi_i}}{I_{\hat{f}_{\cup_{i \le k} \chi_i}} + \hat{r}_{\cup_{i \le k} \chi_i}}, \tag{4.1}$$

   and the rejection constant with the formula:

$$M_{k+1} = I_{\hat{f}_{\cup_{i \le k} \chi_i}} + \hat{r}_{\cup_{i \le k} \chi_i}, \tag{4.2}$$

   where $\hat{r}_{\cup_{i \le k} \chi_i}$ is defined in Equation (4.3) below, in Definition 19. Note that g_{k+1} and M_{k+1} are indexed here by the number of the round, unlike in the last section where the index was the current time.

4. If $k < K - 1$, set N_{k+1} as $2N_k = 2^k N$. Otherwise $N_K = n - (2^{K-1} - 1)N$.

Finally, the algorithm outputs $\mathcal{S}$, the set of accepted samples that have been collected.

Definition 19 (Approximate Nearest Neighbor Estimator applied to χ)**.**
Let f be a positive density satisfying Assumption 2. We consider a set of $\tilde{N}$ points and their images by f, $\chi = \{(X_1, f(X_1)), \dots, (X_{\tilde{N}}, f(X_{\tilde{N}}))\}$. Let us define a set of centers of cells constituting a uniform grid of $[0,1]^d$, namely

$$\mathcal{C}_{\tilde{N}} = \left\{ 2^{-1} (\lfloor \tilde{N}^{\frac{1}{d}} \rfloor + 1)^{-1} u, u \in \{1, \dots, 2(\lfloor \tilde{N}^{\frac{1}{d}} \rfloor + 1) - 1\}^d \right\}.$$

The cells are of side-length $1/(\lfloor N^{\frac{1}{d}} \rfloor + 1)$. For $x \in [0,1]^d$, write

$$C_{\tilde{N}}(x) = \arg\min_{u \in \mathcal{C}_{\tilde{N}}} \|x - u\|_\infty.$$

We define the approximate nearest neighbour estimator, related to the estimator presented in Equation (1.2) in Section 1.3.2, as the piecewise-constant estimator $\hat{f}_\chi$ *of* f *by* $\forall x \in [0,1]^d,\ \hat{f}_\chi(x) = \hat{f}_\chi(C_{\tilde{N}}(x)) = f\left(X_{i(C_{\tilde{N}}(x))}\right)$, *where* $i(x) = \arg\min_{i \leq \tilde{N}}(\|x - X_i\|_\infty)$.

We also define a confidence term as

$$\hat{r}_\chi = H\left(\max_{u \in \mathcal{C}_{\tilde{N}}} \min_{i \leq \tilde{N}} \|u - X_i\|_\infty + \frac{1}{2(\lfloor \tilde{N}^{\frac{1}{d}} \rfloor + 1)}\right)^s. \tag{4.3}$$

Remarks on the proposal densities and rejection constants. At each step, the envelope is made up of evaluations of f summed with a positive constant which stands for a confidence term of the estimation. It provides an upper bound for f. Furthermore, the use of nearest neighbour estimation in a noiseless setting implies that this bound is optimal. Besides, the approximate construction of the estimator builds proposal densities which are simple to sample from.

As explained in Lemma 30, an important remark is that the proposal density g_k from Equation (4.1) multiplied by the rejection constant M_k from Equation (4.2) is an envelope of f. This means $M_k g_k \geq f$ for all $k \leq K$. So by Theorem 33, NNARS is a perfect sampler.

The algorithm is illustrated in Figure 4.1.

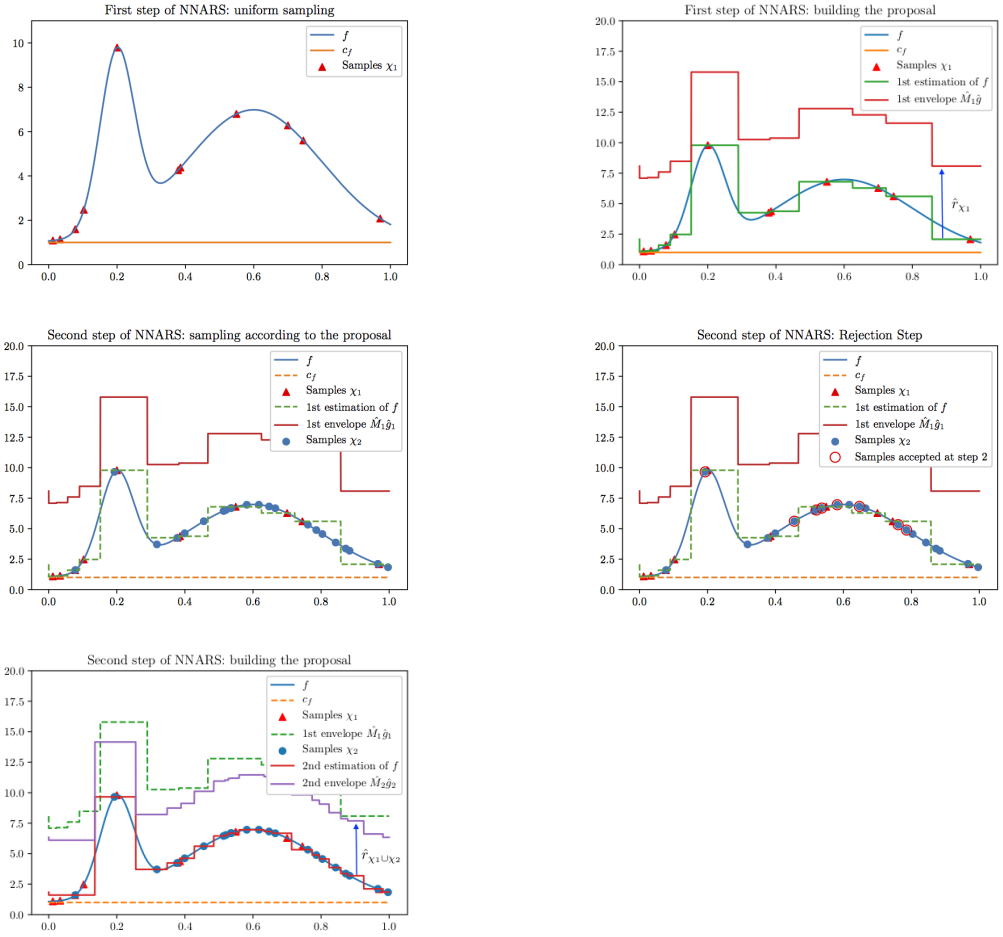

FIGURE 4.1: NNARS' first steps on a mixture of Gaussians (ordered in the natural reading direction)

Algorithm 3 Nearest Neighbor Adaptive Rejection Sampling

Input: the budget n; the constants H, s and c_f; the dimension d.
Output: the set $\mathcal{S}$ of i.i.d. samples from f.
Initialize $\mathcal{S} = \emptyset$, $\chi_k = \emptyset \ \forall k$.
Set $N_1 = N$, $g_1 = \mathcal{U}_{[0,1]^d}$, $M_1 = 1 + H$.
for $k = 1$ **to** K **do**
  for $i = 1$ **to** N_k **do**
    Perform a Rejection Sampling Step RSS(f, g_k, M_k).
    Add the output of RSS to $\mathcal{S}$.
    Add to χ_k both the sample from g_k collected in the RSS, and its image by f.
  end for
  Estimate $\hat{f}_{\cup_{i\leq k}\chi_k}$ according to Definition 19.
  Compute g_{k+1} and M_{k+1} as in Equations (4.1) and (4.2).
  if $k < K - 1$ **then**
    set $N_{k+1} = 2N_k$.
  else
    Set $N_K = n - (2^{K-1} - 1)N$.
  end if
end for

Remark on sampling from the proposal densities g_k in NNARS. The number of rounds is of order $\lfloor \log(n) \rfloor$. The construction of the proposal in NNARS involves at each round k the storage of $|\cup_{i\leq k} \chi_i| \propto p^{k+1}\lfloor \log(n) \rfloor$ values. So the total number of values stored is upper bounded by the budget. At each round, each value corresponds to a hypercube of side-length $1/|\cup_{i\leq k} \chi_i|^{1/d}$ splitting $[0,1]^d$ equally. Partitioning the space in this way allows us to efficiently assign a value to every $x \in [0,1]^d$, depending on which cell of the grid x belongs to. Besides, sampling from the proposal amounts to sampling from a multinomial convolved with a uniform distribution on a hypercube. In other words, a cell is chosen multinomially, then a point is sampled uniformly inside that cell, because the proposal is piecewise constant.

The process to sample according to g_k is the following: given $\cup_{i\leq k}\chi_i$,

1. Each center of the cells from the grid $u \in \mathcal{C}_{|\cup_{i\leq k}\chi_i|}$ is mapped to a value $g_k(u)$.
2. One of the centers $\tilde{C} \in \mathcal{C}_{|\cup_{i\leq k}\chi_i|}$ is sampled with probability $g_k(\tilde{C})$.
3. The sample point is drawn according to the uniform distribution on the hypercube of center $\tilde{C}$ and side-length $1/|\cup_{i\leq k} \chi_i|^{1/d}$.

4.3.2 Upper bound on the loss of NNARS

In this section, we present an upper bound for the expectation of the loss of the NNARS sampler. This bound holds under Assumption 2, that only requires n to be large enough in comparison with constants depending on d, s, c_f and H. Related conditions about the sample size are in most theoretical works on Rejection Sampling (Gilks and Wild, 1992, Meyer, Cai, and Perron, 2008, Görür and Teh, 2011, Erraqabi et al., 2016).

Assumption 3 (Assumption on n)**.**
Assume that $n \geq 8$ and $N/n \leq 1/(2K^2)$, i.e.,

$$n \geq \left\lceil 2(10H)^{d/s} \log(n) c_f^{-1-d/s} \right\rceil \frac{4\log(n)^2}{\log(2)^2} = O(\log(n)^3).$$

Theorem 34. *Let $0 < s \leq 1$, $H \geq 0$ and $c_f > 0$. If f satisfies Assumption 2 with (s, H, c_f) such that $f \in \mathcal{F}_0(s, H, c_f, d)$, then NNARS is a perfect sampler according to f.*

Besides if n satisfies Assumption 3, then

$$\begin{aligned}\mathbb{E}_f L_n(NNARS) \leq & \frac{20}{2^{1-s/d}-1} {c_f}^{-2}(1+\sqrt{2\log 3n})\log^{s/d}(5n)n^{1-s/d} \\ & + \quad (25+40+2(10Hc_f^{-1})^{d/s})c_f^{-1}\log^2(n) = O(\log^2(n)n^{1-s/d}),\end{aligned}$$

where $\mathbb{E}_f L_n(NNARS)$ is the expected loss of NNARS on the problem defined by f. The expectation is taken over the randomness of the algorithm. This result is uniform over $\mathcal{F}_0(s, H, c_f, d)$.

The proof of this theorem is in Section 4.8. The loss presented here divided by n is to be interpreted as an upper bound for the expected rejection rate obtained by the NNARS algorithm.

Sketch of the proof. The average number of rejected samples is $\sum_k N_k(1 - 1/M_k)$, since a sample is accepted at round k with probability $1 - 1/M_k$. In order to bound the average number of rejected samples, we bound M_k at each round k with high probability.

By Hölder regularity and the definition of $\hat{r}_{\cup_{i\leq k}\chi_i}$ in Equation (4.3) (in Definition 19), we always have $|\hat{f}_{\cup_{i\leq k}\chi_i} - f| \leq \hat{r}_{\cup_{i\leq k}\chi_i}$, as shown in the proof of Proposition 16. So $M_k = I_{\hat{f}_{\cup_{i\leq k-1}\chi_i}} + \hat{r}_{\cup_{i\leq k-1}\chi_i} \leq I_f + 2\hat{r}_{\cup_{i\leq k-1}\chi_i}$ with $I_f = 1$. Then, we consider the event $\mathcal{A}_{k,\delta} = \{\forall j \leq k,\ \hat{r}_{\cup_{i\leq j}} \leq C_0 H(\log(N_j/\delta)/N_j)^{s/d}\}$, where C_0 is a constant. Now, for each k, on $\mathcal{A}_{k-1,\delta}$, M_k is bounded from above, with a bound of the order of $(\log(N_{k-1}/\delta)/N_{k-1})^{s/d}$. So, on $\mathcal{A}_{K,\delta}$, the average number of rejected samples has an upper bound of the order of $\log(n)^2 n^{1-s/d}$, as presented in Theorem 34.

Now, we prove by induction that the event $\mathcal{A}_{k,\delta}$ has high probability, as in the proof of Lemma 31. More precisely, $\mathcal{A}_{k,\delta}$ has probability larger than $1 - 2k\delta$. At every step k, we verify that g_k is positively lower bounded conditionally on $\mathcal{A}_{k-1,\delta}$. Hence, the probability of having drawn at least one point in each hypercube of the grid with centers $\mathcal{C}_{|\cup_{i\leq k}\chi_i|}$ is high, as shown in the proof of Proposition 17. So the distance from any center to its closest drawn point is upper bounded with high probability. And this implies that $\mathcal{A}_{k,\delta}$ has high probability if $\mathcal{A}_{k-1,\delta}$ has high probability, which gets the induction going. On the other hand, the number of rejected samples is always bounded by n on the small probability event where $\mathcal{A}_{K,\delta}$ does not hold. This concludes the proof.

4.4 Minimax Lower Bound on the Rejection Rate

It is now essential to get an idea of how much it is possible to reduce the loss obtained in Theorem 34. That is why we apply the framework of minimax optimality and complement the upper bound with a lower bound. The minimax lower bound on this problem is the infimum of the supremum of the loss of algorithm A on the problem defined by f; the infimum is taken over all adaptive rejection sampling algorithms A and the supremum over all functions f satisfying Assumption 2. It characterizes the difficulty of the rejection sampling problem. And it provides the best rejection rate that can possibly be achieved by such an algorithm in a worst-case sense over the class $\mathcal{F}_0(s, H, c_f, d)$.

Theorem 35. *For $0 < s \leq 1$, there exists a constant $N(s,d)$ that depends only on s, d and such that for any $n \geq N(s,d)$:*

$$\inf_{A \in \mathcal{A}} \sup_{f \in \mathcal{F}_0(s,1,1/2,d) \cap \{f : I_f = 1\}} \mathbb{E}_f(L_n(A)) \geq 3^{-1} 2^{-1-3s-2d} 5^{-s/d} n^{1-s/d} = O(n^{1-s/d}),$$

where $\mathbb{E}_f(L_n(A))$ is the expectation of the loss of A on the problem defined by f. It is taken over the randomness of the algorithm A.

The proof of this theorem is in Section 4.9, but the following discussion outlines its main arguments.

Sketch of the proof in dimension 1. Consider the setup where firstly n points from f are chosen and evaluated. Secondly, n other points are sampled using rejection sampling with a proposal based only on the n first points. This is related to Definition 1. This setting is easier than that of adaptive rejection sampling, as proven in Lemma 32. Consequently a lower bound for this simpler problem also constitutes a lower bound for adaptive rejection sampling. Now, $\mathcal{F}_0(1,1,1/2,1)$ corresponds to one-dimensional $(1,1)$-Hölder functions which are bounded from below by $1/2$. We consider a subset of $\mathcal{F}_0(1,1,1/2,1)$ satisfying Assumption 2. Set $V_n = \{\nu = (\nu_i)_{0 \leq i \leq 4n-1} \mid \nu_i \in \{-1,1\}, ; \sum_{i=0}^{4n-1} \nu_i = 0\}$.

Let us define the bump function $b : [0, 1/(4n)] \to \mathbb{R}^+$ such that for any $\nu \in V_n$:

$$b(x) = \begin{cases} x, & \text{for } x \leq 1/(8n). \\ 1/(4n) - x, & \text{otherwise.} \end{cases}$$

We will consider the following functions $f_\nu : [0,1] \to \mathbb{R}_+^*$ such that for any $\nu \in V_n$:

$$f_\nu(x) = 1 + \nu_i b(x - i/(4n)), \text{ if } i/(4n) \leq x \leq (i+1)/(4n),$$

We note that $f_\nu \in \mathcal{F}_0(1,1,1/2,1)$, for n large enough ensuring that $f_\nu \geq 1/2$.

An upward bump at position i corresponds to $\nu_i = 1$ and a downward bump to $\nu_i = -1$. The construction presented here is analog to the one in the proof of Lemma 34. The function f_ν is entirely determined by the knowledge of ν. It is only possible to determine a ν_i by evaluating f at some $x \in (i/(4n), (i+1)/(4n))$. So with a budget of n, we observe at most n of the $4n$ signs in ν. Among the unobserved ν_i, at least n are positive and n are negative, because $\sum_{i=0}^{4n-1} \nu_i = 0$. Now, we compute the loss. In the case when Mg is not an envelope, the loss simply is n. Now let us consider the case where Mg is an envelope. The loss is $n(1 - 1/I_{Mg})$. Mg has to account for at least n upward bumps at unknown positions; and the available information is insufficient to distinguish between upward and downward bumps. This results in an envelope that is not tight for the negative ν_i with unknown positions. So a necessary loss is incurred at the downward bumps corresponding to those negative ν_i. This translates as $I_{Mg} - 1 \geq n c_s n^{-(1+s)}$, where c_s is a constant only dependent on s, with $s = 1$ in our case. Finally, we obtain a risk $n(1 - 1/I_{Mg})$ which is of order n^{1-s}, as seen in Lemma 33.

In a nutshell, we first made a setup with more available information than in the problem of adaptive rejection sampling, from Definition 18. Then we restricted the setting to some subspace of $\mathcal{F}_0(1,1,1/2,1)$. This led to our obtaining of a lower bound on the risk for an easier setting. This implies we have displayed a lower bound for the problem of adaptive rejection sampling over $\mathcal{F}_0(1,1,1/2,1)$, too.

This theorem gives a lower bound on the minimax risk of all possible adaptive rejection sampling algorithms. Up to a $\log(n)$ factor, NNARS is minimax-optimal and the rate in the lower bound is then the minimax rate of the problem. It is remarkable that this problem admits such a fast minimax rate; the same rate as a standard rejection sampling scheme with an envelope built using the knowledge of n evaluations of the target density (see Setting 1).

4.5 Discussion

Theorem 35 asserts that NNARS provides a minimax near-optimal solution in expectation, up to a multiplicative factor of the order of $\log(n)^{s/d}$. This result holds for all adaptive rejection sampling algorithms and densities in $\mathcal{F}_0(s, H, c_f, d)$. To the best of our knowledge, this is the first time a lower bound is proved on adaptive rejection samplers; or that an adaptive rejection sampling algorithm that achieves near-optimal performance is presented. In order to ensure the theoretical rates mentioned in this work, the algorithm requires to know c_f, a positive lower bound for f, and the regularity constants of f: s, and H. Note that to achieve a near-optimal rejection rate, the precise knowledge of s is required. Indeed, replacing the exponent s by a smaller number will result in adding a confidence term $\hat{r}_{\cup_{i\leq k}\chi_i}$ to the estimator which is too large. Finally, it will result in a higher rejection rate than if one had set s to the exact Hölder exponent of f. The assumption on c_f implies in particular that f does not vanish. However, as long as it remains positive, c_f can be chosen arbitrarily small, and n has to be taken large enough to ensure that c_f is approximately larger than $\frac{1}{\log\log n}$. When c_f is not available, asymptotically taking c_f of this order will offer a valid algorithm, which outputs independent samples drawn according to f. Moreover taking c_f of this order will still result in a minimax near-optimal rejection rate. Indeed it will approximately boil down to multiplying the rejection rate by a $\log\log n$. Similarly H can be taken of order $\log n$ without hindering the minimax near-optimality. Extending NNARS to non lower-bounded densities is still an open question.

The algorithm NNARS is a perfect sampler. Since our objective is to maximize the number of i.i.d. samples generated according to f, we cannot compare the algorithm with MCMC methods, which provide non-i.i.d. samples. In our setting, they have a loss of n. The same argument is valid for other adaptive rejection samplers that produce correlated samples, like e.g., Gilks, Best, and Tan (1995), Martino, Read, and Luengo (2012), and Meyer, Cai, and Perron (2008).
Considering other perfect adaptive rejection samplers, like the ones in e.g., Gilks (1992), Martino and Míguez (2011), Hörmann (1995), and Görür and Teh (2011), their assumptions differ in nature from ours. Instead of shape constraint assumptions, like log-concavity, which are often assumed in the quoted literature, we only assume Hölder regularity. Note that log-concavity implies Hölder regularity of order two almost everywhere. Moreover no theoretical results on the proportion of rejected samples are available for most samplers, except possibly asymptotic convergence to 0, which is induced by our result.

Pliable rejection sampling (PRS) from Erraqabi et al. (2016) is the only algorithm with a theoretical guarantee on the rate with the proportion of rejected samples decreasing to 0. But it is not optimal, as explained in Section 4.1. So the near-optimal rejection rate is a major asset of the NNARS algorithm compared to the PRS algorithm. Besides, PRS only provides an envelope with high probability, whereas NNARS provides it with probability 1 at any time. The improved performance of

NNARS compared to PRS may be attributed to the use of an estimator more adapted to noiseless evaluations of f, and to the multiple updates of the proposal.

4.6 Experiments

Let us compare NNARS numerically with Simple Rejection Sampling (SRS), PRS (Erraqabi et al., 2016), OS* (Dymetman, Bouchard, and Carter, 2012) and A* sampling (Maddison, Tarlow, and Minka, 2014). The value of interest is the sampling rate corresponding to the proportion of samples produced with respect to the number of evaluations of f. This is equivalent to the acceptance rate in rejection sampling. Every result is an average over 10 runs with its standard deviation. The implementation of the code of NNARS can be found on the following webpage: `https://github.com/jlamweil/NNARS`.

4.6.1 Presentation of the experiments

EXP1. We first consider the following target density from Maddison, Tarlow, and Minka (2014): $f(x) \propto e^{-x}/(1+x)^a$, where a is the peakiness parameter. Increasing a also increases the sampling difficulty. In Figure 4.2a, PRS and NNARS both give good results for low peakiness values, but their sampling rates fall drastically as the peakiness increases. So their results are similar to SRS after a peakiness of 5.0. On the other hand, the rates of A* and OS* sampling decrease more smoothly.

EXP2. For the next experiment, we are interested in how the method scale when the dimension increases and consider a distribution that is related to the one in Erraqabi et al. (2016): $f(x_1, \ldots, x_d) \propto \prod_{i \in [|0,1|]^d} \left(2 + \sin\left(4\pi x_i - \frac{\pi}{2}\right)\right)$, where $(x_1, \ldots, x_d) \in [0,1]^d$. In Figure 4.2b, we present the results for d between 1 and 7. NNARS scales the best in dimension. A* and OS* have the same behaviour, while PRS and SRS share very similar results. A* and OS* start with good sampling rates, which however decrease radically when the dimension increases.

EXP3. Then, we focus on how the efficiency scales with respect to the budget. The distribution tested is: $f(x) \propto \exp(\sin(x))$, with x in $[0, 1]$. In Figure 4.3a, NNARS, A* and OS* give the best performance, reaching the asymptotic regime after 20,000 function evaluations. So NNARS is applicable in a reasonable number of evaluations. Coupled with the study of the evolution of the standard deviations in Figure 4.3b, we conclude that the results in the transition regime may vary, but the time to the asymptotic region is not initialization-sensitive.

EXP4. Finally, we show the efficiency of NNARS on non-synthetic data from the set in Cortez and Morais (2007). It consists of 517 observations of meteorological data used in order to predict forest fires in the north-eastern part of Portugal. The goal is to enlarge the data set. So we would like to sample artificial data points from a distribution which is close to the one which generated the data set. This target distribution is obtained in a non-parametric way, using the Epanechnikov kernel which creates a non-smooth f. We then apply samplers which do not use the decomposition of f described in Maddison, Tarlow, and Minka (2014). That is why A* and OS* sampling will not be applied. From the 13 dimensions of the dataset we work with those corresponding to Duff Moisture Code (DMC) and Drought Code (DC) and we get the sampling rates in Table 4.1. NNARS clearly offers the best performance.

(A) Sampling rate vs. Peakiness [Exp1] (B) Sampling rate vs. Dimension [Exp2]

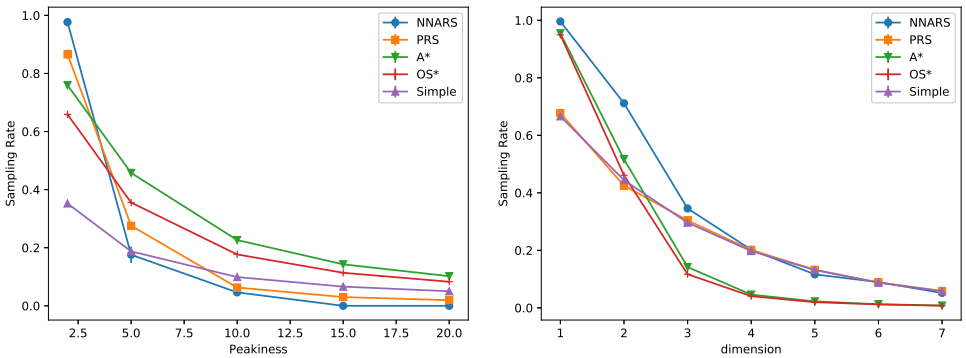

FIGURE 4.2: Empirical sampling rates for [Exp1] and [Exp2]

(A) Sampling rate vs. Budget [Exp3]

(B) Standard deviation of the sampling rate vs. budget [Exp3]

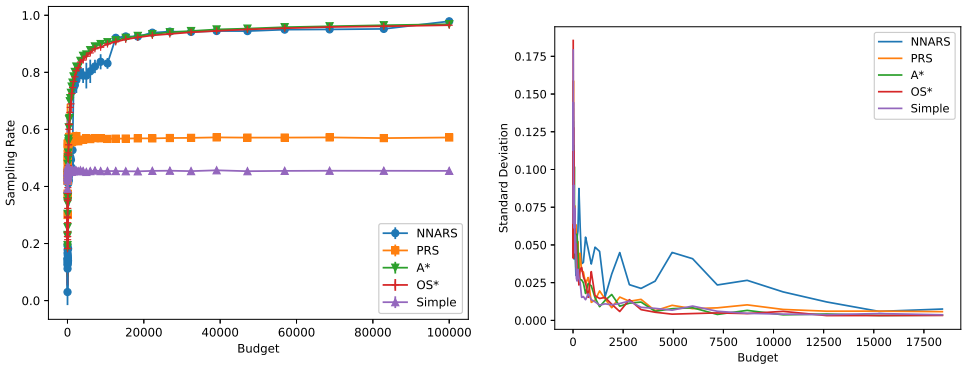

FIGURE 4.3: Empirical sampling rates and their standard deviations for [Exp3]

n=10^5, 2D	sampling rate
NNARS	$45.7\% \pm 0.1\%$
PRS	$16.0\% \pm 0.1\%$
SRS	$15.5\% \pm 0.1\%$

TABLE 4.1: Sampling rates for forest fires data [Exp4]

4.6.2 Synthesis on the numerical experiments

The essential features of NNARS have been brought to light in the experiments presented in Figures 4.2, 4.3 and using the non-synthetic data from Cortez and Morais (2007). In particular, Figure 4.3a gives the evidence that the algorithm reaches good sampling rates in a relatively small number of evaluations of the target distribution. Furthermore, Figure 4.2b illustrates the possibility of applying the algorithm in a multidimensional setting. In Figure 4.2a, we observe that A* and OS* sampling benefit from the knowledge of the specific decomposition of f needed in Maddison, Tarlow, and Minka (2014). We highlight the fact that this assumption is not true in general. Besides, A* sampling requires relevant bounding and splitting strategies. We note that tuning NNARS only requires the choice of a few numerical hyperparameters. They might be chosen thanks to generic strategies like grid search. Finally, the application to forest fire data generation illustrates the great potential of NNARS for applications reaching beyond the scope of synthetic experiments.

4.7 Conclusion

In this work, we introduced an adaptive rejection sampling algorithm, which is a perfect sampler according to f. It offers a rejection rate of order $(\log(n)/n)^{s/d}$, if $s \leq 1$. This rejection rate is near-optimal, in the minimax sense over the class of s-Hölder smooth densities. Indeed, we provide the first lower bound for the adaptive rejection sampling problem, which provides a measure of the difficulty of the problem. Our algorithm matches this bound up to logarithmic terms.

In the experiments, we test our algorithm in the context of synthetic target densities and of a non-synthetic dataset. A first set of experiments shows that the behavior of the sampling rate of our algorithm is similar to that of state of the art methods, as the dimension and the budget increase. Two of the methods used in this set of experiments require the target density to allow a specific decomposition. Therefore, these methods are neglected for the experiment which aims at generating forest fire data. In this experiment, NNARS clearly performs better than its competitors.

The extension of the NNARS algorithm to non lower-bounded densities is still an open question, as well as the development of an optimal adaptive rejection sampler, when the density's derivative is Hölder regular instead. We leave these interesting open questions for future work.

4.8 Proof of Theorem 34

In the following sections, we do not assume that f is a density. In fact ARS samplers could be given evaluations of the density multiplied by a positive constant. We prove in the sequel that as long as the resulting function satisfies Assumption 2, the upper bound presented in Theorem 34 holds in this case as well as in the case when f is a density. The lower bound is also proved without the assumption that f is a density.

4.8.1 Approximate Nearest Neighbor Estimator

In this subsection, we study the characteristics of the Approximate Nearest Neighbor Estimator. First, we prove a bound on the distance between the image of x by the Approximate Nearest Neighbor Estimator of f and $f(x)$, under the condition that f satisfies Assumption 2. More precisely, we prove that $\hat{f}(x)$ lies within a radius of $\hat{r}_\chi$ away from $f(x)$. Then we prove a high probability bound on the radius $\hat{r}_\chi$ under the same assumptions. This bound only depends on the probability, the number of samples, and constants of the problem. These propositions will be of use in the proof of Theorem 34.

Let $\widetilde{N} > 0$, we write $C := C_{\widetilde{N}}$ (as in Definition 19) for simplicity.

Proposition 16. *Let f be a positive function satisfying Assumption 2. Consider $\widetilde{N}$ points $\chi = \{(X_1, f(X_1)), \dots, (X_{\widetilde{N}}, f(X_{\widetilde{N}}))\}$.*
If $\hat{f}_\chi$ is the Approximate Nearest Neighbor Estimate of f, as defined in Definition 19, then

$$\forall x \in [0,1]^d,\ |\hat{f}_\chi(x) - f(x)| \leq \hat{r}_\chi,$$

where $\hat{r}_\chi$ is defined in Equation (4.3) (in Definition 19).

Proof of Proposition 16. We have that $\forall x \in [0,1]^d$,

$$\|x - X_{i(C(x))}\|_\infty \leq \|x - C(x)\|_\infty + \|C(x) - X_{i(C(x))}\|_\infty,$$

where the set $\mathcal{C}_{\widetilde{N}}$ and the function i are defined in Definition 19.
Now, $\|x - C(x)\|_\infty \leq \frac{1}{2\widetilde{N}^{\frac{1}{d}}}$ and $\|C(x) - X_{i(C(x))}\|_\infty \leq \max_{u \in \mathcal{C}_{\widetilde{N}}} \|u - X_{i(u)}\|_\infty, \forall x \in [0,1]^d$ and where $\mathcal{C}_{\widetilde{N}}$ is defined in Definition 19.
Thus $\forall x \in [0,1]^d$,

$$\|x - X_{i(C(x))}\|_\infty \leq \frac{1}{2\widetilde{N}^{\frac{1}{d}}} + \max_{u \in \mathcal{C}_{\widetilde{N}}} \|u - X_{i(u)}\|_\infty, \tag{4.4}$$

and from Assumption 2

$$\forall x \in [0,1]^d,\ |\hat{f}_\chi(x) - f(x)| \leq \hat{r}_\chi.$$

Proposition 17. *Consider the same notations and assumptions as in Proposition 16. Let g be a density on $[0,1]^d$ such that:*

$$\exists\, 1 \geq c > 0 \text{ such that } \forall x \in [0,1]^d,\ c < g(x),$$

and assume that the points X_i in χ are sampled in an i.i.d. fashion according to g. Defining $\delta_0 = \frac{1}{\widetilde{N}} \exp(-\widetilde{N})$, it holds for any $\delta > \delta_0$, that with probability larger than $1 - \delta$:

$$\hat{r}_\chi \leq 2^s r_{\widetilde{N},\delta,c}.$$

where we write $r_{\widetilde{N},\delta,c} = H\left(\frac{\log(\widetilde{N}/\delta)}{c\widetilde{N}}\right)^{\frac{s}{d}}$.

Proof of Proposition 17. Let ϵ be a positive number smaller than 1 such that ϵ^{-d} is an integer. We split $[0,1]^d$ in $\frac{1}{\epsilon^d}$ hypercubes of side-length ϵ and of centers in $\mathcal{C}_{\epsilon^{-d}}$. Let I be one of these hypercubes, we have $\mathbb{P}(X_1 \dots X_{\widetilde{N}} \notin I) \leq (1 - c\epsilon^d)^{\widetilde{N}} \leq \exp(-c\epsilon^d \widetilde{N})$.

So with probability larger than $1 - \exp(-c\epsilon^d \widetilde{N})$, at least one point has been drawn in I.
Thus $\forall x \in [0,1]^d$, with probability larger than $1 - \exp(-c\epsilon^d \widetilde{N})$, it holds:

$$\|x - X_{i(x)}\|_\infty \leq \epsilon,$$

where $i(x) = \arg\min_{i \in \{1,\dots,N\}}(\|x - X_i\|_\infty)$.
Thus $\forall x \in [0,1]^d$, with probability larger than $1 - \delta'$,

$$\|x - X_{i(x)}\|_\infty \leq \left(\frac{\log(1/\delta')}{c\widetilde{N}}\right)^{\frac{1}{d}},$$

where $\delta' = \exp(-c\epsilon^d \widetilde{N})$ (observe $\delta' > \exp(-\widetilde{N})$).
Thus with probability larger than $1 - \frac{1}{\epsilon^d}\delta'$, it holds

$$\forall x \in [0,1]^d,\ \|x - X_{i(x)}\|_\infty \leq \left(\frac{\log(1/\delta')}{c\widetilde{N}}\right)^{\frac{1}{d}}.$$

With probability larger than $1 - c\widetilde{N}\delta' > 1 - \frac{c\widetilde{N}}{\log(1/\delta')}\delta'$, it holds

$$\forall x \in [0,1]^d,\ \|x - X_{i(x)}\|_\infty \leq \left(\frac{\log(1/\delta')}{c\widetilde{N}}\right)^{\frac{1}{d}}.$$

Hence, by letting $\delta = (c\widetilde{N})\delta'$, with probability larger than $1 - \delta$,

$$\begin{aligned}\forall x \in [0,1]^d,\ \|x - X_{i(x)}\|_\infty &\leq \left(\frac{\log(c\widetilde{N}/\delta)}{c\widetilde{N}}\right)^{\frac{1}{d}} \\ &\leq \left(\frac{\log(\widetilde{N}/\delta)}{c\widetilde{N}}\right)^{\frac{1}{d}}.\end{aligned}$$

Thus $\forall \delta > c\widetilde{N}\exp(-\widetilde{N})$, with probability larger than $1 - \delta$,

$$\forall x \in [0,1]^d,\ \|x - X_{i(x)}\|_\infty \leq \left(\frac{\log(\widetilde{N}/\delta)}{c\widetilde{N}}\right)^{\frac{1}{d}},$$

and in particular, with probability larger than $1 - \delta$,

$$\max_{u \in \mathcal{C}_{\tilde{N}}} \|u - X_{i(u)}\|_\infty \leq \left(\frac{\log(\widetilde{N}/\delta)}{c\widetilde{N}}\right)^{\frac{1}{d}}.$$

Furthermore we have since $|\chi_{\tilde{N}}| = \tilde{N}$

$$\frac{1}{2\widetilde{N}^{\frac{1}{d}}} \leq \max_{x \in [0,1]^d} \|x - X_i(x)\|_\infty.$$

So we also have (since $c \leq 1$ and $\log(1/\delta) \geq 1$)

$$\frac{1}{2\widetilde{N}^{\frac{1}{d}}} \leq \left(\frac{\log(\widetilde{N}/\delta)}{c\widetilde{N}}\right)^{\frac{1}{d}}.$$

Finally, from Equation (4.4), with probability larger than $1-\delta$, $\forall x \in [0,1]^d$,

$$\begin{aligned}\|x - X_{i(C(x))}\|_\infty &\leq \frac{1}{2\widetilde{N}^{\frac{1}{d}}} + \max_{u \in \mathcal{C}_{\tilde{N}}} \|u - X_{i(u)}\|_\infty \\ &\leq 2\left(\frac{\log(\widetilde{N}/\delta)}{c\widetilde{N}}\right)^{\frac{1}{d}},\end{aligned}$$

and with probability larger than $1-\delta$,

$$\begin{aligned}\hat{r}_\chi &= H\left(\frac{1}{2\widetilde{N}^{\frac{1}{d}}} + \max_{u \in \mathcal{C}_{\tilde{N}}} \|u - X_{i(u)}\|_\infty\right)^s \\ &\leq H\left(2\left(\frac{\log(\widetilde{N}/\delta)}{c\widetilde{N}}\right)^{\frac{1}{d}}\right)^s = 2^s r_{\widetilde{N},\delta,c}.\end{aligned}$$

4.8.2 Proof of Theorem 34

In this subsection, we prove Theorem 34 by first proving a high probability bound on $n - \hat{n}$. We prove this high probability bound thanks to Proposition 18, and Lemma 31. Proposition 18 claims that the algorithm provides independent samples drawn according to f/I_f, under Assumption 2. The proof of Proposition 18 uses Lemma 30 which states that $\forall x \in [0,1]^d$, $f(x) \leq M_k g_k(x)$, under the relevant assumptions. We define two events: one on which every proposal envelope until time $k+1$ is bounded from below by $\frac{6}{10}c_f$: $\mathcal{W}_{k+1}$, and the other one on which every confidence radius $\hat{r}_{\cup_{i\leq k}\chi_k}$ until time k is upper bounded by a quantity $r_{\widetilde{N},\delta,6c_f/10}$: $\mathcal{A}_{k,\delta}$, where δ is a confidence term (designed to be used in the high probability bound on $n - \hat{n}$). Lemma 31 states that the probability of the event $\mathcal{W}_{k+1}$ conditional to the event $\mathcal{A}_{k,\delta}$ is equal to 1, and that that the probability of the event $\mathcal{A}_{k,\delta}$ is larger than $1 - 2k\widetilde{\delta}$ when $\delta = N/nK$. The proof of Theorem 34 uses the fact that the number of rejected samples at step k on $\mathcal{A}_{k,\delta}$ is a sum of Bernoulli variables of parameter smaller than a known quantity that depends on $\widehat{\delta}$, and by applying the Bernstein inequality on this sum. The proof is then concluded by summing on k.

In this subsection, we write

$$\hat{f}_k = \hat{f}_{\cup_{i\leq k}\chi_i}, \quad \text{and} \quad \hat{r}_k = \hat{r}_{\cup_{i\leq k}\chi_i},$$

to ensure the simplicity of notations. We also write

$$r_{\widetilde{N},\delta} := r_{\widetilde{N},\delta,6c_f/10} = H\left(\frac{10\log(\widetilde{N}/\delta)}{6c_f\widetilde{N}}\right)^{\frac{s}{d}}.$$

Let us also define the events

$$\begin{cases} \mathcal{W}_k = \{\forall j \leq k,\ \forall x \in [0,1]^d,\ g_j(x) > \frac{6}{10} c_f\}, \\ \mathcal{A}_{k,\delta} = \{\forall j \leq k,\ \hat{r}_j \leq 2^s r_{N_j,\delta}\}. \end{cases}$$

Proposition 18. *If Assumption 2 holds, the algorithm provides independent samples drawn according to the density f/I_f.*

Lemma 30. *Consider any $k \leq K$. Under the assumptions made in Proposition 18,*

$$\forall x \in [0,1]^d,\ f(x) \leq M_k g_k(x).$$

Proof of Lemma 30. g_1 is the uniform density and M_1 is taken as an upper bound on f. So $\forall x \in [0,1]^d$:

$$M_1 g_1(x) \geq f(x).$$

Let $k \in \{2, \dots, K\}$. From Proposition 16:

$$\forall x \in [0,1]^d,\ |\hat{f}_{k-1}(x) - f(x)| \leq \hat{r}_{k-1}.$$

Thus, $\forall x \in [0,1]^d$:

$$g_k(x) = \frac{\hat{f}_{k-1}(x) + \hat{r}_{k-1}}{I_{\hat{f}_{k-1}} + \hat{r}_{k-1}} \geq \frac{f(x)}{I_{\hat{f}_{k-1}} + \hat{r}_{k-1}} \geq \frac{f(x)}{M_k}.$$

Hence

$$\forall x,\ M_k g_k(x) \geq f(x).$$

Proof of Proposition 18. We have that $\forall j \leq k,\ \forall x \in [0,1]^d,\ f(x) \leq M_k g_k(x)$. Theorem 33 proves that the algorithm provides independent samples drawn according to the density f/I_f.

Lemma 31. *Let $\tilde{\delta} = N/(nK)$. If Assumption 2 and 3 hold for n, then*

$$\begin{cases} \mathbb{P}(\mathcal{W}_1) = 1, \quad \mathbb{P}(\mathcal{W}_{k+1} | \mathcal{A}_{k,\tilde{\delta}}) = 1, \\ \mathbb{P}(\mathcal{A}_{k,\tilde{\delta}}) \geq 1 - 2k\tilde{\delta}. \end{cases}$$

Proof of Lemma 31. Since $g_1(x) = 1$, $s \leq 1, c_f \leq 1$, the event $\mathcal{W}_1 = \{\forall x \in [0,1]^d,\ g_1(x) > \frac{6}{10} c_f\}$ has probability 1. Also by Proposition 16 and Proposition 17, the event $\mathcal{A}_{1,\tilde{\delta}}$ has probability larger than $1 - \tilde{\delta}$.
Consider now that the event $\mathcal{A}_{k,\tilde{\delta}}$ holds for a given $k \leq K$. Then by Proposition 16 and Proposition 17, it holds that for all $j \leq k$ and for all $x \in [0,1]^d$

$$|\hat{f}_j(x) - f(x)| \leq 2^s r_{N_j,\tilde{\delta}}.$$

This implies that

$$
\begin{aligned}
g_{j+1}(x) &\geq \frac{f(x)}{M_{j+1}} \geq \frac{f(x)}{I_f + 2^{s+1} r_{N_j,\tilde{\delta}}} \\
&\geq \frac{f(x)/I_f}{1 + 2^{s+1} r_{N_j,\tilde{\delta}}/I_f} \geq \frac{f(x)}{I_f}\left(1 - 2^{s+1}\frac{r_{N_j,\tilde{\delta}}}{I_f}\right) \\
&\geq c_f\left(1 - 2^{s+1}\frac{r_{N_j,\tilde{\delta}}}{I_f}\right).
\end{aligned}
$$

Hence,

$$
g_{j+1}(x) \geq c_f\left(1 - \frac{2^{s+1} r_{N,\tilde{\delta}}}{c_f}\right) \geq \frac{6}{10} c_f,
$$

where we have used $r_{N_j,\tilde{\delta}} \leq r_{N,\tilde{\delta}} \leq c_f/10$ (see Assumption 3). So $\mathbb{P}(\mathcal{W}_{k+1}|\mathcal{A}_{k,\tilde{\delta}}) = 1$ and we have proved the first part of the lemma.

Moreover, conditional to $\mathcal{A}_{k,\tilde{\delta}}$ we have that $g_{k+1}(x) \geq \frac{6}{10} c_f$. Then we apply Proposition 16 and Proposition 17. With probability larger than $1 - \tilde{\delta}$ on χ_k only, and conditional to $\mathcal{A}_{k,\tilde{\delta}}$, it holds that for all $x \in [0,1]^d$:

$$
|\hat{f}_{k+1}(x) - f(x)| \leq 2^s r_{N_{k+1},\tilde{\delta}},
$$

where we use that $\hat{r}_{k+1} \leq \hat{r}_{\chi_{k+1}}$. This implies that $\mathbb{P}(\mathcal{A}_{k+1,\tilde{\delta}}|\mathcal{A}_{k,\tilde{\delta}}) \geq 1 - \tilde{\delta}$, and so for any $k \leq K$

$$
\mathbb{P}(\mathcal{A}_{k,\tilde{\delta}}) \geq (1 - \tilde{\delta})^k.
$$

This concludes the proof since $(1 - \tilde{\delta})^k \geq 1 - 2k\tilde{\delta}$ for $\tilde{\delta} \leq 1/(2K)$.

Proof of Theorem 34. Let $\tilde{\delta} = N/(nK)$ and $\delta = K\tilde{\delta}$ and let $\hat{n}_k$ denote the number of accepted samples at round k.
From Lemma 30, we know that $\forall k \leq K$,

$$
\forall x,\ M_k g_k(x) \geq f(x).
$$

Hence, the samples accepted at step $k+1$ are independently sampled according to f/I_f, and $N_{k+1} - \hat{n}_k$, the number of rejected samples, is a sum of Bernoulli variables of parameter $1 - I_f/M_k$.
On $\mathcal{A}_{k,\tilde{\delta}} \cap \mathcal{W}_k$,

$$
\begin{aligned}
\frac{I_f}{M_{k+1}} &\geq \frac{I_f}{I_f + 2^{s+1} r_{N_k,\tilde{\delta}}} \\
&\geq 1 - \frac{2^{s+1} r_{N_k,\tilde{\delta}}}{I_f}.
\end{aligned}
$$

Thus, $1 - \frac{I_f}{M_{k+1}} \leq \frac{2^{s+1} r_{N_k,\widetilde{\delta}}}{I_f}$.
On $\mathcal{A}_{K,\widetilde{\delta}} \cap \mathcal{W}_K$, according to the Bernstein inequality, $\forall k \leq K$ the event

$$\mathcal{V}_k = \left\{ N_{k+1} - \hat{n}_k - \left(1 - \frac{I_f}{M_{k+1}}\right) N_{k+1} \leq \sqrt{2N_{k+1}\left(1 - \frac{I_f}{M_{k+1}}\right) \log\left(\frac{1}{\widetilde{\delta}}\right)} + \log\left(\frac{1}{\widetilde{\delta}}\right) \right\}$$

has probability larger than $1 - \widetilde{\delta}$.
Hence on $\mathcal{A}_{K,\widetilde{\delta}} \cap \mathcal{W}_K$, $\bigcap_{k\in\{1,\dots,K\}} \mathcal{V}_k$ has probability $1 - K\widetilde{\delta}$.
Consequently, since $\mathcal{A}_{K,\widetilde{\delta}} \cap \mathcal{W}_K$ has probability larger than $1 - K\widetilde{\delta}$ according to Lemma 30, $\bigcap_{k\in\{1,\dots,K\}} \mathcal{V}_k \cap \mathcal{A}_{K,\widetilde{\delta}} \cap \mathcal{W}_K$ has probability larger than $1 - 2K\widetilde{\delta}$.
On $\mathcal{V}_k \cap \mathcal{A}_{K,\widetilde{\delta}} \cap \mathcal{W}_K$,

$$N_{k+1} - \hat{n}_k - \frac{2^{s+1} r_{N_k,\widetilde{\delta}}}{I_f} N_{k+1} \leq \sqrt{2N_{k+1} \frac{2^s r_{N_k,\widetilde{\delta}}}{I_f} \log\left(\frac{1}{\widetilde{\delta}}\right)} + \log\left(\frac{1}{\widetilde{\delta}}\right)$$

(and we know from Proposition 18, that on $\mathcal{V}_k \cap \mathcal{A}_{K,\widetilde{\delta}} \cap \mathcal{W}_K$, we also have that the drawn samples are independently drawn according to f/I_f).
Hence on $\bigcap_k \mathcal{V}_k \cap \mathcal{A}_{K,\widetilde{\delta}} \cap \mathcal{W}_K$, which has probability larger than $1 - 2K\widetilde{\delta} := 1 - 2\delta$:

$$\begin{aligned}
&\sum_1^{K-1} \left(N_{k+1} - \hat{n}_k - \frac{2^{s+1} r_{N_k,\widetilde{\delta}}}{I_f} N_{k+1} \right) \\
&\leq \sum_1^{K-1} \left(\sqrt{2N_{k+1} \frac{2^{s+1} r_{N_k,\widetilde{\delta}}}{I_f} \log(\frac{1}{\widetilde{\delta}})} \right) + K \log\left(\frac{1}{\widetilde{\delta}}\right),
\end{aligned}$$

i.e:

$$\begin{aligned}
&n - \hat{n} \\
&\leq \underbrace{\frac{2^{s+1}}{I_f} \sum_1^{K-1} \left(r_{N_k,\widetilde{\delta}} N_{k+1} \right) + 4\sqrt{\frac{\log(\frac{1}{\widetilde{\delta}})}{I_f}} \sum_1^{K-1} \left(\sqrt{N_{k+1} r_{N_k,\widetilde{\delta}}} \right)}_{(1)} + \underbrace{K \log\left(\frac{1}{\widetilde{\delta}}\right)}_{(2)} + N.
\end{aligned}$$

Hence,

$$\begin{aligned}
(2) &= K \log\left(\frac{K}{\delta}\right) \\
&= \log_2\left(\frac{n}{N}\right) \log\left(\frac{\log_2(n/N)}{\delta}\right),
\end{aligned}$$

and if $\beta = \frac{2^{s+1}}{I_f} + 4\sqrt{\frac{\log(\frac{1}{\widetilde{\delta}})}{I_f}}$, and $\tilde{C} = H(10/(6c_f))^{s/d}$,

$$\begin{aligned}
(1) &\leq \beta \Big[\sum_1^{K-1} \left(r_{N_k,\widetilde{\delta}} N_{k+1} \right) + K \Big] \\
&\leq \beta \sum_1^{K-1} \left(\widetilde{C} \log\left(\frac{2^k N}{\widetilde{\delta}}\right)^{s/d} (2^k N)^{1-s/d} \right) + K\beta.
\end{aligned}$$

Now, assume $s/d < 1$, then

$$(1) \leq \beta\widetilde{C}\left(\log\left(\frac{n}{\tilde{\tilde{\delta}}}\right)\right)^{s/d} N^{1-s/d}\left(\frac{2^{(1-s/d)K}-1}{2^{(1-s/d)}-1}\right) + K\beta$$
$$\leq \frac{\beta\widetilde{C}}{2^{(1-s/d)}-1}\left(\log\left(\frac{n}{\tilde{\tilde{\delta}}}\right)\right)^{s/d} n^{1-s/d} + K\beta.$$

And if $s = d = 1$, then

$$(1) \leq K\beta\widetilde{C}\left(\log\left(\frac{n}{\tilde{\tilde{\delta}}}\right)\right) + K\beta.$$

We have proved that if the assumptions of Theorem 34 are satisfied, with probability $1 - 2\delta$,

$$n - \hat{n} \leq \left(\frac{2^{s+1}}{I_f} + 4\sqrt{\frac{\log(\frac{1}{\tilde{\tilde{\delta}}})}{I_f}}\right) C_{s,d}\left(\log\left(\frac{n}{\tilde{\tilde{\delta}}}\right)\right)^{s/d} n^{1-s/d}$$
$$+ \quad \log_2\left(\frac{n}{N}\right)\log\left(\frac{\log_2(n/N)}{\delta}\right) + N + \left(\frac{2^{s+1}}{I_f} + 4\sqrt{\frac{\log(\frac{1}{\tilde{\tilde{\delta}}})}{I_f}}\right)\log_2\left(\frac{n}{N}\right),$$

where $C_{s,d}$ is a constant dependent on s and d. Finally, the proof is finished following a few strings of inequalities and taking the expected value. The following reminders may help,
$d \geq 1$; $s \leq 1$; $c_f \leq 1$, $\tilde{C} = H(10/(6c_f))^{s/d}$; $\delta = N/n = K\tilde{\delta}$; $K = \lceil \log_2(n/N) \rceil$ and $N = \lceil 2(10H)^{d/s}\log(n)c_f^{-1-d/s} \rceil$. In particular, we have $I_f \geq c_f$ and $1/\tilde{\delta} \leq 5n^2$.

4.9 Proof of Theorem 35.

4.9.1 Setting

Let us introduce two different settings:

Setting 1. (Class of Rejection Samplers with Access to Multiple Evaluations of the density (RSAME))
A sampler belongs to the RSAME class if it follows the following steps:

- *For each step $t \in \{1 \dots n\}$:*
  Choose a distribution $\mathcal{D}_t$ on $\mathbb{R}$, depending on $\big((Y_1, f(Y_1)) \dots (Y_{t-1}, f(Y_{t-1}))\big)$.
  Draw Y_t according to $\mathcal{D}_t$.

- *Choose a density g and a positive constant M depending on $\big((Y_1, f(Y_1)) \dots (Y_n, f(Y_n))\big)$, and sample Z by performing one Rejection Sampling Step(f, M, g).*

Objective : *The objective of a RSAME sampler is to sample one point according to a normalized version of f.*
Loss : *The loss of a RSAME sampler is defined as follows :*

$$L'_n = n(1 - \mathbf{1}\{Z \text{ is accepted }\}\mathbf{1}\{f \leq Mg\}).$$

Strategy : *A strategy* $\mathfrak{s}'$ *consists of the choice of* $\mathcal{D}_t$ *depending on* $\big((Y_1, f(Y_1)) \dots (Y_{t-1}, f(Y_{t-1}))\big)$*, and of the choice of* M, g *depending on* $\big((Y_1, f(Y_1)) \dots (Y_n, f(Y_n))\big)$*. Denote* $\mathfrak{S}'$ *the set of strategies for this setting.*

Setting 2. (Class of Adaptive Rejection Samplers (ARS))
A sampler belongs to the ARS class if, at each step $t \in \{1 \dots n\}$*: it*

- *Chooses a density* g_t*, and a positive constant* M_t*, depending only on* $\Big\{(X_1, f(X_1)), \dots, (X_{t-1}, f(X_{t-1}))\Big\}$.
- *Samples* X_t *by performing rejection sampling on the target function* f *using* M_t *and* g_t *as the rejection constant and the proposal. Store* X_t *in* $\mathcal{S}$ *if it is accepted.*

Objective : *The objective of an ARS sampler is to sample i.i.d. points according to a normalized version of* f
Loss : *The loss of an ARS sampler is defined as follows :* $L_n = n - \#\mathcal{S}\mathbf{1}\{\forall t \leq n, f \leq M_t g_t\}$.
Strategy : *A strategy* $\mathfrak{s}$ *consists of the choice of* M_t, g_t *depending on* $\big((X_1, f(X_1)) \dots (X_{t-1}, f(X_{t-1}))\big)$*. Denote* $\mathfrak{S}$ *the set of strategies for this setting.*

For the class of samplers defined in Setting 2 (and similarly for Setting 1) we call value of the class the quantity $\inf_{\mathfrak{s}\in\mathfrak{S}} \sup_{f\in\mathcal{F}_0} \mathbb{E}^{(\mathfrak{s},f)}(L_n)$, where the symbol $\mathbb{E}^{(\mathfrak{s},f)}$ denotes the expectation with respect to all relevant random variables, when those are generated by a sampler of the relevant class, using function f and strategy $\mathfrak{s}$; and $\mathcal{F}_0$ denotes the set of functions satisfying Assumption 2.

4.9.2 Setting Comparison

Lemma 32. *The value of the class defined in Setting 1 is smaller than the value of the class defined in Setting 2:*

$$\inf_{\mathfrak{s}'\in\mathfrak{S}'} \sup_{f\in\mathcal{F}_0} \mathbb{E}^{(\mathfrak{s}',f)}(L'_n) \leq \inf_{\mathfrak{s}\in\mathfrak{S}} \sup_{f\in\mathcal{F}_0} \mathbb{E}^{(\mathfrak{s},f)}(L_n).$$

In other terms, Setting 1 is easier than Setting 2

Proof of Lemma 32. For any given strategy $\mathfrak{s}$ designed for Setting 2 that chooses (g_i, M_i) to generate X_i, consider the associated strategies $\mathfrak{s}'_1, \dots, \mathfrak{s}'_n$ for Setting 1 consisting of:

1. Generating $Y_1, \dots, Y_{n-1}$ from the same probability distributions as $X_1, \dots, X_{n-1}$ generated for Setting 2 using strategy $\mathfrak{s}$; this is a valid choice since the distribution $\mathcal{D}_t$ of X_t only depends on $\big((X_1, f(X_1), \dots, (X_{t-1}, f(X_{t-1}))\big)$.
2. Using (g_i, M_i), where it is obtained at step i by application of strategy $\mathfrak{s}$ to the known values of $\big((Y_1, f(Y_1), \dots, (Y_{i-1}, f(Y_{i-1}))\big)$, in order to sample Z by rejection sampling. It is still a valid choice, which actually discards the information of $\big((Y_i, f(Y_i), \dots, (Y_{n-1}, f(Y_{n-1}))\big)$.

Then, we have for any $f \in \mathcal{F}_0$:

$$\begin{aligned}
\mathbb{E}^{(\mathfrak{s},f)}(L_n) &= n - \mathbb{E}^{(\mathfrak{s},f)}(\#\mathcal{S}\mathbf{1}\{\forall t \leq n, f \leq M_t g_t\}) \\
&= n - \mathbb{E}^{(\mathfrak{s},f)}\Big(\sum_{i=1}^{n} \mathbf{1}\{X_i \text{ is accepted }\}\mathbf{1}\{\forall t \leq n, f \leq M_t g_t\}\Big) \\
&\geq n - \mathbb{E}^{(\mathfrak{s},f)}\Big(\sum_{i=1}^{n} \mathbf{1}\{X_i \text{ is accepted }\}\mathbf{1}\{f \leq M_i g_i\}\Big) \\
&= \sum_{i=1}^{n} \mathbb{E}^{(\mathfrak{s}_i,f)}\Big(1 - \mathbf{1}\{X_i \text{ is accepted }\}\mathbf{1}\{f \leq M_i g_i\}\Big) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}^{(\mathfrak{s}_i,f)}(L'_n).
\end{aligned}$$

Hence, there exists at least one strategy amongst $\mathfrak{s}'_1, \ldots, \mathfrak{s}'_n$ that reaches an expected loss in Setting 1 lower than that of strategy $\mathfrak{s}$ in Setting 2.

4.9.3 Lower Bound for Setting 1

Lemma 33.

$$\inf_{\mathfrak{s}' \in \mathfrak{S}'} \sup_{f \in \mathcal{F}_0} \mathbb{E}_f(L'_n(\mathfrak{s}')) \geq 3^{-1}2^{-1-3s-2d}5^{-s/d}n^{1-s/d},$$

for n large enough.

The Theorem is a direct consequence of Lemmas 32 and 33. We use Lemma 34 to prove Lemma 33.

Lemma 34. *Let*

$$\begin{aligned}
\mathcal{F}'_1 = \Bigg\{ & f \text{ s.t. } \forall u = (k_1, \ldots k_d) \in \{0, 1, \ldots, a_{n,d} - 1\}^d, \\
& \text{either } \forall x \in H_u := \Big[\frac{k_1}{a_{n,d}}, \frac{k_1+1}{a_{n,d}}\Big] \ldots \Big[\frac{k_d}{a_{n,d}}, \frac{k_d+1}{a_{n,d}}\Big], \\
& f(x) = \phi^+\Big(x - \frac{u}{a_{n,d}}\Big), \\
& \text{or } \forall x \in H_u, f(x) = \phi^-\Big(x - \frac{u}{a_{n,d}}\Big)\Bigg\},
\end{aligned} \tag{4.5}$$

where:

$$\begin{aligned}
& a_{n,d} = \min\{2p \in \mathbb{N}; 2p \geq (4n)^{\frac{1}{d}}\}, \\
& \phi^+(x) = 1 + (2a_{n,d})^{-s} - \Big\|x - \frac{1}{2a_{n,d}}\mathbf{I}\Big\|_\infty^s, \\
& (\textit{with } \mathbf{I} \textit{ denoting the unit vector}), \\
& \phi^-(x) = 2 - \phi^+(x).
\end{aligned}$$

Then any function in $\mathcal{F}'_1$ is s-Hölder-smooth.

Remark 38. *If* $d = 1$,

$$\mathcal{F}_1' = \Bigg\{ f \text{ s.t. } \forall i \in \{0, 1, \dots, 4n-1\}, \\ \text{either } \forall x \in H_i := \Big[\frac{i}{4n}, \frac{i+1}{4n}\Big],\ f(x) = \phi^+\Big(x - \frac{i}{4n}\Big), \\ \text{or } \forall x \in H_i, f(x) = \phi^-\Big(x - \frac{i}{4n}\Big) \Bigg\}, \tag{4.6}$$

with $\forall x \in [0, 1/(4n)]$

$$\phi^+(x) = 1 + (8n)^{-s} - \left| x - \frac{1}{8n} \right|^s,$$

and

$$\phi^-(x) = 1 - (8n)^{-s} + \left| x - \frac{1}{8n} \right|^s.$$

Proof of Lemma 34. Let us first prove that ϕ^+ is s-smooth. $|\phi^+(x) - \phi^+(y)| = |\|y - \frac{1}{2a_{n,d}}\mathbf{I}\|_\infty^s - \|x - \frac{1}{2a_{n,d}}\mathbf{I}\|_\infty^s| \leq \|x - y\|_\infty^s$.
It is straightforward to see that ϕ^- is also s-smooth and that all $f \in \mathcal{F}_1'$ are also s-smooth.

Proof of Lemma 33. Let us consider Setting 1 on a subset of functions of $\mathcal{F}_0$. Let

$$\mathcal{F}_1 = \mathcal{F}_{int} \cap \mathcal{F}_1',$$

where $\mathcal{F}_1'$ is defined in Equation (4.5) and

$$\mathcal{F}_{int} = \left\{ f, \int_0^1 f = 1 \right\}.$$

And $\mathcal{F}_1$ is not empty since $a_{n,d}$ defined in equation (4.5) is even. Since $\mathcal{F}_1 \subset \mathcal{F}_0$ by application of Lemma 34,

$$\inf_{\mathfrak{s}' \in \mathfrak{S}'} \sup_{f \in \mathcal{F}_0} \mathbb{E}_f(L_n'(\mathfrak{s}')) \geq \inf_{\mathfrak{s}' \in \mathfrak{S}'} \sup_{f \in \mathcal{F}_1} \mathbb{E}_f(L_n'(\mathfrak{s}')).$$

We first note that:

$$\inf_{\mathfrak{s}' \in \mathfrak{S}'} \sup_{f \in \mathcal{F}_0} \mathbb{E}(L_n'(\mathfrak{s}')) \geq \inf_{\mathfrak{s}' \in \mathfrak{S}'} \mathbb{E}_{f \sim \mathcal{D}_{\mathcal{F}_1}}(L_n'(\mathfrak{s}')),$$

where $\mathcal{D}_{\mathcal{F}_1}$ is the distribution such that for any F, $\mathbb{P}_{f \sim \mathcal{D}_{\mathcal{F}_1}}(f = F) = \frac{\mathbf{1}\{F \in \mathcal{F}_1\}}{\#\mathcal{F}_1}$. A hypercube will refer to a H_u as defined in Equation (4.5).

We also note that the choice of M, g where M is a multiplicative constant and g is a density is equivalent to the choice of a positive function G, where $G = Mg$, or $M = I_G$ and $g = \frac{G}{I_G}$.
Furthermore a strategy $\mathfrak{s}'$ for this setting is the combination of three strategies:

1. $\mathfrak{s}_1'$: The strategy to choose $Y_1 \dots Y_n$,
2. $\mathfrak{s}_2'$: The strategy to choose G.

For the first step, let us fix a strategy $\mathfrak{s}_1'$. Let f_1 be a realization of $\mathcal{D}_{\mathcal{F}_1}$. Then by application of strategy $\mathfrak{s}_1'$, $Y_1, \dots, Y_n$ are drawn. Then the evaluations $f_1(Y_1), \dots, f_1(Y_n)$ are

obtained. Now let $u_1, \dots u_n$ be the indices such that $H_{u_1} \dots H_{u_n}$ are the hypercubes where $Y_1 \in H_{u_1}, \dots, Y_n \in H_{u_n}$.

Let us define the restricted set $\mathcal{F}_{1|f_1} = \{f \in \mathcal{F}_1 \text{ and } f(Y_1) = f_1(Y_1), \dots, f(Y_n) = f_1(Y_n)\}$. And we consider the distribution $\mathcal{D}_{\mathcal{F}_{1|f_1}}$ such that $\mathbb{P}_{f\sim\mathcal{D}_{\mathcal{F}_{1|f_1}}}(f = F) = \frac{\mathbf{1}\{F\in\mathcal{F}_1\}}{\#\mathcal{F}_1}$.
In a second step, let us fix a strategy $\mathfrak{s}'_2$. This defines a distribution $\mathcal{D}_G$ corresponding to the choice of G. By the law of total expectation, we have:

$$\begin{aligned}\mathbb{E}_{f\sim\mathcal{D}_{\mathcal{F}_1}}(L'_n(\mathfrak{s}')) &= \mathbb{E}_{f_1\sim\mathcal{D}_{\mathcal{F}_1}}\mathbb{E}_{G\sim\mathcal{D}_G}\Big[\mathbb{E}_{f\sim\mathcal{D}_{\mathcal{F}_{1|f_1}}}\Big(L'_n(\mathfrak{s}')|(Y_1, f_1(Y_1)), \dots (Y_n, f_1(Y_n)), G\Big)\\ &\qquad\qquad |(Y_1, f_1(Y_1)), \dots (Y_n, f_1(Y_n))\Big]\\ &= \mathbb{E}_{f_1\sim\mathcal{D}_{\mathcal{F}_1}}\mathbb{E}_{G\sim\mathcal{D}_G}\Big[\mathbb{E}_{f\sim\mathcal{D}_{\mathcal{F}_{1|f_1}}}\Big(L'_n(\mathfrak{s}')|(Y_1, f(Y_1)), \dots (Y_n, f(Y_n)), G\Big)\\ &\qquad\qquad |(Y_1, f_1(Y_1)), \dots (Y_n, f_1(Y_n))\Big].\end{aligned}$$

We can write:

$$\begin{aligned}&\mathbb{E}_{f\sim\mathcal{D}_{\mathcal{F}_{1|f_1}}}\left(L'_n\left(\mathfrak{s}'\right)|(Y_1, f(Y_1)), \dots (Y_n, f(Y_n)), G\right)\\ &= \mathbb{E}_{f\sim\mathcal{D}_{\mathcal{F}_{1|f_1}}}\Bigg[\mathbf{1}\left\{\exists u \notin \{u_1 \dots u_n\},\ \exists x \in H_u : G(x) < f(x)\right\} n\\ &\quad + \mathbf{1}\left\{\forall u \notin \{u_1 \dots u_n\},\ \forall x \in H_u : G(x) \geq f(x)\right\} n \left(1 - \frac{1}{1 + \|f - G\|_1}\right)\\ &\qquad \Big| (Y_1, f(Y_1)) \dots (Y_n, f(Y_n)), G\Bigg]\\ &\geq \mathbb{E}_{f\sim\mathcal{D}_{\mathcal{F}_{1|f_1}}}\left(1 - \frac{1}{1 + \|f - G\|_1}\Big|G\right) n.\end{aligned}$$

Now, since for any $x \geq 0$,

$$1 - \frac{1}{1+x} \geq \frac{1}{2}(1 \wedge x),$$

we have

$$\begin{aligned}&\mathbb{E}_{f\sim\mathcal{D}_{\mathcal{F}_{1|f_1}}}\left(L'_n\left(\mathfrak{s}'\right)|(Y_1, f(Y_1)), \dots (Y_n, f(Y_n)), G\right)\\ &\qquad\geq \frac{1}{2}\mathbb{E}_{f\sim\mathcal{D}_{\mathcal{F}_{1|f_1}}}\left(\|f - G\|_1 \wedge 1\Big|G\right) n\\ &\qquad\geq \frac{1}{2}\mathbb{E}_{f\sim\mathcal{D}_{\mathcal{F}_{1|f_1}}}\left(\big\||f - G| \wedge 1\big\|_1\Big|G\right) n\\ &\qquad\geq \frac{1}{2}\left\|\mathbb{E}_{f\sim\mathcal{D}_{\mathcal{F}_{1|f_1}}}\left(|f - G| \wedge 1\Big|G\right)\right\|_1 n\\ &\qquad\geq \frac{1}{2}\left\|\mathbb{E}_{f\sim\mathcal{D}_{\mathcal{F}_{1|f_1}}}\left([|f - G| \wedge 1](1 - \mathbf{1}\{\cup_{i=1}^n H_{u_i}\})\Big|G\right)\right\|_1 n.\end{aligned}$$

For any $u \neq u_1 \dots u_n$, $\forall x \in H_u$, since any realization from $\mathcal{D}_{\mathcal{F}_{1|f_1}}$ is in $\mathcal{F}_{int}$ almost surely,

$$\mathbb{E}_{f\sim\mathcal{D}_{\mathcal{F}_{1|f_1}}}\left(|f(x)-G(x)|\wedge 1\right) \geq \frac{1}{3}\Bigg[\left(\left|\phi^+\left(x-\frac{u}{a_{n,d}}\right)-G(x)\right|\wedge 1\right) + \left(\left|\phi^-\left(x-\frac{u}{a_{n,d}}\right)-G(x)\right|\wedge 1\right)\Bigg]$$

And, for any $u \notin \{u_1,\dots,u_n\}$, and for any $x \in H_u$:

$$\begin{aligned}
&\left(\left|\phi^+\left(x-\frac{u}{a_{n,d}}\right)-G(x)\right|\wedge 1\right)+\left(\left|\phi^-\left(x-\frac{u}{a_{n,d}}\right)-G(x)\right|\wedge 1\right)\\
&\quad\geq \left(\left|\phi^+\left(x-\frac{u}{a_{n,d}}\right)-G(x)\right|\wedge 1\right)\vee\left(\left|\phi^-\left(x-\frac{u}{a_{n,d}}\right)-G(x)\right|\wedge 1\right)\\
&\quad\geq \min_{\theta}\left[\left(\left|\phi^+\left(x-\frac{u}{a_{n,d}}\right)-\theta\right|\wedge 1\right)\vee\left(\left|\phi^-\left(x-\frac{u}{a_{n,d}}\right)-\theta\right|\wedge 1\right)\right]\\
&\quad= \left(\left|\phi^+\left(x-\frac{u}{a_{n,d}}\right)-1\right|\wedge 1\right)\vee\left(\left|\phi^-\left(x-\frac{u}{a_{n,d}}\right)-1\right|\wedge 1\right).
\end{aligned}$$

And since $|\phi^+-1|=|\phi^--1|$, we end up with:

$$\begin{aligned}
&\left(\left|\phi^+\left(x-\frac{u}{a_{n,d}}\right)-G(x)\right|\wedge 1\right)+\left(\left|\phi^-\left(x-\frac{u}{a_{n,d}}\right)-G(x)\right|\wedge 1\right)\\
&\geq \left|\phi^+\left(x-\frac{u}{a_{n,d}}\right)-1\right|\wedge 1=\phi^+\left(x-\frac{u}{a_{n,d}}\right)-1.
\end{aligned}$$

So

$$\begin{aligned}
&\mathbb{E}_{f\sim\mathcal{D}_{\mathcal{F}_{1|f_1}}}\left(L_n'\left(\mathfrak{s}'\right)|(Y_1,f(Y_1)),\dots(Y_n,f(Y_n)),G\right)\\
&\qquad\geq \frac{1}{6}\sum_{u\neq u_1,\dots u_n}\int_{H_u}\left(\phi^+\left(x-\frac{u}{a_{n,d}}\right)-1\right)dx.
\end{aligned}$$

And

$$\begin{aligned}
\sum_{u\neq u_1,\dots u_n}\int_{H_u}\left(\phi^+\left(x-\frac{u}{a_{n,d}}\right)-1\right)dx &\geq (a_{n,d}^d-n)\int_{[0,1/a_{n,d}]^d}\left(\phi^+(x)-1\right)dx\\
&\geq (a_{n,d}^d-n)\int_{[1/(4a_{n,d}),3/(4a_{n,d})]^d}\left(\phi^+(x)-1\right)dx.
\end{aligned}$$

Now, for any $x\in[1/(4a_{n,d}),3/(4a_{n,d})]^d$, we have $\phi^+(x)-1\geq(4a_{n,d})^{-s}$.
Then

$$\begin{aligned}
\sum_{u\neq u_1,\dots u_n}\int_{H_u}\left(\phi^+\left(x-\frac{u}{a_{n,d}}\right)-1\right)dx &\geq (a_{n,d}^d-n)(4a_{n,d})^{-s}(2a_{n,d})^{-d}\\
&\geq 2^{-3s-2d}5^{-s/d}n^{-s/d},
\end{aligned}$$

where the second inequality used the fact that $a_{n,d}\leq 2(5n)^{1/d}$.
Hence, there exists $N(s,d)$, such that for n larger than $N(s,d)$,

$$\mathbb{E}_{f\sim\mathcal{D}_{\mathcal{F}_1}}(L_n'(\mathfrak{s}'))\geq 3^{-1}2^{-1-3s-2d}5^{-s/d}n^{1-s/d}.$$

Bibliography

Acharya, Jayadev, Clément L Canonne, Cody Freitag, and Himanshu Tyagi (2018). "Test without trust: Optimal locally private distribution testing". In: *arXiv preprint arXiv:1808.02174*.

Acharya, Jayadev, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Shengjun Pan, and Ananda Suresh (2012). "Competitive classification and closeness testing". In: *Conference on Learning Theory*, pp. 22–1.

Acharya, Jayadev, Ziteng Sun, and Huanyu Zhang (2018). "Differentially private testing of identity and closeness of discrete distributions". In: *Advances in Neural Information Processing Systems*, pp. 6878–6891.

Agrawal, Dakshi and Charu C Aggarwal (2001). "On the design and quantification of privacy preserving data mining algorithms". In: *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, pp. 247–255.

Agrawal, Rakesh and Ramakrishnan Srikant (2000). "Privacy-preserving data mining". In: *ACM Sigmod Record*. Vol. 29. 2. ACM, pp. 439–450.

Agrawal, Shipra and Jayant R Haritsa (2005). "A framework for high-accuracy privacy-preserving mining". In: *21st International Conference on Data Engineering (ICDE'05)*. IEEE, pp. 193–204.

Aliakbarpour, Maryam, Ilias Diakonikolas, Daniel Kane, and Ronitt Rubinfeld (2019). "Private testing of distributions via sample permutations". In: *Advances in Neural Information Processing Systems*, pp. 10878–10889.

Aliakbarpour, Maryam, Ilias Diakonikolas, and Ronitt Rubinfeld (2018). "Differentially private identity and equivalence testing of discrete distributions". In: *International Conference on Machine Learning*, pp. 169–178.

Andrieu, Christophe, Nando De Freitas, Arnaud Doucet, and Michael I Jordan (2003). "An introduction to MCMC for machine learning". In: *Machine learning* 50.1-2, pp. 5–43.

Arias-Castro, Ery, Bruno Pelletier, and Venkatesh Saligrama (2018). "Remember the curse of dimensionality: the case of goodness-of-fit testing in arbitrary dimension". In: *Journal of Nonparametric Statistics* 30.2, pp. 448–471.

Assouad, Bin Yu (1996). "Fano, and Le Cam". In: *Festschrift for Lucien Le Cam*, pp. 423–435.

Auer, Peter, Nicolo Cesa-Bianchi, and Paul Fischer (2002). "Finite-time analysis of the multiarmed bandit problem". In: *Machine learning* 47.2, pp. 235–256.

Balakrishnan, Sivaraman and Larry Wasserman (2017a). "Hypothesis testing for densities and high-dimensional multinomials: Sharp local minimax rates". In: *arXiv preprint arXiv:1706.10003*.

— (2017b). "Hypothesis Testing for High-Dimensional Multinomials: A Selective Review". In: *arXiv preprint arXiv:1712.06120*.

— (2018). "Hypothesis testing for high-dimensional multinomials: A selective review". In: *Annals of Applied Statistics* 12.2, pp. 727–749.

Baraud, Yannick (2002a). "Non-asymptotic minimax rates of testing in signal detection". In: *Bernoulli* 8.5, pp. 577–606.

Baraud, Yannick (2002b). "Non-asymptotic minimax rates of testing in signal detection". In: *Bernoulli* 8.5, pp. 577–606.

Batu, Tugkan, Lance Fortnow, Ronitt Rubinfeld, Warren D Smith, and Patrick White (2000). "Testing that distributions are close". In: *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*. IEEE, pp. 259–269.

Bauer, Benedikt, Luc Devroye, Michael Kohler, Adam Krzyżak, and Harro Walk (2017). "Nonparametric estimation of a function from noiseless observations at random points". In: *Journal of Multivariate Analysis* 160, pp. 93–104.

Berend, Daniel and Aryeh Kontorovich (2013). "A sharp estimate of the binomial mean absolute deviation with applications". In: *Statistics & Probability Letters* 83.4, pp. 1254–1259.

Berrett, Thomas B and Cristina Butucea (2020). "Locally private non-asymptotic testing of discrete distributions is faster using interactive mechanisms". In: *arXiv preprint arXiv:2005.12601*.

Berthier, Raphaël, Francis Bach, and Pierre Gaillard (2020). "Tight Nonparametric Convergence Rates for Stochastic Gradient Descent under the Noiseless Linear Model". In: *arXiv preprint arXiv:2006.08212*.

Bhattacharya, Bhaswar and Gregory Valiant (2015). "Testing closeness with unequal sized samples". In: *Advances in Neural Information Processing Systems*, pp. 2611–2619.

Bickel, P. J. and Y. Ritov (1992). "Testing for goodness-of-fit : A new approach". In: *In Nonparametric statistics and related topics- North-Holland, Amsterdam* 23.5, pp. 51–57.

Butucea, Cristina, Amandine Dubois, Martin Kroll, and Adrien Saumard (2020). "Local differential privacy: Elbow effect in optimal density estimation and adaptation over Besov ellipsoids". In: *Bernoulli* 26.3, pp. 1727–1764.

Butucea, Cristina, Angelika Rohde, and Lukas Steinberger (2020). "Interactive versus non-interactive locally, differentially private estimation: Two elbows for the quadratic functional". In: *arXiv preprint arXiv:2003.04773*.

Butucea, Cristina and Karine Tribouley (2006). "Nonparametric homogeneity tests". In: *Journal of statistical planning and inference* 136.3, pp. 597–639.

Cai, Bryan, Constantinos Daskalakis, and Gautam Kamath (2017). "Priv'it: Private and sample efficient identity testing". In: *arXiv preprint arXiv:1703.10127*.

Canévet, Olivier, Cijo Jose, and François Fleuret (2016). "Importance Sampling Tree for Large-scale Empirical Expectation". In: *Proceedings of the International Conference on Machine Learning (ICML)*. EPFL-CONF-218848.

Canonne, Clément L, Gautam Kamath, Audra McMillan, Jonathan Ullman, and Lydia Zakynthinou (2019). "Private identity testing for high-dimensional distributions". In: *arXiv preprint arXiv:1905.11947*.

Cappé, Olivier, Randal Douc, Arnaud Guillin, Jean-Michel Marin, and Christian P Robert (2008). "Adaptive importance sampling in general mixture classes". In: *Statistics and Computing* 18.4, pp. 447–459.

Chan, Siu-On, Ilias Diakonikolas, Paul Valiant, and Gregory Valiant (2014). "Optimal algorithms for testing closeness of discrete distributions". In: *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*. SIAM, pp. 1193–1203.

Chaudhuri, Kamalika and Sanjoy Dasgupta (2014). "Rates of convergence for nearest neighbor classification". In: *arXiv preprint arXiv:1407.0067*.

Cortez, Paulo and Aníbal de Jesus Raimundo Morais (2007). "A data mining approach to predict forest fires using meteorological data". In.

Delyon, Bernard and François Portier (2018). "Efficiency of adaptive importance sampling". In: *arXiv preprint arXiv:1806.00989*.

Devroye, Luc (1986). "Sample-based non-uniform random variate generation". In: *Proceedings of the 18th conference on Winter simulation*. ACM, pp. 260–265.

Diakonikolas, Ilias, Themis Gouleakis, John Peebles, and Eric Price (2017). "Sample-optimal identity testing with high probability". In: *arXiv preprint arXiv:1708.02728*.

Diakonikolas, Ilias and Daniel M Kane (2016). "A new approach for testing properties of discrete distributions". In: *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*. IEEE, pp. 685–694.

Diakonikolas, Ilias, Daniel M Kane, and Vladimir Nikishkin (2015). "Optimal algorithms and lower bounds for testing closeness of structured distributions". In: *arXiv preprint arXiv:1508.05538*.

— (2017). "Near-optimal closeness testing of discrete histogram distributions". In: *arXiv preprint arXiv:1703.01913*.

Dinur, Irit and Kobbi Nissim (2003). "Revealing information while preserving privacy". In: *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 202–210.

Donoho, David L and Iain M Johnstone (1998). "Minimax estimation via wavelet shrinkage". In: *The annals of Statistics* 26.3, pp. 879–921.

Donoho, David L, Iain M Johnstone, Gérard Kerkyacharian, and Dominique Picard (1996). "Density estimation by wavelet thresholding". In: *The Annals of Statistics*, pp. 508–539.

Duchi, John C, Michael I Jordan, and Martin J Wainwright (2013a). "Local privacy and minimax bounds: Sharp rates for probability estimation". In: *arXiv preprint arXiv:1305.6000*.

— (2013b). "Local privacy and statistical minimax rates". In: *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*. IEEE, pp. 429–438.

— (2013c). "Local privacy, data processing inequalities, and statistical minimax rates". In: *arXiv preprint arXiv:1302.3203*.

— (2018). "Minimax optimal procedures for locally private estimation". In: *Journal of the American Statistical Association* 113.521, pp. 182–201.

Duchi, John C, Martin J Wainwright, and Michael I Jordan (2013). "Local privacy and minimax bounds: Sharp rates for probability estimation". In: *Advances in Neural Information Processing Systems*, pp. 1529–1537.

Duncan, George T and Diane Lambert (1986). "Disclosure-limited data dissemination". In: *Journal of the American statistical association* 81.393, pp. 10–18.

— (1989). "The risk of disclosure for microdata". In: *Journal of Business & Economic Statistics* 7.2, pp. 207–217.

Dwork, Cynthia, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor (2006a). "Our data, ourselves: Privacy via distributed noise generation". In: *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, pp. 486–503.

Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith (2006b). "Calibrating noise to sensitivity in private data analysis". In: *Theory of cryptography conference*. Springer, pp. 265–284.

Dwork, Cynthia and Aaron Roth (2014). "The algorithmic foundations of differential privacy." In: *Foundations and Trends in Theoretical Computer Science* 9.3-4, pp. 211–407.

Dymetman, Marc, Guillaume Bouchard, and Simon Carter (2012). "The OS* algorithm: a joint approach to exact optimization and sampling". In: *arXiv preprint arXiv:1207.0742*.

Erlingsson, Úlfar, Vasyl Pihur, and Aleksandra Korolova (2014). "Rappor: Randomized aggregatable privacy-preserving ordinal response". In: *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pp. 1054–1067.

Erraqabi, Akram, Michal Valko, Alexandra Carpentier, and Odalric Maillard (2016). "Pliable rejection sampling". In: *International Conference on Machine Learning*, pp. 2121–2129.

Evans, Michael and Timothy Swartz (1998). "Random variable generation using concavity properties of transformed densities". In: *Journal of Computational and Graphical Statistics* 7.4, pp. 514–528.

Evfimievski, Alexandre, Johannes Gehrke, and Ramakrishnan Srikant (2003). "Limiting privacy breaches in privacy preserving data mining". In: *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, pp. 211–222.

Fienberg, Stephen E and Russell J Steele (1998). "Disclosure limitation using perturbation and related methods for categorical data". In: *Journal of Official Statistics* 14.4, p. 485.

Fill, James Allen (1997). "An interruptible algorithm for perfect sampling via Markov chains". In: *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*. ACM, pp. 688–695.

Fromont, Magalie and Béatrice Laurent (2006a). "Adaptive goodness-of-fit tests in a density model". In: *The annals of statistics* 34.2, pp. 680–720.

— (2006b). "Adaptive goodness-of-fit tests in a density model". In: *The Annals of Statistics* 34.2, pp. 680–720.

Fromont, Magalie, Béatrice Laurent, and Patricia Reynaud-Bouret (2011). "Adaptive tests of homogeneity for a Poisson process". In: *Annales de l'IHP Probabilités et statistiques*. Vol. 47. 1, pp. 176–213.

Gaboardi, Marco, Hyun-Woo Lim, Ryan M Rogers, and Salil P Vadhan (2016). "Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing". In: *ICML'16 Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*. JMLR.

Gaboardi, Marco and Ryan Rogers (2017). "Local private hypothesis testing: Chi-square tests". In: *arXiv preprint arXiv:1709.07155*.

Gilks, Walter R (1992). *Derivative-free adaptive rejection sampling for Gibbs sampling, Bayesian Statistics 4*.

Gilks, Walter R, NG Best, and KKC Tan (1995). "Adaptive rejection Metropolis sampling within Gibbs sampling". In: *Applied Statistics*, pp. 455–472.

Gilks, Walter R and Pascal Wild (1992). "Adaptive rejection sampling for Gibbs sampling". In: *Applied Statistics*, pp. 337–348.

Giné, Evarist and Richard Nickl (2016). *Mathematical foundations of infinite-dimensional statistical models*. Vol. 40. Cambridge University Press.

— (2021). *Mathematical foundations of infinite-dimensional statistical models*. Cambridge university press.

Goldenshluger, Alexander and Oleg Lepski (2014). "On adaptive minimax density estimation on R^d". In: *Probability Theory and Related Fields* 159.3-4, pp. 479–543.

Goldreich, Oded, Shari Goldwasser, and Dana Ron (1998). "Property testing and its connection to learning and approximation". In: *Journal of the ACM (JACM)* 45.4, pp. 653–750.

Görür, Dilan and Yee Whye Teh (2011). "Concave-convex adaptive rejection sampling". In: *Journal of Computational and Graphical Statistics* 20.3, pp. 670–691.

Han, Yanjun, Jiantao Jiao, and Tsachy Weissman (2015). "Minimax estimation of discrete distributions". In: *2015 IEEE International Symposium on Information Theory (ISIT)*. IEEE, pp. 2291–2295.

Härdle, Wolfgang, Gerard Kerkyacharian, Dominique Picard, and Alexander Tsybakov (2012). *Wavelets, approximation, and statistical applications*. Vol. 129. Springer Science & Business Media.

Hörmann, Wolfgang (1995). "A rejection technique for sampling from T-concave distributions". In: *ACM Transactions on Mathematical Software (TOMS)* 21.2, pp. 182–193.

Inglot T., Ledwina T. (1996). "Asymptotic optimality of data-driven Neyman's tests for uniformity". In: *The Annals of Statistics* 24.5, pp. 1982–2019.

Ingster, Yu I (1987). "Minimax testing of nonparametric hypotheses on a distribution density in the L_p metrics". In: *Theory of Probability & Its Applications* 31.2, pp. 333–337.

— (2000a). "Adaptive chi-square tests". In: *Journal of Mathematical Sciences* 99.2, pp. 1110–1119.

— (2000b). "Adaptive chi-square tests". In: *Journal of Mathematical Sciences* 99.2, pp. 1110–1119.

Ingster, Yuri and Irina A Suslina (2012). *Nonparametric goodness-of-fit testing under Gaussian models*. Vol. 169. Springer Science & Business Media.

Ingster, Yuri I (1993). "Asymptotically minimax hypothesis testing for nonparametric alternatives. I, II, III". In: *Math. Methods Statist* 2.2, pp. 85–114.

Jank, Wolfgang, Galit Shmueli, and Shanshan Wang (2008). *Statistical methods in e-commerce research*. Wiley Online Library.

Joseph, Matthew, Jieming Mao, Seth Neel, and Aaron Roth (2019). "The role of interactivity in local differential privacy". In: *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, pp. 94–105.

Juditsky, Anatoli and Sophie Lambert-Lacroix (2004). "On minimax density estimation on $\mathbb{R}$". In: *Bernoulli* 10.2, pp. 187–220.

Kairouz, Peter, Keith Bonawitz, and Daniel Ramage (2016). "Discrete distribution estimation under local privacy". In: *International Conference on Machine Learning*. PMLR, pp. 2436–2444.

Kallenberg W., Ledwina T. (1995). "Consistency and Monte Carlo simulation of a data driven version of smooth goodness-of-fit tests". In: *The Annals of Statistics* 23.5, pp. 1594–1608.

Kasiviswanathan, Shiva Prasad, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith (2011). "What can we learn privately?" In: *SIAM Journal on Computing* 40.3, pp. 793–826.

Kerkyacharian, Gérard and Dominique Picard (1992). "Density estimation in Besov spaces". In: *Statistics & probability letters* 13.1, pp. 15–24.

— (1993). "Density estimation by kernel and wavelets methods: optimality of Besov spaces". In: *Statistics & Probability Letters* 18.4, pp. 327–336.

Kim, Ilmun (2020). "Multinomial goodness-of-fit based on u-statistics: High-dimensional asymptotic and minimax optimality". In: *Journal of Statistical Planning and Inference* 205, pp. 74–91.

Kim, Ilmun, Sivaraman Balakrishnan, and Larry Wasserman (2018). "Robust Multivariate Nonparametric Tests via Projection-Pursuit". In: *arXiv preprint arXiv:1803.00715*.

Kohler, Michael (2014). "Optimal global rates of convergence for noiseless regression estimation problems with adaptively chosen design". In: *Journal of Multivariate Analysis* 132, pp. 197–208.

Kohler, Michael and Adam Krzyżak (2013). "Optimal global rates of convergence for interpolation problems with random design". In: *Statistics & Probability Letters* 83.8, pp. 1871–1879.

Kpotufe, Samory (2011). "k-NN regression adapts to local intrinsic dimension". In: *arXiv preprint arXiv:1110.4300*.

Lam-Weil, Joseph, Béatrice Laurent, and Jean-Michel Loubes (2020). "Minimax optimal goodness-of-fit testing for densities under a local differential privacy constraint". In: *arXiv preprint arXiv:2002.04254*.

Lattimore, Tor and Csaba Szepesvári (2020). *Bandit algorithms*. Cambridge University Press.

Le Cam, Lucien M (1986). "Asymptotic methods in statistical theory". In.

Ledwina, Teresa (1994). "Data-driven version of Neyman's smooth test of fit". In: *Journal of the American Statistical Association* 89.427, pp. 1000–1005.

Lehmann, Erich L and Joseph P Romano (2006). *Testing statistical hypotheses*. Springer Science & Business Media.

Maddison, Chris J, Daniel Tarlow, and Tom Minka (2014). "A* sampling". In: *Advances in Neural Information Processing Systems*, pp. 3086–3094.

Martino, Luca (2017). "Parsimonious adaptive rejection sampling". In: *Electronics Letters* 53.16, pp. 1115–1117.

Martino, Luca and Francisco Louzada (2017). "Adaptive Rejection Sampling with fixed number of nodes". In: *Communications in Statistics-Simulation and Computation*, pp. 1–11.

Martino, Luca and Joaquín Míguez (2011). "A generalization of the adaptive rejection sampling algorithm". In: *Statistics and Computing* 21.4, pp. 633–647.

Martino, Luca, Jesse Read, and David Luengo (2012). "Improved adaptive rejection Metropolis sampling algorithms". In: *arXiv preprint arXiv:1205.5494*.

Metropolis, Nicholas and Stanislaw Ulam (1949). "The monte carlo method". In: *Journal of the American statistical association* 44.247, pp. 335–341.

Meyer, Renate, Bo Cai, and François Perron (2008). "Adaptive rejection Metropolis sampling using Lagrange interpolation polynomials of degree 2". In: *Computational Statistics & Data Analysis* 52.7, pp. 3408–3423.

Meyer, Yves (1990). "Ondelettes et opérateurs". In: *I: Ondelettes*.

Meynaoui, A., M. Albert, B. Laurent, and A. Marrel (2019). "Adaptive test of independence based on HSIC measures". In: *hal-02020084*.

Mishra, Nina and Mark Sandler (2006). "Privacy via pseudorandom sketches". In: *Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, pp. 143–152.

Naesseth, Christian, Francisco Ruiz, Scott Linderman, and David Blei (2017). "Reparameterization gradients through acceptance-rejection sampling algorithms". In: *Artificial Intelligence and Statistics*, pp. 489–498.

Naesseth, Christian A, Francisco JR Ruiz, Scott W Linderman, and David M Blei (2016). "Rejection Sampling Variational Inference". In: *arXiv preprint arXiv:1610.05683*.

Neyman, J. (1937). "Smooth test for goodness of fit". In: *Scandinavian Actuarial Journal* 1937.3-4, pp. 149–199.

Neyman, Jerzy and Egon S Pearson (1933). "IX. On the problem of the most efficient tests of statistical hypotheses". In: *Phil. Trans. R. Soc. Lond. A* 231.694-706, pp. 289–337.

Oh, Man-Suk and James O Berger (1992). "Adaptive importance sampling in Monte Carlo integration". In: *Journal of Statistical Computation and Simulation* 41.3-4, pp. 143–168.

Orabona, Francesco (2019). "A modern introduction to online learning". In: *arXiv preprint arXiv:1912.13213*.

Paninski, Liam (2008). "A coincidence-based test for uniformity given very sparsely sampled discrete data". In: *IEEE Transactions on Information Theory* 54.10, pp. 4750–4755.

Propp, James and David Wilson (1998). "Coupling from the past: a user's guide". In: *Microsurveys in Discrete Probability* 41, pp. 181–192.

Raskutti, Garvesh, Martin J Wainwright, and Bin Yu (2011). "Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls". In: *IEEE transactions on information theory* 57.10, pp. 6976–6994.

Reeve, Henry WJ and Gavin Brown (2017). "Minimax rates for cost-sensitive learning on manifolds with approximate nearest neighbours". In: *International Conference on Algorithmic Learning Theory*. PMLR, pp. 11–56.

Rubinfeld, Ronitt and Madhu Sudan (1996). "Robust characterizations of polynomials with applications to program testing". In: *SIAM Journal on Computing* 25.2, pp. 252–271.

Ryu, Ernest K and Stephen P Boyd (2014). "Adaptive importance sampling via stochastic convex programming". In: *arXiv preprint arXiv:1412.4845*.

Serfling, Robert J (2009). *Approximation theorems of mathematical statistics*. Vol. 162. John Wiley & Sons.

Shalev-Shwartz, Shai and Shai Ben-David (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Sheffet, Or (2018). "Locally private hypothesis testing". In: *arXiv preprint arXiv:1802.03441*.

Tsybakov, Alexandre B (2004). "Introduction to nonparametric estimation, 2009". In: *URL https://doi. org/10.1007/b13794. Revised and extended from the*.

— (2008). *Introduction to nonparametric estimation*. Springer Science & Business Media.

Valiant, Gregory and Paul Valiant (2017). "An automatic inequality prover and instance optimal identity testing". In: *SIAM Journal on Computing* 46.1, pp. 429–455.

Valiant, Paul (2011). "Testing symmetric properties of distributions". In: *SIAM Journal on Computing* 40.6, pp. 1927–1968.

van den Hout, Ardo and Peter GM van der Heijden (2002). "Randomized response, statistical disclosure control and misclassificatio: a review". In: *International Statistical Review* 70.2, pp. 269–288.

Van der Vaart, Aad W (2000). *Asymptotic statistics*. Vol. 3. Cambridge university press.

Von Neumann, John (1951). "13. Various Techniques Used in Connection With Random Digits". In: *Appl. Math Ser* 12, pp. 36–38.

Warner, Stanley L (1965). "Randomized response: A survey technique for eliminating evasive answer bias". In: *Journal of the American Statistical Association* 60.309, pp. 63–69.

Wasserman, Larry and Shuheng Zhou (2010). "A statistical framework for differential privacy". In: *Journal of the American Statistical Association* 105.489, pp. 375–389.

Yang, Yuhong and Andrew Barron (1999). "Information-theoretic determination of minimax rates of convergence". In: *Annals of Statistics*, pp. 1564–1599.

Yu, Bin (1997). "Assouad, fano, and le cam". In: *Festschrift for Lucien Le Cam*. Springer, pp. 423–435.

Zhang, Ping (1996). "Nonparametric importance sampling". In: *Journal of the American Statistical Association* 91.435, pp. 1245–1253.

Zhao, Jun, Teng Wang, Tao Bai, Kwok-Yan Lam, Zhiying Xu, Shuyu Shi, Xuebin Ren, Xinyu Yang, Yang Liu, and Han Yu (2019). "Reviewing and improving the Gaussian mechanism for differential privacy". In: *arXiv preprint arXiv:1911.12060*.