Journal of Machine Learning Research 26 (2025) 1-52 Submitted 10/24; Revised 10/25; Published 11/25

Generalized multi-view model: Adaptive density estimation under low-rank constraints

Julien Chhor JULIEN.CHHOR@TSE-FR.EU
Toulouse School of Economics, University of Toulouse Capitole, Toulouse, France

Olga Klopp KLOPP@ESSEC.EDU
ESSEC Business School
Cergy-Pontoise, France

Alexandre B. Tsybakov ALEXANDRE.TSYBAKOV@ENSAE.FR
CREST, ENSAE
Institut Polytechnique de Paris, France
Palaiseau, France

Editor: Gabor Lugosi

Abstract

We study the problem of bivariate discrete or continuous probability density estimation under low-rank constraints. For discrete distributions, we assume that the two-dimensional array to estimate is a low-rank probability matrix. In the continuous case, we assume that the density with respect to the Lebesgue measure satisfies a generalized multi-view model, meaning that it is β-Hölder and can be decomposed as a sum of K components, each of which is a product of one-dimensional functions. In both settings, we propose estimators that achieve, up to logarithmic factors, the minimax optimal convergence rates under such low-rank constraints. In the discrete case, the proposed estimator is adaptive to the rank K. In the continuous case, our estimator converges with the L_1 rate $\min((K/n)^{\beta/(2\beta+1)}, n^{-\beta/(2\beta+2)})$ up to logarithmic factors, and it is adaptive to the unknown support as well as to the smoothness β and to the unknown number of separable components K. We present efficient algorithms to compute our estimators.

Keywords: density estimation, multi-view model, low-rank models, minimax rate of convergence, adaptive estimation

1. Introduction

Estimating discrete and continuous probability distributions is one of the fundamental problems in statistics and machine learning. A classical density estimator for both discrete and continuous data is the histogram, while for continuous densities the most popular method is kernel density estimator (KDE) see, e.g., Silverman (1986); Devroye and Györfi (1985); Scott (1992); Klemelä (2009); Tsybakov (2009); Gramacki (2017); Wang and Scott (2019).

Asymptotically, these estimators can consistently recover any probability density on $\mathbb{R}^m$ in the total variation (L_1) distance based on n independent identically distributed (iid) observations, if $nh^m \to \infty$, where h is the tuning parameter (bin width for the histogram in the continuous case and bandwidth for the KDE), and we assume that $h \to 0$, $n \to \infty$,

see, e.g., (Devroye and Györfi, 1985, Theorems 1 and 2). On the other hand, smoothness assumptions on the underlying density are not enough to grant good accuracy of these estimators when m is large, even for compactly supported densities. Their rate of convergence drastically deteriorates with the dimension if h is chosen optimally, and it remains true for any estimators under only smoothness assumptions. This gives rise to suggestions of density estimators that overcome the curse of dimensionality under more assumptions on the underlying density than only smoothness. An early suggestion is Projection Pursuit density estimation (PPDE) Friedman et al. (1984), which is an iterative algorithm to estimate the density by finding a subspace spanned by a small number of significant components. One may consider PPDE as being adapted to the setting where there exists a linear map that transforms the underlying random vector to a smaller dimensional random vector with independent components. This technique is rather popular but, to the best of our knowledge, theoretical guarantees on the performance of PPDE are not available, see a recent survey on PPDE in Wang and Scott (2019). A related dimension reduction model for density estimation arises from independent component analysis (ICA), where one assumes the existence of a linear bijection of the underlying random vector to a random vector of the same dimension with independent components Samarov and Tsybakov (2004). It is shown that under this model there is no curse of dimensionality in the sense that there exist estimators achieving one-dimensional rates Samarov and Tsybakov (2004, 2007); Amato et al. (2010); Lepski and Rebelles (2020). Finally, a recent line of work starting from Song et al. (2014) and further developed in Amiridi et al. (2022b,a); Kargas and Sidiropoulos (2019); Song and Dai (2013); Vandermeulen and Ledent (2021); Vandermeulen (2023) deals with the multi-view model for density estimation, that is, a finite mixture model whose components are products of one-dimensional probability densities. Densities $f : \mathbb{R}^m \to \mathbb{R}$ satisfying the multi-view model are the form

$$f(x) = \sum_{i=1}^{K} w_i \prod_{j=1}^{m} f_{ij}(x^T e_j) \quad \text{with} \quad \sum_{i=1}^{K} w_i = 1, w_i \geq 0, \tag{1}$$

where e_j's are the canonical basis vectors in $\mathbb{R}^m$ and f_{ij}'s are one-dimensional probability densities. Weights w_i and f_{ij}'s are unknown. The fact that this model is free from the curse of dimensionality when f_{ij}'s are Lipschitz continuous and supported of $[0, 1]$ is demonstrated in Vandermeulen and Ledent (2021) through a theoretical analysis of its sample complexity. However, Vandermeulen and Ledent (2021) does not develop computationally tractable estimators.

In this paper, we focus on two-dimensional density estimation and consider a new model that generalizes the multi-view model in two aspects. We call it the *generalized multi-view model*. First, in contrast to the usual multi-view model, we do not assume the additive components to be products of densities but rather products of arbitrary functions. Second, we do not assume these functions to be Lipschitz continuous. We only need a β-Hölder continuity for some $\beta \in (0, 1]$ of the overall two-dimensional density, which is a sum of K such products. We propose a new estimator that is both computationally tractable and offers improved statistical guarantees by achieving the one-dimensional estimation rate to within a logarithmic factor. Our analysis deals with the L_1 risk of density estimators. As argued in Devroye and Györfi (1985), using the L_1 norm to characterize the error of

density estimation has several advantages. Indeed, the L_1 distance between two densities is equivalent, up to a multiplicative constant factor, to the total variation distance between the corresponding probability measures. Moreover, the L_1 risk for density estimation is transformation invariant, which is not the case for the L_2 risk most often studied in the literature. From the technical point of view, it is more difficult to deal with the L_1 risk than with the L_2 risk.

Our approach to tackling generalized multi-view models is based on a reduction to the problem of estimating multivariate discrete distributions under a low-rank structure that we also consider in detail. This problem is of independent interest and it arises in many applications, in particular, if the aim is to explore correlations between categorical random variables, which is a particularly relevant subject Johndrow et al. (2017); Dunson and Xing (2009); Diakonikolas et al. (2019); Tahmasebi et al. (2018). Without structural assumptions, estimating a discrete distribution on a set of cardinality D in total variation distance is associated with the minimax risk of the order of $\sqrt{D/n}$, where n is the number of observations (see Kamath et al. (2015); Han et al. (2015) and Corollary 2 below). However, by imposing certain structure assumptions, it is possible to reduce the estimation risk. Several assumptions, including monotonicity, unimodality, t-modality, convexity, log-concavity or t-piecewise degree k-polynomial structure have been considered in the literature (see, for example, Canonne et al. (2018); Diakonikolas et al. (2015); Durot et al. (2013)). In the present paper, we deal with a different setting where a multivariate discrete distribution is estimated under a low-rank structure. Specifically, for two integers $d_1, d_2 \geq 2$, we consider the problem of estimating a discrete distribution on a set of cardinality $D = d_1 d_2$ defined by a matrix of probabilities $P = (P_{ij})_{i\in[d_1],j\in[d_2]}$ with rank at most $K \geq 1$. This setting arises, for example, in the analysis of data represented as a table with n rows and 2 columns. Each row corresponds to one individual, while each column contains information about one feature of the individual, for instance, (1) eye color, and (2) hair color. We assume that the value of the cell in the table is a categorical random variable, for example, that hair color can only take 6 possible values: black, brown, red, blond, gray, white. Thus, each element of the ℓ-th column has a discrete distribution over $\{1, \dots, d_\ell\}$, for some $d_\ell \geq 2$, $\ell = 1, 2$. The individuals are assumed to be iid but the columns can be correlated. For instance, hair color can be correlated with eye color. In other words, each row can be viewed as a realization of a pair of correlated discrete random variables. If we are interested in possible associations between the two variables, we are lead to estimating their joint distribution. Assuming that P has low rank means that there exists a reduced representation of the correlation structure. A basic unbiased estimator of P is a histogram Y/n where Y is a $d_1 \times d_2$ matrix such that its (i, j)-th entry Y_{ij} is the number of individuals whose row in the data table equals (i, j). However, the histogram does not take advantage of low-rank structure and it only attains the slow rate $\sqrt{d_1 d_2/n}$ in the total variation distance. We will suggest an estimator attaining a faster rate and show that it is minimax optimal up to logarithmic factors.

2. Motivation

In many applications, one needs to explore relations between two objects that may have a complex structure, yet are linked via a low-dimensional latent space. This situation

can be often described by mixture models and low-rank matrix models. For the problem with discrete distributions, one of the important examples is given by the *probabilistic Latent Semantic Indexing* framework for topic models Hofmann (1999). It assumes that co-occurrences of words and documents are independent given one of K latent topic classes. Then the joint probability matrix of words and documents is a mixture of at most K matrices and its rank does not exceed K, which is typically a small number. Another example of low-rank probability matrix estimation is provided by the Stochastic Block Model Holland et al. (1983); Abbe (2018). In this case, the problem is to estimate the matrix of connection probabilities of a random graph under the assumption that its nodes fall into K groups with constant connection probabilities within and between each two groups. Such a probability matrix is of rank at most K. Low-rank probability matrix estimation problems also arise in the context of collaborative filtering, matrix completion and other Candès and Recht (2009); Rennie and Srebro (2005).

For the problems characterized by continuous probability densities, multi-view models provide a nonparametric analog of classical mixture models. In contrast to these classical models, they do not assume that the components of the mixture depend on finite number of parameters but rather consider them as functions satisfying some general constraints, such as smoothness or just integrability. In model (1), the resulting function f is the probability density of a random vector $X = (x_1, \dots, x_m) \in [0,1]^m$ with entries $x_1, \dots, x_m$ that are independent conditional on a latent variable that can take K distinct values. The generalized multi-view model considered in this paper is broader than the basic multi-view model (1) as it allows the functions f_{ij} to be integrable real-valued functions rather than densities. Also, we only assume Hölder smoothness of f and not of all the components f_{ij}. Nevertheless, we will demonstrate that estimation over the class of generalized multi-view models is, up to logarithmic factors, not harder than estimation over the subclass described by (1). In this paper, we focus on the setting, where the aim is to explore relations between two variables ($m = 2$) and we explicitly construct polynomial-time estimators achieving the optimal rates for such models.

A relevant question is to check whether the multi-view model holds for a given particular problem in practice. We address this issue by providing estimators that are adaptive to the unknown number of components K varying on a wide scale of values. Very large values of K correspond to the absence of low-rank structure. For such K, our estimator achieves the same rate as the usual nonparametric density estimator of a smooth density (with no additional structure), and we show that this is optimal. In other words, our adaptive estimator achieves the minimax optimal rate regardless of whether the multi-view model holds or not. Thus, adaptation guarantees that checking the low-rank assumption is not necessary in practice.

3. Summary of contributions and related work

The importance of low-rank structures in nonparametric density estimation has been discussed in several papers suggesting and analyzing various estimation methods Amiridi et al. (2022b,a); Kargas and Sidiropoulos (2019); Song et al. (2014); Song and Dai (2013); Vandermeulen and Ledent (2021). In Song et al. (2014), the authors introduced the multi-view model and proposed a kernel method for learning it but did not study whether it can lead

to an improvement in nonparametric density estimation. The paper Song and Dai (2013) provided empirical evidence that hierarchical low-rank decomposition of kernel embeddings can lead to improved performance in density estimation. More recently, Amiridi et al. (2022b,a) used a low-rank characteristic function to improve nonparametric density estimation. Finally, Vandermeulen and Ledent (2021) proved that there exists an estimator that converges at a rate $O\big(n^{-1/3}\big)$ in the L_1 norm to any density satisfying the multi-view model with Lipschitz continuous marginals. This estimator is not constructed explicitly and is not computationally tractable in general. Furthermore, Vandermeulen and Ledent (2021) shows that the standard histogram estimator can converge at a rate slower than $O\big(n^{-1/m}\big)$ on the same class of densities, where m is the dimension of the data.

A related problem of estimating a low-rank probability matrix from discrete counts has been considered in Jain and Orlitsky (2020), where the authors propose a polynomial time algorithm (called the curated SVD) for the case of square matrix, $d_1 = d_2 = d$, and requiring the exact knowledge of the rank K. They prove an upper bound on the total variation error of the curated SVD that scales as $\psi(K, d, n) := \sqrt{\frac{Kd}{n}} \wedge 1$ (to within a logarithmic factor in K, n) with probability at least $1 - d^{-2}$. They also state a lower bound with the rate $\psi(K, d, n)$ *in expectation* by referring to the lower bounds of Han et al. (2015); Kamath et al. (2015) for general discrete distributions on a set of cardinality $D = Kd$. Based on that, the authors in Jain and Orlitsky (2020) claim minimax optimality of the rate $\psi(K, d, n)$, up to logarithmic factors. However, the lower bound obtained by this argument only holds under significant restrictions on K, d, n that are not specified in Jain and Orlitsky (2020). Indeed, the lower bounds of Han et al. (2015); Kamath et al. (2015) for general discrete distributions are only meaningful under some specific conditions on D, n. For example, the lower bound of (Han et al., 2015, Theorem 1) is vacuous for $D \asymp 1$ and the lower bound of (Kamath et al., 2015, Lemma 8) is vacuous for $D \asymp \sqrt{n}$. These lower bounds are established only in expectation and it is not legitimate to compare them directly with upper bounds in probability derived in Jain and Orlitsky (2020). The upper bound in probability in (Jain and Orlitsky, 2020, Theorem 2) can be transformed into a bound in expectation at the expense of adding a $O(d^{-2})$ term to the rate. This imposes one more restriction $d^{-2} \lesssim \psi(K, d, n)$ to match the lower bound up to a logarithmic factor. The algorithm suggested in Jain and Orlitsky (2020) is based on normalization of the probability matrix by rescaling each row and column. Similar ideas have been developed in the literature on topic models, where topic matrices are estimated using an SVD-based technique on a rescaled corpus matrix Ke and Wang (2022). Our approach is different and based on novel techniques that we call a *localized SVD denoising*. The localized SVD algorithm that we suggest does not need to rescale the input matrix and relies on splitting the matrix into sub-matrices with entries of similar order of magnitude. It may be of interest in other contexts as well.

The contributions of the present work are as follows.

- We prove minimax lower bounds in the total variation distance for general discrete distributions on a set of cardinality D. We generalize Han et al. (2015); Kamath et al. (2015) in the sense that we derive lower bounds not only in expectation but also in probability and, in contrast to those works, we obtain the lower rate $\sqrt{\frac{D}{n}} \wedge 1$ for all $D, n \geq 1$ with no restriction. Next, under the low-rank matrix structure, we prove

lower bounds of the order of $\psi(K, d, n)$ both in expectation and in probability, with no restriction on K, d, n, where $d = d_1 \vee d_2$. Moreover, we propose a computationally efficient algorithm to estimate a low-rank probability matrix P and show that it attains the same rate $\psi(K, d, n)$ up to a logarithmic factor. Thus, we prove the minimax optimality of this rate and of our algorithm, up to a logarithmic factor. Unlike the curated SVD of Jain and Orlitsky (2020), our algorithm applies to non-square matrices and is adaptive to the unknown rank K.

- We propose a method to estiamte β-Hölder densities for $\beta \in (0, 1]$ under the generalized multi-view model. Our algorithm achieves the rate of convergence $\big(K/n\big)^{\beta/(2\beta+1)} \wedge n^{-\beta/(2\beta+2)}$ up to a logarithmic factor on the class of densities that are (i) β-Hölder over an *unknown* sub-rectangle of $[0, 1]^2$ and (ii) represented as a sum of K separable components. In the two-dimensional case that we consider, we improve upon the prior work Vandermeulen and Ledent (2021) in the following aspects:
  - Our estimator is computationally tractable.
  - The study in Vandermeulen and Ledent (2021) was devoted to the case $\beta = 1$ and the standard multi-view model while we provide an extension to the generalized multi-view model described above and to any $\beta \in (0, 1]$.
  - We prove a lower bound showing that the above convergence rate is minimax optimal up to a logarithmic factor on the class of densities satisfying the generalized multi-view model. We establish the explicit dependence of the minimax rate on K, n, β revealing, in particular, that it exhibits an elbow at $K \asymp n^{1/(2\beta+2)}$. We note that our lower bound is stronger since we prove it for the smaller class $\mathcal{G}^{\circ}_{K,\beta}$ of densities satisfying the standard multi-view model (1) with $m = 2$, where f_{ij}'s are probability densities on $[0, 1]$ that are β-Hölder on their support.
  - We propose an estimator that is adaptive to the unknown number of separable components K, to the unknown smoothness β, and to the unknown support of the density. As shown by our lower bound, this estimator also reaches the minimax optimal convergence rate, up to a logarithmic factor, on the class $\mathcal{G}^{\circ}_{K,\beta}$. It can therefore be employed for learning mixture models from the class $\mathcal{G}^{\circ}_{K,\beta}$ while guaranteeing robustness to the model misspecification since it attains a comparable rate over the substantially larger class of generalized multi-view models.
- We provide a package for computation available at `https://github.com/hi-paris/Lowrankdensity`. We run a numerical experiment demonstrating the efficiency of our estimators both in discrete and continuous settings.

4. Notation

For two real numbers x, y, we define $x \wedge y := \min(x, y)$ and $x \vee y := \max(x, y)$. For $d \in \mathbb{N}$, we set $[d] = \{1, \dots, d\}$. For any probability vector or probability matrix P and for any $n \in \mathbb{N}^*$, we denote by $\mathcal{M}(P, n)$ the multinomial distribution with probability parameter P and sample size n.

For any matrix Λ, we denote by Λ_{ij} its (i, j)th entry and by $\mathrm{rk}(\Lambda)$ its rank. We denote by $\Lambda_{IJ} = (\Lambda_{ij})_{i \in I, j \in J}$ an extraction of matrix $\Lambda \in \mathbb{R}^{d_1 \times d_2}$ corresponding to the sets of

indices $I \subseteq [d_1]$ and $J \subseteq [d_2]$. We will use several matrix norms, namely, the operator norm denoted by $\|\Lambda\|$, the nuclear norm $\|\Lambda\|_*$, the Frobenius norm $\|\Lambda\|_F$, the entry-wise ℓ_1-norm $\|\Lambda\|_1$, and the norms

$$\|\Lambda\|_{1,\infty} = \max_{j\in[d_2]} \sum_{i=1}^{d_1} |\Lambda_{ij}|,$$
$$\big\|\Lambda\big\|_{\square} = \|\Lambda\|_{1,\infty} \vee \|\Lambda^\top\|_{1,\infty}.$$

The notation $\|\cdot\|_{L_1}$ is used for the norms in $L_1([0,1], \mathrm{Leb})$ and in $L_1([0,1]^2, \mathrm{Leb})$, where Leb denotes the Lebesgue measure.

We denote by $\|\cdot\|$ the Euclidean norm in $\mathbb{R}^2$, by $\mathrm{Supp}(f) = f^{-1}\big(\mathbb{R}\setminus\{0\}\big)$ the support of a real-valued function f, by $\mathbb{1}_A(\cdot)$ the indicator function of set A, and by $|A|$ the cardinality of finite set A. Throughout the paper, the absolute positive constants are denoted by C and may take different values on each appearance and we assume, unless otherwise stated, that $n, d_1, d_2 \geq 2$.

5. Discrete distributions

Consider first the setting with discrete distributions, which provides a base for studying continuous distributions in Section A. Let $\mathcal{T}_K$ be the class of all probability matrices of rank at most K:

$$\mathcal{T}_K = \left\{ P \in \mathbb{R}^{d_1\times d_2} \;\middle|\; \mathrm{rk}(P) \leq K, \sum_{(i,j)\in[d_1]\times[d_2]} P_{ij} = 1 \;\text{ and } P_{ij} \geq 0,\ \forall (i,j) \in [d_1]\times[d_2] \right\}, \quad (2)$$

where $K \leq d_1 \wedge d_2$ is a positive integer. Assume that for some unknown $P \in \mathcal{T}_K$ we are given iid observations $X_1, \ldots, X_n$ with distribution P, that is, $\mathbb{P}(X_k = (i,j)) = P_{ij}$ for all $k \in [n], (i,j) \in [d_1]\times[d_2]$.

We can assume, for example, that $P \in \mathbb{R}^{d_1\times d_2}$ denotes the joint probability matrix of two categorical variables $X \in [d_1]$ and $Y \in [d_2]$. A parsimonious and widely applicable generative mechanism assumes that there exists a latent class $Z \in [K]$ such that X and Y are independent given Z, where K is small. Then

$$P_{ij} = \mathbb{P}(X = i, Y = j) = \sum_{k=1}^{K} \pi_k\, a_{k,i}\, b_{k,j} \quad \Longleftrightarrow \quad P = \sum_{k=1}^{K} \pi_k\, a_k b_k^\top,$$

so that $\mathrm{rk}(P) \leq K$. This low-rank structure encodes dependence through a low-dimensional latent space and yields substantial statistical savings. Without structure, estimating an arbitrary $d_1 \times d_2$ table in total variation is only possible with the rate $\sqrt{d_1 d_2/n}$, whereas under the condition $\mathrm{rk}(P) \leq K$, as we will prove, the rate improves to $\sqrt{K(d_1 \vee d_2)/n}$ (up to logarithmic factors).

Low rank assumption arises naturally whenever two high-cardinality categorical views are conditionally independent given a small latent variable:

- **Topic models / word–document co-occurrence (pLSI).** Words and documents are independent given a topic; the word × document probability matrix has rank at most K when there are K topics.

- **Networks / stochastic block models (bipartite or directed).** Edges are independent given community labels; the inter-block connection table has rank at most K, which is the number of communities.
- **Collaborative filtering.** User and item categories interact through a few latent "taste × style" factors, yielding a rank-K interaction matrix.
- **Latent-class contingency tables.** Cross-tabs such as eye color × hair color, education × political preference, or customer segment × product category are well modeled by mixtures of a few product tables with rank at most K).

These examples motivate treating P as (exactly or approximately) low rank and justify estimators that explicitly exploit this structure. In this section, we consider the problem of minimax estimation of P on the class $\mathcal{T}_K$ with respect to the norm $\|\cdot\|_1$. We derive the minimax optimal rate and propose a computationally efficient estimator achieving this rate, up to a logarithmic factor.

5.1 The localized SVD estimator

We start by formally describing the localized SVD algorithm. In the next subsection, we will provide some intuition regarding its construction and sketch the ideas of proving the upper bounds on its performance.

Without loss of generality, assume that the total number of observations is even and equal to $2n$. We use sample splitting to define $H^{(1)} \in \mathbb{R}^{d_1 \times d_2}$ and $H^{(2)} \in \mathbb{R}^{d_1 \times d_2}$ as the matrices of empirical frequencies (the histograms) corresponding to the sub-samples $(X_1, \dots, X_n)$ and $(X_{n+1}, \dots, X_{2n})$ respectively. In what follows, it will useful to express the matrices P, $H^{(1)}$ and $H^{(2)}$ in terms of their columns and rows:

$$P = \Big[C_1, \dots, C_{d_2}\Big] = \begin{bmatrix} L_1 \\ \vdots \\ L_{d_1} \end{bmatrix} \quad \text{and} \quad H^{(\ell)} = \Big[\widehat{C}_1^{(\ell)}, \dots, \widehat{C}_{d_2}^{(\ell)}\Big] = \begin{bmatrix} \widehat{L}_1^{(\ell)} \\ \vdots \\ \widehat{L}_{d_1}^{(\ell)} \end{bmatrix}, \text{ for } \ell = 1, 2.$$

We set $T = \lfloor \log_2 d \rfloor - 1$, where $d \geq 2$ is a suitably chosen parameter. In this section, we take $d = d_1 \vee d_2$. Other choices of d will be used when applying Algorithm 1 as a building block for estimation of continuous densities. For any $t \in \{0, \dots, T\}$ we define

$$I_t = \left\{ i \in [d_1] : \big\|\widehat{L}_i^{(1)}\big\|_1 \in \left(\frac{1}{2^{t+1}}, \frac{1}{2^t}\right] \right\}, \qquad J_t = \left\{ j \in [d_2] : \big\|\widehat{C}_j^{(1)}\big\|_1 \in \left(\frac{1}{2^{t+1}}, \frac{1}{2^t}\right] \right\} \quad (3)$$

and set

$$I_{T+1} = \left\{ i \in [d_1] : \big\|\widehat{L}_i^{(1)}\big\|_1 \leq \frac{1}{2^{T+1}} \right\}, \qquad J_{T+1} = \left\{ j \in [d_2] : \big\|\widehat{C}_j^{(1)}\big\|_1 \leq \frac{1}{2^{T+1}} \right\}. \quad (4)$$

Next, for any $k = (t, t') \in \{0, \dots, T+1\}^2$ we define

$$U_k = I_t \times J_{t'}, \quad (5)$$

$$M^{(k)} = \Big(H^{(2)}_{ij}\, \mathbb{1}_{\{(i,j)\in U_k\}}\Big)_{i,j}, \qquad P^{(k)} = \Big(P_{ij}\, \mathbb{1}_{\{(i,j)\in U_k\}}\Big)_{i,j}. \tag{6}$$

Algorithm 1: Estimation procedure

Input: $\alpha > 1$, $d \geq 2$, $N > 1$, integer $n \geq 1$, matrices $H^{(1)}, H^{(2)}$.

Output: Estimator $\widehat{P}^*$ of P.

If $n < 14\alpha d \log N$ **return** $\widehat{P}^* = \frac{1}{2}(H^{(1)} + H^{(2)})$

Else: $T \leftarrow \lfloor \log_2(d) \rfloor - 1$

For $t, t' = 0, \ldots, T+1$:

1. $k \leftarrow (t,t')$; $\tau_k \leftarrow 12\sqrt{\alpha \frac{\log N}{n} 2^{-t\wedge t'}}$ **and define** $M^{(k)}$ **as in** (6)
2. $\widehat{P}^{(k)} \leftarrow \underset{A\in\mathbb{R}^{d_1\times d_2}}{\operatorname{argmin}} \Big(\|M^{(k)} - A\|_F^2 + \tau_k \|A\|_*\Big)$
3. $\widehat{P} \leftarrow \sum_{k\in\{0,\ldots,T+1\}^2} \widehat{P}^{(k)}$
4. $\widehat{P}_+ \leftarrow \big(\widehat{P}_{ij} \vee 0\big)_{ij}$
5. **If** $\widehat{P}_+ = 0_{\mathbb{R}^{d_1\times d_2}}$ **return** $\widehat{P}^* = \frac{1}{2}(H^{(1)} + H^{(2)})$

Return $\widehat{P}^* = \frac{\widehat{P}_+}{\|\widehat{P}_+\|_1}$.

In Algorithm 1 above, the parameter N is a user-specified input of the algorithm. It controls the probability of deviations, see Theorem 4 below. In what follows, the notation Alg1$(\alpha, d, N, n, H^{(1)}, H^{(2)})$ stands for Algorithm 1 with input parameters $(\alpha, d, N, n, H^{(1)}, H^{(2)})$.

5.2 Intuition underlying the localized SVD algorithm

Standard approaches to estimate a low-rank matrix from noisy data consist in using methods based on global SVD on the underlying matrices. The main drawback of such methods is that they can be sub-optimal under non-isotropic noise. In particular, it is the case for the multinomial noise that we are dealing with in our setting since $nH^{(\ell)} \sim \mathcal{M}(P, n)$ for $\ell = 1, 2$. Any entry Y_{ij} of a multinomial matrix $Y \sim \mathcal{M}(P, n)$ has a binomial distribution, $Y_{ij} \sim \mathrm{Bin}(n, P_{ij})$, with variance $nP_{ij}(1 - P_{ij})$ that varies across the indices (i, j). To overcome this difficulty, Algorithm 1 splits the multinomial matrix into a logarithmic number of sub-matrices, on which the multinomial noise can be more carefully controlled. Then, each sub-matrix is de-noised separately using a nuclear norm penalized estimator. Recall that this estimator is based on soft thresholding of the singular values.

To appreciate why do we split the multinomial matrix as in equations (3) - (4), assume that $Y \sim \mathcal{M}(P, n)$ and for any two subsets $I \subset [d_1]$ and $J \subset [d_2]$, consider the extractions according to I and J:

$$Y_{IJ} = (Y_{ij})_{(i,j)\in I\times J}, \quad P_{IJ} = (P_{ij})_{(i,j)\in I\times J}, \quad \text{and} \quad W_{IJ} = \frac{Y_{IJ}}{n} - P_{IJ}.$$

By Lemma 23 the operator norm of W_{IJ} is controlled, with high probability and ignoring the logarithmic factors and smaller order terms, by a function of column-wise and row-wise sums of entries of P_{IJ}:

$$\|W_{IJ}\|^2 \lesssim \frac{1}{n}\big\|P_{IJ}\big\|_{\square} = \frac{1}{n}\max\Big(\max_{i\in I}\sum_{j\in J}P_{ij},\ \max_{j\in J}\sum_{i\in I}P_{ij}\Big) \tag{7}$$
$$\leq \frac{1}{n}\max\Big(\max_{i\in I}\|L_i\|_1,\ \max_{j\in J}\|C_j\|_1\Big).$$

This bound is accurate enough if we take "balanced" subsets I, J such that

$$\forall i \in I:\ \|L_i\|_1 \asymp \max_{i'\in I}\|L_{i'}\|_1 \qquad \text{and} \qquad \forall j \in J:\ \|C_j\|_1 \asymp \max_{j'\in J}\|C_{j'}\|_1,$$

that is, we split the multinomial matrix according to similar values of $\|L_i\|_1$'s and $\|C_j\|_1$'s in order for the multinomial noise to be almost isotropic over the considered sub-matrices.

Since the values $(\|L_1\|_1, \dots, \|L_{d_1}\|_1)$ and $(\|C_j\|_1, \dots, \|C_{d_2}\|_1)$ are unknown, we use sample splitting to obtain good enough estimators $(\|\hat{L}_1\|_1, \dots, \|\hat{L}_{d_1}\|_1$ and $(\|\hat{C}_j\|_1, \dots, \|\hat{C}_{d_2}\|_1)$ from the first half of the data. To simplify the argument, here we do not discuss these pilot estimators and assume that an oracle gives us the exact values $(\|L_1\|_1, \dots, \|L_{d_1}\|_1)$ and $(\|C_j\|_1, \dots, \|C_{d_2}\|_1)$.

Set $d = d_1 \vee d_2$, and for any $t \in \{0, \dots, \lfloor \log_2(d) - 1 \rfloor\}$, define the following index sets

$$\widetilde{I}_t = \left\{ i : \|L_i\|_1 \in \left(\frac{1}{2^{t+1}}, \frac{1}{2^t}\right] \right\}, \quad \widetilde{J}_t = \left\{ j : \|C_j\|_1 \in \left(\frac{1}{2^{t+1}}, \frac{1}{2^t}\right] \right\}. \tag{8}$$

For $T = \lfloor \log_2(d) \rfloor - 1$ define

$$\widetilde{I}_T = [d_1] \setminus \bigcup_{t<T} \widetilde{I}_t \text{ and } \widetilde{J}_T = [d_2] \setminus \bigcup_{t<T} \widetilde{J}_t.$$

Fix $t, t' \in \{0, \dots, T\}$ and set $\widetilde{M}_{t,t'} = \frac{1}{n} Y_{\widetilde{I}_t \widetilde{J}_{t'}}$ and $P_{t,t'} = P_{\widetilde{I}_t \widetilde{J}_{t'}}$. Then $\mathrm{rk}(P_{t,t'}) \leq K$ and $\big\|P_{t,t'}\big\|_{\square} \leq \frac{1}{2^{t\wedge t'}}$ by construction. The idea is now to take an estimator $\widehat{P}_{t,t'}$ of $P_{t,t'}$ obtained from $\widetilde{M}_{t,t'}$ by performing a soft-thresholding of its singular values, with a threshold based on Lemma 24 below, which states guarantees for SVD soft-thresholding (nuclear norm penalized) estimators, see also Giraud (2021). This strategy leads to the following bound on the error in the Frobenius norm, which holds with high probability:

$$\Big\|\widehat{P}_{t,t'} - P_{t,t'}\Big\|_F^2 \lesssim \frac{K}{n}\big\|P_{t,t'}\big\|_{\square} \lesssim \frac{K}{2^{t\wedge t'} n}.$$

Here and below, the sign $\lesssim$ hides a logarithmic factor. This implies the following bound

$$\Big\|\widehat{P}_{t,t'} - P_{t,t'}\Big\|_1 \lesssim \sqrt{\frac{K}{2^{t\wedge t'} n}|\widetilde{I}_t|\,|\widetilde{J}_{t'}|}. \tag{9}$$

Denoting by $\widehat{P}$ the $d_1 \times d_2$ matrix obtained by concatenation of all the cells $\tilde{I}_t \times \tilde{J}_{t'}$, and summing (9) over (t, t') we obtain

$$\|\widehat{P} - P\|_1 \lesssim \sum_{t,t'=0}^{T} \sqrt{\frac{K}{2^{t\wedge t'} n}|\widetilde{I}_t|\,|\widetilde{J}_{t'}|}.$$

By the Cauchy-Schwarz inequality and the fact that $T = \lfloor \log_2(d) \rfloor - 1$ this bound can be simplified as follows:

$$\begin{aligned}\sum_{t,t'=0}^{T} \sqrt{\frac{K}{2^{t\wedge t'}n}|\widetilde{I}_t|\,|\widetilde{J}_{t'}|} &\le \sqrt{\sum_{t,t'=0}^{T} \frac{K}{2^{t\wedge t'}n}|\widetilde{I}_t||\widetilde{J}_{t'}| \sum_{t,t'=0}^{T} 1} \\ &\le \log_2(d) \sqrt{\sum_{t,t'=0}^{T} \frac{K}{n}\left(\frac{1}{2^t}+\frac{1}{2^{t'}}\right)|\widetilde{I}_t|\,|\widetilde{J}_{t'}|} \\ &\lesssim \sqrt{\frac{K(d_1+d_2)}{n}},\end{aligned}$$

where the last inequality uses the relations

$$\sum_{t=0}^{T} |\widetilde{I}_t| = d_1 \quad \text{and} \quad \sum_{t=0}^{T} 2^{-t}|\widetilde{I}_t| = \sum_{t=0}^{T}\sum_{i\in\widetilde{I}_t} 2^{-t} \le \sum_{t=0}^{T}\sum_{i\in\widetilde{I}_t} 2p_i \le 2$$

(by the definition of $\widetilde{I}_t$), and the analogous relations for the family $(\widetilde{J}_t)_t$.

The above argument outlines a strategy for proving a bound of order $\sqrt{\frac{Kd}{n}}$ (up to a logarithmic factor) for the estimator defined by Algorithm 1. The exact statement of the result is given in Theorems 4 and 5 below.

5.3 Results for discrete distributions

We first provide minimax optimal rates for estimating a general discrete distribution on a finite set of size D. Without loss of generality, assume that this set is $\{1,\dots,D\}$. Let $\Delta_D = \left\{p \in \mathbb{R}_+^D : \sum_{j=1}^D p_j = 1\right\}$ be the set of all probability distributions on $\{1,\dots,D\}$. We denote by $\mathbb{P}_p$ the probability measure of n iid observations drawn from p, by $\mathbb{E}_p$ the expectation with respect to $\mathbb{P}_p$, and by $\inf_{\widehat{p}}$ the infimum over all $\mathbb{R}^D$-valued estimators.

Theorem 1 *Let $D \ge 1, n \ge 1$. There exist two absolute positive constants c, c' such that*

$$\inf_{\widehat{p}} \sup_{p\in\Delta_D} \mathbb{P}_p\left(\|\widehat{p}-p\|_1 \ge c\Big\{\sqrt{\frac{D}{n}} \wedge 1\Big\}\right) \ge c', \tag{10}$$

and

$$\inf_{\widehat{p}} \sup_{p\in\Delta_D} \mathbb{E}_p\|\widehat{p}-p\|_1 \ge c\Big\{\sqrt{\frac{D}{n}} \wedge 1\Big\}. \tag{11}$$

The lower bound for the expectation (11) improves upon the bounds in Han et al. (2015); Kamath et al. (2015) that provide the same lower rate $\sqrt{\frac{D}{n}}$ under some additional conditions on D, n. The lower bound in probability (10) is new.

As a corollary of Theorem 1, we get the minimax optimal rate of convergence for the problem of estimating general discrete distributions. Given $\delta > 0$ and a class of discrete

distributions $\mathcal{P} \subseteq \Delta_D$, we define the *minimax in probability rate* of estimation of $p \in \mathcal{P}$ based on an iid sample $(X_1, \dots, X_n)$ from p as

$$\psi^*_\delta(n, \mathcal{P}) = \inf\left\{\psi > 0 \;\middle|\; \inf_{\widehat{p}} \sup_{p\in\mathcal{P}} \mathbb{P}_p\Big(\big\|\widehat{p}(X_1, \dots, X_n) - p\big\|_1 > \psi\Big) \le \delta\right\}.$$

Corollary 2 *Let $D \ge 1, n \ge 1$. There exist two absolute positive constants c, c' such that, for all $\delta \in (0, c')$ we have*

$$c\Big\{\sqrt{\frac{D}{n}} \wedge 1\Big\} \le \psi^*_\delta(n, \Delta_D) \le c_\delta\Big\{\sqrt{\frac{D}{n}} \wedge 1\Big\}, \tag{12}$$

where $c_\delta > 0$ depends only on δ. Furthermore,

$$\inf_{\widehat{p}} \sup_{p\in\Delta_D} \mathbb{E}_p\|\widehat{p} - p\|_1 \asymp \sqrt{\frac{D}{n}} \wedge 1. \tag{13}$$

The proof of Corollary 2 follows immediately by combining Theorem 1 with the standard upper bound for the empirical frequency estimator (see, for example, Lemma 22 below).

Next, we obtain a lower bound for the class of discrete distributions $\mathcal{T}_K$ defined by probability matrices of rank at most K, see (2).

Theorem 3 (Lower bounds for $\mathcal{T}_K$) *Let n, K, d_1, d_2 be positive integers such that $K \le d_1 \wedge d_2$. Set $d = d_1 \vee d_2$. There exist two absolute positive constants c, c' such that*

$$\inf_{\widetilde{P}} \sup_{P\in\mathcal{T}_K} \mathbb{P}_P\left(\|\widetilde{P} - P\|_1 \ge c\Big\{\sqrt{\frac{Kd}{n}} \wedge 1\Big\}\right) \ge c', \tag{14}$$

and

$$\inf_{\widetilde{P}} \sup_{P\in\mathcal{T}_K} \mathbb{E}_P\|\widetilde{P} - P\|_1 \ge c\Big\{\sqrt{\frac{Kd}{n}} \wedge 1\Big\}. \tag{15}$$

The next two theorems give upper bounds matching the lower rate of Theorem 3 up to a logarithmic factor.

Theorem 4 (Upper bound for $\mathcal{T}_K$ in probability) *Let $\alpha > 1$, $d = d_1 \vee d_2$, $N \ge 2$, and let the estimator $\widehat{P}^*$ be obtained by* $\mathrm{Alg1}(\alpha, d, N, n, H^{(1)}, H^{(2)})$. *Then there exist constants $C_0, C_1 > 0$ depending only on α such that*

$$\sup_{P\in\mathcal{T}_K} \mathbb{P}_P\left(\|\widehat{P}^* - P\|_1 > C_1\Big\{\sqrt{\frac{Kd}{n}}\log(d)\,\log^{1/2}(N) \wedge 1\Big\}\right) \le C_0(\log d)^2 dN^{-\alpha}.$$

Note that Theorem 4 can be used for several meaningful choices of N, such as $N = n$, $N = d$ or $N = d \vee n$.

Theorem 5 (Upper bound for $\mathcal{T}_K$ in expectation) *Let $\alpha > 3/2$, $d = d_1 \vee d_2$, and let the estimator $\widehat{P}^*$ be obtained by* Alg1$(\alpha, d, d \vee n, H^{(1)}, H^{(2)})$. *Then there exists a constant $C > 0$ depending only on α such that*

$$\sup_{P \in \mathcal{T}_K} \mathbb{E}_P \|\widehat{P}^* - P\|_1 \leq C\Big\{\sqrt{\frac{Kd}{n}}(\log n)^{3/2} \wedge 1\Big\}.$$

Theorem 3, Theorem 4 with $N = d$ and Theorem 5 lead to the following corollary, which provides the minimax rate for the class $\mathcal{T}_K$.

Corollary 6 *Let $\gamma > 0$ and $\delta = O(d^{-\gamma})$. There exist two constants $c, C > 0$ depending only on γ such that*

$$c\Big\{\sqrt{\frac{Kd}{n}} \wedge 1\Big\} \leq \psi^*_\delta(n, \mathcal{T}_K) \leq C\Big\{\sqrt{\frac{Kd}{n}}(\log d)^{3/2} \wedge 1\Big\} \tag{16}$$

and

$$c\Big\{\sqrt{\frac{Kd}{n}} \wedge 1\Big\} \leq \inf_{\widetilde{P}} \sup_{P \in \mathcal{T}_K} \mathbb{E}_P \|\widetilde{P} - P\|_1 \leq C\Big\{\sqrt{\frac{Kd}{n}}(\log n)^{3/2} \wedge 1\Big\}. \tag{17}$$

If K is substantially smaller than $d_1 \wedge d_2$ the rate of convergence $\sqrt{\frac{Kd}{n}}$ provided by Theorems 4 and 5 is much faster than the estimation rate $\sqrt{\frac{D}{n}}$ for general $D = d_1 d_2$-dimensional discrete distributions. This characterizes the gain that is achieved due to the low-rank structure.

6. Continuous distributions

In this section, we use the ideas developed for discrete distributions in Section 5 to derive estimators of probability densities under the *generalized multi-view model* with known support $[0,1]^2$. The case of generalized multi-view model with unknown support and adaptation to the unknown parameters β and K is deferred to the Appendix.

We start by defining the class of considered densities. Let $L > 0$ be a constant. For any $\beta \in (0, 1]$, we say that $f : [0,1]^2 \to \mathbb{R}$ is a β-Hölder function if $|f(z) - f(z')| \leq L\|z - z'\|_\infty^\beta$ for all $z, z' \in [0,1]^2$. We denote by $\mathcal{L}^L_{\beta,[0,1]^2}$ the set of all β-Hölder densities supported on $[0,1]^2$:

$$\mathcal{L}^L_{\beta,[0,1]^2} = \Big\{ f : [0,1]^2 \to \mathbb{R} \,\Big|\, f \text{ is } \beta\text{-Hölder probability density supported on } [0,1]^2 \Big\} \tag{18}$$

For integer $K \geq 1$, we define $\mathcal{F}_K$ as the set of functions on $[0,1]^2$ that are sums of K separable functions:

$$\mathcal{F}_K = \Big\{ (x,y) \in [0,1]^2 \longmapsto \sum_{k=1}^K u_k(x) v_k(y) \in \mathbb{R} \;\Big|\; u_k, v_k \in L_1[0,1],\ \forall k \in [K] \Big\}.$$

We consider the following set of β-Hölder probability densities

$$\mathcal{G}^L_{K,\beta,[0,1]^2} := \mathcal{L}_{\beta,[0,1]^2} \cap \mathcal{F}_K.$$

If $f \in \mathcal{G}^L_{K,\beta,[0,1]^2}$ we will say that f follows the *generalized multi-view model* with known support $[0,1]^2$. We emphasize that the functions u_k and v_k appearing in the decomposition $f(x,y) = \sum_{k=1}^K u_k(x)v_k(y)$ may take negative values and need not be β-Hölder or continuous, as long as f is a β-Hölder probability density over $[0,1]^2$. It is clear that the set $\mathcal{G}^L_{K,\beta,[0,1]^2}$ contains all the β-Hölder densities on $[0,1]^2$ that can be expressed as mixtures of product densities.

Assume that for some integer $K \geq 1$ and some unknown $f \in \mathcal{G}^L_{K,\beta,[0,1]^2}$, we are given n iid observations $X_1, \dots, X_n$ distributed with probability density f. The goal is to estimate f based on $X_1, \dots, X_n$. The minimax rate for estimation of β-Hölder densities on $[0,1]^2$ under the L_1 risk is known to be $n^{-\beta/(2\beta+2)}$ and this rate is attained by KDE and other basic density estimators, see Devroye and Györfi (1985); Devroye and Lugosi (2001). We show that the minimax rate of convergence for the class $\mathcal{G}^L_{K,\beta,[0,1]^2}$ with a rank-K structure is faster than $n^{-\beta/(2\beta+2)}$ and we propose a computationally simple estimator that achieves this optimal rate up to a logarithmic factor.

Our estimator is defined in Algorithm 2. The main steps can be summarized as follows.

- We partition the domain $[0,1]^2$ into disjoint rectangular cells $(C_{j,j'})_{j,j'} := (A_j \times A_{j'})_{j,j'}$ with side length of order $h^* \asymp n^{-1/(2\beta+2)} \wedge (K/n)^{1/(2\beta+1)}$ in both dimensions, up to logarithmic factors. We define

$$A_j = \big[jh, (j+1)h\big) \qquad \text{where} \qquad h = \frac{1}{\lfloor 1/h^* \rfloor}.$$

- Dividing the sample into two equal parts, we construct two independent histogram matrices N and N' based on this partition of $[0,1]^2$ into $(C_{j,j'})_{j,j'}$. We apply Algorithm 1 with $H^{(1)} = N$ and $H^{(2)} = N'$, which outputs a matrix with entries corresponding to the cells of the partition. We define our density estimator as a function that takes a constant value in each cell, proportional to the output of Algorithm 1 in the cell.

In Algorithm 2 we assume without loss of generality that n is a multiple of 2.

In what follows we denote by $\mathbb{P}_f$ the probability measure induced by $(X_1, \dots, X_n)$ when X_i's are iid distributed with density f, and by $\mathbb{E}_f$ the corresponding expectation.

Theorem 7 *There exist constants $C_0' > 0$, $C_1' > 0$ depending only on α and L such that for the estimator $\widehat{f}$ defined by Algorithm 2 with $\alpha > 1$ we have*

$$\sup_{f \in \mathcal{G}^L_{K,\beta,[0,1]^2}} \mathbb{P}_f \left(\|\widehat{f} - f\|_{L_1} > C_1' \left\{ \left(\frac{K}{n}\right)^{\beta/(2\beta+1)} \log^{3/2} n \wedge n^{-\frac{\beta}{2\beta+2}} \right\} \right) \leq C_0' (\log n)^2 n^{1/(2\beta+1)-\alpha}, \tag{19}$$

and for the estimator $\widehat{f}$ defined by Algorithm 2 with $\alpha > 4/3$ we have

$$\sup_{f \in \mathcal{G}^L_{K,\beta,[0,1]^2}} \mathbb{E}_f \|\widehat{f} - f\|_{L_1} \leq C_1' \left\{ \left(\frac{K}{n}\right)^{\beta/(2\beta+1)} \log^{3/2} n \wedge n^{-\frac{\beta}{2\beta+2}} \right\}. \tag{20}$$

Algorithm 2: Two-dimensional density estimator

Input: $X_1, \dots, X_n \in \mathbb{R}^2$ with $n = 2k$ for an integer $k \geq 1$; $\alpha > 0$; $K \in \mathbb{N}$; $\beta \in (0, 1]$

If $\left(\frac{K}{n}\right)^{\frac{\beta}{2\beta+1}} \log^{3/2}(n) \leq n^{-\frac{\beta}{2\beta+2}}$: set $K' = K$ **else:** set $K' = n^{\frac{1}{2\beta+2}}$.

$h^* \leftarrow (K'/n)^{1/(2\beta+1)}$ and $h = \frac{1}{\lfloor 1/h^* \rfloor}$.

For $(j, j') \in \{0, \dots, \lfloor \frac{1}{h^*} \rfloor - 1\}$:

$C_{jj'} \leftarrow \big[jh, (j+1)h\big] \times \big[j'h, (j'+1)h\big]$.

$G_{jj'} \leftarrow \frac{2}{n} \sum_{1 \leq i \leq n/2} \mathbb{1}_{X_i \in C_{jj'}}$; $\quad G'_{jj'} \leftarrow \frac{2}{n} \sum_{n/2+1 \leq i \leq n} \mathbb{1}_{X_i \in C_{jj'}}$;

If $\left(\frac{K}{n}\right)^{\frac{\beta}{2\beta+1}} \log^{\frac{3}{2}}(n) > n^{-\frac{\beta}{2\beta+2}}$: set $\widehat{P}^* \leftarrow \frac{G+G'}{2}$ where $G = (G_{jj'})_{j,j'}$, $G' = (G'_{jj'})_{j,j'}$

Else: set $\widehat{P}^* \leftarrow \mathrm{Alg1}\Big(\alpha, \lfloor \frac{1}{h^*} \rfloor, n, \frac{n}{2}, G, G'\Big)$

$\widehat{\phi}(x, y) := \frac{1}{h^2} \sum_{j,j'} \widehat{P}^*_{jj'} \, \mathbb{1}_{(x,y) \in C_{jj'}}$

Return $\widehat{f} = \frac{\widehat{\phi}}{\|\widehat{\phi}\|_{L_1}}$

Theorem 7 guarantees that the estimator $\widehat{f}$ adapts to the best rate between $(K/n)^{\beta/(2\beta+1)}$, which is a "one-dimensional" rate as function of n but deteriorates as K grows, and $n^{-\beta/(2\beta+2)}$, which is the standard rate of estimating a β-Hölder two-dimensional density. This demonstrates a dimension reduction property. The elbow between the two rates occurs at $K \asymp n^{1/(2\beta+2)}$.

The lower bound below shows that the rate obtained in Theorem 7 cannot be improved up to a logarithmic factor. We derive even a stronger lower bound that holds for the subclass of $\mathcal{G}^L_{K,\beta,[0,1]^2}$ containing densities with support $[0, 1]^2$ that can be decomposed as *mixtures* of separable densities that are Hölder smooth over their support. Specifically, let $\mathcal{G}^{\circ L}_{K,\beta,[0,1]^2}$ be the set of all probability densities f with support $[0, 1]^2$ and such that

$$f(x, y) = \sum_{k=1}^{K} w_k \, u_k(x) v_k(y), \quad \forall \, (x, y) \in [0, 1]^2,$$

where u_k, v_k are are β-Hölder probability densities on $[0, 1]$ for all k, and $\sum_{k=1}^K w_k = 1, \; w_k \geq 0, \forall k$. Clearly, $\mathcal{G}^{\circ L}_{K,\beta,[0,1]^2} \subset \mathcal{G}^L_{K,\beta,[0,1]^2}$.

Theorem 8 *Let $L > 0$, $\beta \in (0, 1]$. There exist two positive constants c, c' that can depend only on L and β such that*

$$\inf_{\widetilde{f}} \sup_{f \in \mathcal{G}^{\circ L}_{K,\beta,[0,1]^2}} \mathbb{P}_f \left(\|\widetilde{f} - f\|_{L_1} \geq c \Big\{ (K/n)^{\beta/(2\beta+1)} \wedge n^{-\beta/(2\beta+2)} \Big\} \right) \geq c', \tag{21}$$

and

$$\inf_{\widetilde{f}} \sup_{f \in \mathcal{G}^{\circ L}_{K,\beta,[0,1]^2}} \mathbb{E}_f \|\widetilde{f} - f\|_{L_1} \geq c \Big\{ (K/n)^{\beta/(2\beta+1)} \wedge n^{-\beta/(2\beta+2)} \Big\}, \tag{22}$$

where $\inf_{\tilde{f}}$ *denotes the infimum over all density estimators.*

For $\delta > 0$, we define the minimax estimation rate on $\mathcal{G}^L_{K,\beta,[0,1]^2}$ as follows:

$$\overline{\psi}_\delta(n, \mathcal{G}^L_{K,\beta,[0,1]^2}) = \inf\left\{ t > 0 \;\middle|\; \inf_{\tilde{f}} \sup_{f \in \mathcal{G}^L_{K,\beta,[0,1]^2}} \mathbb{P}_f\Big(\big\| \tilde{f}(X_1, \dots, X_n) - f \big\|_{L_1} > t \Big) \le \delta \right\}. \quad (23)$$

Theorems 7 and 8 immediately imply the following corollary.

Corollary 9 *For any* $\gamma > 0$, $\beta \in (0, 1]$, $L > 0$, *there exist two constants* $c, C > 0$ *depending only on* γ, β *and* L *such that*

$$c\big\{(K/n)^{\beta/(2\beta+1)} \wedge n^{-\beta/(2\beta+2)}\big\} \le \overline{\psi}_{n^{-\gamma}}(n, \mathcal{G}^L_{K,\beta,[0,1]^2}) \le C\big\{(K/n)^{\beta/(2\beta+1)} (\log n)^{3/2} \wedge n^{-\beta/(2\beta+2)}\big\}.$$

Remark 10 (Adaptation) *We refer the reader to Appendix B for a procedure that adapts to unknown* $\beta \in [0, 1)$ *and* $K \in \mathbb{N}$.

7. Numerical experiments

We present the results of numerical experiments on synthetic data. We have performed simulations with different values of parameters d, n and the number of components K. For the experiments, we use the Python implementation of our algorithm[1].

Figures 1 and 2 present numerical experiments with discrete distributions. We compare the total variation error of our estimator with that of the classical histogram estimator. In Figure 1, we fix $K = 1$ and $n = 10^5$, and apply the two estimators on square matrices of size ranging from $d = 10$ to $d = 1600$. To better appreciate the dependency on d, we also represent the same experiment on a logarithmic scale in Figure 2. We can see that the total variation error of the histogram estimator is approximately proportional to d, whereas the error of our estimator is approximately proportional to $\sqrt{d}$ for the ranges of values represented in this figure.

Figure 3 presents the dependence of the total variation error on the rank K for fixed dimension $d = 100$ and fixed number of observations $n = 10^5$. We also provide in Figure 4 a representation of this error on a logarithmic scale, which shows that it grows as $\sqrt{K}$. These two figures are obtained for low-rank matrices close to the set used in the lower bound.

Finally, we provide simulations for the problem of density estimation (Figures 5 and 6). We compare the standard histogram density estimator with bin width $n^{-1/4}$ and our estimator defined by Algorithm 3 with $\beta = 1$, $K = 1$. We let n vary from 1000 to 10^6. Again, we observe that our estimator performs better than the classical estimator and allows us to recover the one-dimensional estimation rate $n^{-1/3}$.

8. Conclusion

In this paper, we obtained minimax near-optimal estimators for the problem of multinomial estimation under low-rank matrix constraints and for density estimation under the

1. The code of Lowrankdensity algorithm is available at https://github.com/hi-paris/Lowrankdensity

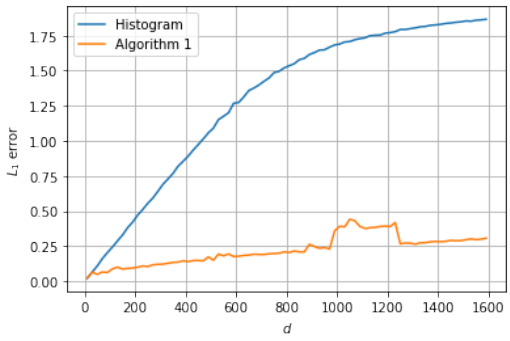

Figure 1: Total variation error of probability matrix estimators as a function of dimension d. Here, $n = 10^5$, $K = 1$.

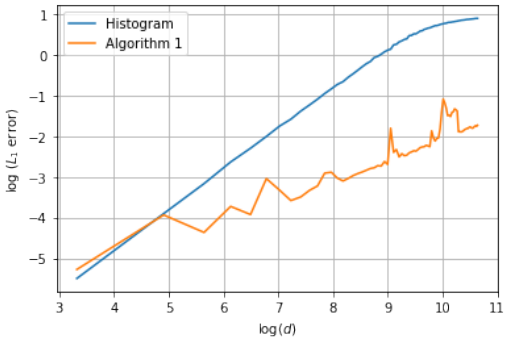

Figure 2: Total variation error of probability matrix estimators as a function of dimension d (on a log-log scale). Here, $n = 10^5$, $K = 1$.

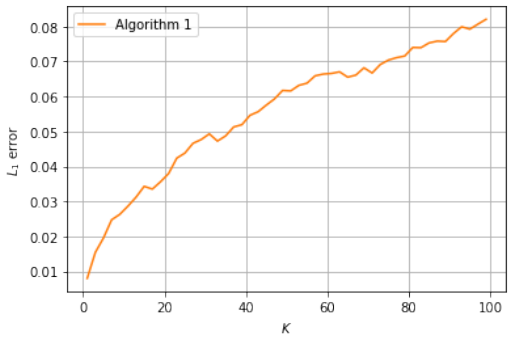

Figure 3: Total variation error of Algorithm 1 as a function of rank K. Here, $d = 100$ and $n = 10^5$.

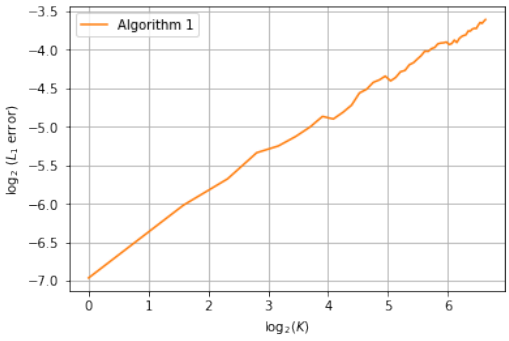

Figure 4: Total variation error of Algorithm 1 as a function of rank K (on a log-log scale). Here, $d = 100$, $n = 10^5$.

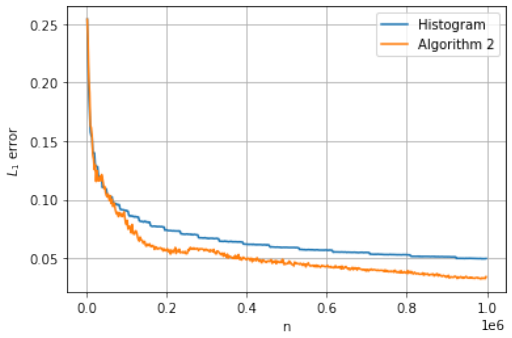

Figure 5: Total variation error of density estimators as a function of n for $K = 1$.

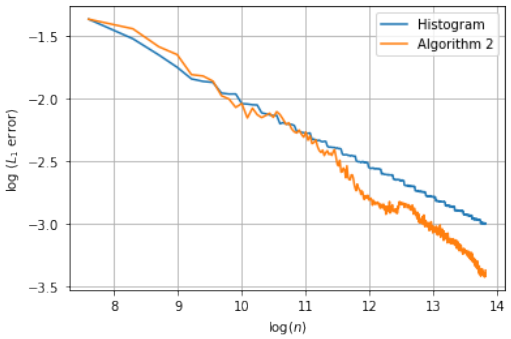

Figure 6: Total variation error of density estimators as a function of n for $K = 1$ (on a log-log scale).

generalized multi-view model. In both cases, we demonstrated that the rank constraint can substantially reduce the worst-case estimation error. Our findings suggest that low-rank matrix model can be a powerful tool for dimension reduction in the context of density estimation. It would be interesting to extend our results to larger dimensions, in particular, to low-rank multinomial tensors, which presumably needs developing different tools.

Acknowledgments

The work of Julien Chhor is supported by the Agence Nationale de la Recherche under the grant ANR-17-EURE-0010 (Investissements d'Avenir program). The work of Olga Klopp was funded by CY Initiative (grant "Investissements d'Avenir" ANR-16-IDEX-0008) and Labex MME-DII (ANR11-LBX-0023-01). The work of A.B. Tsybakov was supported by the grant of French National Research Agency (ANR) "Investissements d'Avenir" LabEx Ecodec/ANR-11-LABX-0047. The authors are grateful to Laurène David, Gaëtan Brison, and Shreshtha Shaurya from Hi!PARIS for their help in implementing the algorithms of this paper.

Appendix A. Continuous distributions with unknown support

In this section, we extend the results for probability distribution with support $[0,1]^2$ to the case of unknown support. More precisely, we consider the class of densities whose respective supports are unknown sub-rectangles of $[0,1]^2$ and assume no lower bound on their edge lengths. This framework is more challenging than the classical one (see Remark 14). To the best of our knowledge, it was not considered in the existing literature on nonparametric density estimation.

We start by defining the class of considered densities. Let $L > 0$ be a constant. For any $\beta \in (0,1]$, we say that $f : [0,1]^2 \to \mathbb{R}$ is a β-Hölder function if $|f(z) - f(z')| \leq L\|z - z'\|_\infty^\beta$ for all $z, z' \in \mathrm{Supp}(f)$. We denote by $\mathcal{L}_\beta^L$ the set of all β-Hölder densities supported on rectangles contained in $[0,1]^2$:

$$\mathcal{L}_\beta^L = \left\{ f : [0,1]^2 \to \mathbb{R} \,\middle|\, \exists\, r_1, r_2, R_1, R_2 \in [0,1] \text{ s.t. } \begin{cases} \mathrm{Supp}(f) = [r_1, R_1] \times [r_2, R_2], \\ f \text{ is } \beta\text{-Hölder on } \mathrm{Supp}(f), \\ \int_{\mathrm{Supp}(f)} f = 1 \text{ and } f \geq 0. \end{cases} \right\} \tag{24}$$

For integer $K \geq 1$, we define $\mathcal{F}_K$ as the set of functions on $[0,1]^2$ that are sums of K separable functions:

$$\mathcal{F}_K = \left\{ (x,y) \in [0,1]^2 \longmapsto \sum_{k=1}^K u_k(x) v_k(y) \in \mathbb{R} \;\middle|\; u_k, v_k \in L_1[0,1],\ \forall k \in [K] \right\}.$$

We consider the following set of β-Hölder probability densities

$$\mathcal{G}_{K,\beta}^L := \mathcal{L}_\beta \cap \mathcal{F}_K.$$

If $f \in \mathcal{G}_{K,\beta}^L$ we will say that f follows the *generalized multi-view model*. We emphasize that any function $f \in \mathcal{G}_{K,\beta}^L$ is only assumed to be β-Hölder on an unknown rectangle of the form $[r_1, R_1] \times [r_2, R_2]$ and not necessarily over the whole domain $[0,1]^2$. In particular, $f \in \mathcal{G}_{K,\beta}^L$ can have jumps at the boundary of its support $[r_1, R_1] \times [r_2, R_2]$. Moreover, the functions u_k and v_k appearing in the decomposition $f(x,y) = \sum_{k=1}^K u_k(x) v_k(y)$ may take negative values and need not be β-Hölder or continuous. Clearly, the set $\mathcal{G}_{K,\beta}^L$ contains all densities that are β-Hölder on their support and can be expressed as mixtures of product densities.

Since we consider density estimation under the L_1 risk it is not restrictive to assume that the support of the density is a compact set. Indeed, it has been highlighted in Ibragimov and Khas'minskii (1984), Juditsky and Lambert-Lacroix (2004), Goldenshluger and Lepski (2014), Chhor and Carpentier (2021) that there exist no *uniformly* consistent estimators with respect to the L_1 norm on classes of Hölder continuous densities with unbounded support. On such classes, the minimax L_1 rate is of trivial order 1 regardless of the number of observations. Note also that in our setting the support of f is an *unknown* set. It allows us to handle densities that are not necessarily Hölder continuous on the whole domain $[0,1]^2$. To the best of our knowledge, this setting was not explored in the prior work.

Assume that for some integer $K \geq 1$ and some unknown $f \in \mathcal{G}^L_{K,\beta}$, we are given n iid observations $X_1, \dots, X_n$ distributed with probability density f. The goal is to estimate f based on $X_1, \dots, X_n$. The minimax rate for estimation of β-Hölder densities on $[0,1]^2$ under the L_1 risk is known to be $n^{-\beta/(2\beta+2)}$ and this rate is attained by KDE and other basic density estimators Devroye and Györfi (1985); Devroye and Lugosi (2001). We show that the minimax rate of convergence for the class $\mathcal{G}^L_{K,\beta}$ with a rank-K structure is faster than $n^{-\beta/(2\beta+2)}$ and we propose a computationally simple estimator that achieves this optimal rate up to a logarithmic factor.

Our estimator is defined in Algorithm 3.
The main steps of Algorithm 3 can be summarized as follows.

- We divide the data into two subsamples of equal size. We estimate the support $[r_1, R_1] \times [r_2, R_2]$ by the smallest rectangle that contains all data points in the first subsample and denote this estimated support as $[\widehat{r}_1, \widehat{R}_1] \times [\widehat{r}_2, \widehat{R}_2]$. In fact, we have

$$\widehat{r}_m = \min\Big\{\Pi_m(X_i) : i \in \Big\{1, \dots, \frac{n}{2}\Big\}\Big\}, \qquad \forall m \in \{1,2\}$$
$$\widehat{R}_m = \max\Big\{\Pi_m(X_i) : i \in \Big\{1, \dots, \frac{n}{2}\Big\}\Big\}, \qquad \forall m \in \{1,2\}$$

  where for $X \in \mathbb{R}^2$, we denote by $\Pi_1(X)$ and $\Pi_2(X)$ its first and second coordinates, respectively: $X = (\Pi_1(X), \Pi_2(X))$.

- Assuming first that $\widehat{R}_1 - \widehat{r}_1 \geq h^*$ and $\widehat{R}_2 - \widehat{r}_2 \geq h^*$, we partition the domain $[0,1]^2$ into disjoint rectangular cells

$$(C_{j,j'})_{j,j'} := (A_j \times B_{j'})_{j,j'},$$

  so that the estimated support $[\widehat{r}_1, \widehat{R}_1] \times [\widehat{r}_2, \widehat{R}_2]$ is exactly covered by a finite subcollection of the cells $(C_{j,j'})$.

  More precisely, the cells have side lengths of order $h^* \asymp n^{-1/(2\beta+2)} \wedge (K/n)^{1/(2\beta+1)}$ in both dimensions, up to logarithmic factors. To this end, we construct a family of intervals $(A_j)_j$ covering $[0,1]$, with

$$A_j = \Big[\widehat{r}_1 + jh_1,\, \widehat{r}_1 + (j+1)h_1\Big), \qquad j \in E_1 := \Big\{ -\Big\lceil \tfrac{\widehat{r}_1}{h_1} \Big\rceil, \dots, \Big\lfloor \tfrac{1-\widehat{r}_1}{h_1} \Big\rfloor \Big\},$$

  where

$$h_m = \frac{\widehat{R}_m - \widehat{r}_m}{\lfloor (\widehat{R}_m - \widehat{r}_m)/h^* \rfloor} \asymp h^*.$$

  The family $(B_{j'})_{j'}$ is defined analogously. The index set E_1 is chosen so that $(A_j)_j$ covers $[0,1]$ entirely.

  Finally, the partition $(A_j)_j$ is aligned so that the estimated interval $[\widehat{r}_1, \widehat{R}_1]$ is exactly decomposed into some of the A_j's. In particular,

$$A_0 = [\widehat{r}_1, \widehat{r}_1 + h_1], \qquad A_{\lfloor (\widehat{R}_1 - \widehat{r}_1)/h_1 \rfloor - 1} = [\widehat{R}_1 - h_1, \widehat{R}_1],$$

  so the endpoints $\widehat{r}_1$ and $\widehat{R}_1$ coincide with endpoints of certain A_j's.

- Dividing the second subsample in two equal parts, we construct two independent histogram matrices N and N' based on this partition of $[0,1]^2$ into $(C_{j,j'})_{j,j'}$. We apply Algorithm 1 with $H^{(1)} = N$ and $H^{(2)} = N'$, which outputs a matrix with entries corresponding to the cells of the partition. We define our density estimator as a function that takes a constant value in each cell, proportional to the output of Algorithm 1 in the cell.

Note also that not knowing the support requires modifying this scheme in some degenerate situations. Indeed, the partition of the rectangle $[\widehat{r}_1, \widehat{R}_1] \times [\widehat{r}_2, \widehat{R}_2]$ degenerates if either $\widehat{R}_1 - \widehat{r}_1$ or $\widehat{R}_2 - \widehat{r}_2$ is smaller than the size of the cell h^*. If, for example, $\widehat{R}_1 - \widehat{r}_1$ is too small then, due to the β-Hölder property, f does not vary much in the first coordinate direction and we are lead to a one-dimensional density estimation problem over $[\widehat{r}_2, \widehat{R}_2]$, for which we use Algorithm 4.

In Algorithm 3 we assume without loss of generality that n is a multiple of 4. We denote by $\mathrm{Alg4}(Z_1, \dots, Z_n, K')$ the output of Algorithm 4 with input $(Z_1, \dots, Z_n, K')$.

Remark 11 (Choice of E_1 and E_2 in Algorithm 3) *In order to apply the results of Section 5, we need the union of cells $C_{jj'}$ in Algorithm 3 to be such that*

$$\sum_{j \in E_1} \sum_{j' \in E_2} G_{jj'} = \sum_{j \in E_1} \sum_{j' \in E_2} G'_{jj'} = 1. \tag{25}$$

This condition cannot be guaranteed if we only take cells that form a partition of $[\widehat{r}_1, \widehat{R}_1] \times [\widehat{r}_2, \widehat{R}_2]$ since some data points from the second subsample $\{X_{n/2+1}, \dots, X_n\}$ may fall outside of the estimated support $[\widehat{r}_1, \widehat{R}_1] \times [\widehat{r}_2, \widehat{R}_2]$. Taking the sets of indices E_1 and E_2 as in Algorithm 3 ensures that the union of cells $(C_{jj'})_{j \in E_1, j' \in E_2}$ contains the whole domain $[0,1]^2$ if $\widehat{R}_m - \widehat{r}_m > h^*$ for $m \in \{1,2\}$. In fact, under this condition, E_1 and E_2 are the sets of indices of smallest cardinality such that the union of $(C_{jj'})_{j \in E_1, j' \in E_2}$ contains $[0,1]^2$. Of course, some of these cells $C_{jj'}$ may fall beyond $[0,1]^2$. However, it does not affect the sums in (25) since $f = 0$ on such cells, so that the associated $G_{jj'}, G'_{jj'}$ vanish almost surely. Moreover, the order of magnitude of $|E_1|$ and $|E_2|$ remains controlled as needed. Indeed, we have $|E_1| \vee |E_2| \leq C/h^*$, which is sufficient for our purposes.

In what follows we denote by $\mathbb{P}_f$ the probability measure induced by $(X_1, \dots, X_n)$ when X_i's are iid distributed with density f, and by $\mathbb{E}_f$ the corresponding expectation.

Theorem 12 *There exist constants $C'_0 > 0$, $C'_1 > 0$ depending only on α and L such that for the estimator $\widehat{f}$ defined by Algorithm 3 with $\alpha > 1$ we have*

$$\sup_{f \in \mathcal{G}^L_{K,\beta}} \mathbb{P}_f \left(\|\widehat{f} - f\|_{L_1} > C'_1 \left\{ \left(\frac{K}{n} \right)^{\beta/(2\beta+1)} \log^{3/2} n \wedge n^{-\frac{\beta}{2\beta+2}} \right\} \right) \leq C'_0 (\log n)^2 n^{1/(2\beta+1)-\alpha}, \tag{26}$$

and for the estimator $\widehat{f}$ defined by Algorithm 3 with $\alpha > 4/3$ we have

$$\sup_{f \in \mathcal{G}^L_{K,\beta}} \mathbb{E}_f \|\widehat{f} - f\|_{L_1} \leq C'_1 \left\{ \left(\frac{K}{n} \right)^{\beta/(2\beta+1)} \log^{3/2} n \wedge n^{-\frac{\beta}{2\beta+2}} \right\}. \tag{27}$$

Algorithm 3: Two-dimensional density estimator

Input: $X_1, \dots, X_n \in \mathbb{R}^2$ with $n = 4k$ for an integer $k \geq 1$; $\alpha > 0$; $K \in \mathbb{N}$; $\beta \in (0, 1]$

For $m \in \{1, 2\}$, set $\widehat{r}_m \leftarrow \min\big\{\Pi_m(X_i) : i \in \{1, \dots, \frac{n}{2}\}\big\}$ and $\widehat{R}_m \leftarrow \max\big\{\Pi_m(X_i) : i \in \{1, \dots, \frac{n}{2}\}\big\}$.

If $\big(\frac{K}{n}\big)^{\frac{\beta}{2\beta+1}} \log^{3/2}(n) \leq n^{-\frac{\beta}{2\beta+2}}$: set $K' = K$ **else:** set $K' = n^{\frac{1}{2\beta+2}}$.

$h^* \leftarrow (K'/n)^{1/(2\beta+1)}$.

If $\widehat{R}_1 - \widehat{r}_1 < h^*$:

  $\widehat{g} \leftarrow \mathrm{Alg4}\big(\Pi_2(X_{n/2+1}), \dots, \Pi_2(X_n), K'\big)$

  $\widehat{\phi}(x, y) := \frac{1}{\widehat{R}_1 - \widehat{r}_1} \mathbb{1}_{x \in [\widehat{r}_1, \widehat{R}_1]}\, \widehat{g}(y)$.

Else If $\widehat{R}_2 - \widehat{r}_2 < h^*$:

  $\widehat{g} \leftarrow \mathrm{Alg4}\big(\Pi_1(X_{n/2+1}), \dots, \Pi_1(X_n), K'\big)$

  $\widehat{\phi}(x, y) := \frac{1}{\widehat{R}_2 - \widehat{r}_2} \mathbb{1}_{y \in [\widehat{r}_2, \widehat{R}_2]}\, \widehat{g}(x)$.

Else:

  For $m \in \{1, 2\}$, set $h_m = \ell_m^{-1}(\widehat{R}_m - \widehat{r}_m)$, where $\ell_m = \big\lfloor (\widehat{R}_m - \widehat{r}_m)/h^* \big\rfloor$

  For $m \in \{1, 2\}$, set $E_m = \big\{ -\big\lceil \widehat{r}_m/h_m \big\rceil, \dots, \big\lfloor (1 - \widehat{r}_m)/h_m \big\rfloor \big\}$

  For $(j, j') \in E_1 \times E_2$:

    $A_j \leftarrow \Big[\widehat{r}_1 + jh_1,\, \widehat{r}_1 + (j+1)h_1\Big)$.

    $B_{j'} \leftarrow \Big[\widehat{r}_2 + j'h_2,\, \widehat{r}_2 + (j'+1)h_2\Big)$.

    $C_{jj'} \leftarrow A_j \times B_{j'}$.

    $G_{jj'} \leftarrow \frac{4}{n} \sum_{n/2+1 \leq i \leq 3n/4} \mathbb{1}_{X_i \in C_{jj'}}$; $\quad G'_{jj'} \leftarrow \frac{4}{n} \sum_{3n/4+1 \leq i \leq n} \mathbb{1}_{X_i \in C_{jj'}}$;

  If $\big(\frac{K}{n}\big)^{\frac{\beta}{2\beta+1}} \log^{3/2}(n) > n^{-\frac{\beta}{2\beta+2}}$:

    $\widehat{P}^* \leftarrow (G + G')/2$, where $G = (G_{jj'})_{(j,j') \in E_1 \times E_2}$, $G' = (G'_{jj'})_{(j,j') \in E_1 \times E_2}$

  Else:

    $\widehat{P}^* \leftarrow \mathrm{Alg1}\Big(\alpha, |E_1| \vee |E_2|, n, \frac{n}{4}, G, G'\Big)$

$\widehat{\phi}(x, y) := \frac{1}{h_1 h_2} \sum_{j \in E_1} \sum_{j' \in E_2} \widehat{P}^*_{jj'}\, \mathbb{1}_{(x,y) \in C_{jj'}}$

If $\widehat{\phi} = 0$:

  Return $\widehat{f} = \mathbb{1}_{[0,1]^2}$

Return $\widehat{f} = \frac{\widehat{\phi}}{\|\widehat{\phi}\|_{L_1}}$

Algorithm 4: One-dimensional density estimator

Input: $Z_1, \dots, Z_n$ and $K' \geq 1$.

$\widehat{r} \leftarrow \min\big\{Z_i : i \in \{1, \dots, \lfloor \frac{n}{2} \rfloor\}\big\}$; $\widehat{R} \leftarrow \max\big\{Z_i : i \in \{1, \dots, \lfloor \frac{n}{2} \rfloor\}\big\}$.

$h^* \leftarrow (K'/n)^{1/(2\beta+1)}$.

If $\widehat{R} - \widehat{r} < h^*$, **then return** $\frac{1}{\widehat{R}-\widehat{r}} \mathbb{1}_{[\widehat{r},\widehat{R}]}$.

Else:

$h = \ell^{-1}(\widehat{R} - \widehat{r})$, where $\ell = \big\lfloor (\widehat{R} - \widehat{r})/h^* \big\rfloor$

$A_j = \big[\widehat{r} + (j-1)h,\ \widehat{r} + jh\big)$, for $j = 1, \dots, \ell - 1$, and $A_\ell = \big[\widehat{r} + (\ell - 1)h,\ \widehat{R}\big]$

For $j \in \big\{1, \dots, \ell\big\}$:

$N_j \leftarrow \sum_{i=\lfloor \frac{n}{2} \rfloor + 1}^{n} \mathbb{1}_{Z_i \in A_j}$

Return $\widehat{g} = \frac{1}{n - \lfloor \frac{n}{2} \rfloor} \sum_{j=1}^{\ell} \frac{1}{h} N_j \, \mathbb{1}_{A_j}$.[2]

Theorem 12 guarantees that the estimator $\widehat{f}$ adapts to the best rate between $(K/n)^{\beta/(2\beta+1)}$, which is a "one-dimensional" rate as function of n but deteriorates as K grows, and $n^{-\beta/(2\beta+2)}$, which is the standard rate of estimating a β-Hölder two-dimensional density. This demonstrates a dimension reduction property. The elbow between the two rates occurs at $K \asymp n^{1/(2\beta+2)}$. Note that, in our setting with unknown support of f, the possibility to estimate the density even with the slow rate $n^{-\beta/(2\beta+2)}$ does not follow from the results on nonparametric density estimation developed in the prior work (see Goldenshluger and Lepski (2014) and the references therein).

Remark 13 *The lower bound for the problem of estimation over the class $\mathcal{G}^L_{\beta,K}$ follows directly from Theorem 8.*

Remark 14 (Need for support estimation) *Recall that in the case of support $[0,1]^2$ we used Algorithm 2 that consists in partitioning the domain $[0,1]^2$ into equal-sided cells with side length $h \asymp n^{-1/(2\beta+2)} \wedge (K/n)^{1/(2\beta+1)}$ (to within logarithms) and then applying Algorithm 1 on the histogram of the data points generated by this binning. However, in the case of an unknown support, such a procedure does not lead to a reasonable outcome, whatever is the choice of h. To appreciate why, we can consider the following example. Let C_0 be a given cell of the binning and write $C_0 = [a_1, a_1 + h] \times [b_1, b_1 + h]$. Consider the unknown probability density f, which is constant on its support, and assume the support is $[a_1, a_1 + h] \times [b_1, b_1 + h/2]$. Note that the support of f is a strict subset of the cell C_0, and f takes on it the value $2/h^2$. If we use the binning described above, the resulting histogram will contain all the observations in the corresponding cell C_0 and zero observations in any other cell. The output of Algorithm 1 on this discretized data set will yield an estimator $\hat{f}$ that is equal to $1/h^2$ on the cell C_0 and zero outside. The L_1 distance between this estimator*

and the true density is constant: $\|\hat{f} - f\|_1 = 1$, which yields a worst-case risk that does not go to zero as $n \to \infty$. Note that this argument is valid for any h and not only for h defined above.

To address this difficulty, we first estimate the support, which allows us to adapt the domain binning in a more precise way. In the case above, our procedure would detect that, along the y axis, the domain is narrower than the side-length of a cell and call Algorithm 3 on the estimated support rather than Algorithm 1.

Appendix B. Adaptive density estimation

The density estimators proposed in Section A require knowledge of the number of separable components K and of the smoothness β. In this section, we provide an estimator that is adaptive to both K and β.

Throughout this section, we assume that $n \geq 3$. We consider a grid $\{\beta_1, \dots \beta_{\lceil \log(n) \log\log(n) \rceil}\}$ on the values of $\beta \in (0, 1]$, where

$$\beta_j = \Big(1 + \frac{1}{\log n}\Big)^{-j+1}.$$

Let $K_{\max} \geq 2$ be an integer. Set $m = K_{\max} \lceil \log(n) \log\log(n) \rceil$ and fix $\alpha > 4/3$. The adaptive estimator is obtained by the minimum distance choice from the family of m estimators $(\widehat{f}_{(K,j)})_{K \in [K_{\max}], 1 \leq j \leq \lceil \log(n) \log\log(n) \rceil}$, where $\widehat{f}_{(K,j)}$ is an output of Algorithm 3 with parameters K, $\beta = \beta_j$, α when the input sample is $X_1, \dots, X_n$. The minimum distance estimator (see, e.g., Devroye and Lugosi (2001)) is defined as follows:

$$\widehat{f}^* = \widehat{f}_{(K^*, j^*)} \text{ where } (K^*, j^*) \in \operatorname*{argmin}_{(K,j)} \max_{B \in \mathcal{B}} \Big| \int_B \widehat{f}_{(K,j)} - \mathbb{P}_n(B) \Big|, \tag{28}$$

and $\mathcal{B} = \{B_{ii'}, i, i' = 1, \dots, m, i \neq i'\}$ with $B_{ii'} = \{x : \widehat{f}_{i'}(x) > \widehat{f}_i(x)\}$ for $i, i' \in \{(K, j) : K \in [K_{\max}], 1 \leq j \leq \lceil \log(n) \log\log(n) \rceil\}$. For a set $B \in \mathcal{B}$, the notation $\mathbb{P}_n(B)$ stands for the empirical probability measure of this set computed from the sample $(X_1, \dots, X_n)$.

Note that each $\widehat{f}_{(K,j)}$ is a piece-wise constant function, so that the integrals in the definition of the adaptive estimator $\widehat{f}^*$ can be easily computed. To get $\widehat{f}^*$ we need $O(m^3)$ computations of such integrals Devroye and Lugosi (2001), where m is logarithmic in n.

In what follows, we set $K_{\max} = \lceil \sqrt{n} \rceil$, which is essentially the smallest sufficient choice. Indeed, $\sqrt{n} = \sup_{\beta \in (0,1]} n^{\frac{1}{2\beta+2}}$, while (as discussed in Section A) choosing K over the threshold $n^{\frac{1}{2\beta+2}}$ makes no sense since it does not bring any improvement compared to the standard two-dimensional rate $n^{-\frac{\beta}{2\beta+2}}$. The next theorem gives a bound on the L_1-risk of the adaptive estimator $\widehat{f}^*$.

Theorem 15 *There exists a constant $C > 0$ such that estimator $\widehat{f}^*$ defined in (28) with $K_{\max} = \lceil \sqrt{n} \rceil$ satisfies*

$$\sup_{f \in \mathcal{G}^L_{K,\beta}} \mathbb{E}_f \|\widehat{f}^* - f\|_{L_1} \leq C \left\{ (K/n)^{\beta/(2\beta+1)} \log^{3/2} n \wedge n^{-\frac{\beta}{2\beta+2}} \right\}, \quad \forall K \geq 1, \beta \in (0, 1].$$

Note that the result of Theorem 15 holds for all $K \geq 1$ while the selection leading to $\widehat{f}^*$ is made only over the estimators $\widehat{f}_{(K,j)}$ with $K \leq K_{\max} = \lceil \sqrt{n} \rceil$. This is because for $K > \lceil \sqrt{n} \rceil$ the estimators $\widehat{f}_{(K,j)}$ do not depend on K, cf. the definition of Algorithm 3, and they achieve the same rate as $\widehat{f}_{(\lceil \sqrt{n} \rceil, j)}$.

Appendix C. Proofs

C.1 Upper bounds for discrete distributions

Proof of Theorem 4.

Let $d = d_1 \vee d_2$, split the sample in two halves and form the histograms $H^{(1)}, H^{(2)}$ as in Algorithm 1. Write the target probability matrix P by rows $L_i^\top$ and columns C_j:

$$P = \big[\, C_1, \ldots, C_{d_2} \,\big] = \begin{bmatrix} L_1^\top \\ \vdots \\ L_{d_1}^\top \end{bmatrix}.$$

Step 0: It suffices to control the unnormalized estimator. Algorithm 1 first builds a matrix $\widehat{P}$ (by blockwise denoising) and then projects it to the simplex by truncation and renormalization:

$$\widehat{P}^* = \begin{cases} \frac{1}{2}\big(H^{(1)} + H^{(2)}\big), & \text{if all entries of } \widehat{P} \text{ are } \leq 0, \\ \dfrac{\widehat{P}_+}{\|\widehat{P}_+\|_1}, & \text{otherwise,} \end{cases} \qquad \widehat{P}_+ = \big(\widehat{P}_{ij} \vee 0\big)_{i,j}.$$

For any probability matrix P one has

$$\|P - \widehat{P}^*\|_1 \leq 2\, \|P - \widehat{P}\|_1. \tag{29}$$

Indeed, in the first branch $\|P - \widehat{P}^*\|_1 \leq 2$ while $\|P - \widehat{P}\|_1 \geq \|P\|_1 = 1$; in the second branch $\|P - \widehat{P}^*\|_1 \leq \|P - \widehat{P}_+\|_1 + \big\|\widehat{P}_+ - \widehat{P}^*\big\|_1 \leq \|P - \widehat{P}\|_1 + \big|\|\widehat{P}_+\|_1 - 1\big| \leq 2\|P - \widehat{P}\|_1$. Hence it is enough to prove the theorem for $\widehat{P}$.

Step 1: Small-n case. If $n < 14\alpha\, d \log N$, Algorithm 1 returns $\widehat{P}^* = \frac{1}{2}(H^{(1)} + H^{(2)})$, so $\|P - \widehat{P}^*\|_1 \leq 2$. This already matches the stated bound (for a larger constant), so we henceforth assume

$$n \geq 14\alpha\, d \log N.$$

Step 2: Data–driven partition of rows/columns and a high–probability event. Let $T = \lfloor \log_2 d \rfloor - 1$ and define row/column buckets from $H^{(1)}$:

$$I_t = \Big\{ i : \ \|L_i^{(1)}\|_1 \in (2^{-(t+1)}, 2^{-t}] \Big\}, \qquad J_{t'} = \Big\{ j : \ \|C_j^{(1)}\|_1 \in (2^{-(t'+1)}, 2^{-t'}] \Big\},$$

for $t, t' \in \{0, \ldots, T\}$, with the "remainder" buckets $I_{T+1} = \{i : \|L_i^{(1)}\|_1 \leq 2^{-(T+1)}\}$ and $J_{T+1} = \{j : \|C_j^{(1)}\|_1 \leq 2^{-(T+1)}\}$. For $k = (t, t')$ set $U_k = I_t \times J_{t'}$ and let

$$M^{(k)} = \big(H_{ij}^{(2)} \mathbf{1}_{(i,j) \in U_k}\big)_{i,j}, \qquad P^{(k)} = \big(P_{ij} \mathbf{1}_{(i,j) \in U_k}\big)_{i,j}.$$

We will work on the event

$$\mathcal{A} = \Big\{ \ \|L_i^{(1)}\|_1 \geq \tfrac{1}{4}\|L_i\|_1 \ \text{ and } \ \|C_j^{(1)}\|_1 \geq \tfrac{1}{4}\|C_j\|_1 \quad \text{whenever} \quad \|L_i\|_1 \vee \|C_j\|_1 \geq 14\alpha\tfrac{\log N}{n} \ \Big\}.$$

Let us evaluate the probability $\mathbb{P}(\mathcal{A})$. We will use the following lemma proved in Section D.

Lemma 16 *Let $Y \sim \mathcal{M}(P, n)$ be a multinomial $d_1 \times d_2$ random matrix. To any column $j \in [d_2]$, we associate a subset of row indices $V_j \subseteq [d_1]$ and a random variable $Z_j = \sum_{i \in V_j} Y_{ij}$. We define $\lambda_j = \sum_{i \in V_j} P_{ij} = \frac{1}{n}\mathbb{E}(Z_j)$. Let $\alpha > 0$, $N > 1$ be such that $n > 14\alpha \log(N)$. If $\lambda_j \in \big[\frac{14\alpha \log(N)}{n}, 1\big)$ for any $j \in J$, where $J \subseteq [d_2]$, then*

$$\mathbb{P}\left(\forall j \in J : \frac{Z_j}{n} \geq \frac{\lambda_j}{4}\right) \geq 1 - \frac{|J|+1}{N^\alpha}.$$

By applying Lemma 16 with $V_j = [d_1]$ for all j, $Z_j = n\|\widehat{C}_j^{(1)}\|_1$, $\lambda_j = \|C_j\|_1 = \sum_{i=1}^{d_1} P_{ij}$, and $J = \Big\{ j \in [d_2] : \|C_j\|_1 \geq 14\,\alpha\frac{\log N}{n} \Big\}$ we obtain

$$\mathbb{P}\left(\|\widehat{C}_j^{(1)}\|_1 \geq \frac{1}{4}\|C_j\|_1 \ \text{ for all } j \in [d_2] \ \text{ such that } \|C_j\|_1 \geq 14\,\alpha\frac{\log N}{n}\right) \geq 1 - \frac{d_2+1}{N^\alpha}.$$

Quite analogously,

$$\mathbb{P}\left(\|\widehat{L}_i^{(1)}\|_1 \geq \frac{1}{4}\|L_i\|_1 \ \text{ for all } i \in [d_1] \ \text{ such that } \|L_i\|_1 \geq 14\,\alpha\frac{\log N}{n}\right) \geq 1 - \frac{d_1+1}{N^\alpha}.$$

Therefore, since $d = d_1 \vee d_2 \geq 2$,

$$\mathbb{P}(\mathcal{A}) \geq 1 - \frac{d_1 + d_2 + 2}{N^\alpha} \geq 1 - 3\,dN^{-\alpha}. \tag{30}$$

Step 3: Noise level on each block and choice of the penalty. Let $W^{(k)} = M^{(k)} - P^{(k)}$. On $\mathcal{A}$,

$$\|P^{(k)}\|_\square := \max\Big\{ \ \max_{i \in I_t} \sum_{j \in J_{t'}} P_{ij} \ , \ \max_{j \in J_{t'}} \sum_{i \in I_t} P_{ij} \ \Big\} \ \leq \ 2^{\,2-(t \wedge t')}. \tag{31}$$

Applying Lemma 19 to the extraction on U_k gives, with probability at least $1 - 2d\,N^{-\alpha}$ (for fixed $H^{(1)}$),

$$\|W^{(k)}\| \ \leq \ c_1\Big(\sqrt{\tfrac{\alpha \log N}{n}\ \|P^{(k)}\|_\square} \ \vee \ \tfrac{\alpha \log N}{n}\Big) \ \leq \ c_2\sqrt{\tfrac{\alpha \log N}{n}}\ 2^{\,1-\frac{t \wedge t'}{2}}, \tag{32}$$

where the last inequality uses (31) together with $n \geq 14\alpha d \log N$ to let the square–root term dominate (constants $c_1, c_2 > 0$ depend only on α). Choose the blockwise penalty

$$\tau_k = c_3\sqrt{\tfrac{\alpha \log N}{n}}\ 2^{\,1-\frac{t \wedge t'}{2}}$$

with a large enough numerical constant c_3 so that $\|W^{(k)}\| \leq \tau_k/2$.

Step 4: Blockwise denoising error. On U_k, Algorithm 1 computes the nuclear–norm penalized estimator

$$\widehat{P}^{(k)} \in \arg\min_{A} \ \big\|M^{(k)} - A\big\|_F^2 + \tau_k \|A\|_*.$$

Since $\mathrm{rk}(P^{(k)}) \le K$ and $\|W^{(k)}\| \le \tau_k/2$, Lemma 20 yields

$$\big\|\widehat{P}^{(k)} - P^{(k)}\big\|_F^2 \ \le \ c_4\, K\, \tau_k^2 \ \le \ c_5\, \frac{K\,\alpha \log N}{n}\ 2^{\,2-(t\wedge t')}. \tag{33}$$

Moreover, by Cauchy–Schwarz and the block size bound $|U_k| \le (2^{t+1}\wedge d)\,(2^{t'+1}\wedge d)$ (because the entries of $H^{(1)}$ are nonnegative and sum to 1),

$$\begin{aligned}\big\|\widehat{P}^{(k)} - P^{(k)}\big\|_1 &\le \sqrt{|U_k|}\ \big\|\widehat{P}^{(k)} - P^{(k)}\big\|_F \\ &\le c_6 \sqrt{\frac{K\,\alpha\log N}{n}}\ 2^{\,1-\frac{t\wedge t'}{2}}\ \sqrt{(2^{t+1}\wedge d)(2^{t'+1}\wedge d)} \ \le \ c_7 \sqrt{\frac{K\,\alpha\log N}{n}}\ (2^{\frac{t\vee t'}{2}} \wedge \sqrt{d}).\end{aligned} \tag{34}$$

Step 5: Summing the blocks. Set $\widehat{P} = \sum_k \widehat{P}^{(k)}$. Using (34) and a union bound over the $(T+2)^2 \lesssim (\log d)^2$ blocks,

$$\|\widehat{P} - P\|_1 \ \le \ \sum_{t,t'=0}^{T+1} \big\|\widehat{P}^{(t,t')} - P^{(t,t')}\big\|_1 \ \le \ c_7 \sqrt{\tfrac{K\,\alpha\log N}{n}} \sum_{t,t'=0}^{T+1} (2^{\frac{t\vee t'}{2}} \wedge \sqrt{d}). \tag{35}$$

In view of (30), this bound holds with probability over the joint distribution of $(H^{(1)}, H^{(2)})$, which is at least $1 - (2(T+2)^2+3)dN^{-\alpha} \ge 1 - C_0(\log d)^2 dN^{-\alpha}$ for an absolute constant $C_0 > 0$. Note that

$$\begin{aligned}\sum_{(t,t')\in\{0,\dots,T+1\}^2} 2^{(t\vee t')/2} &\le 2 \sum_{(t,t')\in\{0,\dots,T+1\}^2 : t\ge t'} 2^{t/2} = 2\sum_{t=0}^{T+1} (t+1)\, 2^{t/2} \\ &\le 2\int_0^{T+2} (x+1)\, 2^{x/2} dx = 2\left[\frac{2}{\ln 2}\, x\, 2^{x/2} - \Big(\frac{2}{\ln 2}\Big)^2 2^{x/2} + \frac{2}{\ln 2}\, 2^{x/2}\right]_0^{T+2} \\ &= 2^{(T+2)/2}\left(\frac{4}{\ln 2}\,(T+3) - \frac{8}{(\ln 2)^2}\right) + \frac{8}{(\ln 2)^2} - \frac{4}{\ln 2}.\end{aligned}$$

Using $T = \lfloor \log_2 d \rfloor - 1$ gives $2^{(T+2)/2} \le \sqrt{2d}$ and $T + 3 \le \log_2 d + 2$, hence

$$\sum_{t,t'=0}^{T+1} 2^{(t\vee t')/2} \le \frac{4\sqrt{2}}{\ln 2}\,\sqrt{d}\,(\log_2 d + 2) + \Big(\frac{8}{(\ln 2)^2} - \frac{4}{\ln 2}\Big) \le \frac{C\sqrt{d}\log_2(d)}{\log 2}.$$

It follows that $\|\widehat{P} - P\|_1 \le C_1 \sqrt{\frac{Kd}{n}}\log(d)\log^{1/2}(N)$ with probability at least $1 - C_0(\log d)^2 dN^{-\alpha}$, where $C_0 > 0$ is an absolute constant and $C_1 > 0$ depends only on α.

Step 6: Return to the simplex. Finally, (29) transfers the same bound (up to a factor 2) to $\|\widehat{P}^* - P\|_1$, completing the proof. ■

Proof of Theorem 5.

We follow the same argument as in the proof of Theorem 4 with the only difference that now we set $N = d \vee n$ in Lemmas 16 and 23. This leads to the bound

$$\|\widehat{P}^* - P\|_1 \leq C_1 \sqrt{\frac{Kd}{n}} \log^{3/2}(d \vee n),$$

which holds with probability at least $1 - C_0(\log(d \vee n))^2 d(d \vee n)^{-\alpha}$. Therefore, since $\|\widehat{P}^* - P\|_1 \leq 2$ and $\alpha > 3/2$ we have

$$\mathbb{E}_P \|\widehat{P}^* - P\|_1 \leq C_1 \sqrt{\frac{Kd}{n}} \log^{3/2}(d \vee n) + 2C_0(\log(d \vee n))^2 n^{1-\alpha} \leq C \sqrt{\frac{Kd}{n}} \log^{3/2}(d \vee n)$$

for some constant $C > 0$ depending only on α. Noticing that $\sqrt{\frac{Kd}{n}} \log^{3/2}(d \vee n) \wedge 1 = \sqrt{\frac{Kd}{n}} \log^{3/2}(n) \wedge 1$ for $d, n \geq 2$ completes the proof. ■

C.2 Lower bounds for discrete distributions

The aim of this subsection is to prove Theorems 1 and 3. Note that it suffices to prove Theorem 3. Indeed, Theorem 1 is obtained as a corollary of Theorem 3 by taking $K = d_1 = 1$ and $d_2 = D$.

Proof of Theorem 3. We first prove the lower bound in expectation (15) and then combine it with Lemma 21 (see Section D) to deduce the bound in probability (14).

Proof of (15).

Note first that the result is trivial if $d_1 = d_2 = 1$. Therefore, assume that $d_2 \geq 2$ and, without loss of generality, $d_2 \geq d_1$. Set $D_2 = \lfloor d_2/2 \rfloor$ and $D = 2KD_2$. Define $\gamma = \left(\frac{1}{4\sqrt{nD}} \wedge \frac{1}{2D} \right)$. For any $\epsilon = (\epsilon_{ij}) \in \{-1, 1\}^{K \times D_2}$, define a $d_1 \times d_2$ matrix P_ϵ with the following entries:

$$\forall (i,j) \in \big[d_1\big] \times \big[d_2\big] : \; P_\epsilon(i,j) = \begin{cases} \frac{1}{D} + \epsilon_{ij}\gamma & \text{if } i \leq K \text{ and } j \leq D_2, \\ \frac{1}{D} - \epsilon_{i(j-D_2)}\gamma & \text{if } i \leq K \text{ and } D_2 < j \leq 2D_2, \\ 0 & \text{otherwise.} \end{cases} \tag{36}$$

Consider the set of $d_1 \times d_2$ matrices

$$\mathcal{P} := \left\{ P_\epsilon \;\middle|\; \epsilon \in \{-1, 1\}^{K \times D_2} \right\}. \tag{37}$$

This set consists of 2^m matrices, where $m = KD_2$. Note also that $\mathcal{P} \subset \mathcal{T}_K$. Indeed, all matrices $P \in \mathcal{P}$ are of rank K and have non-negative entries summing up to 1. For any $\epsilon, \epsilon' \in \{-1, 1\}^{K \times D_2}$ we have

$$\|P_\epsilon - P_{\epsilon'}\|_1 = 2\gamma\rho(\epsilon, \epsilon'), \tag{38}$$

where $\rho(\epsilon,\epsilon') = \sum_{i=1}^{K}\sum_{j=1}^{D_2} \mathbb{1}_{\epsilon_{i,j}\neq\epsilon'_{i,j}}$ denotes the Hamming distance between ϵ and ϵ'.

We now apply Assouad's lemma (see Theorem 2.12(iv) in Tsybakov (2009)). Let $\epsilon,\epsilon' \in \{-1,1\}^{K\times D_2}$ be such that $\rho(\epsilon,\epsilon')=1$. Denote by (i_0,j_0), where $i_0\in[K]$ and $j_0\in[D_2]$, the unique pair of indices such that $\epsilon_{i_0,j_0} = -\epsilon'_{i_0,j_0}$. Then the χ^2-divergence between P_ϵ and $P_{\epsilon'}$ satisfies

$$\begin{aligned}
\chi^2\left(P_\epsilon, P_{\epsilon'}\right) &= \sum_{i=1}^{K}\sum_{j=1}^{2D_2} \frac{\big(P_\epsilon(i,j) - P_{\epsilon'}(i,j)\big)^2}{P_{\epsilon'}(i,j)} \\
&= \frac{\big(P_\epsilon(i_0,j_0) - P_{\epsilon'}(i_0,j_0)\big)^2}{P_{\epsilon'}(i_0,j_0)} + \frac{\big(P_\epsilon(i_0,j_0+D_2) - P_{\epsilon'}(i_0,j_0+D_2)\big)^2}{P_{\epsilon'}(i_0,j_0+D_2)} \\
&\le \frac{8\gamma^2}{\frac{1}{D}-\gamma} \le 16\gamma^2 D \quad \Big(\text{since } \gamma \le \frac{1}{2D}\Big) \\
&= \frac{1}{n}\wedge\frac{4}{D} \le \frac{1}{n}.
\end{aligned}$$

The χ^2-divergence between the corresponding product measures satisfies, cf. (Tsybakov, 2009, page 86),

$$\chi^2\left(P_\epsilon^{\otimes n}, P_{\epsilon'}^{\otimes n}\right) = \Big(1+\chi^2\left(P_\epsilon, P_{\epsilon'}\right)\Big)^n - 1 \le e-1 \tag{39}$$

for all $\epsilon,\epsilon' \in \{-1,1\}^{K\times D_2}$ such that $\rho(\epsilon,\epsilon')=1$. Taking into account (38), (39), and applying (Tsybakov, 2009, Theorem 2.12(iv)) we obtain

$$\begin{aligned}
\inf_{\widetilde{P}}\max_{P\in\mathcal{P}} \mathbb{E}_P\|\widetilde{P}-P\|_1 &\ge \frac{\gamma m}{4}\exp(1-e) = \frac{D}{8}\exp(1-e)\left(\frac{1}{4\sqrt{nD}}\wedge\frac{1}{2D}\right) \\
&\ge \frac{\exp(1-e)}{32}\Big\{\sqrt{\frac{D}{n}}\wedge 1\Big\} \ge c\Big\{\sqrt{\frac{Kd_2}{n}}\wedge 1\Big\},
\end{aligned} \tag{40}$$

where $c>0$ is an absolute constant. This proves (15). ■

Proof of (14).

We apply Lemma 21, where we take $\mathcal{P}_0$ as the set of all $d_1\times d_2$ matrices, $\mathcal{P}$ as the set of matrices defined in (37), and we consider the metric $v(P,P') = \|P-P'\|_1$. Notice that, under these definitions, assumption (61) of Lemma 21 is satisfied with $U = \Big[\frac{1}{D}\,\mathbb{1}_{i\le K, j\le 2D_2}\Big]_{i,j}$ and $s = \gamma D = \Big(\frac{1}{4}\sqrt{\frac{D}{n}}\wedge\frac{1}{2}\Big)$, where D, D_2 are defined in the proof of (15). Moreover, due to (40) there exists an absolute constant $a>0$ such that assumption (62) of Lemma 21 is satisfied with the same s. Thus, we can apply Lemma 21, which yields the desired lower bound in probability. ■

C.3 Upper bounds for continuous distributions

Proof of Theorem 7.

Recall that in Algorithm 2 we assume that n is a multiple of 2.

Since $f \in \mathcal{G}_{K,\beta}$ we have the representation $f(x,y) = \sum_{k=1}^{K} u_k(x) v_k(y)$ with some functions $u_k \in L_1[0,1]$, $v_k \in L_1[0,1]$ for $k \in [K]$. Recall the definitions of the cells $C_{ij} = A_i \times A_j$, for any $i, j \in \{0, \dots, \lfloor 1/h^* \rfloor - 1\}$, where $A_i = [ih, (i+1)h]$ and $h = \lfloor 1/h^* \rfloor^{-1}$. Introduce the matrix $P = (P_{ij})_{i,j}$ with entries

$$\begin{aligned} P_{ij} &= \int_{C_{ij}} f(x,y) dx dy \\ &= \sum_{k=1}^{K} \int_{A_i} u_k(x) dx \int_{A_j} v_k(y) dy \\ &= \sum_{k=1}^{K} U_k(i) V_k(j), \qquad i, j \in \{0, \dots, \lfloor 1/h^* \rfloor - 1\}, \end{aligned}$$

where $U_k(i) = \int_{A_i} u_k(x) dx$ and $V_k(j) = \int_{A_j} v_k(y) dy$ for any $i, j \in \{0, \dots, \lfloor 1/h^* \rfloor - 1\}$ and $k \in [K]$. Set $U_k = (U_k(i))_{i \in E_1}$, $V_k = (V_k(j))_{j \in E_2}$. Then we can write

$$P = \sum_{k=1}^{K} U_k V_k^\top.$$

The matrix P has rank at most K. Consider now the histogram matrices $G = (G_{ij})_{i,j}$ and $G' = (G'_{ij})_{i,j}$ defined in Algorithm 2 with entries

$$G_{ij} = \frac{2}{n} \sum_{\ell=1}^{n/2} \mathbb{1}_{X_\ell \in C_{ij}} \quad \text{and} \quad G'_{ij} = \frac{2}{n} \sum_{\ell=n/2+1}^{n} \mathbb{1}_{X_\ell \in C_{ij}}, \qquad i, j \in \{0, \dots, \lfloor 1/h^* \rfloor - 1\}.$$

The matrices G and G' are mutually independent, and both $nG/2$ and $nG'/2$ follow the multinomial distribution $\mathcal{M}(P, n/4)$.

To alleviate the notation, we define the following two quantities

$$\psi_{\text{low-rank}} = (K/n)^{\frac{\beta}{2\beta+1}} \log^{3/2}(n) \quad \text{and} \quad \psi_{2D} = n^{-\frac{\beta}{2\beta+2}}.$$

The estimator $\widehat{\phi}$ in Algorithm 2 has the form

$$\widehat{\phi}(x,y) = \frac{1}{h^2} \sum_{j,j'} \widehat{P}^*_{jj'} \, \mathbb{1}_{(x,y) \in C_{jj'}}.$$

By the definition of Algorithm 2, the matrix $\widehat{P}^*$ is the output of $\text{Alg1}(\alpha, b, n, \frac{n}{2}, G, G')$ with $b = \lfloor 1/h^* \rfloor$, and $\alpha > 1$ if $\psi_{\text{low-rank}} \le \psi_{2D}$ and $\widehat{P}^* = (G+G')/2$ if $\psi_{\text{low-rank}} > \psi_{2D}$. Therefore, if $\psi_{\text{low-rank}} \le \psi_{2D}$, then Theorem 4 implies that, for some constants $C > 0$ depending only on α,

$$\|\widehat{P}^* - P\|_1 \le C \sqrt{\frac{Kb}{n}} \log(b) \log^{1/2}(n) \le C \left(\frac{K}{n}\right)^{\beta/(2\beta+1)} (\log n)^{3/2}$$

with probability at least $1 - C_0 (\log b)^2 b n^{-\alpha}$, where $C_0 > 0$ depends only on α.

Now, if $\psi_{\text{low-rank}} > \psi_{2D}$, then the relation

$$b := \Big\lfloor \frac{1}{h^*} \Big\rfloor \le C(n/K')^{1/(2\beta+1)}, \tag{41}$$

implies that $b \le Cn^{\frac{1}{2\beta+2}}$. Using this fact and Lemma 22 we obtain that, conditionally on $\mathcal{D}_1 = \{X_1, \dots, X_{n/2}\}$,

$$\mathbb{P}\bigg(\|\widehat{P}^* - P\|_1 > Cn^{\frac{\beta}{2\beta+2}} \;\Big|\; \mathcal{D}_1 \bigg) \le \mathbb{P}\bigg(\|\widehat{P}^* - P\|_1 > \sqrt{\frac{b^2}{n/2}} + \sqrt{\frac{2\alpha\log(n)}{n/2}} \;\Big|\; \mathcal{D}_1 \bigg) \le n^{-\alpha},$$

where the constant $C > 0$ depends only on α.

Combining the cases $\psi_{\text{low-rank}} \le \psi_{2D}$ and $\psi_{\text{low-rank}} > \psi_{2D}$ and using the fact that $\|\widehat{P}^* - P\|_1 \le 2$ we obtain the bound

$$\mathbb{P}\bigg(\|\widehat{P}^* - P\|_1 > C\big(\psi_{\text{low-rank}} \wedge \psi_{2D} \wedge 1\big) \bigg) \le C_0 (\log b)^2 b n^{-\alpha}, \tag{42}$$

where the constant $C > 0$ depends only on α. This bound will be used to control the stochastic component $\|\widehat{\phi} - \bar{f}\|_{L_1}$ of the L_1-error of the estimator $\widehat{\phi}$, where $\bar{f}$ is piecewise constant function defined as follows:

$$\bar{f}(x,y) = \frac{P_{ij}}{h^2} \qquad \text{if } (x,y) \in C_{ij}.$$

We have $\int_{C_{ij}} \bar{f} = P_{ij}$. On the other hand, by the definition of Algorithm 2, $\widehat{\phi}(x,y) = \widehat{P}^*_{ij}/h^2$ for $(x,y) \in C_{i,j}$, and $\int_{C_{ij}} \widehat{\phi} = \widehat{P}^*_{ij}$.

The bias component of the error is $\|f - \bar{f}\|_{L_1}$. For (x,y) in a given cell C_{ij} we have

$$\big|f(x,y) - \bar{f}(x,y)\big| = \left| \frac{1}{h^2} \int_{C_{ij}} (f(x,y) - f(x',y'))dx'dy' \right| \le CL(h^*)^\beta,$$

which yields that

$$\|f - \bar{f}\|_{L_1} = \sum_{i,j} \int_{C_{ij}} \big|f(x,y) - \bar{f}(x,y)\big|\, dxdy \le CL(h^*)^\beta. \tag{43}$$

Combining (42)–(43) we obtain that

$$\begin{aligned} \|f - \widehat{\phi}\|_{L_1} &\le \|f - \bar{f}\|_{L_1} + \|\bar{f} - \widehat{\phi}\|_{L_1} \\ &\le CL(h^*)^\beta + \sum_{i,j} |\widehat{P}^*_{ij} - P_{ij}| \\ &\le C\left(\left(\frac{K}{n}\right)^{\beta/(2\beta+1)} \log^{3/2} n \wedge n^{-\frac{\beta}{2\beta+2}} \right) \end{aligned} \tag{44}$$

with probability at least $1 - C_0(\log b)^2 b n^{-\alpha} - 4\exp(-n^{2/3}/2)$. Note also that $b \le Cn^{1/(2\beta+1)}$. This implies that there exist constants $C_0' > 0, C > 0$ depending only on α and L such that

the bound (44) holds with probability at least $1-C_0'(\log n)^2 n^{1/(2\beta+1)-\alpha}$. Next, since $\|\widehat{\phi}\|_{L_1} \leq 1$ we have $\|f-\widehat{\phi}\|_{L_1} \leq 2$. Thus, we can replace the bound in (44) by a stronger bound $C\left\{\left(\frac{K}{n}\right)^{\beta/(2\beta+1)} \log^{3/2} n \wedge n^{-\frac{\beta}{2\beta+2}}\right\}$ that holds with the same probability. Furthermore, a bound of the same order is satisfied for $\|f-\widehat{f}\|_{L_1}$ with the same probability. Indeed, we have $\|f-\widehat{f}\|_{L_1} \leq 2\|f-\widehat{\phi}\|_{L_1}$. Thus, the bound (19) of Theorem 7 is proved. Next, if $\alpha > 4/3$ then the bound (20) for the expectation follows easily from (19) and the inequality $\|f-\widehat{f}\|_{L_1} \leq 2$. ∎

C.4 Upper bounds for continuous distributions: unknown support

Proof of Theorem 12.

Recall that in Algorithm 3 we assume that n is a multiple of 4. Note first that it suffices to consider the case $\beta > -\frac{2\log((8L)^{-3/2}\wedge 1)}{\log(n)}$. Indeed, if the interval $(0, -\frac{2\log((8L)^{-3/2}\wedge 1)}{\log(n)}]$ is non-empty and β belongs to this interval then the desired rate from equation (26) satisfies, for $n \geq 4$,

$$\left(\frac{K}{n}\right)^{\beta/(2\beta+1)} \log^{3/2} n \wedge n^{-\frac{\beta}{2\beta+2}} \geq \left(\frac{1}{n}\right)^{\beta} \log^{3/2} n \wedge n^{-\frac{\beta}{2}}$$
$$\geq ((8L)^{-3/2} \wedge 1)^2 \log^{3/2} 4 \wedge ((8L)^{-3/2} \wedge 1) =: a(L).$$

On the other hand, since $\widehat{f}$ is a probability density we have the trivial bound $\|f-\widehat{f}\|_{L_1} \leq 2$ for all probability densities f. Thus, we immediately get (26) with any $C_1' > 2/a(L)$.

In the rest of this proof, we assume that $\beta > -\frac{2\log((8L)^{-3/2}\wedge 1)}{\log(n)}$.

Fix a density $f \in \mathcal{G}_K$, and consider the marginal densities $g_1 : [0,1] \longrightarrow \mathbb{R}$ and $g_2 : [0,1] \longrightarrow \mathbb{R}$ defined by

$$g_1(x) = \int_0^1 f(x,y)\,dy \quad \text{and} \quad g_2(y) = \int_0^1 f(x,y)\,dx, \qquad \forall x, y \in [0,1].$$

The functions g_1 and g_2 are β-Hölder on $[r_1, R_1]$ and $[r_2, R_2]$, respectively. They are densities of random variables $\Pi_1(X_i)$ and $\Pi_2(X_i)$, respectively, where $\Pi_j(\cdot)$ denotes the projector onto the j-th coordinate. Let $q_-^{(j)}$ and $q_+^{(j)}$ be the quantiles of order $n^{-1/3}$ and $1-n^{-1/3}$ of the probability measure induced by g_j:

$$\int_{-\infty}^{q_-^{(j)}} g_j = n^{-1/3} \quad \text{and} \quad \int_{q_+^{(j)}}^{+\infty} g_j = n^{-1/3}.$$

Since $f \in \mathcal{G}_K$ it follows from the definition in (24) that there exist real numbers $r_1, r_2, R_1, R_2 \in [0,1]$ (depending on f) such that $\Delta_j := R_j - r_j > 0$, $j = 1, 2$, and

$$\begin{cases} \mathrm{Supp}(f) = [r_1, R_1] \times [r_2, R_2], \\ f \text{ is } \beta\text{-Hölder over } \mathrm{Supp}(f), \\ \int_{\mathrm{Supp}(f)} f = 1 \text{ and } f \geq 0. \end{cases}$$

For $j = 1, 2$, let $\widehat{g}_j$ denote the output of Alg 4$(\Pi_j(X_{n/2+1}), \dots, \Pi_j(X_n), K')$. Let $\widehat{r}_j$ and $\widehat{R}_j$ be the estimators of r_j and R_j defined in Algorithm 3. For the rest of this proof, we place ourselves on the event $\mathcal{E} \cap \mathcal{F}$, where

$$\mathcal{E} = \left\{ \widehat{r}_j < q_-^{(j)} < q_+^{(j)} < \widehat{R}_j \text{ for } j = 1, 2 \right\},$$

$$\mathcal{F} = \left\{ \|g_j - \widehat{g}_j\|_{L_1} \le C(K'/n)^{\beta/(2\beta+1)} \text{ for } j = 1, 2 \right\}.$$

Here, $C > 0$ is the constant from Lemma 17 and $K' = K$ if $\left(\frac{K}{n}\right)^{\frac{\beta}{2\beta+1}} \log^{3/2}(n) \le n^{-\frac{\beta}{2\beta+2}}$, otherwise $K' = n^{\frac{1}{2\beta+2}}$. By Lemma 17, if $n \ge 64$ and $K'/n \le (8L)^{-3/2} \wedge 1$ then $\mathbb{P}(\mathcal{F}) \ge 1 - 10\exp\left(-n^{1/3}\right)$, and $\mathbb{P}(\mathcal{E}) \ge 1 - 4\exp(-n^{2/3}/2)$, so that $\mathbb{P}(\mathcal{E} \cap \mathcal{F}) \ge 1 - 14\exp\left(-n^{1/3}\right)$. Here, the condition $K'/n \le (8L)^{-3/2} \wedge 1$ is satisfied because, by the definition of K', we have $K' \le n^{\frac{1}{2\beta+2}} \le n^{1/2} \le n((8L)^{-3/2} \wedge 1)$, where the last inequality is due to the fact that $1 \ge \beta > -\frac{2\log((8L)^{-3/2} \wedge 1)}{\log(n)}$.

Let $\widehat{\Delta}_j = \widehat{R}_j - \widehat{r}_j$, $j = 1, 2$. We distinguish between the following two cases.

First case: $\widehat{\Delta}_1 \wedge \widehat{\Delta}_2 \le h^*$. It suffices to assume that $\widehat{\Delta}_1 \le h^*$ since the case $\widehat{\Delta}_2 \le h^*$ is treated in the same way. If $\widehat{\Delta}_1 \le h^*$ the estimator $\widehat{\phi}$ in Algorithm 3 has the form $\widehat{\phi}(x, y) = \frac{1}{\widehat{\Delta}_1} \mathbb{1}_{x \in [\widehat{r}_1, \widehat{R}_1]} \widehat{g}_2(y)$ and we get

$$\|f - \widehat{\phi}\|_{L_1} \le \int_{y=0}^{1} \int_{x=r_1}^{R_1} \left| f(x, y) - \frac{g_2(y)}{\Delta_1} \right| dxdy + \int_{y=0}^{1} \int_{x=r_1}^{R_1} \left| \frac{g_2(y)}{\Delta_1} - \widehat{\phi}(x, y) \right| dxdy$$

$$\le \int_{y=0}^{1} \int_{x=r_1}^{R_1} \left\{ \frac{1}{\Delta_1} \int_{r_1}^{R_1} \underbrace{\left| f(x, y) - f(x', y) \right|}_{\le L\Delta_1^\beta} dx' \right\} dxdy + \int_{y=0}^{1} \int_{x=r_1}^{R_1} \left| \frac{g_2(y)}{\Delta_1} - \frac{\widehat{g}_2(y)}{\widehat{\Delta}_1} \mathbb{1}_{x \in [\widehat{r}_1, \widehat{R}_1]} \right| dxdy.$$

Next we will use that $\int_{x=\widehat{r}_1}^{\widehat{R}_1} \mathbb{1}_{x \in [\widehat{r}_1, \widehat{R}_1]} = \widehat{\Delta}_1$ and split the integral over $[r_1, R_1]$ into two integrals: one over $[\widehat{r}_1, \widehat{R}_1]$ and the other over $[r_1, R_1] \setminus [\widehat{r}_1, \widehat{R}_1]$:

$$\|f - \widehat{\phi}\|_{L_1} \le \int_{y=0}^{1} \int_{x=r_1}^{R_1} L\Delta_1^\beta \, dxdy \; + \int_{y=0}^{1} \left| g_2(y) \frac{\widehat{\Delta}_1}{\Delta_1} - \widehat{g}_2(y) \right| dy + \int_{y=0}^{1} \int_{x \in [r_1, R_1] \setminus [\widehat{r}_1, \widehat{R}_1]} \frac{g_2(y)}{\Delta_1} dxdy$$

$$\le L\Delta_1^{\beta+1} + \|g_2 - \widehat{g}_2\|_{L_1} + \|\widehat{g}_2\|_{L_1} \left| 1 - \frac{\widehat{\Delta}_1}{\Delta_1} \right| + \|g_2\|_{L_1} \frac{\Delta_1 - \widehat{\Delta}_1}{\Delta_1}$$

$$\le L(h^*)^{\beta+1} \left(1 + 16n^{-1/(2\beta+1)}\right)^{\beta+1} + C\left(\frac{K'}{n}\right)^{\beta/(2\beta+1)} + 32n^{-1/(2\beta+1)}$$

$$\le C' \left(\frac{K'}{n}\right)^{\frac{\beta}{2\beta+1}}$$

$$= C' \begin{cases} (K/n)^{\frac{\beta}{2\beta+1}} & \text{if } (K/n)^{\frac{\beta}{2\beta+1}} \log^{3/2}(n) \le n^{-\frac{\beta}{2\beta+2}} \\ n^{-\frac{\beta}{2\beta+2}} & \text{otherwise} \end{cases}$$

$$\leq C'\Big(\frac{K}{n}\Big)^{\beta/(2\beta+1)} \log^{3/2}(n) \wedge n^{-\frac{\beta}{2\beta+2}}.$$

where the constant $C' > 0$ depends only on L. Here, we have used the facts that, by Lemma 17, if $\mathcal{E}$ holds and $\widehat{\Delta}_1 \leq h^*$, then $\Delta_1 - \widehat{\Delta}_1 \leq 16\widehat{\Delta}_1 n^{-1/(2\beta+1)} \leq 16\Delta_1 n^{-1/(2\beta+1)}$ and on the event $\mathcal{F}$ we have $\|g_2 - \widehat{g}_2\|_{L_1} \leq C(K'/n)^{\beta/(2\beta+1)}$. We conclude that, in the first case, the bound (26) of Theorem 12 is satisfied.

Second case: $\widehat{\Delta}_1 > h^*$ and $\widehat{\Delta}_2 > h^*$. In this case the estimator $\widehat{\phi}$ in Algorithm 3 has the form

$$\widehat{\phi}(x,y) = \frac{1}{h_1 h_2} \sum_{j\in E_1} \sum_{j'\in E_2} \widehat{P}^*_{jj'} \, \mathbb{1}_{(x,y)\in C_{jj'}}.$$

Recalling the notation of Algorithm 3 we have

$$b := |E_1| \vee |E_2| \leq 1 + 2\lceil h_1^{-1} \rceil \vee 2\lceil h_2^{-1} \rceil \leq C(n/K')^{1/(2\beta+1)}, \tag{45}$$

where $C > 0$ is an absolute constant. Since $f \in \mathcal{G}_{K,\beta}$ we have the representation $f(x,y) = \sum_{k=1}^K u_k(x) v_k(y)$ with some functions $u_k \in L_1[0,1]$, $v_k \in L_1[0,1]$ for $k \in [K]$. Introduce the matrix $P = (P_{ij})_{i\in E_1, j\in E_2}$ with entries

$$P_{ij} = \int_{C_{ij}} f(x,y)dxdy = \sum_{k=1}^K \int_{A_i} u_k(x)dx \int_{B_j} v_k(y)dy = \sum_{k=1}^K U_k(i)V_k(j), \qquad (i,j) \in E_1 \times E_2,$$

where $U_k(i) = \int_{A_i} u_k(x)dx$ and $V_k(j) = \int_{B_j} v_k(y)dy$ for any $(i,j) \in E_1 \times E_2$ and $k \in [K]$. Set $U_k = (U_k(i))_{i\in E_1}$, $V_k = (V_k(j))_{j\in E_2}$. Then we can write

$$P = \sum_{k=1}^K U_k V_k^\top.$$

Matrix P has rank at most K. Consider now the histogram matrices $G = (G_{ij})_{i\in E_1, j\in E_2}$ and $G' = (G'_{ij})_{i\in E_1, j\in E_2}$ defined in Algorithm 3 with entries

$$G_{ij} = \frac{4}{n} \sum_{\ell=n/2+1}^{3n/4} \mathbb{1}_{X_\ell \in C_{ij}} \quad \text{and} \quad G'_{ij} = \frac{4}{n} \sum_{\ell=3n/4+1}^{n} \mathbb{1}_{X_\ell \in C_{ij}}, \qquad (i,j) \in E_1 \times E_2.$$

The matrices G and G' are mutually independent, and both $nG/4$ and $nG'/4$ follow the multinomial distribution $\mathcal{M}(P, n/4)$.

To alleviate the notation, we define the following two quantities

$$\psi_{\text{low-rank}} = (K/n)^{\frac{\beta}{2\beta+1}} \log^{3/2}(n) \quad \text{and} \quad \psi_{2D} = n^{-\frac{\beta}{2\beta+2}}.$$

By the definition of Algorithm 3, matrix $\widehat{P}^*$ is the output of $\text{Alg1}(\alpha, b, n, \frac{n}{4}, G, G')$ with $b = |E_1| \vee |E_2|$, and $\alpha > 1$ if $\psi_{\text{low-rank}} \leq \psi_{2D}$ and $\widehat{P}^* = (G + G')/2$ if $\psi_{\text{low-rank}} > \psi_{2D}$. Therefore, if $\psi_{\text{low-rank}} \leq \psi_{2D}$, then Theorem 4 implies that, for some constants $C > 0$ depending only on α,

$$\|\widehat{P}^* - P\|_1 \leq C\sqrt{\frac{Kb}{n}} \log(b) \log^{1/2}(n) \leq C\left(\frac{K}{n}\right)^{\beta/(2\beta+1)} (\log n)^{3/2}$$

with probability at least $1 - C_0(\log b)^2 b n^{-\alpha}$, where $C_0 > 0$ depends only on α.

If $\psi_{\text{low-rank}} > \psi_{2D}$, then (45) implies that $b \le C n^{\frac{1}{2\beta+2}}$. Using this fact and Lemma 22 we obtain that, conditionally on $\mathcal{D}_1 = \{X_1, \dots, X_{3n/4}\}$,

$$\mathbb{P}\Big(\|\widehat{P}^* - P\|_1 > C n^{\frac{\beta}{2\beta+2}} \;\Big|\; \mathcal{D}_1 \Big) \le \mathbb{P}\Big(\|\widehat{P}^* - P\|_1 > \sqrt{\frac{b^2}{n/2}} + \sqrt{\frac{2\alpha \log(n)}{n/2}} \;\Big|\; \mathcal{D}_1 \Big) \le n^{-\alpha},$$

where the constant $C > 0$ depends only on α.

Combining the cases $\psi_{\text{low-rank}} \le \psi_{2D}$ and $\psi_{\text{low-rank}} > \psi_{2D}$ and using the fact that $\|\widehat{P}^* - P\|_1 \le 2$ we obtain the bound

$$\mathbb{P}\Big(\|\widehat{P}^* - P\|_1 > C\big(\psi_{\text{low-rank}} \wedge \psi_{2D} \wedge 1\big) \Big) \le C_0 (\log b)^2 b n^{-\alpha}, \tag{46}$$

where the constant $C > 0$ depends only on α. This bound will be used to control the stochastic component $\|\widehat{\phi} - \bar{f}\|_{L_1}$ of the L_1-error of the estimator $\widehat{\phi}$, where $\bar{f}$ is piecewise constant function defined as follows:

$$\bar{f}(x,y) = \frac{P_{ij}}{h_1 h_2} \qquad \text{if } (x,y) \in C_{ij}.$$

We have $\int_{C_{ij}} \bar{f} = P_{ij}$. On the other hand, by the definition of Algorithm 3, $\widehat{\phi}(x,y) = \widehat{P}^*_{ij}/(h_1 h_2)$ for $(x,y) \in C_{i,j}$, and $\int_{C_{ij}} \widehat{\phi} = \widehat{P}^*_{ij}$.

The bias component of the error is $\|f - \bar{f}\|_{L_1}$. In order to control it, we need to distinguish between two cases. Indeed, f can be discontinuous at the boundaries of its rectangular support, which requires separately analyzing the behavior of $\widehat{f}$ on the cells C_{ij} that intersect the boundary of $\text{Supp}(f)$. Let $i_0, i_1 \in E_1$ be the indices such that $r_1 \in A_{i_0}$ and $R_1 \in A_{i_1}$ respectively. Similarly, let $j_0, j_1 \in E_2$ be the indices such that $r_2 \in B_{i_0}$ and $R_2 \in B_{i_1}$ respectively. We note that $i_0, j_0 \le -1$ and that $i_1 \ge \ell_1$ and $j_1 \ge \ell_2$. We let $\mathcal{B}$ denote the indices (i,j) of the cells C_{ij} intersecting the boundary of $\text{Supp}(f)$:

$$\begin{aligned}
\mathcal{B} &= \big\{ (i,j) \in E_1 \times E_2 : C_{i,j} \cap \partial\, \text{Supp}(f) \neq \emptyset \big\} \\
&= \left\{ (i,j) \in E_1 \times E_2 : \begin{cases} i \in \{i_0, i_1\} \text{ and } j \in [j_0, j_1] \\ \text{or} \\ j \in \{j_0, j_1\} \text{ and } i \in [i_0, i_1] \end{cases} \right\}.
\end{aligned}$$

We define $\mathcal{C} = \bigcup_{(i,j)\in\mathcal{B}} C_{ij}$.

We first consider a cell C_{ij} that does not intersect the boundary of $\text{Supp}(f)$, which means that $C_{ij} \in \mathcal{C}^c$, that is, $(i,j) \in (E_1 \times E_2) \setminus \mathcal{B}$. For (x,y) in such cells C_{ij} we have

$$\big| f(x,y) - \bar{f}(x,y) \big| = \left| \frac{1}{h_1 h_2} \int_{C_{ij}} (f(x,y) - f(x',y'))\, dx' dy' \right| \le C L (h^*)^\beta,$$

which yields that

$$\|f - \bar{f}\|_{L_1(\mathcal{C}^c)} = \sum_{(i,j)\in(E_1\times E_2)\setminus\mathcal{B}} \int_{C_{ij}} \big| f(x,y) - \bar{f}(x,y) \big|\, dxdy \le C L (h^*)^\beta. \tag{47}$$

We consider now the opposite case $(i,j) \in \mathcal{B}$, and we analyze the behavior of $\widehat{\phi}$ on $\mathcal{C}$. Note that, by construction, the sets C_{ij} with $(i,j) \in \mathcal{B}$ cannot belong to the rectangle $[\widehat{r}_1, \widehat{R}_1] \times [\widehat{r}_2, \widehat{R}_2]$, which is included in the interior of $\mathrm{Supp}(f)$ and represents a union of sets C_{ij} with some (i,j)'s. Therefore, on the event $\mathcal{E}$, we have $\int_{\mathcal{C}} f \le Cn^{-1/3} \le n^{-\beta/(2\beta+1)}$ for an absolute constant $C > 0$. It follows that, on the event $\mathcal{E}$,

$$\|f - \bar{f}\|_{L_1(\mathcal{C})} \le \|f\|_{L_1(\mathcal{C})} + \|\bar{f}\|_{L_1(\mathcal{C})} = 2\|f\|_{L_1(\mathcal{C})} \le Cn^{-\beta/(2\beta+1)}. \tag{48}$$

Combining (46)–(48) we obtain that

$$\begin{aligned}
\|f - \widehat{\phi}\|_{L_1} &\le \|f - \bar{f}\|_{L_1} + \|\bar{f} - \widehat{\phi}\|_{L_1} \\
&\le \|f - \bar{f}\|_{L_1(\mathcal{C}^c)} + \|f - \bar{f}\|_{L_1(\mathcal{C})} + \sum_{(i,j)\in E_1\times E_2} \int_{C_{ij}} |\bar{f} - \widehat{\phi}| \\
&\le CL(h^*)^\beta + Cn^{-\beta/(2\beta+1)} + \sum_{(i,j)\in E_1\times E_2} |\widehat{P}^*_{ij} - P_{ij}| \\
&\le C\left(\left(\frac{K}{n}\right)^{\beta/(2\beta+1)} \log^{3/2} n \wedge n^{-\frac{\beta}{2\beta+2}}\right)
\end{aligned} \tag{49}$$

with probability at least $1 - C_0(\log b)^2 b n^{-\alpha} - 4\exp(-n^{2/3}/2)$. Note also that $b \le Cn^{1/(2\beta+1)}$. This implies that there exist constants $C_0' > 0, C > 0$ depending only on α and L such that the bound (49) holds with probability at least $1 - C_0'(\log n)^2 n^{1/(2\beta+1)-\alpha}$. Next, since $\|\widehat{\phi}\|_{L_1} \le 1$ we have $\|f - \widehat{\phi}\|_{L_1} \le 2$. Thus, we can replace the bound in (49) by a stronger bound $C\left\{\left(\frac{K}{n}\right)^{\beta/(2\beta+1)} \log^{3/2} n \wedge n^{-\frac{\beta}{2\beta+2}}\right\}$ that holds with the same probability. Furthermore, a bound of the same order is satisfied for $\|f - \widehat{f}\|_{L_1}$ with the same probability. Indeed, we have $\|f - \widehat{f}\|_{L_1} \le 2\|f - \widehat{\phi}\|_{L_1}$. Thus, the bound (26) of Theorem 12 is proved. Next, if $\alpha > 4/3$ then the bound (27) for the expectation follows easily from (26) and the inequality $\|f - \widehat{f}\|_{L_1} \le 2$. Finally, note that the condition $n \ge 64$ used to apply Lemma 17 can be dropped since for $n < 64$ the result of the theorem follows from the trivial bound $\|f - \widehat{f}\|_{L_1} \le 2$. ∎

Lemma 17 *Let $f : \mathbb{R} \longrightarrow \mathbb{R}_+$ be a probability density. Assume that for some $r, R \in [0,1]$, the function f is β-Hölder on $[r,R]$ and that $f = 0$ on $\mathbb{R} \setminus [r,R]$. Let $Z_1, \dots, Z_n$ be iid random variables distributed with probability density f, where $n \ge 64$ is an even integer. Define $\widehat{r} = \min\left\{Z_i : i \in \{1, \dots, \frac{n}{2}\}\right\}$ and $\widehat{R} = \max\left\{Z_i : i \in \{1, \dots, \frac{n}{2}\}\right\}$. Let also $\Delta = R - r$ and $\widehat{\Delta} = \widehat{R} - \widehat{r}$. Let q_- and q_+ be the quantiles of order $n^{-1/3}$ and $1 - n^{-1/3}$ of the probability measure induced by f:*

$$\int_{-\infty}^{q_-} f(x)dx = n^{-1/3}, \qquad \int_{q_+}^{+\infty} f(x)dx = n^{-1/3}.$$

Define the event $\mathcal{E} = \left\{\widehat{r} < q_- < q_+ < \widehat{R}\right\}$ and set $h^* = (K'/n)^{1/(2\beta+1)}$, where $K' \ge 1$ is such that $K'/n \le (8L)^{-3/2} \wedge 1$. Let $\widehat{g}$ be an output of $\mathrm{Alg4}(Z_1, \dots, Z_n, K')$. Then the following holds.

1. $\mathbb{P}(\mathcal{E}) \geq 1 - 2e^{-n^{2/3}/2}$.
2. *On the event* $\mathcal{E}$, *if* $\widehat{\Delta} \leq h^*$ *then* $\Delta < \widehat{\Delta}\left[1 + 16n^{-1/(2\beta+1)}\right]$.
3. *There exists a constant* $C > 0$ *depending only on* L *such that*

$$\mathbb{P}\left(\|f - \widehat{g}\|_{L_1} \leq C(K'/n)^{\beta/(2\beta+1)}\right) \geq 1 - 4\exp\left(-n^{1/3}\right).$$

Proof The first assertion of the lemma follows from the inequalities

$$\mathbb{P}(\mathcal{E}^c) \leq \mathbb{P}(\widehat{r} \geq q_-) + \mathbb{P}(\widehat{R} \leq q_+) = \mathbb{P}\left(Z_1 \geq q_-\right)^{\frac{n}{2}} + \mathbb{P}\left(Z_1 \leq q_+\right)^{\frac{n}{2}} = 2\left(1 - n^{-1/3}\right)^{\frac{n}{2}} \leq 2e^{-n^{2/3}/2}. \tag{50}$$

We now prove the second assertion of the lemma. On the event $\mathcal{E}$ and if $\widehat{\Delta} < h^*$, using the β-Hölder continuity of f we have

$$\begin{aligned}
\frac{1}{2} &\leq 1 - 2n^{-1/3} \leq \int_{\widehat{r}}^{\widehat{R}} f(x)dx \leq \widehat{\Delta} \max_{[\widehat{r},\widehat{R}]} f \\
&\leq \widehat{\Delta} \min_{[\widehat{r},\widehat{R}]} f + L\widehat{\Delta}^{\beta+1} \leq \widehat{\Delta} \min_{[\widehat{r},\widehat{R}]} f + L(h^*)^{\beta+1} \\
&\leq \widehat{\Delta} \min_{[\widehat{r},\widehat{R}]} f + \frac{1}{8},
\end{aligned}$$

where the last inequality follows from the fact that $L(h^*)^{\beta+1} = L(K'/n)^{(\beta+1)/(2\beta+1)} \leq \frac{1}{8}$ due to the assumption $K'/n \leq (8L)^{-3/2} \wedge 1$. Therefore, $f(\widehat{r}) \geq \min_{[\widehat{r},\widehat{R}]} f \geq \frac{3}{8\widehat{\Delta}}$.

Set $\overline{r} := \widehat{r} - \frac{2}{f(\widehat{r})} n^{-1/(2\beta+1)}$ and let us prove that $r \geq \overline{r}$. Indeed, assume that, on the contrary, $r < \overline{r}$. Then, on the event $\mathcal{E}$ and if $\widehat{\Delta} \leq h^*$, using the β-Hölder continuity of f we get

$$\begin{aligned}
n^{-1/3} &\geq \int_r^{\widehat{r}} f(x)dx \geq \int_{\overline{r}}^{\widehat{r}} f(x)dx \geq (\widehat{r} - \overline{r})f(\widehat{r}) - L(\widehat{r} - \overline{r})^{\beta+1}/(\beta+1) \\
&= 2n^{-1/(2\beta+1)} - \frac{L}{\beta+1}\Big(\frac{2}{f(\widehat{r})}\Big)^{\beta+1} n^{-(\beta+1)/(2\beta+1)} \\
&\geq 2n^{-1/3} - \frac{L}{\beta+1}\Big(\frac{16\widehat{\Delta}}{3}\Big)^{\beta+1} n^{-(\beta+1)/(2\beta+1)} \\
&\geq 2n^{-1/3} - \frac{L(h^*)^{\beta+1}}{\beta+1}\Big(\frac{16}{3}\Big)^{\beta+1} n^{-(\beta+1)/(2\beta+1)} \\
&\geq 2n^{-1/3} - \frac{1}{8(\beta+1)}\Big(\frac{16}{3}\Big)^{\beta+1} n^{-2/3} > n^{-1/3}
\end{aligned}$$

for all $n \geq 64$ and $\beta \in (0, 1]$, which is a contradiction. Thus, we have $r \geq \overline{r} \geq \widehat{r} - 8\widehat{\Delta}n^{-1/(2\beta+1)}$ and, similarly, $R \leq \widehat{R} + 8\widehat{\Delta}n^{-1/(2\beta+1)}$, which implies the desired inequality $\Delta \leq \widehat{\Delta}(1 + 16n^{-1/(2\beta+1)})$. This concludes the proof of the second assertion of the lemma.

Finally, we prove the third assertion of the lemma. We have, for any $t > 0$,

$$\mathbb{P}(\|f - \widehat{g}\|_{L_1} > t) \le \mathbb{P}\left(\{\|f - \widehat{g}\|_{L_1} > t\} \cap \mathcal{E} \cap \{\widehat{\Delta} \le h^*\}\right) + \mathbb{P}\left(\{\|f - \widehat{g}\|_{L_1} > t\} \cap \mathcal{E} \cap \{\widehat{\Delta} > h^*\}\right) + \mathbb{P}(\mathcal{E}^c). \tag{51}$$

Consider the first probability on the right hand side of (51). Recall that if $\widehat{\Delta} < h^*$ the estimator of Algorithm 4 is $\widehat{g} = \frac{1}{\widehat{R}-\widehat{r}}\mathbb{1}_{[\widehat{r},\widehat{R}]}$. By continuity of f over $[r, R]$, there exists $x_0 \in [r, R]$ such that $f(x_0)(R - r) = \int_r^R f = 1$. Thus, using the second assertion of the lemma we obtain that, on the event $\mathcal{E}$ and if $\widehat{\Delta} \le h^*$,

$$\begin{aligned}
\|f - \widehat{g}\|_{L_1} &= \int_r^R |f(x) - \widehat{g}(x)|dx = \int_r^{\widehat{r}} f(x)dx + \int_{\widehat{r}}^{\widehat{R}} |f(x) - \widehat{g}(x)|dx + \int_{\widehat{R}}^R f(x)dx \\
&\le \int_{\widehat{r}}^{\widehat{R}} |f(x) - f(x_0)|dx + \int_{\widehat{r}}^{\widehat{R}} |f(x_0) - \widehat{g}(x)|dx + 2n^{-1/3} \\
&\le L|R - r|^{\beta+1} + \int_{\widehat{r}}^{\widehat{R}} |f(x_0) - \widehat{g}(x)|dx + 2n^{-1/3} \\
&\le L\left(\widehat{\Delta} + 16\widehat{\Delta}n^{-1/(2\beta+1)}\right)^{\beta+1} + (\widehat{R} - \widehat{r})\left|\frac{1}{R-r} - \frac{1}{\widehat{R}-\widehat{r}}\right| + 2n^{-1/3} \\
&\le 25L(h^*)^{\beta+1} + \frac{\Delta - \widehat{\Delta}}{\Delta} + 2n^{-1/3} \\
&\le 25L(h^*)^{\beta+1} + \frac{\widehat{\Delta}16n^{-1/(2\beta+1)}}{\Delta} + 2n^{-1/3} \\
&\le 25L(K'/n)^{(\beta+1)/(2\beta+1)} + 18n^{-1/3} \quad (\text{since} \quad \widehat{\Delta}/\Delta \le 1) \\
&\le (25L + 18)(K'/n)^{\beta/(2\beta+1)}
\end{aligned}$$

for all $n \ge 64$ and $\beta \in (0, 1]$. We conclude that the first probability on the right hand side of (51) vanishes for all $t > (25L + 18)(K'/n)^{\beta/(2\beta+1)}$.

Next, consider the second probability on the right hand side of (51). Fix the subsample $\mathcal{D}_1 = (Z_1, \dots, Z_{n/2})$ such that $\widehat{\Delta} > h^*$. Then the partition $(A_j)_{j\in E}$ defined in Algorithm 4 is also fixed, where $E = \{1, \dots, \ell\}$. By continuity of f over A_j, we define $x_j \in A_j$ such that $f(x_j)\, h = \int_{A_j} f$. For any $j \in E$, we define

$$p_j = \int_{A_j} f(x)dx, \quad \text{and} \quad \widehat{p}_j = \int_{A_j} \widehat{g}(x)dx \;=\; \frac{2N_j}{n},$$

where $N_j = \sum_{i=\lfloor n/2\rfloor+1}^n \mathbb{1}_{Z_i \in A_j}$, see the definition of Algorithm 4. Note that $\widehat{g}$ is supported on $[\widehat{r}, \widehat{R}]$ and $(A_j)_{j\in E}$ is a partition of $[\widehat{r}, \widehat{R}]$. Then we have the following bound on the estimation error of $\widehat{g}$, which is valid for any fixed subsample $\mathcal{D}_1$ such that event $\mathcal{E}$ holds and $\widehat{\Delta} > h^*$:

$$\|f - \widehat{g}\|_{L_1} = \int_r^{\widehat{r}} f(x)dx + \int_{\widehat{R}}^R f(x)dx + \sum_{j\in E} \int_{A_j} \big|f(x) - \widehat{g}(x)\big|dx$$

$$
\begin{aligned}
&\leq 2n^{-1/3} + \sum_{j\in E} \int_{A_j} \big|f(x) - \widehat{g}(x)\big| dx \quad \text{(since } \mathcal{E} \text{ holds)} \\
&\leq 2n^{-1/3} + \sum_{j\in E} \int_{A_j} \big(\big|f(x) - f(x_j)\big| + \big|f(x_j) - \widehat{g}(x)\,\big|\big)\, dx \qquad (52) \\
&\leq 2n^{-1/3} + \sum_{j\in E} \Big\{ Lh^{\beta+1} + |p_j - \widehat{p}_j| \Big\} \\
&\leq 2n^{-1/3} + Lh^{\beta} + \sum_{j=1}^{\ell} |p_j - \widehat{p}_j| \quad \text{(since } h = \widehat{\Delta}/\ell \leq 1/\ell) \\
&\leq 2n^{-1/3} + 2L(K'/n)^{\beta/(2\beta+1)} + \sum_{j=1}^{\ell} |p_j - \widehat{p}_j|,
\end{aligned}
$$

where the last inequality uses the fact that $h = \frac{\widehat{\Delta}}{\lfloor \widehat{\Delta}/h^* \rfloor} = h^* \frac{\widehat{\Delta}/h^*}{\lfloor \widehat{\Delta}/h^* \rfloor} \leq 2h^* = 2(K'/n)^{1/(2\beta+1)}$ for $\widehat{\Delta} > h^*$. Lemma 22 stated below yields that, for any $\delta \in (0,1)$,

$$
\mathbb{P}\left(\sum_{j=1}^{\ell} |p_j - \widehat{p}_j| > 2\sqrt{\frac{2(\ell + \log(1/\delta))}{n}} \;\Big|\; \mathcal{D}_1 \right) \leq \delta.
$$

Set $\delta = \exp\big(-n^{1/(2\beta+1)}\big)$. Then combining this bound with (52) and the fact that $\ell = \lfloor \widehat{\Delta}/h^* \rfloor \leq 1/h^* \leq n^{1/(2\beta+1)}$ we obtain that there exists a constant $C_* > 0$ depending only on L such that

$$
\mathbb{P}\left(\|f - \widehat{g}\|_{L_1} > C_*(K'/n)^{\beta/(2\beta+1)} \;\Big|\; \mathcal{D}_1 \right) \leq \exp\big(-n^{1/(2\beta+1)}\big) \quad \text{if } \mathcal{D}_1 \text{ is such that } \mathcal{E} \text{ holds and } \widehat{\Delta} > h^*.
$$

It follows that

$$
\mathbb{P}\left(\{\|f - \widehat{g}\|_{L_1} > C_*(K'/n)^{\beta/(2\beta+1)}\} \cap \mathcal{E} \cap \{\widehat{\Delta} > h^*\} \right) \leq \exp\big(-n^{1/(2\beta+1)}\big).
$$

Plugging this bound and (50) in (51) and recalling that the first probability on the right hand side of (51) vanishes for $t > (25L+18)(K'/n)^{\beta/(2\beta+1)}$ we find that, for $C > \max(25L+18, C_*)$,

$$
\mathbb{P}(\|f - \widehat{g}\|_{L_1} > C(K'/n)^{\beta/(2\beta+1)}) \leq \exp\big(-n^{1/(2\beta+1)}\big) + \mathbb{P}(\mathcal{E}^c) \leq 4\exp\big(-n^{1/3}\big),
$$

where we have used the first assertion of the lemma and the fact that $\exp\big(-n^{2/3}/2\big) \leq \exp\big(-n^{1/3}\big)$ for $n \geq 8$. ∎

C.5 Lower bounds for continuous distributions

Proof of Theorem 8.

We define $f_0 = \mathbb{1}_{[0,1]^2}$ and set $K' = \lfloor n^{1/(2\beta+2)} \rfloor \wedge K$, $H = \lceil (n/K')^{1/(2\beta+1)} \rceil$, $h_x = 1/K'$, $h_y = 1/H$. Let $\varphi : \mathbb{R} \to [0,1]$ be a non identically zero infinitely many times differentiable

function with support $(-1/2, 1/2)$ satisfying the β-Hölder condition with Hölder constant $1/2$ (see (2.34) in Tsybakov (2009) for an example of such function). For $i, j \in \mathbb{Z}$, set $U_i(x) := \mathbb{1}_{[x_i \pm h_x/2]}(x)$ where $x_i = \big(i - \frac{1}{2}\big)h_x$ and $V_{j,h_y}(y) = \varphi\big(\frac{y-y_j^-}{h_y/2}\big) - \varphi\big(\frac{y-y_j^+}{h_y/2}\big)$, where $y_j^- = \big(j - \frac{3}{4}\big)h_y$ and $y_j^+ = \big(j - \frac{1}{4}\big)h_y$. The supports of the functions $\varphi\big(\frac{\cdot-y_j^-}{h_y/2}\big)$ and $\varphi\big(\frac{\cdot-y_j^+}{h_y/2}\big)$ are disjoint and each of these functions is β-Hölder with Hölder constant $(1/2)(h_x/2)^{-\beta}$. Hence, $V_{j,h_y}(\cdot)$ is β-Hölder with Hölder constant $(h_y/2)^{-\beta}$. Note also that the functions V_{j,h_y} have disjoint supports in $[0,1]$ for different j's and that the functions V_{j,h_y} integrate to 0. Similarly, the functions U_i have disjoint supports in $[0,1]$ and are β-Hölder on their support with Hölder constant 0 since they are constant over their support.

Set $c_* = 1/4 \wedge 1/(2L)$. For $\omega = (\omega_{ij})$, where $\omega_{ij} \in \{0,1\}$ for all i, j, we define the functions f_ω as follows:

$$f_\omega(x,y) = \sum_{i=1}^{K'} U_i(x)\bigg(1 + \sum_{j=1}^{H} \omega_{ij} c_* L h_y^\beta V_{j,h_y}(y)\bigg). \tag{53}$$

The functions f_ω belong to the class $\mathcal{F}_{K,\beta}^{\text{mixtures}}$ since they can be rewritten as mixtures of separable densities that are products of Hölder-smooth 1-dimensional densities. Indeed, it holds that

$$f_\omega(x,y) = \frac{1}{K'}\sum_{k=1}^{K'} K' U_i(x)\bigg(1 + \sum_{j=1}^{H} \omega_{ij} c_* L h_y^\beta V_{j,h_y}(y)\bigg).$$

The functions $K'U_i$ are densities over $[0,1]$ and are β-Hölder over their support since U_i is a constant function over an interval. Moreover, it suffices to take c_* small enough to guarantee that $1 + \frac{1}{C_0}\sum_{j=1}^{H} \omega_{ij} c_* L h_y^\beta V_{j,h_y}$ is non-negative, ensuring that $1 + \frac{1}{C_0}\sum_{j=1}^{H} \omega_{ij} c_* L h_y^\beta V_{j,h_y}$ is a β-Hölder density over $[0,1]$ since the V_{j,h_y}'s have a zero integral and are β-Hölder.

First, we check that each f_ω belongs to $\mathcal{G}_{K',\beta} \subseteq \mathcal{G}_{K,\beta}$.

- We have $\int_{[0,1]^2} f_\omega = 1$ by construction.
- We have $f_\omega(x,y) \geq 1/2$ for all $(x,y) \in [0,1]^2$ since $c_* \leq (2L)^{-1}$, $h_y \leq 1$, and the functions $(x,y) \mapsto U_i(x)V_{j,h_y}(y)$ have disjoint supports and take values in $[-1,1]$.
- We now check the Hölder condition. Let $C_{ij} = [x_i \pm h_x/2] \times [y_j \pm h_y/2]$ where $y_j = (j - 1/2)h_y$ for any $j \in \mathbb{Z}$. Note that for each $i \in [K']$ and $j \in [H]$, the cell C_{ij} contains the support of the function $(x,y) \mapsto U_i(x)V_{j,h_y}(y)$. Since the cells C_{ij} are disjoint and $f_\omega = 1$ at the boundary of each cell C_{ij}, it suffices to show that f_ω is β-Hölder with Hölder constant $L/2$ on each cell C_{ij} to obtain that f_ω is β-Hölder with Hölder constant L on $[0,1]^2$. Fix any two points $z, z' \in [0,1]^2$, belonging to the same cell C_{ij}. Then, writing $z = (x,y)$, $z' = (x',y')$ and recalling that the function φ is β-Hölder with Hölder constant $1/2$ we obtain:

$$\begin{aligned}
\big|f_\omega(z) - f_\omega(z')\big| &= c_* L h_y^\beta \Big|U_i(x)V_{j,h_y}(y) - U_i(x')V_{j,h_y}(y')\Big| \\
&= c_* L h_y^\beta \bigg|\varphi\Big(\frac{y-y_j^+}{h_y/2}\Big) - \varphi\Big(\frac{y'-y_j^+}{h_y/2}\Big) - \Big[\varphi\Big(\frac{y-y_j^-}{h_y/2}\Big) + \varphi\Big(\frac{y'-y_j^-}{h_y/2}\Big)\Big]\bigg|
\end{aligned}$$

$$\leq c_* L h_y^\beta \left(\frac{|y - y'|}{h_y/2} \right)^\beta$$
$$\leq \frac{L}{2} \|z - z'\|_\infty^\beta,$$

where we have used the facts that $h_y \leq h_x$ and $c_* \leq 1/4$.

- Function f_ω belongs to $\mathcal{F}_{K'} \subseteq \mathcal{F}_K$ since it admits a separable representation given in (53).

Thus, $\mathcal{G}' := \{f_\omega : \omega \in \{0,1\}^{K' \times H}\}$ is a subset of $\mathcal{G}_{K',\beta}$, and it suffices to prove the lower bound with the required rate on this subset. We first prove the lower bound in expectation (22). We use the version of Assouad's lemma given in (Tsybakov, 2009, Theorem 2.12(*iv*)). For any $\omega, \omega' \in \{0,1\}^{K' \times H}$ that only differ in one entry, that is, for exactly one (i_0, j_0), we have $\omega_{i_0,j_0} \neq \omega'_{i_0,j_0}$ and $\omega_{i,j} = \omega'_{i,j}$ for all other i, j, the χ^2-divergence between the densities f_ω and $f_{\omega'}$ is bounded as follows:

$$\chi^2 \left(f_{\omega'}, f_\omega \right) = \int \frac{\left(U_{i_0}(x) c_* L h_y^\beta V_{j_0,h_y}(y) \right)^2}{f_\omega(x,y)} \leq \int_{C_{i_0,j_0}} 2(c_* L h_y^\beta)^2 = 2(c_* L h_y^\beta)^2 h_x h_y \leq \frac{\widetilde{c}}{n},$$

where $\widetilde{c} > 0$ depends only on L, β. Thus, the χ^2-divergence between the corresponding product densities satisfies

$$\chi^2 \left(f_{\omega'}^{\otimes n}, f_\omega^{\otimes n} \right) = \left(1 + \chi^2 \left(f_{\omega'}, f_\omega \right) \right)^n - 1 \leq e^{\widetilde{c}} - 1,$$

cf. (Tsybakov, 2009, page 86). Moreover, the functions f_ω and $f_{\omega'}$ are separated in the L_1 norm as follows

$$\|f_\omega - f_{\omega'}\|_{L_1} = 2 c_* L h_y^\beta h_x \int \varphi\Big(\frac{y}{h_y/2} \Big) dy = c_* L h_x h_y^{\beta+1} \int \varphi(t) dt.$$

Therefore, applying (Tsybakov, 2009, Theorem 2.12(*iv*)) we obtain

$$\inf_{\widehat{f}} \sup_{f_\omega \in \mathcal{G}'} \mathbb{E} \|\widehat{f} - f_\omega\|_{L_1} \geq \frac{K'H}{2} (c_* L h_y^\beta) c'' h_x h_y = \frac{c_* c''}{2} L h_y^\beta \geq c \Big((K/n)^{\beta/(2\beta+1)} \wedge n^{-\beta/(2\beta+2)} \Big), \tag{54}$$

where $c'' = \exp(1 - e^{\widetilde{c}}) \int \varphi(x) dx$ and $c > 0$ is a constant depending only on L and β. This implies the lower bound in expectation (22).

To obtain the lower bound in probability (21) we apply Lemma 21 with $\mathcal{P} = \mathcal{P}_0 = \mathcal{G}'$, $v(f,g) = \|f - g\|_{L_1}$ for any $f, g \in \mathcal{G}'$, and $U = \mathbb{1}_{[0,1]^2} \in \mathcal{G}'$. Note that there exists a constant $C > 0$ such that for any $f \in \mathcal{G}'$ we have $v(U, f) \leq s$, where $s = C\big(\big(K/n\big)^{\beta/(2\beta+1)} \wedge n^{-\beta/(2\beta+2)} \big)$. Therefore, taking into account the bound in expectation (54) we can apply Lemma 21 with a small enough constant $a > 0$ to get (21). ■

C.6 Upper bound for the adaptive estimator

Lemma 18 *Let $0 < \delta < 1$. For any probability density f we have that, with $\mathbb{P}_f$-probability at least $1 - \delta$,*

$$\|\widehat{f}^* - f\|_{L_1} \le 3 \min_{(K,j)} \|\widehat{f}_{(K,j)} - f\|_{L_1} + 2\sqrt{\frac{2\log(2m^2/\delta)}{n}}, \tag{55}$$

and

$$\mathbb{E}_f\|\widehat{f}^* - f\|_{L_1} \le 3 \min_{(K,j)} \mathbb{E}_f\|\widehat{f}_{(K,j)} - f\|_{L_1} + 2\sqrt{\frac{2\log(2m^2)}{n}}. \tag{56}$$

Proof This lemma is a simple corollary of Theorem 6.3 in Devroye and Lugosi (2001), which states that

$$\|\widehat{f}^* - f\|_{L_1} \le 3 \min_{(K,j)} \|\widehat{f}_{(K,j)} - f\|_{L_1} + 4\Delta,$$

where $\Delta = \max_{B\in\mathcal{B}} \Big| \int_B f - \mathbb{P}_n(B)\Big|$. Since $|\mathcal{B}| = m(m-1)$ it follows from Hoeffding's inequality and the union bound that, for any $t > 0$,

$$\mathbb{P}_f(\Delta > t) \le 2m(m-1)\exp(-2nt^2).$$

This proves (55). Inequality (56) follows from the fact that $\mathbb{E}_f \max_{B\in\mathcal{B}} \Big| \int_B f - \mathbb{P}_n(B)\Big| \le \sqrt{\frac{\log(2|\mathcal{B}|)}{2n}}$ (see (Devroye and Lugosi, 2001, Lemma 2.2)). ■

Proof of Theorem 15.

To establish the adaptivity of $\widehat{f}^*$, we will use the inclusions between the Hölder classes:

$$\forall\, 0 < \beta \le \beta' \le 1: \quad \mathcal{L}_{\beta'} \subseteq \mathcal{L}_\beta.$$

Indeed, for any $0 < \beta \le \beta' \le 1$, any $f \in \mathcal{L}_{\beta'}$ and any $z, z' \in \mathrm{Supp}(f)$, we have $\|z - z'\|_\infty \le 1$, and consequently

$$|f(z) - f(z')| \le L\|z - z'\|_\infty^{\beta'} \le L\|z - z'\|_\infty^{\beta}.$$

The embedding of the classes $(\mathcal{G}^L_{K,\beta})_{\beta\in(0,1]}$ immediately follows:

$$\forall\, 0 < \beta \le \beta' \le 1,\ \forall K \in \mathbb{N}: \quad \mathcal{G}_{K,\beta'} \subseteq \mathcal{G}^L_{K,\beta}.$$

Assume first that $\beta \in [\beta_j, \beta_{j-1}]$ for some $j \in \{2, \dots, \lceil \log(n)\log\log(n)\rceil\}$. Using the inclusion $\mathcal{G}^L_{K,\beta} \subseteq \mathcal{G}^L_{K,\beta_j}$ and (56) we can bound from above the risk of $\widehat{f}^*$ over the class $\mathcal{G}^L_{K,\beta}$ as follows:

$$\begin{aligned}\sup_{f\in\mathcal{G}^L_{K,\beta}} \mathbb{E}_f\|\widehat{f}^* - f\|_{L_1} &\le \sup_{f\in\mathcal{G}^L_{K,\beta}} \left\{3 \min_{(K,i)} \mathbb{E}_f\|\widehat{f}_{(K,i)} - f\|_{L_1} + C\sqrt{\frac{\log(m)}{n}}\right\} \\ &\le \sup_{f\in\mathcal{G}^L_{K,\beta}} 3\,\mathbb{E}_f\|\widehat{f}_{(K,j)} - f\|_{L_1} + C\sqrt{\frac{\log\left(K_{\max}\log(n)\log\log(n)\right)}{n}}\end{aligned}$$

$$\leq \sup_{f\in\mathcal{G}^L_{K,\beta_j}} 3\,\mathbb{E}_f\|\widehat{f}_{(K,j)} - f\|_{L_1} + C\sqrt{\frac{\log(n)}{n}} \qquad \text{since } K_{\max} = \lceil\sqrt{n}\rceil$$

$$\leq C\left\{\left(\frac{K}{n}\right)^{\beta_j/(2\beta_j+1)} \log^{3/2} n \wedge n^{-\frac{\beta_j}{2\beta_j+2}}\right\} \qquad \text{by Theorem 12.} \quad (57)$$

Note that, for $c \in \{1, 2\}$,

$$\frac{\beta}{2\beta+c} - \frac{\beta_j}{2\beta_j+c} \leq \frac{c(\beta_{j-1}-\beta_j)}{(2\beta+c)(2\beta_j+c)} = \frac{c\beta_j}{(2\beta+c)(2\beta_j+c)\log(n)} \leq \frac{1}{\log(n)}.$$

Combining this with the bound (57) we obtain

$$\sup_{f\in\mathcal{G}^L_{K,\beta}} \mathbb{E}\|\widehat{f}^* - f\|_{L_1} \leq C\Big\{(n/K)^{-\beta/(2\beta+1)+\frac{1}{\log(n)}}(\log n)^{3/2} \wedge n^{-\frac{\beta}{2\beta+2}+\frac{1}{\log(n)}}\Big\}$$

$$\leq C\Big\{(n/K)^{-\beta/(2\beta+1)}(\log n)^{3/2} \wedge n^{-\frac{\beta}{2\beta+2}}\Big\}.$$

This proves the theorem for $\beta \in [\beta_M, \beta_1] = [\beta_M, 1]$, where $M = \lceil\log(n)\log\log(n)\rceil$. Finally, consider the values $\beta \in (0, \beta_M)$. We have

$$\frac{\beta}{2\beta+2} \leq \frac{\beta}{2\beta+1} \leq \frac{\beta_M}{2\beta_M+1} \leq \beta_M = \left(1+\frac{1}{\log(n)}\right)^{-M} \leq \exp\left(-\frac{\log(n)\log\log(n)}{\log(n)}\right) = \frac{1}{\log(n)}.$$

Thus, for $\beta \in (0, \beta_M)$,

$$(K/n)^{\beta/(2\beta+1)} \wedge n^{-\beta/(2\beta+2)} \geq n^{-\frac{1}{\log(n)}} = 1/e.$$

The desired result follows immediately from this inequality and the fact that $\sup_{f\in\mathcal{G}^L_{K,\beta}} \mathbb{E}_f\|\widehat{f}^* - f\|_{L_1} \leq 2$.

■

Appendix D. Auxiliary results

First, we recall the definition of the multinomial distribution. Given a finite set Ω and a probability distribution $P = (P_\omega)_{\omega\in\Omega}$ on Ω, we say that a random vector $Y = (Y_\omega)_{\omega\in\Omega}$ follows the multinomial distribution $\mathcal{M}(P, n)$, $n \in \mathbb{N}^*$, if for all $(n_\omega)_{\omega\in\Omega} \in \mathbb{N}^\Omega$ satisfying $\sum_{\omega\in\Omega} n_\omega = n$, we have:

$$\mathbb{P}\big(Y_\omega = n_\omega, \forall\omega\in\Omega\big) = \frac{n!}{\prod_{\omega\in\Omega} n_\omega!}\prod_{\omega\in\Omega} P_\omega^{n_\omega}.$$

We denote by $\mathrm{Poi}(\lambda)$ the Poisson distribution with mean λ.

Lemma 19 (Shorack and Wellner (2009), p. 486) *Let $\zeta \sim \mathrm{Poi}(\lambda)$ be a Poisson random variable with mean λ. Then*

$$\mathbb{P}(\zeta \leq \lambda - x) \leq \exp\Big(-x - (\lambda-x)\log\Big(\frac{\lambda-x}{\lambda}\Big)\Big), \quad \forall x \in (0,\lambda), \qquad (58)$$

$$\mathbb{P}(\zeta \geq \lambda + x) \leq \exp\Big(x - (\lambda+x)\log\Big(\frac{\lambda+x}{\lambda}\Big)\Big), \quad \forall x > 0. \qquad (59)$$

Lemma 20 (Poissonization) *Let $d \geq 2$, $\lambda > 0$, and let $p = (p_1, \ldots, p_d)$ be a probability distribution on $\Omega = \{1, \ldots, d\}$. Let $\widetilde{n} \sim \mathrm{Poi}(\lambda)$ and let $\widetilde{Y} = (\widetilde{Y}_1, \ldots, \widetilde{Y}_d)$ be a random vector such that $\widetilde{Y} \,|\, \widetilde{n} \sim \mathcal{M}(p, \widetilde{n})$. Then the entries $\widetilde{Y}_j$ are mutually independent and $\widetilde{Y}_j \sim \mathrm{Poi}(\lambda p_j)$ for all $j \in \{1, \ldots, d\}$.*

Proof For any $(n_1, \ldots, n_d) \in \mathbb{N}^d$ we have

$$\begin{aligned}
\mathbb{P}(\widetilde{Y}_1 = n_1, \ldots, \widetilde{Y}_d = n_d) &= \sum_{k=0}^{\infty} \mathbb{P}(\widetilde{n} = k)\ \mathbb{P}\big(\widetilde{Y}_1 = n_1, \ldots, \widetilde{Y}_d = n_d \mid \widetilde{n} = k\big) \\
&= \sum_{k=0}^{\infty} \Big(\frac{k!}{n_1! \cdots n_d!} \prod_{j=1}^{d} p_j^{n_j}\Big) \frac{e^{-\lambda}\lambda^k}{k!}\, \mathbb{1}_{n_1+\cdots+n_d=k} \\
&= \Big(\frac{\lambda^{n_1} \cdots \lambda^{n_d}}{n_1! \cdots n_d!} \prod_{j=1}^{d} p_j^{n_j}\Big) e^{-\lambda} \sum_{k=0}^{\infty} \mathbb{1}_{n_1+\cdots+n_d=k} \\
&= \prod_{j=1}^{d} \frac{e^{-\lambda p_j}(\lambda p_j)^{n_j}}{n_j!}.
\end{aligned}$$

■

Proof [Proof of Lemma 16] The elements of the sum $Z_j = \sum_{i \in V_j} Y_{ij}$ are not mutually independent. To overcome this difficulty, we use poissonization. Let $\widetilde{n} \sim \mathrm{Poi}(n/2)$ and let $\widetilde{Y} | \widetilde{n} \sim \mathcal{M}(P, \widetilde{n})$. We define $\widetilde{Z}_j = \sum_{i \in V_j} \widetilde{Y}_{ij}$ for any $j \in J$. Lemma 20 implies that $\widetilde{Y}_{ij} \sim \mathrm{Poi}(nP_{ij}/2)$ for all i, j, and the random variables $(\widetilde{Y}_{ij})_{i,j}$ are mutually independent. It follows that $\widetilde{Z}_j$ has a Poisson distribution: $\widetilde{Z}_j \sim \mathrm{Poi}(\sum_{i \in V_j} nP_{ij}/2)$, and the random variables $(\widetilde{Z}_j)_j$ are mutually independent. At the same time, for any $k \geq 1$ the conditional distribution of $\widetilde{Z}_j$ given $\widetilde{n} = k$ coincides with the distribution of Z_j under $Y \sim \mathcal{M}(P, k)$.

It follows from (58) with $x = \lambda/2$ and $\lambda = n\lambda_j/2$ that

$$\mathbb{P}\left(\widetilde{Z}_j \leq \frac{n\lambda_j}{4}\right) \leq \exp\Big(-\frac{n\lambda_j}{4}(1 - \log 2)\Big) \leq \exp\Big(-\frac{7}{2}(1 - \log 2)\alpha \log(N)\Big) < N^{-\alpha}.$$

Hence,

$$\mathbb{P}\left(\forall j \in J : \widetilde{Z}_j \geq \frac{n\lambda_j}{4}\right) \geq 1 - \frac{|J|}{N^{\alpha}}.$$

On the other hand,

$$\begin{aligned}
\mathbb{P}\left(\forall j \in J : \widetilde{Z}_j \geq \frac{n\lambda_j}{4}\right) &\leq \mathbb{P}\left(\widetilde{n} > n\right) + \mathbb{P}\left(\forall j \in J : \widetilde{Z}_j \geq \frac{n\lambda_j}{4}, \text{ and } \widetilde{n} \leq n\right) \\
&= \mathbb{P}\left(\widetilde{n} > n\right) + \sum_{k \leq n} \mathbb{P}\left(\forall j \in J : \widetilde{Z}_j \geq \frac{n\lambda_j}{4} \;\Big|\; \widetilde{n} = k\right) \mathbb{P}\left(\widetilde{n} = k\right) \\
&= \mathbb{P}\left(\widetilde{n} > n\right) + \sum_{k \leq n} \mathbb{P}_{Z \sim \mathcal{M}(P,k)}\left(\forall j \in J : Z_j \geq \frac{n\lambda_j}{4}\right) \mathbb{P}\left(\widetilde{n} = k\right)
\end{aligned}$$

$$\leq \mathbb{P}\left(\widetilde{n} > n\right) + \mathbb{P}_{Y\sim\mathcal{M}(P,n)}\left(\forall j \in J : Z_j \geq \frac{n\lambda_j}{4}\right), \tag{60}$$

where we have used the inequality $\mathbb{P}_{Y\sim\mathcal{M}(P,k)}\left(\forall j \in J : Z_j \geq \frac{n\lambda_j}{4}\right) \leq \mathbb{P}_{Y\sim\mathcal{M}(P,n)}\left(\forall j \in J : Z_j \geq \frac{n\lambda_j}{4}\right)$ that holds for all $k \leq n$ due to stochastic dominance since, under $Y \sim \mathcal{M}(P,k)$, each Y_{ij} has a binomial distribution with parameters (P_{ij}, k). It follows that

$$\mathbb{P}_{Y\sim\mathcal{M}(P,n)}\left(\forall j \in J : Z_j \geq \frac{n\lambda_j}{4}\right) \geq 1 - \frac{|J|}{N^\alpha} - \mathbb{P}\left(\widetilde{n} > n\right).$$

Applying (59) with $\zeta = \widetilde{n}$ and $\lambda = x = n/2$ we get

$$\mathbb{P}(\widetilde{n} > n) \leq \exp(n(1/2 - \log 2)) \leq \exp(14(1/2 - \log 2)\alpha \log(N)) \leq \frac{1}{N^\alpha}.$$

Combining the last two displays yields the lemma. ■

The following lemma, that may be of independent interest, provides a tool for deducing lower bounds in probability from lower bounds in expectation.

Lemma 21 (Deducing lower bound in probability from lower bound in expectation) *Let $\mathcal{P}_0$ be a metric space with metric $v : \mathcal{P}_0 \times \mathcal{P}_0 \to \mathbb{R}_+$, and let $\mathcal{P}$ be a subset of $\mathcal{P}_0$ with the property that there exists $U \in \mathcal{P}_0$ such that*

$$v(U, P) \leq s, \quad \forall P \in \mathcal{P}, \tag{61}$$

where $s > 0$. Let $\{\mathbb{P}_P, P \in \mathcal{P}\}$ be a family of probability measures indexed by $\mathcal{P}$. Assume that

$$\inf_{\widetilde{P}} \sup_{P\in\mathcal{P}} \mathbb{E}_P\, v(\widetilde{P}(Y), P) \geq as, \tag{62}$$

where $a > 0$, $\mathbb{E}_P$ denotes the expectation with respect to random variable Y distributed according to $\mathbb{P}_P$, and $\inf_{\widetilde{P}}$ is the infimum over all estimators $\widetilde{P}$ that take values in $\mathcal{P}_0$. Then

$$\inf_{\widetilde{P}} \sup_{p\in\mathcal{P}} \mathbb{P}_P\big(v(\widetilde{P}(Y), P) \geq as/2\big) \geq a/6.$$

Proof Consider the set $\mathcal{P}' = \{P' \in \mathcal{P}_0 :\ v(P', P) \leq 3s,\ \forall P \in \mathcal{P}\}$. This set is not empty since it contains U. Note that it is sufficient to consider estimators taking values in $\mathcal{P}'$, that is, for any estimator $\widetilde{P}$ there exists an estimator $\bar{P}$ with values in $\mathcal{P}'$ such that

$$v(\bar{P}, P) \leq v(\widetilde{P}, P), \quad \forall P \in \mathcal{P}. \tag{63}$$

In fact, let $\widetilde{P}$ be any estimator. Define another estimator

$$\bar{P} = \begin{cases} \widetilde{P} & \text{if } \ v(\widetilde{P}, U) \leq 2s, \\ U & \text{if } \ v(\widetilde{P}, U) > 2s. \end{cases}$$

Let us check that this estimator $\bar{P}$ satisfies (63) and $\bar{P} \in \mathcal{P}'$. Indeed, if $v(\widetilde{P}, U) \leq 2s$ then (63) obviously holds, and we have $v(\bar{P}, P) \leq v(\widetilde{P}, U) + v(P, U) \leq 3s$ for all $P \in \mathcal{P}$.

Otherwise, if $v(\widetilde{P}, U) > 2s$ then $v(\bar{P}, P) = v(U, P) \le 2s - s < v(\widetilde{P}, U) - v(U, P) \le v(\widetilde{P}, P)$ for all $P \in \mathcal{P}$.

It follows that

$$\inf_{\widetilde{P}} \sup_{p\in\mathcal{P}} \mathbb{E}_P\, v(\widetilde{P}, P) = \inf_{\widetilde{P}\in\mathcal{P}'} \sup_{p\in\mathcal{P}} \mathbb{E}_P\, v(\widetilde{P}, P),$$
$$\inf_{\widetilde{P}} \sup_{p\in\mathcal{P}} \mathbb{P}_P\big(v(\widetilde{P}, P) \ge as/2\big) = \inf_{\widetilde{P}\in\mathcal{P}'} \sup_{p\in\mathcal{P}} \mathbb{P}_P\big(v(\widetilde{P}, P) \ge as/2\big). \tag{64}$$

Thus, we have

$$\begin{aligned} as &\le \inf_{\widetilde{P}\in\mathcal{P}'} \sup_{p\in\mathcal{P}} \mathbb{E}_P\, v(\widetilde{P}, P) \\ &= \inf_{\widetilde{P}\in\mathcal{P}'} \sup_{p\in\mathcal{P}} \left\{ \mathbb{E}_P\left[v(\widetilde{P}, P)\ \mathbb{1}_{\left\{v(\widetilde{P},P)\ge as/2\right\}}\right] \ +\ \mathbb{E}_P\left[v(\widetilde{P}, P)\ \mathbb{1}_{\left\{v(\widetilde{P},P)< as/2\right\}}\right] \right\} \\ &\le 3s \inf_{\widetilde{P}\in\mathcal{P}'} \sup_{p\in\mathcal{P}} \mathbb{P}_P\big(v(\widetilde{P}, P) \ge as/2\big) + as/2, \end{aligned}$$

which together with (64) implies the lemma. ■

Lemma 22 *Let $Z_1, \dots, Z_m$ be iid random variables on a measurable space $(\mathcal{Z}, \mathcal{U})$. Let $\widehat{p}_j = \frac{1}{m}\sum_{i=1}^m \mathbb{1}_{Z_i\in A_j}$ and $p_j = \mathbb{P}(Z_1 \in A_j)$, where A_j, $j = 1, \dots, \ell$, are disjoint subsets of $\mathcal{X}$. Then, for any $\delta \in (0, 1)$,*

$$\mathbb{P}\left(\sum_{j=1}^{\ell} |p_j - \widehat{p}_j| > \sqrt{\frac{\ell}{m}} + \sqrt{\frac{2\log(1/\delta)}{m}}\right) \le \delta.$$

Proof Set $G(Z_1, \dots, Z_m) := \sum_{j=1}^{\ell} |p_j - \widehat{p}_j|$. We have

$$\mathbb{E}G(Z_1, \dots, Z_m) = \mathbb{E}\sum_{j=1}^{\ell} |p_j - \widehat{p}_j| \le \sum_{j=1}^{\ell} \Big(\mathbb{E}[|p_j - \widehat{p}_j|^2]\Big)^{1/2} \le \sum_{j=1}^{\ell} \sqrt{\frac{p_j}{m}} \le \sqrt{\frac{\ell}{m}}.$$

Note that $G(Z_1, \dots, Z_m)$ changes its value by at most $1/m$ if we replace any single Z_i by another Z_i'. Therefore, by the bounded difference inequality (see, e.g., (Devroye and Lugosi, 2001, Theorem 2.2)),

$$\mathbb{P}\left(G(Z_1, \dots, Z_m) > \mathbb{E}G(Z_1, \dots, Z_m) + t\right) \le e^{-2t^2 m}, \quad \forall t > 0,$$

which yields the result. ■

Lemma 23 (Control of multinomial noise) *Let $\alpha > 1$, $N > 1$, $n \in \mathbb{N}^*$, let $P \in \mathbb{R}_+^{d_1\times d_2}$ be a matrix such that $\sum_{i,j} P_{ij} = 1$, and let $Y \sim \mathcal{M}(P, n)$. Consider an extraction Q of P corresponding to two sets of indices $I \subseteq [d_1]$ and $J \subseteq [d_2]$, that is $Q = (P_{ij})_{i\in I, j\in J}$. Let $W_Q = (W_{ij})_{i\in I, j\in J}$, where $W = \frac{Y}{n} - P$ is the multinomial noise. Then*

$$\mathbb{P}\left(\|W_Q\|^2 \le 9\max\left\{\frac{\alpha\|Q\|_\square \log N}{n}, \left(\frac{\alpha\log N}{n}\right)^2\right\}\right) \ge 1 - \frac{|I| + |J|}{N^\alpha}.$$

Proof We use matrix Bernstein inequality (see, for example, (Vershynin, 2018, Exercise 5.4.15)), which yields that for any independent zero mean matrices $M_1, \dots, M_n \in \mathbb{R}^{|I|\times|J|}$ such that, almost surely, $\|M_i\| \le K, \forall i \in [n]$, we have

$$\mathbb{P}\left(\left\|\sum_{i=1}^n M_i\right\| > t\right) \le (|I|+|J|)\exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3}\right), \quad \forall t > 0, \tag{65}$$

where

$$\sigma^2 = \max\left(\left\|\sum_{i=1}^n \mathbb{E}(M_i M_i^\top)\right\|, \left\|\sum_{i=1}^n \mathbb{E}(M_i^\top M_i)\right\|\right).$$

We apply (65) with $M_i = \big(X_i - P\big)_{I,J}$, where X_i's are independent random matrices with distribution $\mathcal{M}(P,1)$. In this case, the inequality $\|M_i\| \le K$ holds with $K = 2$. To prove this, notice first that if Q is an extraction of a probability matrix P, then $\|Q\|^2 \le \big\|Q\big\|_\square$. Indeed, letting $p_i = \sum_{j\in J} P_{ij},\ \forall i \in I$, and $p_{\max} = \max_{i\in I} p_i$, we obtain

$$\|Q\|^2 \le \|Q\|_F^2 = \sum_{i\in I, j\in J} P_{ij}^2 \le \sum_{i\in I} p_i^2 \le p_{\max}\sum_{i\in I} p_i \le p_{\max} \le \big\|Q\big\|_\square.$$

It follows that almost surely for all $i \in [n]$ we have

$$\|M_i\| \le \|(X_i)_{IJ}\| + \|Q\| \le 1 + \sqrt{\big\|P\big\|_\square} \le 2.$$

Next,

$$\begin{aligned}\left\|\sum_{i=1}^n \mathbb{E}(M_i M_i^\top)\right\| &= \left\|\sum_{i=1}^n \mathbb{E}[(X_i)_{IJ}(X_i)_{IJ}^\top] - QQ^\top\right\| = \Big\|n\,\mathrm{Diag}(p_k)_{k\in I} - QQ^\top\Big\| \\ &\le np_{\max} \le n\big\|Q\big\|_\square,\end{aligned}$$

and controlling $\left\|\sum_{i=1}^n \mathbb{E}(M_i^\top M_i)\right\|$ analogously yields

$$\sigma^2 \le n\big\|Q\big\|_\square.$$

Now, we use the fact that nW_Q has the same distribution as $\sum_{i=1}^n M_i$. Therefore, applying inequality (65) and the above bounds on σ^2 and K we obtain that, for all $t > 0$,

$$\mathbb{P}\left(\|nW_Q\| > t\right) \le (|I|+|J|)\exp\left(-\frac{t^2/2}{n\big\|Q\big\|_\square + 2t/3}\right).$$

Now, set

$$t = 3\max\Big(\sqrt{\alpha n\big\|Q\big\|_\square \log(N)},\ \alpha\log(N)\Big).$$

If $n\|Q\|_\square \geq \alpha \log(N)$ then $t = 3\sqrt{\alpha n \|Q\|_\square \log(N)} \leq 3n\|Q\|_\square$ and

$$\mathbb{P}\left(\|nW_Q\| > t\right) \leq (|I|+|J|)\exp\left(-\frac{t^2/2}{3n\|Q\|_\square}\right) \leq (|I|+|J|)\exp\left(-\frac{3}{2}\alpha\log(N)\right) \leq (|I|+|J|)N^{-\alpha}.$$

Otherwise, if $n\|Q\|_\square < \alpha\log(N)$ then $t = 3\alpha\log(N)$ and the same bound holds:

$$\mathbb{P}\left(\|nW_Q\| > t\right) \leq (|I|+|J|)\exp\left(-\frac{t^2/2}{3\alpha\log(N)}\right) \leq (|I|+|J|)N^{-\alpha}.$$

■

Lemma 24 (Deterministic bound for nuclear–norm penalized denoising) *Let* $Y = M + W \in \mathbb{R}^{d_1\times d_2}$ *with* $\operatorname{rank}(M) = r$*, and fix* $\tau > 0$*. Consider*

$$\widehat{M} \in \arg\min_{A\in\mathbb{R}^{d_1\times d_2}} \left\{ \|Y - A\|_F^2 + \tau\|A\|_* \right\}.$$

If $\|W\| \leq \tau/2$*, then every minimizer* $\widehat{M}$ *satisfies*

$$\|\widehat{M} - M\|_F^2 \;\leq\; 8\,r\,\tau^2.$$

Proof Set $\Delta := \widehat{M} - M$. By optimality of $\widehat{M}$,

$$\|W - \Delta\|_F^2 + \tau\|\widehat{M}\|_* \;\leq\; \|W\|_F^2 + \tau\|M\|_*, \tag{66}$$

which rearranges to the basic inequality

$$\|\Delta\|_F^2 \;\leq\; 2\langle W, \Delta\rangle \;-\; \tau\big(\|M+\Delta\|_* - \|M\|_*\big). \tag{67}$$

Let $M = U\Sigma V^\top$ be a compact SVD with $\operatorname{rank}(M) = r$. Define the tangent space

$$T := \{\, UA^\top + BV^\top :\; A \in \mathbb{R}^{d_2\times r},\; B \in \mathbb{R}^{d_1\times r} \,\},$$

and decompose $\Delta = \Delta_T + \Delta_{T^\perp}$.

(i) Decomposability of the nuclear norm. For any matrix B,

$$\|M + B\|_* - \|M\|_* \;\geq\; \|B_{T^\perp}\|_* - \|B_T\|_*. \tag{68}$$

Proof of (68). Take $G \in \partial\|\cdot\|_*(M)$ of the form $G = UV^\top + Z$ with $Z \in T^\perp$ and $\|Z\| \leq 1$. Choosing Z so that $\langle Z, B_{T^\perp}\rangle = \|B_{T^\perp}\|_*$ gives

$$\|M+B\|_* \;\geq\; \|M\|_* + \langle G, B\rangle = \|M\|_* + \langle UV^\top, B_T\rangle + \|B_{T^\perp}\|_* \;\geq\; \|M\|_* - \|B_T\|_* + \|B_{T^\perp}\|_*.$$

This proves (68).

(ii) Stochastic term. By duality of operator and nuclear norms and the assumption $\|W\| \le \tau/2$,

$$2\langle W, \Delta\rangle \;\le\; 2\|W\|\,\|\Delta\|_* \;\le\; \tau\big(\|\Delta_T\|_* + \|\Delta_{T^\perp}\|_*\big). \tag{69}$$

(iii) Combine (67)–(69)**.** Plugging (68) and (69) into (67) yields

$$\|\Delta\|_F^2 \;\le\; \tau\big(\|\Delta_T\|_* + \|\Delta_{T^\perp}\|_*\big) - \tau\big(\|\Delta_{T^\perp}\|_* - \|\Delta_T\|_*\big) \;=\; 2\tau\,\|\Delta_T\|_*.$$

Since $\operatorname{rank}(\Delta_T) \le 2r$, we have $\|\Delta_T\|_* \le \sqrt{2r}\,\|\Delta_T\|_F \le \sqrt{2r}\,\|\Delta\|_F$, and therefore

$$\|\Delta\|_F^2 \;\le\; 2\tau\sqrt{2r}\,\|\Delta\|_F \quad\Longrightarrow\quad \|\Delta\|_F \;\le\; 2\sqrt{2r}\,\tau.$$

Squaring both sides gives $\|\widehat{M} - M\|_F^2 = \|\Delta\|_F^2 \le 8\,r\,\tau^2$, as claimed. ∎

References

Emmanuel Abbe. Community detection and stochastic block models: Recent developments. Journal of Machine Learning Research, 18(177):1–86, 2018. URL `http://jmlr.org/papers/v18/16-480.html`.

Umberto Amato, Anestis Antoniadis, Alexander Samarov, and Alexandre B. Tsybakov. Noisy independent factor analysis model for density estimation and classification. Electronic Journal of Statistics, 4:707 – 736, 2010. doi: 10.1214/09-EJS498. URL `https://doi.org/10.1214/09-EJS498`.

Magda Amiridi, Nikos Kargas, and Nicholas D. Sidiropoulos. Low-rank characteristic tensor density estimation part II: Compression and latent density estimation. IEEE Transactions on Signal Processing, 70:2669–2680, 2022a. doi: 10.1109/TSP.2022.3158422.

Magda Amiridi, Nikos Kargas, and Nicholas D. Sidiropoulos. Low-rank characteristic tensor density estimation part I: Foundations. IEEE Transactions on Signal Processing, 70:2654–2668, 2022b. doi: 10.1109/TSP.2022.3175608.

Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. Found. Comput. Math., 9(6):717–772, 2009. URL `http://dblp.uni-trier.de/db/journals/focm/focm9.html#CandesR09`.

Clément L. Canonne, Ilias Diakonikolas, Themis Gouleakis, and Ronitt Rubinfeld. Testing shape restrictions of discrete distributions. Theory of Computing Systems, 62(1):4–62, 2018.

Julien Chhor and Alexandra Carpentier. Goodness-of-fit testing for Hölder-continuous densities: Sharp local minimax rates. arXiv preprint arXiv:2109.04346, 2021.

Luc Devroye and Laszlo Györfi. Nonparametric Density Estimation. The L_1 View. J. Wiley, 1985.

Luc Devroye and Gábor Lugosi. Combinatorial Methods in Density Estimation. Springer, 2001. ISBN 978-0-387-95117-1.

Ilias Diakonikolas, Moritz Hardt, and Ludwig Schmidt. Differentially private learning of structured discrete distributions. Advances in Neural Information Processing Systems, 28, 2015.

Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. SIAM Journal on Computing, 48(2):742–864, 2019.

David B Dunson and Chuanhua Xing. Nonparametric bayes modeling of multivariate categorical data. Journal of the American Statistical Association, 104(487):1042–1051, 2009.

Cécile Durot, Sylvie Huet, François Koladjo, and Stéphane Robin. Least-squares estimation of a convex discrete distribution. Computational Statistics & Data Analysis, 67:282–298, 2013.

Jerome H. Friedman, Werner Stuetzle, and Anne Schroeder. Projection pursuit density estimation. Journal of the American Statistical Association, 79(387):599–608, 1984.

Christophe Giraud. Introduction to High-dimensional Statistics. CRC Press, 2021.

Alexander Goldenshluger and Oleg Lepski. On adaptive minimax density estimation on R^d. Probability Theory and Related Fields, 159(3):479–543., 2014. doi: 10.1007/s00440-013-0512-1. URL `https://hal.science/hal-01265245`.

Artur Gramacki. Nonparametric Kernel Density Estimation and Its Computational Aspects. Springer, 2017.

Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Minimax estimation of discrete distributions under ℓ_1 loss. IEEE Transactions on Information Theory, 61(11):6343–6354, 2015.

Thomas Hofmann. Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pages 50–57, 1999.

Paul Holland, Kathryn B. Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. Social Networks, 5:109–137, 1983. URL `https://api.semanticscholar.org/CorpusID:34098453`.

Ildar A. Ibragimov and Rafail Z. Khas'minskii. More on the estimation of distribution densities. Journal of Soviet Mathematics, 25(3):1155–1165, 1984.

Ayush Jain and Alon Orlitsky. Linear-sample learning of low-rank distributions. In Proceedings of the 34th International Conference on Neural Information Processing Systems, NeurIPS'20, pages 19201–19211, 2020. ISBN 9781713829546.

James E Johndrow, Anirban Bhattacharya, and David B Dunson. Tensor decompositions and sparse log-linear models. Annals of Statistics, 45(1):1–38, 2017.

Anatoli Juditsky and Sophie Lambert-Lacroix. On minimax density estimation on $\mathbb{R}$. Bernoulli, 10(2):187–220, 2004.

Sudeep Kamath, Alon Orlitsky, Dheeraj Pichapati, and Ananda Theertha Suresh. On learning distributions from their samples. In Conference on Learning Theory, pages 1066–1100. PMLR, 2015.

Nikos Kargas and Nicholas D. Sidiropoulos. Learning mixtures of smooth product distributions: Identifiability and algorithm. In Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, pages 388–396. PMLR, 2019. URL https://proceedings.mlr.press/v89/kargas19a.html.

Zheng Tracy Ke and Minzhe Wang. Using SVD for topic modeling. Journal of the American Statistical Association, pages 1–16, 2022. doi: 10.1080/01621459.2022.2123813.

Jussi Sakari Klemelä. Smoothing of Multivariate Data: Density Estimation and Visualization. John Wiley & Sons, 2009.

Oleg Lepski and Gilles Rebelles. Structural adaptation in the density model. Mathematical Statistics and Learning, 3:345–386, 2020.

Jasson D. M. Rennie and Nathan Srebro. Fast maximum margin matrix factorization for collaborative prediction. In Proceedings of the 22nd International Conference on Machine Learning, ICML '05, page 713–719, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595931805. doi: 10.1145/1102351.1102441. URL https://doi.org/10.1145/1102351.1102441.

Alexander Samarov and Alexandre B. Tsybakov. Nonparametric independent component analysis. Bernoulli, 10(4):565–582, 2004.

Alexander Samarov and Alexandre B. Tsybakov. Aggregation of density estimators and dimension reduction. In Advances in Statistical Modeling and Inference. Essays in Honor of Kjell A. Doksum, pages 233–251. World Scientific, Singapore e.a.,, 2007. ISBN 978-981-270-369-9. doi: 10.1142/9789812708298_0012.

David W. Scott. Multivariate Density Estimation: Theory, Practice, and Visualization. Wiley, 1992.

Galen R Shorack and Jon A Wellner. Empirical Processes with Applications to Statistics. SIAM, 2009.

Bernard W. Silverman. Density Estimation for Statistics and Data Analysis. Chapman & Hall, London, 1986.

Le Song and Bo Dai. Robust low rank kernel embeddings of multivariate distributions. In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS'13, pages 3228–3236, 2013.

Le Song, Animashree Anandkumar, Bo Dai, and Bo Xie. Nonparametric estimation of multi-view latent variable models. In Proceedings of the 31st International Conference on Machine Learning, pages 640–648. PMLR, 2014. URL https://proceedings.mlr.press/v32/songa14.html.

Behrooz Tahmasebi, Seyed Abolfazl Motahari, and Mohammad Ali Maddah-Ali. On the identifiability of finite mixtures of finite product measures. arXiv preprint arXiv:1807.05444, 2018.

Alexandre B Tsybakov. Introduction to Nonparametric Estimation. Springer, 2009.

Robert A. Vandermeulen. Sample complexity using infinite multiview models. arXiv preprint arXiv:2302.04292, 2023.

Robert A. Vandermeulen and Antoine Ledent. Beyond smoothness: Incorporating low-rank analysis into nonparametric density estimation. Advances in Neural Information Processing Systems, 34:12180–12193, 2021.

Roman Vershynin. High-dimensional Probability: An Introduction with Applications in Data Science. Cambridge University Press, 2018.

Zhipeng Wang and David W Scott. Nonparametric density estimation for high-dimensional data—algorithms and applications. Wiley Interdisciplinary Reviews: Computational Statistics, 11(4):e1461, 2019.