Electronic Journal of Statistics
Vol. 14 (2020) 2600–2652
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1729

Optimal rates for estimation of two-dimensional totally positive distributions

Jan-Christian Hütter[1], **Cheng Mao**[2], **Philippe Rigollet**[3] **and Elina Robeva**[4]

[1] *Broad Institute, 415 Main Street, Cambridge, MA, 02142, USA. e-mail:* jhuetter@broadinstitute.org

[2] *School of Mathematics, Georgia Institute of Technology, Suite 117, 686 Cherry Street, Atlanta, GA, 30332-0160, USA. e-mail:* cheng.mao@math.gatech.edu

[3] *Department of Mathematics, Massachusetts Institute of Technology, Building 2, Room 106, 77 Massachusetts Avenue, Cambridge, MA, 02139-4307, USA. e-mail:* rigollet@math.mit.edu

[4] *Department of Mathematics, University of British Columbia, Room 121, 1984 Mathematics Road, Vancouver, BC, V6T 1Z2, Canada. e-mail:* erobeva@math.ubc.ca

Abstract: We study minimax estimation of two-dimensional totally positive distributions. Such distributions pertain to pairs of strongly positively dependent random variables and appear frequently in statistics and probability. In particular, for distributions with β-Hölder smooth densities where $\beta \in (0, 2)$, we observe polynomially faster minimax rates of estimation when, additionally, the total positivity condition is imposed. Moreover, we demonstrate fast algorithms to compute the proposed estimators and corroborate the theoretical rates of estimation by simulation studies.

MSC 2010 subject classifications: Primary 62G05, 62G07.
Keywords and phrases: Totally positive distributions, nonparametric density estimation, shape-constrained estimation.

Received March 2020.

1. Introduction

For a set $\mathcal{X} = \prod_{i=1}^{d} \mathcal{X}_i$ where each $\mathcal{X}_i$ is totally ordered,[1] a function $p : \mathcal{X} \to \mathbb{R}$ is called *multivariate totally positive of order 2* (MTP_2) [22, 24] if

$$p(x \wedge y)p(x \vee y) \geq p(x)p(y)\,, \qquad \forall\, x, y \in \mathcal{X}, \tag{1.1}$$

where $\wedge$ and $\vee$ denote the coordinate-wise min and max operators respectively. The MTP_2 condition is also known as the *FKG lattice condition* because of its central role in the FKG inequality [14]. It is sometimes referred to as

[1] A set $\mathcal{X}$ is totally ordered if it is equipped with a total order, that is, a binary relation which is antisymmetric, transitive and connex. This work is only concerned with $\mathcal{X} \subseteq \mathbb{R}$ equipped with its natural order.

log-supermodularity because of its similarity up to morphism to supermodularity [18, 37]. Throughout, we say that a probability distribution is MTP_2 if it has an MTP_2 density.

A variety of joint distributions are known to be MTP_2, for example, order statistics of i.i.d. variables, eigenvalues of Wishart matrices [24], and ferromagnetic Ising models [28]. Furthermore, Gaussian and binary latent tree models are signed MTP_2, that is, there exists a sign change of each coordinate making the distribution MTP_2 [25, 27]. In particular, all these distributions exhibit positive association, a marked feature of MTP_2 distributions. As opposed to positive association, however, the MTP_2 property is preserved after conditioning or marginalization [24]. As a result of their frequent appearances, MTP_2 distributions have long been studied in statistics and probability [24, 26, 2, 8, 44, 27, 38].

In this paper, we study minimax estimation of an MTP_2 distribution in dimension two[2] from i.i.d. observations. We mainly focus on distributions on the square $[0,1]^2$ for which density functions exist. Since almost surely no four observations from such a distribution form a rectangle, the MTP_2 constraint (1.1) is inactive on the observations and consequently, the maximum likelihood estimator over this class does not exist (see Lemma 14 and Remark 15 in Appendix A). Therefore, we further assume that the distribution has a β-Hölder smooth density, a widely adopted assumption in nonparametric estimation [46].

Smooth MTP_2 distributions have long been studied in the literature. Examples include, but are not limited to, (1) pairwise marginals of Gaussian latent tree models [13], such as Brownian motion tree models and factor analysis models, (2) joint distributions of pairs of time points of a strong Markov process on the real line with continuous paths [23], such as a diffusion process, and (3) MTP_2 transelliptical distributions, such as MTP_2 multivariate t-distributions, which are commonly used in finance [1].

Main contribution Our main results can be stated informally as follows.

Theorem 1 (Informal statement of minimax rates)**.** *Given N i.i.d. observations from a two-dimensional distribution with an* MTP_2 *and β-Hölder smooth density, the minimax rate of estimation in the squared Hellinger distance is (up to a polylogarithmic factor)*

$$\begin{cases} N^{-\frac{2\beta}{2\beta+1}}, & \text{if } 0.5 \le \beta < 1, \\ N^{-2/3}, & \text{if } 1 \le \beta < 2, \\ N^{-\frac{2\beta}{2\beta+2}}, & \text{if } \beta \ge 2. \end{cases}$$

It is well known that without the MTP_2 assumption, the minimax rates for the β-Hölder class in dimension d scale as $N^{-\frac{2\beta}{2\beta+d}}$ up to a polylogarithmic factor, under various comparable models and error metrics (see, for example, [33]). Hence, our results show that for $0.5 \le \beta < 1$, the minimax rate exhibits a one-dimensional behavior thanks to the MTP_2 constraint; for $1 \le \beta < 2$,

[2]In dimension two, MTP_2 is sometimes simply called TP_2 for *totally positive of order 2*.

the rate is polynomially faster than that without the MTP_2 constraint and is independent of the smoothness parameter β; for $\beta > 2$, however, the MTP_2 constraint has no effect on the minimax rate (see Figure 1 for a visualization).

Note that results similar to what we obtain for MTP_2 are expected to arise when p is assumed to be smooth and log-concave. However, MTP_2 only makes assumptions on the behavior of the function along lattice directions. While this is not directly comparable with log-concavity, it is, in essence, a weaker condition in the sense that it imposes a less stringent structure on the density. Our results indicate that when coupled with smoothness, MTP_2 makes up for this deficiency and leads to the same rates of convergence.

Our results for the regime $0 < \beta < 0.5$ are unfortunately inconclusive, but the upper bounds exhibit polynomial improvement in the rates when MTP_2 is assumed; see (3.8) below.

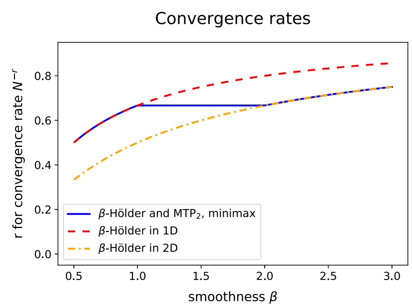


FIG 1. *Visual comparison of the estimation rate for β-Hölder smooth* MTP_2 *distributions in Theorem 1, with estimation rates for β-Hölder smooth distributions (without the* MTP_2 *constraint) in 1D and 2D, suppressing logarithmic factors.*

As a stepping stone to this problem, we also consider the following discrete version of MTP_2. A distribution on the grid $[n_1] \times [n_2]$ is MTP_2 if its probability mass function (PMF) p, which is an $n_1 \times n_2$ matrix, fulfills (1.1). Thus, MTP_2 says that all the 2×2 minors of p are non-negative:

$$p_{ij} p_{k\ell} \geq p_{i\ell} p_{kj}, \qquad \text{for all } 1 \leq i < k \leq n_1, 1 \leq j < \ell \leq n_2. \tag{1.2}$$

We study estimation of the PMF p from N independent observations in this discrete setup.

To obtain upper bounds for estimation of a discrete MTP_2 distribution, we employ a variant of the maximum likelihood estimator (MLE) defined in Section 2.1. For estimating a smooth MTP_2 density, we first discretize the space $[0, 1]^2$ and then apply the discrete MLE to obtain an estimator (defined in Section 3.1) that achieves near-optimal upper bounds. Both estimators are computationally efficient, with the implementations discussed in more detail in Section 4.

Related work There has been a recent surge of interest in the estimation of MTP_2 distributions. The special case of Gaussian MTP_2 distributions has been studied by [44, 27] from the perspective of maximum likelihood estimation and optimization. Maximum likelihood estimation of *log-concave* MTP_2 distributions was also analyzed recently [39, 38]. However, no statistical rate of estimation of MTP_2 distributions is currently known. The present paper establishes the first minimax rates (up to logarithmic factors) of estimation of a smooth MTP_2 density.

More broadly, our work falls into the scope of nonparametric density estimation which is a fundamental problem in nonparametric estimation. As such it has received considerable attention over the years [20, 41, 49, 7, 43]. A central paradigm in this literature is to assume smoothness of the underlying density to be estimated. Such an assumption justifies a variety of statistical methods ranging from kernel density estimation to series expansions. Another approach to nonparametric estimation, and in particular to density estimation, is to use shape constraints whereby the (local) smoothness assumption is dropped and favored by a (global) synthetic constraint such as monotonicity [15, 36], convexity [17, 42] and log-concavity [48, 11, 9, 40] (see [16] for a recent overview). As explained above, the MTP_2 constraint alone does not make the density estimation problem well-defined and it has been combined with another shape constraint, namely log-concavity, in [38]. Instead, the present work combines MTP_2 with smoothness to obtain a faster statistical rate than with smoothness alone, thus demonstrating compatibility of the local and the global approach.

As we have discussed above, MTP_2 is also called log-supermodular. In the recent paper [19], we studied estimation of supermodular matrices (also known as *anti-Monge* matrices) under sub-Gaussian noise. We note that the proof techniques used in [19] are the starting point for the proofs in this paper, but are extended to the context density estimation. In a parallel work [12], the authors study a related but slightly different model under Gaussian noise, and their proof techniques could potentially be extended to yield rates similar to the ones found in this paper.

Organization We present the main results of the paper: upper and lower bounds for the discrete case in Section 2, followed by the continuous case in Section 3. All proofs are postponed to Section 6. The implementation of our estimators is discussed in Section 4. Our theoretical results are complemented by numerical experiments on synthetic data in Section 5. Finally, Section 7 includes a conclusion of the paper and a discussion of questions left for future research.

Notation For a positive integer n, let $[n] = \{1, 2, \dots, n\}$. For a finite set S, we use $|S|$ to denote its cardinality. For two sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$ of real numbers, we write $a_n \lesssim b_n$ if there is a universal constant $C > 0$ such that $a_n \le C b_n$ for all $n \ge 1$. The relation $a_n \gtrsim b_n$ is defined analogously. We use c and C (possibly with subscripts) to denote universal positive constants that may change from line to line. Given a matrix $M \in \mathbb{R}^{n_1 \times n_2}$, we denote its ith row by $M_{i,\cdot}$ and its jth column by $M_{\cdot,j}$. For an entrywise positive vector

$w \in \mathbb{R}^n$ and a vector $v \in \mathbb{R}^n$, we use the notation

$$\|v\|_w := \Big(\sum_{i=1}^n w_i v_i^2\Big)^{1/2}$$

for the w-weighted ℓ_2 norm of the vector v. Similarly, for an entrywise positive matrix $b \in \mathbb{R}^{n_1 \times n_2}$ and a matrix $a \in \mathbb{R}^{n_1 \times n_2}$, we use $\|a\|_b$ to denote the b-weighted Frobenius norm of a. For a reference measure μ on a (continuous or discrete) space $\mathcal{X}$, and two distributions with probability density or mass functions p and q respectively, we let

$$\begin{aligned} \mathsf{h}(p,q) &:= \Big(\int_{\mathcal{X}} \big(\sqrt{p(x)} - \sqrt{q(x)}\,\big)^2 d\mu(x)\Big)^{1/2} \qquad \text{and} \\ \mathsf{KL}(p,q) &:= \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} d\mu(x) \end{aligned}$$

denote the Hellinger distance and the Kullback-Leibler (KL) divergence between the two distributions respectively.

2. $\mathbf{MTP_2}$ distribution estimation on a grid

Let p^* be a probability mass function (PMF) on the grid $[n_1] \times [n_2]$, where we assume without loss of generality that $n_1 \geq n_2$. In the case where $n_1 \leq n_2$, our results and proofs remain valid with the roles of n_1 and n_2 swapped. Suppose that p^* satisfies the MTP_2 condition

$$p^*_{i,j} p^*_{i+1,j+1} \geq p^*_{i,j+1} p^*_{i+1,j} \qquad \text{for all } i \in [n_1 - 1],\ j \in [n_2 - 1]. \tag{2.1}$$

Note that this is equivalent to condition (1.2) by a telescoping sum argument.

Suppose that we are given N i.i.d. observations $\{Z_k\}_{k=1}^N$ from the distribution on $[n_1] \times [n_2]$ with PMF p^*, that is, each $Z_k = (i,j)$ with probability $p^*_{i,j}$ for $(i,j) \in [n_1] \times [n_2]$. Our goal is to estimate p^*. The number of observations at each point (i,j) on the grid $[n_1] \times [n_2]$ is recorded in a matrix $Y = (Y_{i,j})_{i \in [n_1],\, j \in [n_2]}$, defined by

$$Y_{i,j} := \sum_{k=1}^N \mathbb{1}\{Z_k = (i,j)\}. \tag{2.2}$$

Then Y can be viewed as a multinomial random variable with distribution denoted by $\mathsf{Multi}(N, p^*)$.

In addition, we define

$$p^*_{\mathsf{min}} := \min_{i \in [n_1],\, j \in [n_2]} p^*_{i,j} \quad \text{and} \quad p^*_{\mathsf{max}} := \max_{i \in [n_1],\, j \in [n_2]} p^*_{i,j},$$

and assume a mild lower bound on the sample size $N \geq 12 \log(n_1 n_2 / \delta) / p^*_{\mathsf{min}}$. Then $Y_{i,j}$ concentrates around its expectation as indicated by the next lemma. In particular, we have sufficiently many observations per entry on the grid with high probability.

Lemma 2. *For any $\delta \in (0, 1/2]$ and $N \geq 12 \log(n_1 n_2/\delta)/p^*_{\mathsf{min}}$, it holds with probability at least $1 - 2\delta$ that*

$$\frac{1}{2} N p^*_{i,j} \leq Y_{i,j} \leq \frac{3}{2} N p^*_{i,j}$$

for all $(i,j) \in [n_1] \times [n_2]$.

Proof. Note that marginally $Y_{i,j}$ follows the binomial distribution $\mathsf{Bin}(N, p^*_{i,j})$. Hence the result is an immediate consequence of Lemma 16 with $q = p^*_{i,j}/2$, together with a union bound over $(i,j) \in [n_1] \times [n_2]$. □

2.1. Estimator

We begin by describing the MLE of the log-PMF $\theta^* \in (-\infty, 0]^{n_1 \times n_2}$ defined by $\theta^*_{i,j} := \log p^*_{i,j}$. Owing to the fact that p^* is a totally positive PMF, θ^* satisfies the following two constraints:

$$\sum_{i \in [n_1],\, j \in [n_2]} e^{\theta^*_{i,j}} = 1\,, \quad \text{and} \quad D\theta^* \tilde{D}^\top \geq 0\,,$$

where the symbol $\geq$ denotes entrywise inequality and the difference operators $D \in \mathbb{R}^{(n_1-1) \times n_1}$, $\tilde{D} \in \mathbb{R}^{(n_2-1) \times n_2}$ are both of the form

$$\begin{bmatrix} -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ \vdots & & & & \vdots & \\ 0 & 0 & 0 & \dots & -1 & 1 \end{bmatrix}. \tag{2.3}$$

The log-likelihood of a candidate $\theta = \log(p) \in (-\infty, 0]^{n_1 \times n_2}$ is given by

$$\log \prod_{k=1}^{N} p_{Z_k} = \log \prod_{i \in [n_1],\, j \in [n_2]} (p_{i,j})^{Y_{i,j}} = \sum_{i \in [n_1],\, j \in [n_2]} Y_{i,j} \theta_{i,j} = \langle Y, \theta \rangle.$$

Hence the MLE is given by

$$\hat{\theta}^{\mathsf{MLE}} := \operatorname*{argmax}_{\substack{\sum_{i,j} e^{\theta_{i,j}} = 1 \\ D\theta\tilde{D}^\top \geq 0}} \langle Y, \theta \rangle. \tag{2.4}$$

Instead of the MLE, we study a constrained variant which is both amenable to analysis and efficiently computable.[3] Lemma 2 implies that with probability at least $1 - 2\delta$, the true log-PMF θ^* lies in the cube

$$\mathcal{C}(Y) := \Big\{\theta \in (-\infty, 0]^{n_1 \times n_2} : \log \frac{2Y_{i,j}}{3N} \leq \theta_{i,j} \leq \log \frac{2Y_{i,j}}{N} \text{ for all } i \in [n_1], j \in [n_2]\Big\}. \tag{2.5}$$

[3]The MLE itself can also be efficiently computed; see Appendix B.2 and Section 5.

This motivates the constrained optimization problem

$$\tilde{\theta} := \underset{\substack{D\theta\tilde{D}^\top \geq 0 \\ \theta \in \mathcal{C}(Y)}}{\operatorname{argmax}} \frac{1}{N}\langle Y, \theta\rangle - \sum_{i\in[n_1],\, j\in[n_2]} e^{\theta_{i,j}}. \tag{2.6}$$

Note that the objective is concave and there are $O(n_1 n_2)$ inequality constraints, so the program can be solved efficiently. However, the constraint $\sum_{i,j} e^{\theta_{i,j}} = 1$ is replaced by a penalty term, so it is not necessarily true that $\sum_{i,j} e^{\tilde{\theta}_{i,j}} = 1$. Hence we define the estimator of interest $\hat{\theta} \in \mathbb{R}^{n_1 \times n_2}$ by normalizing $\tilde{\theta}$:

$$\hat{\theta}_{i,j} := \tilde{\theta}_{i,j} - \log \sum_{r\in[n_1],\, s\in[n_2]} e^{\tilde{\theta}_{r,s}} \quad \text{for } i \in [n_1],\, j \in [n_2]. \tag{2.7}$$

It is clear then that $\hat{\theta}$ is a supermodular log-PMF. Finally, we define our estimator $\hat{p} = \hat{p}(Y)$ by $\hat{p}_{i,j} := e^{\hat{\theta}_{i,j}}$, which is therefore a properly defined MTP_2 PMF.

2.2. Upper and lower bounds

We measure the performance of our estimator $\hat{p}$ using the Hellinger distance $\mathsf{h}(p^*, \hat{p})$. For any PMF p on the grid, define

$$L(p) := \frac{p_{1,1} p_{n_1,n_2}}{p_{n_1,1} p_{1,n_2}}. \tag{2.8}$$

The quantity $\log\big(L(p)\big)$ is a seminorm of the log-PMF $\theta = \log(p)$ (see (6.17)), which measures the complexity of θ. As a result, the following upper bound for our estimator $\hat{p}$ depends on $\log\big(L(p^*)\big)$.

Theorem 3 (Upper bounds for estimation of discrete MTP_2 distributions)**.** *Fix $\delta \in (0, 1/4]$ and suppose that we are given $N \geq 12\log(n_1 n_2/\delta)/p^*_{\mathsf{min}}$ independent observations from a distribution with an MTP_2 PMF p^* on the grid $[n_1] \times [n_2]$ where $n_1 \geq n_2$. Then the estimator $\hat{p}$ defined above satisfies*

$$\begin{aligned}\mathsf{h}^2(p^*, \hat{p}) \leq \frac{1}{2}\mathsf{KL}(p^*, \hat{p}) \lesssim{}& \frac{n_1 \log(n_1/\delta)}{N} \\ &+ (p^*_{\mathsf{max}}\, n_1 n_2)^{1/3}\Big(\log\big(L(p^*)\big) + 1\Big)^{2/3}\Big(\frac{\log(n_1/\delta)\log(n_2)}{N}\Big)^{2/3}\end{aligned}$$

with probability at least $1 - 4\delta$.

In particular, in the case where $p^*_{\mathsf{max}} \asymp 1/(n_1 n_2)$, the bound in Theorem 3 reduces to

$$\mathsf{h}^2(p^*, \hat{p}) \lesssim \frac{n_1}{N} + \frac{1}{N^{2/3}}$$

up to logarithmic factors. The term $\frac{1}{N^{2/3}}$ results from the MTP_2 shape constraint, while the term $\frac{n_1}{N}$ is present even if the PMF p^* has constant rows.

Technically, the two terms follow from a decomposition of the noise in the proof. The following theorem shows that this upper bound is, in fact, optimal in the minimax sense up to logarithmic factors.

Theorem 4 (Lower bounds for estimation of discrete MTP_2 distributions)**.** *Let $\mathbb{P}_{p^*}$ denote the probability with respect to N independent observations from the distribution with an* MTP_2 *PMF p^* on the grid $n_1 \times n_2$. For $n_1 \le N \le n_1^3 n_2^3$, there exists a universal constant $c > 0$ such that*

$$\inf_{\tilde{p}} \sup_{p^* \ \mathrm{MTP}_2} \mathbb{P}_{p^*}\Big\{ \mathsf{h}^2(p^*, \tilde{p}) \ge c\Big(\frac{n_1}{N} + \frac{1}{N^{2/3}}\Big)\Big\} \ge \frac{1}{3},$$

where the infimum is over all estimators $\tilde{p}$ measurable with respect to the observations. For $N \le n_1$, we have the vacuous lower bound of constant order. For $N \ge n_1^3 n_2^3$, we have the lower bound of order $\frac{n_1 n_2}{N}$, which is the trivial rate of estimation.

Note that in the regime with an enormous sample size $N \ge n_1^3 n_2^3$, the lower bound $\frac{n_1 n_2}{N}$ is achieved by the empirical frequency matrix Y/N (see Appendix B.3), so there is no need to exploit the MTP_2 constraint. In fact, it can be seen from the proof of Theorem 3 in Section 6.1 that the estimator $\hat{p}$ also attains this rate up to logarithmic factors (see Remark 12), a behavior that can be observed in the numerical experiments as well (see Figure 2a in Section 5).

While our upper and lower bounds match in terms of the sample size N and dimensions (n_1, n_2), there are two potential improvements that can be made. First, the assumption $N \ge 12 \log(n_1 n_2/\delta)/p^*_{\mathsf{min}}$ in Theorem 3 is necessary to guarantee that we have sufficient observations at each point on the grid $[n_1] \times [n_2]$, so that the (box-constrained) MLE can be properly defined and efficiently computed. There may exist other estimation procedures that apply in the regime where the sample size is smaller. Second, the upper bound contains the parameter p^*_{max} which is not present in the lower bound. This is likely an artifact of our proof of the upper bound and could potentially be mitigated.

3. Smooth $\mathbf{MTP_2}$ density estimation

We turn to estimation of a probability distribution with a smooth MTP_2 density ρ^* on $[0,1]^2$ with respect to the Lebesgue measure. Recall that MTP_2 requires that for any $x, y \in [0,1]^2$,

$$\rho^*(x \wedge y)\rho^*(x \vee y) \ge \rho^*(x)\rho^*(y). \tag{3.1}$$

In addition, we assume that ρ^* is β-Hölder smooth, defined more precisely as follows.

Definition 5. *For $\beta, R > 0$, we define $\mathcal{D}(\beta, R)$ to be the set of probability densities ρ on $[0,1]^2$ such that ρ is $\lceil \beta - 1 \rceil$ times continuously differentiable with*

$$|\partial^\alpha \rho(x)| \le R, \quad \textit{for all } |\alpha| \le \lceil \beta - 1 \rceil,\ x \in [0,1]^2, \quad \textit{and} \tag{3.2}$$

$$|\partial^\alpha \rho(x) - \partial^\alpha \rho(y)| \leq R \, \|x-y\|_2^{\beta - \lceil \beta - 1 \rceil}, \qquad \text{for all } |\alpha| = \lceil \beta - 1 \rceil, \; x, y \in [0,1]^2. \tag{3.3}$$

Moreover, for $\rho_{\min}, \rho_{\max} > 0$, we define $\tilde{\mathcal{D}}(\beta, R, \rho_{\min}, \rho_{\max})$ to be the subset of $\mathcal{D}(\beta, R)$ consisting of densities ρ such that

$$\rho_{\min} \leq \rho(x) \leq \rho_{\max}, \quad \text{for all } x \in [0,1]^2. \tag{3.4}$$

Equipped with the above definition, we assume that $\rho^* \in \tilde{\mathcal{D}}(\beta, R, \rho_{\min}, \rho_{\max})$. Given N i.i.d. observations from the distribution with density ρ^* where $N > 0$, we aim to estimate the distribution.

3.1. Estimator

To define an estimator of ρ^*, we make use of both smoothness and the MTP_2 assumption. Namely, smoothness allows us to discretize the space $[0,1]^2$ into grid cells and group observations together in each cell, after which we are able to employ the MTP_2 shape constraint.

More precisely, for a positive integer n to be determined later, we consider the equidistant discretization on $[0,1]^2$ with n subdivisions on each dimension, that is, with grid cells

$$S_{i,j} := \Big[\frac{i-1}{n}, \frac{i}{n}\Big) \times \Big[\frac{j-1}{n}, \frac{j}{n}\Big), \quad i, j \in [n].$$

Denote by Y the (unnormalized) histogram estimator with grid $(S_{i,j})_{i,j=1}^n$ for a sample $\{X_1, \dots, X_N\}$, that is,

$$Y_{i,j} := \sum_{k=1}^N \mathbb{1}\{X_k \in S_{i,j}\}. \tag{3.5}$$

Moreover, we define

$$p^*_{i,j} := \int_{S_{i,j}} \rho^*(x) \, \mathrm{d}x. \tag{3.6}$$

Since ρ^* is MTP_2, it is easily verified that the discrete density p^* is MTP_2 in the sense of (2.1).

Given the matrix Y with entries specified by (3.5), we compute the estimator $\hat{p} = \hat{p}(Y)$ defined in Section 2.1, and define an estimator $\hat{\rho}$ of the density ρ^* by

$$\hat{\rho}(x) := n^2 \hat{p}_{i,j}, \quad \text{for } x \in S_{i,j}, \tag{3.7}$$

which is a piecewise constant estimator on the grid $(S_{i,j})_{i,j=1}^n$.

3.2. Upper and lower bounds

The performance of our estimator $\hat{\rho}$ with respect to the Hellinger distance is characterized by the following theorem.

Theorem 6 (Upper bounds for estimation of smooth MTP_2 distributions)**.** *Suppose that we are given N independent observations from an MTP_2 distribution with density $\rho^* \in \tilde{\mathcal{D}}(\beta, R, \rho_{\min}, \rho_{\max})$. With $\tilde{\beta} := \beta \wedge 1$ and the choice*

$$n = \Big\lfloor \Big(\frac{R^2 N}{\rho_{\min}}\Big)^{1/(2\tilde{\beta}+1)} \wedge \Big(\frac{\rho_{\min} N}{\log(\rho_{\min} N)}\Big)^{1/2} \Big\rfloor,$$

we define the estimator $\hat{\rho}$ as in (3.7). Moreover, suppose that N is larger than a constant depending on β, R and $\rho_{\min}$. Then with probability at least $1 - N^{-4}$, the following holds. If $\beta > 0.5$, then

$$\mathsf{h}^2(\hat{\rho}, \rho^*) \lesssim \frac{\log N}{N^{2\tilde{\beta}/(2\tilde{\beta}+1)}} + \frac{(\log N)^{4/3}}{N^{2/3}},$$

and if $0 < \beta \le 0.5$, then

$$\mathsf{h}^2(\hat{\rho}, \rho^*) \lesssim \Big(\frac{\log N}{N}\Big)^{\tilde{\beta}} + \frac{(\log N)^{4/3}}{N^{2/3}},$$

where the suppressed constants depend on the quantities $\beta, R, \rho_{\min}$ and $\rho_{\max}$.

Note that the size of discretization n can be viewed as a tuning parameter in smooth density estimation—the larger n is, the smaller bias and larger variance the piecewise constant estimator has. As the proof of Theorem 6 suggests, the above choice of n achieves the optimal bias-variance trade-off, thereby yielding near-optimal upper bounds.

As in the discrete setting, each of the above upper bounds contains two terms. The term involving $\tilde{\beta}$ in the exponent originates from the smoothness of the density, while the term $\frac{(\log N)^{4/3}}{N^{2/3}}$ is due to the MTP_2 shape constraint.

More precise versions of the bounds in Theorem 6 are given in (6.27) and (6.28) with explicit dependencies on $\rho_{\min}$, $\rho_{\max}$, and R. In particular, treating these quantities as constants, the above bounds yield (up to logarithmic factors) that

$$\mathsf{h}^2(\hat{\rho}, \rho^*) \lesssim \begin{cases} N^{-\beta}, & 0 < \beta < 0.5, \\ N^{-\frac{2\beta}{2\beta+1}}, & 0.5 \le \beta < 1, \\ N^{-2/3}, & \beta \ge 1, \end{cases} \tag{3.8}$$

with high probability. These upper bounds are complemented by the following lower bounds.

Theorem 7 (Lower bounds for estimation of smooth MTP_2 distributions)**.** *Let $\mathbb{P}_{\rho^*}$ denote the probability with respect to N independent observations from the distribution with density $\rho^* \in \mathcal{D}(\beta, R)$, for $\beta > 0$ and $R \ge 1$. Then there exists a universal constant $c > 0$ such that*

$$\inf_{\tilde{\rho}} \sup_{\rho^* \in \mathcal{D}(\beta, R)} \mathbb{P}_{\rho^*}\big(\mathsf{h}^2(\tilde{\rho}, \rho^*) \ge c\phi_\beta(N)\big) \ge \frac{1}{3},$$

where the infimum is taken over all estimators $\tilde{\rho}$ measurable with respect to the observations and

$$\phi_\beta(N) = \begin{cases} N^{-\frac{2\beta}{2\beta+1}}, & \text{if } 0 < \beta < 1, \\ N^{-2/3}, & \text{if } 1 \leq \beta < 2, \\ N^{-\frac{2\beta}{2\beta+2}}, & \text{if } \beta \geq 2. \end{cases}$$

The lower bounds above match the upper bounds up to logarithmic factors in the regime $0.5 \leq \beta \leq 2$. For $0 < \beta < 0.5$, the rate $N^{-\beta}$ coincides with the rate obtained by the nonparametric Hölder-constrained estimator in dimension one [6]. This slow rate is known to be suboptimal, as sieve estimators attain the optimal rate $N^{-2\beta/(2\beta+1)}$. While we conjecture that up to log factors, the latter should also be the optimal rate in our case, we leave the problem of finding the optimal rate in this regime as an open question for future research. For $\beta > 2$, the rate is the same as that for β-Hölder smooth density estimation without the MTP_2 assumption. Therefore, while the lower bound is interesting in our setup, to obtain the upper bound, it suffices to use any existing rate-optimal estimator.

Remark 8. *The construction of the estimator $\hat{\rho}$ depends both on the smoothness parameter β and the Hölder constant R, and it does not match the lower bounds in the case $\beta > 2$. Both of these shortcomings can be remedied by considering an ensemble of estimators that include both our estimators $\hat{\rho}$ for varying parameters $\tilde{\beta}$ and $\tilde{R}$ over a discretization of the set of parameters, and, for example, regular kernel density estimators which are rate-optimal for Hölder smooth density estimation. Over such an ensemble, either selection [32] or aggregation procedures [21] can be used to achieve adaptive rates that match the lower bounds up to logarithmic factors. Since the techniques are standard and yield no new phenomenon, we do not pursue this direction in the current work.*

4. Efficient algorithms

The optimization problem for finding the constrained MLE in (2.6) is a convex problem with a polynomial number of constraints and can thus be solved in polynomial time with a general purpose solver for convex problems such as SCS [34, 35] or ECOS [10]. However, since the number of constraints is of the order $n_1 n_2$, solving the linear systems in each iteration step of these solvers can take a long time without specialized solvers. We address this issue by employing a proximal Newton method, whose main step consists in a projection onto the set of constraints, which in turn can be solved by a variant of Dykstra's algorithm, as discussed in [19]. In this section, in order to emphasize the connection to computing projections, we think about (2.6) as a minimization problem instead of a maximization problem by changing the sign in front of the objective.

First, we derive the outer iteration of our algorithm as a proximal Newton method. These methods are intended to solve nonlinear optimization problems by successively solving local quadratic approximations to the objective functions. For a more thorough introduction to this class of methods, see [29]. Briefly, for $d \in \mathbb{N}$, to minimize a composite function of the form

$$\min_{\theta \in \mathbb{R}^d} f(\theta), \quad f(\theta) = g(\theta) + h(\theta),$$

one starts with an initialization $x^{(0)} = x_0 \in \mathbb{R}^d$ and computes updates by solving

$$\begin{aligned}
\rho^{(k)} = &\operatorname*{argmin}_{\tilde{\rho} \in \mathbb{R}^d} \Big[\nabla g(\theta^{(k-1)})^\top (\tilde{\rho} - \theta^{(k-1)}) \\
&\qquad + \frac{1}{2}(\tilde{\rho} - \theta^{(k-1)})^\top \nabla^2 g(\theta^{(k-1)})(\tilde{\rho} - \theta^{(k-1)}) + h(\tilde{\rho}) \Big], \\
\theta^{(k)} &= \theta^{(k-1)} + t_k(\eta^{(k)} - \theta^{(k-1)}),
\end{aligned}$$

where t_k is usually chosen by a line-search technique. In the case of (2.6), we set

$$\begin{aligned}
g(\theta) &= -\frac{1}{N}\langle Y, \theta \rangle + \sum_{i,j} e^{\theta_{i,j}} \\
h(\theta) &= \begin{cases} 0, & \theta \in \mathcal{M} \cap \mathcal{C}(Y), \\ +\infty, & \theta \notin \mathcal{M} \cap \mathcal{C}(Y), \end{cases}
\end{aligned}$$

where $\mathcal{C}(Y)$ corresponds to the box constraints defined in (2.5), and

$$\mathcal{M} := \{\theta \in \mathbb{R}^{n_1 \times n_2} : D\theta\tilde{D}^\top \geq 0\}.$$

Further computation then shows that the Hessian $\nabla^2 g(\theta)$ has the structure of a diagonal operator, which makes the subproblem of computing $\rho^{(k)}$ equivalent to finding a projection with respect to a weighted Frobenius norm. Namely, computing the first and second derivatives yields

$$\begin{aligned}
(\nabla g(\theta))_{i_1,i_2} &= -\frac{1}{N}Y_{i,j} + \exp(\theta_{i,j}), \\
(\nabla^2 g(\theta))_{(i_1,i_2),(j_1,j_2)} &= \begin{cases} \exp(\theta_{i_1,i_2}), & (i_1,i_2) = (j_1,j_2), \\ 0, & \text{otherwise.} \end{cases}
\end{aligned}$$

Hence, writing

$$\Lambda_{i_1,i_2} = \exp(\theta^{(k)}_{i_1,i_2}), \tag{4.1}$$

computing $\rho^{(k)}$ is equivalent to

$$\begin{aligned}
\rho^{(k)} &= \operatorname*{argmin}_{\tilde{\rho} \in \mathcal{M} \cap \mathcal{C}(Y)} \Big\langle -\frac{1}{N}Y + \Lambda, \tilde{\rho} \Big\rangle + \frac{1}{2}\Big\| \tilde{\rho} - \theta^{(k-1)} \Big\|_\Lambda^2 \\
&= \operatorname*{argmin}_{\tilde{\rho} \in \mathcal{M} \cap \mathcal{C}(Y)} \Big\{ \frac{1}{2}\Big\| \tilde{\rho} - \Big(\theta^{(k-1)} + \frac{1}{N}Y \oslash \Lambda - \mathbb{1}\Big) \Big\|_\Lambda^2
\end{aligned}$$

$$+\left\langle \theta^{(k-1)}, \Lambda - \frac{1}{N}Y \right\rangle - \frac{1}{2}\left\| \Lambda - \frac{1}{N}Y \right\|_{1\oslash\Lambda}^2 \Bigg\}$$

$$= \operatorname*{argmin}_{\tilde{\rho}\in\mathcal{M}\cap\mathcal{C}(Y)} \frac{1}{2}\left\| \tilde{\rho} - \left(\theta^{(k-1)} + \frac{1}{N}Y \oslash \Lambda - \mathbb{1}\right)\right\|_{\Lambda}^2, \tag{4.2}$$

the projection of $\theta^{(k-1)} + \frac{1}{N}Y \oslash \Lambda - \mathbb{1}$ onto $\mathcal{M} \cap \mathcal{C}(Y)$ with respect to the Frobenius norm weighted by Λ.

Second, problem (4.2) can be efficiently solved by a variant of Dykstra's algorithm, as shown in [19]. The idea is to split up the projection onto $\mathcal{M}\cap\mathcal{C}(Y)$ into the projection onto $\mathcal{C}(Y)$ and the sets

$$\mathcal{M}_{i_1,i_2} = \Big\{\theta \in \mathbb{R}^{n_1\times n_2} : \sum_{j_1\in\{0,1\}, j_2\in\{0,1\}} (-1)^{j_1+j_2}\theta_{i_1+j_1,i_2+j_2} \geq 0\Big\}$$

for $i_1 \in [n_1-1], i_2 \in [n_2-1]$, where additional correction terms are applied to the vectors to ensure convergence. The basic Dysktra algorithm for projecting a vector $y \in \mathbb{R}^d$ onto a general collection of sets $\mathcal{M}_1, \dots, \mathcal{M}_m$ is listed as Algorithm 1.

Algorithm 1 Dykstra algorithm

```
Input: y ∈ ℝ^d
Output: θ ≈ Π_M(y)
  function ProjectDykstra(y)
     for i = 1, ..., m do
        p_i = 0_d                                      ▷ Initialize residuals
     end for
     θ_m = y                                           ▷ Initialize iterates
     while not converged do
        for i = 1, ..., m do
           θ_i ← Π_{M_i}(θ_{(i−2)%m+1} + p_i)          ▷ Project shifted iterates
           p_i ← θ_{(i−2)%m+1} + p_i − θ_i             ▷ Compute new residual
        end for
     end while
     return θ
  end function
```

In our case, the projection onto $\mathcal{M}_{i_1,i_2}$ with weight matrix Λ, written as $\Pi_{\mathcal{M}_{i_1,i_2},\Lambda}$, has the following closed form solution. For $i_1 \in [n_1-1], i_2 \in [n_2-1]$, let Λ be given is in (4.1) and set

$$\Gamma_{i_1,i_2} = \left(\sum_{j_1,j_2\in\{0,1\}} \frac{1}{\Lambda_{i_1+j_1,i_2+j_2}}\right)^{-1}.$$

Then, we have for $j_1, j_2 \in \{0,1\}$ that

$$(\Pi_{\mathcal{M}_{i_1,i_2},\Lambda}y)_{i_1+j_1,i_2+j_2} = Z_{i_1+j_1,i_2+j_2} +$$

$$\frac{(-1)^{j_1+j_2}}{\Lambda_{i_1+j_1,i_2+j_2}} \max\left\{-\Gamma_{i_1,i_2} \sum_{k_1\in\{0,1\},k_2\in\{0,1\}} (-1)^{k_1+k_2} Z_{i_1+k_1,i_2+k_2}, 0\right\},$$

and for $(l_1, l_2) \notin (i_1 + \{0,1\}) \times (i_2 + \{0,1\})$ that

$$(\Pi_{\mathcal{M}_{i_1,i_2},\Lambda} y)_{l_1,l_2} = Z_{l_1,l_2}.$$

Together with the closed form solution for projecting onto the box $\mathcal{C}(Y)$,

$$(\Pi_{\mathcal{C}(Y)}(y)_{i_1,i_2}) = \begin{cases} Z_{i_1,i_2}, & Z_{i_1,i_2} \in \left[\log\frac{2Y_{i,j}}{3N}, \log\frac{2Y_{i_1,i_2}}{N}\right] \\ \log\frac{2Y_{i_1,i_2}}{N}, & Z_{i_1,i_2} > \log\frac{2Y_{i_1,i_2}}{N}, \\ \log\frac{2Y_{i_1,i_2}}{3N}, & Z_{i_1,i_2} < \log\frac{2Y_{i_1,i_2}}{3N}, \end{cases}$$

we end up with the iterative projection algorithm given in Algorithm 2, which in turn leads to the proximal Newton method, Algorithm 3. Note that we did not implement a line search but instead chose to directly update our iterates with $\rho^{(k)}$, which seems to not pose any problems in practice.

Algorithm 2 Fast projection onto $\mathcal{M}$

---

function PROJECT$(y, \Lambda, \mathcal{C}(Y))$
  $\theta \leftarrow y, \quad \eta \leftarrow 0 \in \mathbb{R}^{(n_1-1)\times(n_2-1)}, \quad \eta' \leftarrow 0 \in \mathbb{R}^{n_1\times n_2}$ ▷ Initialize θ and residuals η, η'
  for $i_1 = 1, \ldots, n_1 - 1$, $i_2 = 1, \ldots, n_2 - 1$ **do**
    $\Gamma_{i_1,i_2} \leftarrow \left(\sum_{j_1,j_2\in\{0,1\}} (\Lambda_{i_1+j_1,i_2+j_2})^{-1}\right)^{-1}$ ▷ Initialize harmonic mean of weights
  end for
  while not converged **do**
    $\theta' \leftarrow \Pi_{\mathcal{C}(Y)}(\theta + \eta')$ ▷ Project onto $\mathcal{C}(Y)$ ...
    $\eta' \leftarrow \theta + \eta' - \theta', \quad \theta \leftarrow \theta'$ ▷ ... and store corresponding residual
    for $i_1 = 1, \ldots, n_1 - 1$, $i_2 = 1, \ldots, n_2 - 1$ **do** ▷ Project onto $\mathcal{M}$ by projecting onto all $\mathcal{M}_{i_1,i_2}$ in turn
      $\tilde{\eta} \leftarrow \max\left\{\eta_{i_1,i_2} - \Gamma_{i_1,i_2} \sum_{k_1,k_2\in\{0,1\}} (-1)^{k_1+k_2} \theta_{i_1+k_1,i_2+k_2}, 0\right\}$
      for $j_1 \in \{0,1\}$, $j_2 \in \{0,1\}$ **do**
        $\theta_{i_1+j_1,i_2+j_2} \leftarrow \theta_{i_1+j_1,i_2+j_2} + (-1)^{j_1+j_2} \Lambda^{-1}_{i_1+j_1,i_2+j_2} (\tilde{\eta} - \eta_{i_1,i_2})$
      end for
      $\eta_{i_1,i_2} \leftarrow \tilde{\eta}$
    end for
  end while
  return θ
end function

Algorithm 3 Restricted ML solution via proximal Newton method

function RESTRICTEDMAXIMUMLIKELIHOOD(Y)
  $\theta \leftarrow Y/N$
  while not converged **do**
    for $i_1 = 1, \ldots, n_1, i_2 = 1, \ldots, n_2$ **do**
      $\Lambda_{i_1,i_2} \leftarrow \exp(\theta_{i,j})$ ▷ Update weight matrix
    end for
    $y \leftarrow \theta + Y/N \oslash \Lambda - \mathbb{1}$
    $\theta \leftarrow$ PROJECT$(y, \Lambda, \mathcal{C}(Y))$ ▷ perform Newton step
  end while
  return θ.
end function

Similarly, if instead of the box-constrained estimator (2.7), we are interested in calculating the regular MLE over MTP_2, (2.4), we can omit the projection onto $\mathcal{C}(Y)$ in Algorithm 2.

In practice, convergence in Algorithm 2 can be checked by computing a measure of feasibility such as $0 \vee \max\{-(D\theta\tilde{D})_{i,j} : i \in [n_1 - 1], j \in [n_2 - 1]\}$ or stopping when the distance between successive iterates becomes small. Similarly, we stop the proximal Newton method, Algorithm 3, when two successive iterates become very close to each other or the gain in the objective function is very small.

Following these considerations, it is straightforward to compute the density estimator (3.7) by computing a histogram of the samples in $[0, 1]^2$ and applying Algorithm 3 to obtain $\hat{p}$, which yields the piecewise constant approximation in (3.7).

5. Numerical experiments

This section is devoted to simulations which corroborate our theoretical findings. Further details on the underlying implementation can be found in Appendix C.

5.1. *Experiments for the grid estimator*

In this section, we set $n = n_1 = n_2$ for simplicity.

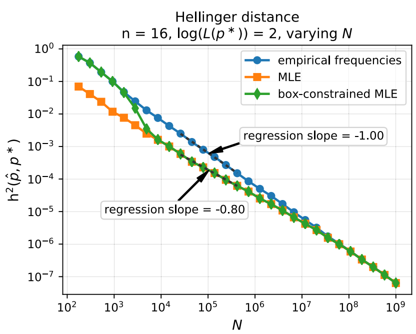
We consider the following family of ground truth probability mass functions: Let $n \in \mathbb{N}$, set

$$\tilde{\theta}_{i,j} = 1 + \log(L)\frac{(i-1)(j-1)}{(n-1)^2}, \quad i, j \in [n],$$

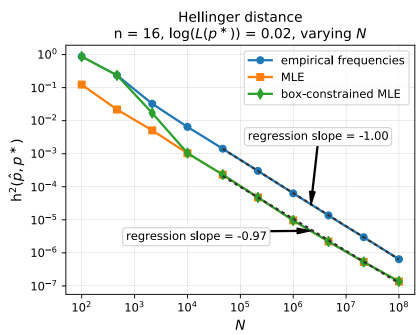
for $L > 1$ that can be varied and

$$p^*_{i,j} = \frac{\exp(\tilde{\theta}_{i,j})}{\sum_{i,j}\exp(\tilde{\theta}_{i,j})}, \quad i, j \in [n]. \tag{5.1}$$

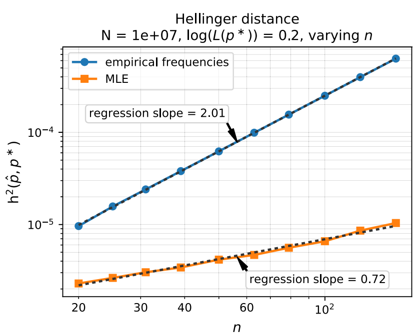
By construction, $\tilde{\theta}$ is supermodular, so p^* is MTP_2, and $\log(L(p^*)) = \log(L)$. We sample N i.i.d. observations $\{Z_k\}_{k=1}^N$ from p and form the matrix Y as in (2.2). We consider three estimators for p^*: the empirical frequency matrix Y/N, the MLE given by (2.4), and the box-constrained estimator in (2.7). The latter two estimators are computed by variants of Algorithm 3. Note that in cases where some of the entries in the frequency matrix are zero, we cannot take logarithms and thus report Y/N as the output of the box-constrained estimator, while the unconstrained MLE can be calculated as in Section 4 above. For the unconstrained MLE, we do observe numerical instabilities when the number of observations is very low, eventually leading to underflows in the calculation. This can be remedied by imposing mild lower bounds on the resulting density, see Appendix C.2.



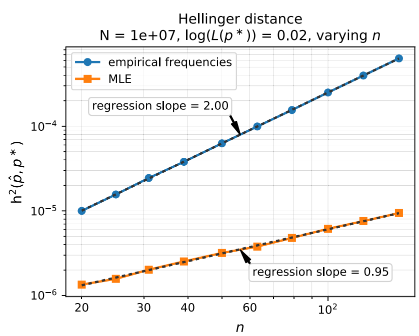
(a) Varying N, $\log(L(p^*)) = 2$



(b) Varying N, $\log(L(p^*)) = 0.02$



(c) Varying n, $\log(L(p^*)) = 0.2$



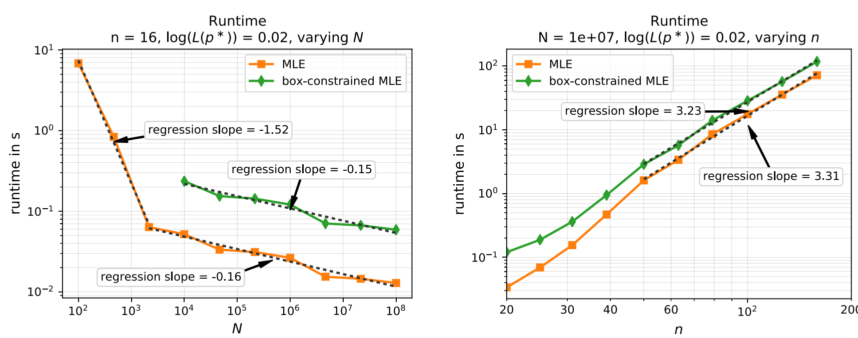
(d) Varying n, $\log(L(p^*)) = 0.02$

FIG 2. *Estimation of a density on a grid*

In Figure 2, we plot the squared Hellinger distance $\mathsf{h}^2(p^*, \hat{p})$ for the three estimators in four setups, averaged over 20 independent replicates. More specifically, we report the results of linearly regressing the logarithms of the distances on the logarithms of the varying parameters over one or more manually selected ranges, corresponding to an estimate of the polynomial dependence on the parameter in question.

In Figures 2a and 2b, we vary the sample size N and keep $n = 16$ fixed for $\log(L(p^*)) \in \{2, 0.02\}$, respectively, while in Figures 2c and 2d, we vary the grid size n and keep $N = 10{,}000{,}000$ fixed for $\log(L(p^*)) \in \{0.2, 0.02\}$, respectively. We observe that all three estimators achieve an N^{-1} asymptotic rate in Figure 2a. Moreover, the box-constrained estimator (2.7) and the regular MLE (2.4) show very similar performance: For small N, the probability for zero entries in Y/N is high and thus Y/N is used instead of the box-constrained estimator as explained in the above paragraph, which explains that its performance coincides with that of the empirical frequency matrix in this regime, while the MLE performs better because the dominant factor in this regime is n/N. For an intermediate regime of N, the estimation performance of the MLE and the box-constraint estimator is consistently better than that of Y/N, but it is attenuated once the $(\log(L(p^*))\, n/N)^{2/3}$ rate becomes active, which can be seen in Figure 2a. On the other hand, in Figure 2b, N is not large enough relative to $\log(L(p^*))$ to capture this regime. Finally, for very large values of N, the performance of all three estimators coincides, which for the box-constraint estimator matches the proof of the upper bound (up to logarithmic factors), see Remark 12.

A similar behavior can be seen in Figures 2c and 2d, where the performance of the frequency matrix scales with n^2, while the regular MLE scales approximately like $n^{2/3}$ (regression coefficient of 0.72) for a larger value of $\log(L(p^*))$ and like n (regression coefficient of 0.95) when $\log(L(p^*))$ is small. Note that the performance of the box-constrained estimator is not plotted here since it mostly coincides with that of the regular MLE.



(a) Varying N

(b) Varying n

Fig 3. *Runtimes for density on grid*

To investigate the practical performance of the proposed algorithms, in Figures 3, we report the runtime averaged over 20 replicates on an AMD 3400G desktop processor. Here, as well as in the previous examples, we stopped Algorithm 2 when a relative change in ℓ^2-norm of less than 10^{-6} was detected. Similarly, Algorithm 3 was stopped at a relative accuracy of 10^{-5}. In Figure 3a, we observe that the conditioning of the problem improves with larger sample sizes N and deteriorates for small values of N, leading to a decay in runtime of approximately $N^{-1.52}$ up to $N \approx 1,000$, followed by a milder dependence of $N^{-0.15}$ for larger values of N. Note that the runtime for the boxed estimator is only plotted for $N \geq 10,000$ because of the presence of zeros in the empirical frequency matrix for smaller values of N.

In Figure 3b, we see that, as expected from a Dykstra-type algorithm, the conditioning of the problem worsens with increasing n, necessitating more iterations and thus leading to an increase of runtime until convergence that is larger than the cost of one iteration, which is of order n^2. However, it is still reasonably mild, scaling roughly like $n^{3.3}$ for larger values of n. Overall, this highlights the practicability of the proposed algorithm for problems of medium size: Instances with $n \approx 50$ can be solved within four seconds, while problems of size $n = 160$ take under two minutes. Moreover, adding the additional box constraint (2.5) only slightly increases the runtime when n is large.

5.2. Experiments for continuous density estimation

We consider the multivariate Gaussian distribution $P^* = N(\mu^*, \Sigma^*)$ with parameters

$$\mu^* = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}, \quad \Sigma^* = \begin{pmatrix} 0.2 & 0.1 \\ 0.1 & 0.2 \end{pmatrix},$$

conditioned on the event that $Z \in [0,1]^2$ where $Z \sim P^*$. In other words, we consider the density

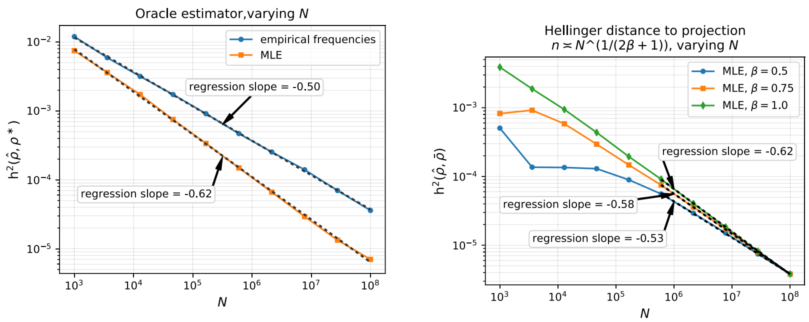
$$\rho^*(x) = \frac{1}{\int_{[0,1]^2} \tilde{\rho}(y)\,\mathrm{d}y} \tilde{\rho}(x), \quad x \in [0,1]^2, \tag{5.2}$$

where

$$\tilde{\rho}(x) = \frac{1}{2\pi\sqrt{0.03}} \exp\left(-\frac{10}{3}\left((x_1 - 0.5)^2 + (x_2 - 0.5)^2 - (x_1 - 0.5)(x_2 - 0.5)\right)\right).$$

Note that on $[0,1]^2$, $\rho^* \in C^\infty$ and that it is MTP_2, which can be easily checked by computing the mixed derivative $\partial_1\partial_2$ of the log-density. Here, we use it to evaluate the performance of the gridding strategy from Section 3.1 for both an oracle choice of n, that is, exploiting the knowledge of the ground truth to pick the best possible value of n from a given list, and a fixed scaling of n in the cases $\beta \in \{0.5, 0.75, 1.0\}$.

First, in Figure 4a, with a varying number of i.i.d. observations $\{Z_k\}_{k=1}^N$ from P^*, we plot the squared Hellinger distance $\mathsf{h}^2(\hat{\rho}, \rho^*)$ for the estimator in (3.7), where n is picked from 10 logarithmically spaced values between $n = 4$ and $n = 200$ according to which yields the smallest Hellinger distance, and for a similarly defined estimator where $\hat{p}$ is replaced by the empirical frequency matrix Y/N. We observe that the empirical frequency matrix achieves a rate of about $N^{-1/2}$, corresponding to the rate for general Hölder functions in 2D with $\beta = 1$, while the MTP_2 MLE comes close to the predicted $N^{-2/3}$ rate that corresponds to the $\beta \in [1,2)$ range (regression coefficient of 0.62).



(a) Comparison between empirical frequencies and MTP_2 estimator, oracle choice of n

(b) Variance part of the error when varying the scaling of n

Fig 4. *Performance of continuous density estimation*

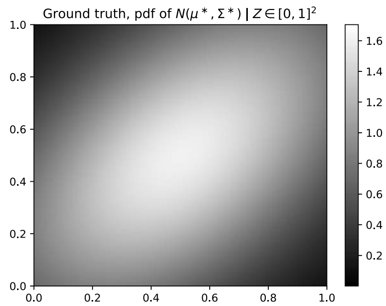
Second, to investigate the effect of different β, in Figure 4b, we use a fixed scaling $n = CN^{1/(2\beta+1)}$. Note that we cannot expect to observe the rates in Theorem 6 when considering the distance $\mathsf{h}^2(\hat{\rho}, \rho^*)$ in this setup since ρ^* is C^∞-smooth. Denoting by $\bar{\rho}$ the piecewise constant approximation to ρ^* (see (6.23) below), this is due to the fact that the bias term $\mathsf{h}^2(\bar{\rho}, \rho^*)$ could dominate the overall error $\mathsf{h}^2(\hat{\rho}, \rho^*)$. Hence, we only plot the Hellinger distance corresponding to the variance part $\mathsf{h}^2(\hat{\rho}, \bar{\rho})$. For computational reasons, C is chosen for each β so that for $N = 10^8$, we have $n = 200$. Due to the similarities between the regular MLE and its box-constrained version observed in the previous section, all calculations were performed using the regular MLE, (2.4), resulting in slightly faster computations.

Performing linear regression on the doubly logarithmic plot for large values of N, we observe rates of 0.53, 0.58, and 0.62 for $\beta = 0.5, 0.75, 1.0$, respectively. These are close to 0.5, 0.6, and 2/3, respectively, as predicted by Theorem 6. Additionally, we present heat maps of the density ρ^* (Figure 5a), as well as an approximation via the frequncy matrix Y/N (Figure 5b) and the MLE (Figure 5c) for $N = 10{,}000$ and $n = 16$. The visual smoothing effect of the MLE is quite obvious in this case.

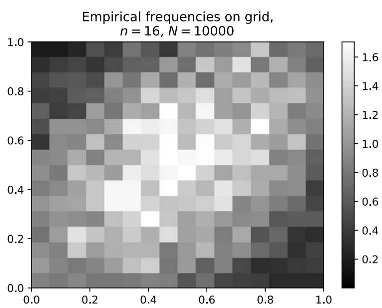
6. Proofs

The proofs of our results are provided in this section. We first prove the upper bounds Theorems 3 and 6 in the discrete and smooth cases respectively, and then the lower bounds Theorems 4 and 7. In the proofs, we make use of the well-known relation [32, Lemma 7.23] that for PMFs p and q on $\mathcal{X} = [n_1] \times [n_2]$ or $[0,1]^2$,

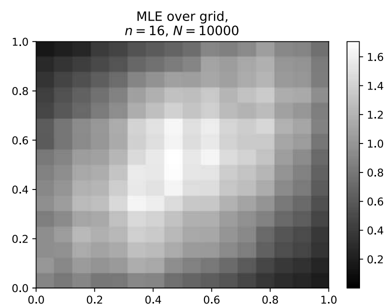
$$2\,\mathsf{h}^2(p,q) \le \mathsf{KL}(p,q) \le 2\Big(2 + \log\big(\max_{x\in\mathcal{X}} \frac{p(x)}{q(x)}\big)\Big)\,\mathsf{h}^2(p,q). \tag{6.1}$$



(a) Ground truth ρ^*



(b) Approximation via frequencies Y/N



(c) Approximation via MLE $\hat{\rho}$

FIG 5. *Visual comparison of continuous density estimation*

6.1. Proof of Theorem 3

6.1.1. Setup of the proof: quadratic approximation

Let us define $\varepsilon := \frac{Y}{N} - p^*$. Denote by $\mathcal{A}_1$ the event of probability $1 - 2\delta$ that the bounds in Lemma 2 hold. On this event, θ^* lies in the cube $\mathcal{C}(Y)$ defined in (2.5), so we have

$$\frac{1}{N}\langle Y, \tilde{\theta}\rangle - \sum_{i,j} e^{\tilde{\theta}_{i,j}} \geq \frac{1}{N}\langle Y, \theta^*\rangle - 1,$$

which is equivalent to

$$\langle p^*, \theta^* - \tilde{\theta}\rangle + \sum_{i,j} e^{\tilde{\theta}_{i,j}} - 1 \leq \langle \varepsilon, \tilde{\theta} - \theta^*\rangle. \tag{6.2}$$

In addition, the definition of $\mathcal{C}(Y)$ yields that $|\tilde{\theta}_{i,j} - \theta^*_{i,j}| \leq \log 2 - \log \frac{2}{3} < 1.1$ for all i, j.

By a quadratic Taylor approximation of e^y, it holds for $x \leq 0$ and $|y-x| \leq 1.1$ that

$$e^x + e^x(y-x) + e^x(y-x)^2/4 \leq e^y \leq e^x + e^x(y-x) + 2\, e^x(y-x)^2.$$

Applying this approximation to the exponential terms of the left-hand side of (6.2), we obtain

$$\begin{aligned}
&\langle p^*, \theta^* - \tilde{\theta}\rangle + \sum_{i,j} e^{\theta^*_{i,j}} + \sum_{i,j} e^{\theta^*_{i,j}}(\tilde{\theta}_{i,j} - \theta^*_{i,j}) + \frac{1}{4}\sum_{i,j} e^{\theta^*_{i,j}}(\tilde{\theta}_{i,j} - \theta^*_{i,j})^2 - 1 \\
&\le \langle p^*, \theta^* - \tilde{\theta}\rangle + \sum_{i,j} e^{\tilde{\theta}_{i,j}} - 1 \\
&\le \langle p^*, \theta^* - \tilde{\theta}\rangle + \sum_{i,j} e^{\theta^*_{i,j}} + \sum_{i,j} e^{\theta^*_{i,j}}(\tilde{\theta}_{i,j} - \theta^*_{i,j}) + 2\sum_{i,j} e^{\theta^*_{i,j}}(\tilde{\theta}_{i,j} - \theta^*_{i,j})^2 - 1,
\end{aligned}$$

or equivalently,

$$\frac{1}{4}\sum_{i,j} p^*_{i,j}(\tilde{\theta}_{i,j} - \theta^*_{i,j})^2 \le \langle p^*, \theta^* - \tilde{\theta}\rangle + \sum_{i,j} e^{\tilde{\theta}_{i,j}} - 1 \le 2\sum_{i,j} p^*_{i,j}(\tilde{\theta}_{i,j} - \theta^*_{i,j})^2. \quad (6.3)$$

The rest of the proof hinges on this quadratic approximation. Particularly, it follows from (6.2) and (6.3) that

$$\frac{1}{4}\sum_{i,j} p^*_{i,j}(\tilde{\theta}_{i,j} - \theta^*_{i,j})^2 \le \langle \varepsilon, \tilde{\theta} - \theta^*\rangle. \quad (6.4)$$

The main task in the sequel is to bound the right-hand side of (6.4). The strategy builds upon a spectral decomposition technique from the paper [19] on Monge matrix estimation.

6.1.2. Spectral decomposition of the difference operator

Recall the difference operator D defined in (2.3) and $\tilde{D}$ defined analogously for dimension n_2. Throughout the proof, whenever we introduce notation in dimension n_1, the analogous one in dimension n_2 is denoted by the same symbol with a tilde. We will decompose the noise ε in (6.4) according to a spectral decomposition of D, so let us recall some basic facts about the matrix D.

Denote the singular value decomposition of D by

$$D = U\Sigma W^\top, \quad U \in \mathbb{R}^{(n_1-1)\times(n_1-1)}, \quad \Sigma \in \mathbb{R}^{(n_1-1)\times n_1}, \quad W \in \mathbb{R}^{n_1\times n_1}$$

where we order the non-zero singular values of D in Σ in ascending magnitude, so that the last column of W spans the null-space of D. In addition, we write $W = \begin{bmatrix} w_1 & \cdots & w_{n_1}\end{bmatrix}$. Let us define a set of double indices

$$J := \big\{(l,r) \in [n_1]\times[n_2] : lr \le k\big\} \cup \big([n_1]\times\{n_2\}\big) \cup \big(\{n_1\}\times[n_2]\big), \quad (6.5)$$

and set $J^c = ([n_1]\times[n_2]) \setminus J$.

We introduce a projection operator $\Pi : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^{n_1\times n_2}$, defined as the projection onto the linear span of $\{w_i\tilde{w}_j^\top : (i,j) \in J^c\}$ that is orthogonal with respect to the inner product

$$\langle A, B\rangle_{1/p^*} := \sum_{i,j} \frac{1}{p^*_{i,j}} A_{i,j} B_{i,j}.$$

In particular, there exists an orthonormal basis $\{V^{(l,r)} \in \mathbb{R}^{n_1\times n_2} : (l,r) \in [n_1]\times[n_2]\}$ of $\mathbb{R}^{n_1\times n_2}$ with respect to the inner product $\langle .,.\rangle_{1/p^*}$ such that

$$\Pi(A) = \sum_{(l,r)\in J^c} V^{(l,r)}\langle V^{(l,r)}, A\rangle_{1/p^*} \quad \text{and}$$
$$(I-\Pi)(A) = \sum_{(l,r)\in J} V^{(l,r)}\langle V^{(l,r)}, A\rangle_{1/p^*}.$$

To characterize these projection operators further, we introduce the following notation. Let $\odot$ and $\oslash$ denote entrywise multiplication and division respectively between matrices. With a slight abuse of notation, we use $\sqrt{p^*}$ and $1/p^*$ to denote the entrywise square root and the entrywise inverse of p^* respectively. Let Λ and Λ^{-1} be the scaling operators from $\mathbb{R}^{n_1\times n_2}$ to itself, defined by

$$\Lambda(A) = A\otimes p^*, \quad A\in\mathbb{R}^{n_1\times n_2},$$
$$\Lambda^{-1}(A) = A\oslash p^*, \quad A\in\mathbb{R}^{n_1\times n_2},$$

respectively. Let $\mathcal{L}$ be the linear operator from $\mathbb{R}^{|J^c|}$ (indexed by $(l,r)\in J^c$) to $\mathbb{R}^{n_1\times n_2}$, defined by

$$\mathcal{L}(B) = \sum_{(l,r)\in J^c} B_{l,r} w_l \tilde{w}_r^\top, \quad B\in\mathbb{R}^{|J^c|},$$

and denote by $\mathcal{L}^\top$ the transpose of $\mathcal{L}$ with respect to the standard inner products in the corresponding spaces. In other words, we have

$$(\mathcal{L}^\top(A))_{l,r} = \langle w_l\tilde{w}_r^\top, A\rangle, \quad (l,r)\in J^c, \quad A\in\mathbb{R}^{n_1\times n_2}. \tag{6.6}$$

The linear operators Π and Λ can be viewed as $n_1n_2\times n_1n_2$ matrices, while the linear operator $\mathcal{L}$ can be seen as an $n_1n_2\times|J^c|$ matrix. Moreover, we have the following lemma whose proof is deferred to Section 6.1.6.

Lemma 9. *The smallest eigenvalue of the operator* $\mathcal{L}^\top\Lambda^{-1}\mathcal{L}$ *satisfies that*

$$\lambda_{\min}(\mathcal{L}^\top\Lambda^{-1}\mathcal{L}) \geq \frac{1}{p^*_{\mathsf{max}}}.$$

Moreover, Π *can be written as*

$$\Pi = \mathcal{L}(\mathcal{L}^\top\Lambda^{-1}\mathcal{L})^{-1}\mathcal{L}^\top\Lambda^{-1}. \tag{6.7}$$

To control $\langle\epsilon, \tilde\theta-\theta^*\rangle$ on the right-hand side of (6.4), we decompose it as

$$\langle\varepsilon, \tilde\theta-\theta^*\rangle = \langle(I-\Pi)(\varepsilon), \tilde\theta-\theta^*\rangle + \langle\Pi(\varepsilon), \tilde\theta-\theta^*\rangle. \tag{6.8}$$

Before proceeding to bound these two terms separately, we state two lemmas whose proofs are deferred to Sections 6.1.7 and 6.1.8, respectively.

Lemma 10. *The image of the projection* Π *is included in the image of the map* $A \mapsto D^\top A \tilde{D}$.

Lemma 11. *For any* $(i,j) \in [n_1] \times [n_2]$*, we have that*

$$\sum_{(l,r)\in J^c} \Sigma_{l,l}^{-2} \tilde{\Sigma}_{r,r}^{-2} U_{i,l}^2 \tilde{U}_{j,r}^2 \lesssim \frac{n_1 n_2}{k} \log(n_2).$$

6.1.3. Bounding the first term in (6.8)

By Hölder's inequality,

$$\begin{aligned} \langle (I-\Pi)(\varepsilon), \tilde{\theta} - \theta^* \rangle &= \big\langle (I-\Pi)(\varepsilon) \oslash \sqrt{p^*}, (\tilde{\theta} - \theta^*) \odot \sqrt{p^*} \,\big\rangle \\ &\le \|(I-\Pi)(\varepsilon)\|_{1/p^*} \|\tilde{\theta} - \theta^*\|_{p^*}. \end{aligned} \tag{6.9}$$

Now we focus on the quantity $\|(I-\Pi)(\varepsilon)\|_{1/p^*}$. By the definition of Π and the orthogonality condition that $\langle V^{(l,r)}, V^{(l',r')} \rangle = 0$ for any $(l,r) \neq (l', r')$, we obtain that

$$\begin{aligned} \|(I-\Pi)(\varepsilon)\|_{1/p^*}^2 &= \Big\| \sum_{(l,r)\in J} V^{(l,r)} \langle V^{(l,r)}, \epsilon \rangle_{1/p^*} \Big\|_{1/p^*}^2 \\ &= \sum_{(l,r)\in J} \langle V^{(l,r)}, \epsilon \rangle_{1/p^*}^2 \le |J| \max_{(l,r)\in J} (\langle V^{(l,r)}, \epsilon \rangle_{1/p^*})^2. \end{aligned} \tag{6.10}$$

Note that we can write

$$\langle V^{(l,r)}, \epsilon \rangle_{1/p^*} = \Big\langle \frac{1}{N} V^{(l,r)} \oslash p^*, N\varepsilon \Big\rangle.$$

Recall that Y has the multinomial distribution $\mathsf{Multi}(N, p^*)$, and $N\varepsilon = Y - Np^*$ is the deviation of Y from its mean. Therefore, Lemma 18 yields that on an event $\mathcal{A}_2$ of probability $1-\delta$,

$$\begin{aligned} \max_{(l,r)\in J} |\langle V^{(l,r)}, \epsilon \rangle_{1/p^*}| &\lesssim \Big(\max_{(l,r)\in J} \Big\| \frac{1}{N} V^{(l,r)} \oslash p^* \Big\|_{Np^*} \Big) \sqrt{\log(|J|/\delta)} \\ &\quad + \Big(\max_{(l,r)\in J} \Big\| \frac{1}{N} V^{(l,r)} \oslash p^* \Big\|_\infty \Big) \log(|J|/\delta). \end{aligned} \tag{6.11}$$

To bound the two norms above, we note that by orthogonality of the $V^{(l,r)}$ with respect to $\langle .,. \rangle_{1/p^*}$,

$$\Big\| \frac{1}{N} V^{(l,r)} \oslash p^* \Big\|_{Np^*}^2 = \Big\| \frac{1}{\sqrt{N}} V^{(l,r)} \Big\|_{1/p^*}^2 = \frac{1}{N}. \tag{6.12}$$

In addition, it holds that

$$\Big\| \frac{1}{N} V^{(l,r)} \Big\|_\infty \le \Big\| \frac{1}{N} V^{(l,r)} \Big\|_F = \Big\| \frac{1}{N\sqrt{N}} V^{(l,r)} \oslash \sqrt{p^*} \Big\|_{Np^*}$$

$$\leq \frac{1}{\sqrt{Np^*_{\mathsf{min}}}}\Big\|\frac{1}{N}V^{(l,r)}\Big\|_{Np^*} = \frac{1}{N\sqrt{p^*_{\mathsf{min}}}} \leq \frac{1}{\sqrt{N\log(n_1/\delta)}}, \quad (6.13)$$

where we used (6.12) and that $N \geq 12\log(n_1 n_2/\delta)/p^*_{\mathsf{min}}$ by assumption. Further, we can control the cardinality of J by

$$n_1 \leq |J| \leq n_1 n_2 \quad \text{and} \quad |J| = \sum_{(l,r)\in J} 1 \leq n_1+n_2-1+\sum_{r=1}^{n_2}\lfloor k/r\rfloor \leq 2n_1+k\log(n_2). \tag{6.14}$$

Combining (6.9), (6.10), (6.11), (6.12), (6.13) and (6.14), we see that on the event $\mathcal{A}_2$,

$$\begin{aligned}
&\langle (I-\Pi)(\varepsilon), \tilde\theta-\theta^*\rangle \\
&\lesssim \|\tilde\theta-\theta^*\|_{p^*}\sqrt{n_1+k\log(n_2)}\Big(\sqrt{\frac{\log(n_1/\delta)}{N}} + \frac{\log(n_1/\delta)}{\sqrt{N\log(n_1/\delta)}}\Big) \\
&\lesssim \|\tilde\theta-\theta^*\|_{p^*}\sqrt{\frac{n_1\log(n_1/\delta)+k\log(n_1/\delta)\log(n_2)}{N}}.
\end{aligned}$$

6.1.4. Bounding the second term in (6.8)

By Lemma 10, the image of Π is included in the image of the adjoint map $A \mapsto D^\top A\tilde D$. Since $A \mapsto D^\top (D^\top)^\dagger A\tilde D^\dagger \tilde D$ is the orthogonal projection onto this image, we thus have

$$\begin{aligned}
\langle \Pi(\varepsilon), \tilde\theta-\theta^*\rangle &= \langle D^\top (D^\top)^\dagger \Pi(\varepsilon)\tilde D^\dagger \tilde D, \tilde\theta-\theta^*\rangle \\
&= \langle (D^\dagger)^\top \Pi(\varepsilon)\tilde D^\dagger, D(\tilde\theta-\theta^*)\tilde D^\top\rangle \\
&\leq \big\|(D^\dagger)^\top \Pi(\varepsilon)\tilde D^\dagger\big\|_\infty \big\|D(\tilde\theta-\theta^*)\tilde D^\top\big\|_1
\end{aligned} \tag{6.15}$$

by Hölder's inequality.

We first consider the term $\big\|(D^\dagger)^\top \Pi(\varepsilon)\tilde D^\dagger\big\|_\infty$. By the formula (6.7) for Π and the singular value decomposition of D, it holds that

$$\begin{aligned}
(D^\dagger)^\top \Pi(\varepsilon)\tilde D^\dagger &= \sum_{(l,r)\in J^c} \Sigma^{-1}_{l,l}\Sigma^{-1}_{r,r} U_{\cdot,l}\tilde U^\top_{\cdot,r}\langle w_l\tilde w_r^\top, \Pi(\epsilon)\rangle \\
&= \sum_{(l,r)\in J^c} \Sigma^{-1}_{l,l}\Sigma^{-1}_{r,r} U_{\cdot,l}\tilde U^\top_{\cdot,r}\langle w_l\tilde w_r^\top, \mathcal{L}(\mathcal{L}^\top\Lambda^{-1}\mathcal{L})^{-1}\mathcal{L}^\top\Lambda^{-1}(\epsilon)\rangle \\
&= \sum_{(l,r)\in J^c} \Sigma^{-1}_{l,l}\Sigma^{-1}_{r,r} U_{\cdot,l}\tilde U^\top_{\cdot,r}\langle \Lambda^{-1}\mathcal{L}(\mathcal{L}^\top\Lambda^{-1}\mathcal{L})^{-1}\mathcal{L}^\top(w_l\tilde w_r^\top), \epsilon\rangle.
\end{aligned}$$

By (6.6) and the orthogonality of the vectors $\{w_l\}_{l\in[n_1]}$ and $\{\tilde w_r\}_{r\in[n_2]}$, we have that $\mathcal{L}^\top(w_l\tilde w_r^\top) = e^{(l,r)}$ if $(l,r)\in J^c$ and zero otherwise, where $e^{(l,r)}$ denotes the coordinate vector in $\mathbb{R}^{|J^c|}$ with a one in the (l,r) th component and zero in all others. Hence, if we define $a^{(i,j)} \in \mathbb{R}^{|J^c|}$ for $(i,j)\in[n_1-1]\times[n_2-1]$ by

$$(a^{(i,j)})_{l,r} := \Sigma^{-1}_{l,l}\Sigma^{-1}_{r,r}U_{i,l}\tilde U_{j,r},$$

then for $(i,j) \in [n_1 - 1] \times [n_2 - 1]$, we obtain

$$\begin{aligned}\left((D^\dagger)^\top \Pi(\varepsilon) \tilde{D}^\dagger\right)_{i,j} &= \sum_{(l,r)\in J^c} (a^{(i,j)})_{l,r} \langle \Lambda^{-1} \mathcal{L} (\mathcal{L}^\top \Lambda^{-1} \mathcal{L})^{-1} e^{(l,r)}, \epsilon\rangle \\ &= \langle \underbrace{\frac{1}{N} \Lambda^{-1} \mathcal{L} (\mathcal{L}^\top \Lambda^{-1} \mathcal{L})^{-1} a^{(i,j)}}_{=:B^{(i,j)}}, N\epsilon\rangle.\end{aligned}$$

As before, Lemma 18 yields that on an event $\mathcal{A}_3$ of probability $1-\delta$,

$$\begin{aligned}\left\|(D^\dagger)^\top \Pi(\varepsilon) \tilde{D}^\dagger\right\|_\infty \lesssim & \left(\max_{i,j} \left\|B^{(i,j)}\right\|_{Np^*}\right) \sqrt{\log(n_1/\delta)} \\ &+ \left(\max_{i,j} \left\|B^{(i,j)}\right\|_\infty\right) \log(n_1/\delta).\end{aligned}$$

We proceed to bound $\|B^{(i,j)}\|_{Np^*}$ and $\|B^{(i,j)}\|_\infty$. First,

$$\begin{aligned}\left\|B^{(i,j)}\right\|_{Np^*}^2 &= \frac{1}{N} \left\|\Lambda^{-1} \mathcal{L} (\mathcal{L}^\top \Lambda^{-1} \mathcal{L})^{-1} a^{(i,j)}\right\|_{p^*}^2 \\ &= \frac{1}{N} (a^{(i,j)})^\top (\mathcal{L}^\top \Lambda^{-1} \mathcal{L})^{-1} \mathcal{L}^\top \Lambda^{-1} \Lambda \Lambda^{-1} \mathcal{L} (\mathcal{L}^\top \Lambda^{-1} \mathcal{L})^{-1} a^{(i,j)} \\ &= \frac{1}{N} \left\|(\mathcal{L}^\top \Lambda^{-1} \mathcal{L})^{-1/2} a^{(i,j)}\right\|_2^2 \\ &\leq \frac{p^*_{\mathsf{max}}}{N} \|a^{(i,j)}\|_2^2\end{aligned}$$

by Lemma 9. Then, by definition,

$$\left\|a^{(i,j)}\right\|_2^2 = \sum_{(l,r)\in J^c} \Sigma_{l,l}^{-2} \tilde{\Sigma}_{r,r}^{-2} U_{i,l}^2 \tilde{U}_{j,r}^2 \lesssim \frac{n_1 n_2}{k} \log(n_2),$$

where the inequality is due to Lemma 11. As for $\|B^{(i,j)}\|_\infty$, we proceed as in (6.13) to obtain

$$\|B^{(i,j)}\|_\infty \leq \frac{1}{\sqrt{N p^*_{\mathsf{min}}}} \|B^{(i,j)}\|_{Np^*} \lesssim \sqrt{\frac{p^*_{\mathsf{max}} n_1 n_2 \log(n_2)}{Nk \log(n_1/\delta)}},$$

where we used again that by assumption, $N \geq 12 \log(n_1 n_2/\delta)/p^*_{\mathsf{min}}$. Combining the above bounds yields that on the event $\mathcal{A}_3$,

$$\left\|(D^\dagger)^\top \Pi(\varepsilon) \tilde{D}^\dagger\right\|_\infty \lesssim \sqrt{\frac{p^*_{\mathsf{max}} n_1 n_2 \log(n_1/\delta) \log(n_2)}{Nk}}. \tag{6.16}$$

Next, we turn to the quantity $\left\|D(\tilde{\theta} - \theta^*)\tilde{D}^\top\right\|_1$. Note that for any θ such that $D\theta\tilde{D}^\top \geq 0$, it holds

$$\|D\theta\tilde{D}^\top\|_1 = \sum_{i=1}^{n_1-1} \sum_{j=1}^{n_2-1} (D\theta\tilde{D}^\top)_{i,j}$$

$$= \sum_{i=1}^{n_1-1}\sum_{j=1}^{n_2-1} (\theta_{i,j} + \theta_{i+1,j+1} - \theta_{i+1,j} - \theta_{i,j+1})$$
$$= \theta_{1,1} + \theta_{n_1,n_2} - \theta_{n_1,1} - \theta_{1,n_2}.$$

Therefore, we obtain

$$\|D\theta^* \tilde{D}^\top\|_1 = \theta^*_{1,1} + \theta^*_{n_1,n_2} - \theta^*_{n_1,1} - \theta^*_{1,n_2} = \log \frac{p^*_{1,1} p^*_{n_1,n_2}}{p^*_{n_1,1} p^*_{1,n_2}} = \log\big(L(p^*)\big). \quad (6.17)$$

Furthermore, recall that on the event $\mathcal{A}_1$, both θ^* and $\tilde{\theta}$ lie in the set $\mathcal{C}(Y)$ defined in (2.5). Hence

$$\begin{aligned}
\|D\tilde{\theta}\tilde{D}^\top\|_1 &= \tilde{\theta}_{1,1} + \tilde{\theta}_{n_1,n_2} - \tilde{\theta}_{n_1,1} - \tilde{\theta}_{1,n_2} \\
&\le \log \frac{2Y_{1,1}}{N} + \log \frac{2Y_{n_1,n_2}}{N} - \log \frac{2Y_{n_1,1}}{3N} - \log \frac{2Y_{1,n_2}}{3N} \\
&= \log \frac{2Y_{1,1}}{3N} + \log \frac{2Y_{n_1,n_2}}{3N} - \log \frac{2Y_{n_1,1}}{N} - \log \frac{2Y_{1,n_2}}{N} + 4\log(3) \\
&\le \theta^*_{1,1} + \theta^*_{n_1,n_2} - \theta^*_{n_1,1} - \theta^*_{1,n_2} + 4\log(3) \\
&= \log\big(L(p^*)\big) + 4\log(3).
\end{aligned}$$

We conclude that on the event $\mathcal{A}_1$,

$$\big\|D(\tilde{\theta} - \theta^*)\tilde{D}^\top\big\|_1 \le \|D\tilde{\theta}\tilde{D}^\top\|_1 + \|D\theta^*\tilde{D}^\top\|_1 \le 2\log\big(L(p^*)\big) + 4\log(3). \quad (6.18)$$

It then follows from (6.15), (6.16), and (6.18) that on the event $\mathcal{A}_1 \cap \mathcal{A}_3$,

$$\langle \Pi(\varepsilon), \tilde{\theta} - \theta^* \rangle \lesssim \sqrt{\frac{p^*_{\mathsf{max}} n_1 n_2 \log(n_1/\delta)\log(n_2)}{Nk}} \Big(\log\big(L(p^*)\big) + 1 \Big).$$

6.1.5. *Finishing the proof of Theorem 3*

Combining the bounds on the two terms of (6.8) and applying (6.4), we obtain that on the event $\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3$ of probability at least $1 - 4\delta$,

$$\begin{aligned}
\|\tilde{\theta} - \theta^*\|^2_{p^*} &\lesssim \|\tilde{\theta} - \theta^*\|_{p^*} \sqrt{\frac{n_1 \log(n_1/\delta) + k\log(n_1/\delta)\log(n_2)}{N}} \\
&\quad + \Big(\log\big(L(p^*)\big) + 1 \Big) \sqrt{\frac{p^*_{\mathsf{max}} n_1 n_2 \log(n_1/\delta)\log(n_2)}{Nk}}. \quad (6.19)
\end{aligned}$$

Finally, by the definitions of $\tilde{\theta}$ and $\hat{\theta}$ in (2.6) and (2.7), it holds that

$$\begin{aligned}
\mathsf{KL}(p^*, \hat{p}) &= \sum_{i,j} p^*_{i,j} \log \frac{p^*_{i,j}}{\hat{p}_{i,j}} \\
&= \sum_{i,j} p^*_{i,j} \big(\theta^*_{i,j} - \hat{\theta}_{i,j}\big)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i,j} p^*_{i,j}\big(\theta^*_{i,j} - \tilde{\theta}_{i,j}\big) + \log \sum_{i,j} e^{\tilde{\theta}_{i,j}} \\
&\le \langle p^*, \theta^* - \tilde{\theta}\rangle + \sum_{i,j} e^{\tilde{\theta}_{i,j}} - 1 \\
&\le 2\|\tilde{\theta} - \theta^*\|^2_{p^*}, \qquad (6.20)
\end{aligned}
$$

where the first inequality holds because $\log x \le x - 1$ and the second holds thanks to (6.3). Therefore, we conclude from (6.19) and (6.20) that

$$
\begin{aligned}
\mathsf{KL}(p^*, \hat{p}) \lesssim{}& \frac{n_1 \log(n_1/\delta) + k \log(n_1/\delta)\log(n_2)}{N} \\
&+ \Big(\log\big(L(p^*)\big) + 1\Big)\sqrt{\frac{p^*_{\mathsf{max}} n_1 n_2 \log(n_1/\delta)\log(n_2)}{Nk}}.
\end{aligned}
$$

Balancing out the terms that depend on k yields the optimal choice

$$
k = \Big(\log\big(L(p^*)\big) + 1\Big)^{2/3}\Big(\frac{p^*_{\mathsf{max}} n_1 n_2 N}{\log(n_2)\log(n_1/\delta)}\Big)^{1/3},
$$

which leads to

$$
\begin{aligned}
\mathsf{KL}(p^*, \hat{p}) \lesssim{}& \frac{n_1 \log(n_1/\delta)}{N} \\
&+ (p^*_{\mathsf{max}} n_1 n_2)^{1/3}\Big(\log\big(L(p^*)\big) + 1\Big)^{2/3}\Big(\frac{\log(n_1/\delta)\log(n_2)}{N}\Big)^{2/3}.
\end{aligned}
$$

Since the KL divergence dominates the Hellinger distance by the first inequality in (6.1), this completes the proof.

Remark 12. *It is not hard to see that, if we choose the set J in (6.5) instead to be the entire grid $[n_1] \times [n_2]$, then the same argument yields the rate*

$$
\mathsf{h}^2(p^*, \hat{p}) \lesssim \frac{n_1 n_2 \log(n_1/\delta)}{N},
$$

which matches the rate of the empirical frequency matrix in Lemma 20 up to a logarithmic factor. In fact, the numerical experiments in Section 5, in particular Figure 2a, suggest that the performance of $\hat{p}$ exactly matches that of the empirical frequency matrix in this regime.

6.1.6. Proof of Lemma 9

Let $B \in \mathbb{R}^{|J^c|}$ with $\|B\|_2 = 1$. Since $\mathcal{L}B$ is a sum of matrices that are orthonormal with respect to the standard inner product, weighted by the entries of B, it holds that $\|\mathcal{L}B\|_2 = 1$. Hence,

$$
B^\top \mathcal{L}^\top \Lambda^{-1} \mathcal{L} B \ge \min_{G : \|G\|_2 = 1} G^\top \Lambda^{-1} G = \lambda_{\min}(\Lambda^{-1}) = \frac{1}{p^*_{\mathsf{max}}},
$$

which yields the first claim.

For the second claim, recall that ΠA is defined to be the orthogonal projection of A onto the image of $\mathcal{L}$ with respect to the inner product $\langle\cdot,\cdot\rangle_{1/p^*}$. Thus $\Pi A = \mathcal{L}B$ where $B \in \mathbb{R}^{|J^c|}$ minimizes

$$\|\mathcal{L}B - A\|_{1/p^*}^2 = \langle \mathcal{L}B - A, \Lambda^{-1}(\mathcal{L}B - A)\rangle.$$

The first-order optimality condition then gives the desired formula for Π.

6.1.7. Proof of Lemma 10

The image of the map $A \mapsto D^\top A\tilde{D}$ is the orthogonal complement of the kernel of $A \mapsto DA\tilde{D}^\top$, which can be characterized as follows.

The matrices Σ, U and W in the singular value decomposition $D = U\Sigma W^\top$ are known [45] to be

$$\Sigma_{i,i} = 2\left|\sin\left(\frac{\pi i}{2n_1}\right)\right|, \quad i \in [n_1 - 1], \tag{6.21}$$

$$U_{i,j} = \sqrt{\frac{2}{n_1}}\sin\left(\frac{\pi i j}{n_1}\right), \quad i, j \in [n_1 - 1], \tag{6.22}$$

$$\text{and} \quad W_{i,j} = \begin{cases} \sqrt{\dfrac{2}{n_1}}\cos\left(\dfrac{\pi j(i - 1/2)}{n_1}\right), & j \in [n_1 - 1],\, i \in [n_1] \\ \dfrac{1}{\sqrt{n_1}}, & j = n_1. \end{cases}$$

Fix a matrix A for which $DA\tilde{D}^\top = 0$. Then we have $\Sigma W^\top A\tilde{W}\tilde{\Sigma}^\top = 0$, so the matrix $W^\top A\tilde{W}$ has all entries equal to zero except on its last row and last column. Consequently, it holds that

$$A = WW^\top A\tilde{W}\tilde{W}^\top = \sum_{i=1}^{n_1} w_i w_i^\top A\tilde{w}_{n_2}\tilde{w}_{n_2}^\top + \sum_{j=1}^{n_2-1} w_{n_1} w_{n_1}^\top A\tilde{w}_j\tilde{w}_j^\top.$$

Hence, the orthogonal complement of the kernel of $A \mapsto DA\tilde{D}^\top$ is spanned by the matrices $\{w_l\tilde{w}_r^\top : (l, r) \in [n_1 - 1] \times [n_2 - 1]\}$. By the definition of Π as the projection onto the span of $\{w_i\tilde{w}_j^\top : (i, j) \in J^c\}$, its image is contained in the kernel of $A \mapsto DA\tilde{D}^\top$.

6.1.8. Proof of Lemma 11

This result can be easily obtained from the proof of Lemma 10 of [19], but we provide a complete proof for the reader's convenience.

We start with the first bound in the lemma. Without loss of generality, assume that n_1 is odd, so $n_1 - 1$ is even. Note that because of the symmetry

$$\sin\left(\frac{\pi i j}{n_1}\right) = \sin\left(\frac{\pi j(n_1 - i)}{n_1}\right), \quad i = 1, \ldots, n_1 - 1,$$

it is enough to consider $i = 1, \ldots, \frac{n_1 - 1}{2}$. We make use of the following inequalities to control the sin terms involved:

$$\begin{aligned} |\sin(x)| &\leq 1, && \text{for all } x \in \mathbb{R}; \\ \sin(x) &\leq x, && \text{for } x \in [0, \infty); \\ \sin(x) &\geq \frac{2}{\pi} x \geq \frac{1}{2} x, && \text{for } x \in [0, \frac{\pi}{2}]. \end{aligned}$$

Plugging in the entries of Σ and U as stated in (6.21) and (6.22), respectively, yields

$$\begin{aligned} \sum_{(l,r) \in J^c} \Sigma_{l,l}^{-2} \tilde{\Sigma}_{r,r}^{-2} U_{i,l}^2 \tilde{U}_{j,r}^2 = {} & \sum_{(l,r) \in J^c} \frac{4 \sin\left(\frac{\pi i l}{n_1}\right)^2 \sin\left(\frac{\pi j r}{n_2}\right)^2}{16 n_1 n_2 \sin\left(\frac{\pi l}{2 n_1}\right)^2 \sin\left(\frac{\pi r}{2 n_2}\right)^2} \\ & \lesssim \frac{1}{n_1 n_2} \sum_{(l,r) \in J^c} \frac{n_1^2 n_2^2}{l^2 r^2} \\ & \lesssim n_1 n_2 \sum_{r=1}^{n_2} \left(\frac{1}{r^2} \sum_{l = \lceil k/r \rceil}^{n_1} \frac{1}{l^2} \right) \\ & \lesssim n_1 n_2 \sum_{r=1}^{n_2} \left(\frac{1}{r^2} \sum_{l = \lceil k/r \rceil + 1}^{n_1} \frac{1}{l^2} \right) + n_1 n_2 \sum_{r=1}^{n_2} \frac{1}{r^2} \frac{1}{\lceil k/r \rceil^2} \\ & \lesssim n_1 n_2 \sum_{r=1}^{n_2} \frac{1}{r^2} \frac{r}{k} + n_1 n_2 \sum_{r=1}^{k} \frac{1}{k^2} + n_1 n_2 \sum_{r=k+1}^{n_2} \frac{1}{r^2} \\ & \lesssim \frac{n_1 n_2}{k} \log(n_2) + \frac{n_1 n_2}{k} + \frac{n_1 n_2}{k} \lesssim \frac{n_1 n_2}{k} \log(n_2), \end{aligned}$$

where we have twice used the bound $\sum_{r=k+1}^{\infty} \frac{1}{r^2} \leq \frac{1}{k}$ for any $k \geq 1$.

6.2. Proof of Theorem 6

With the notation introduced in Section 3.1, we define the piecewise constant density

$$\bar{\rho}(x) := n^2 p_{i,j}^* = n^2 \int_{S_{i,j}} \rho^*(x) \, \mathrm{d}x, \quad x \in S_{i,j}, \, i, j \in [n]. \tag{6.23}$$

By the triangle inequality for the Hellinger distance, we can estimate

$$\mathsf{h}^2(\hat{\rho}, \rho^*) \leq 2 \, \mathsf{h}^2(\hat{\rho}, \bar{\rho}) + 2 \, \mathsf{h}^2(\bar{\rho}, \rho^*), \tag{6.24}$$

and we proceed to bound the two quantities on the right-hand side of (6.24).

For the first term on the right-hand side of (6.24), we have

$$\begin{aligned}\mathsf{h}^2(\hat\rho,\bar\rho) &= \int_{[0,1]^2}\Big(\sqrt{\hat\rho(x)}-\sqrt{\bar\rho(x)}\Big)^2\,\mathrm{d}x\\ &= \sum_{i,j=1}^n \frac{1}{n^2}\Big(\sqrt{n^2\hat p_{i,j}}-\sqrt{n^2 p^*_{i,j}}\Big)^2\\ &= \mathsf{h}^2(\hat p,p^*).\end{aligned}$$

By assumption (3.4) and definition (3.6), we have that $p^*_{\mathsf{min}} \ge \rho_{\min}/n^2$. Hence if $N \ge \frac{12n^2\log(n^2/\delta)}{\rho_{\min}}$, then the results for the estimator $\hat p$ in Theorem 3 lead to

$$\mathsf{h}^2(\hat\rho,\bar\rho) \lesssim \frac{n\log(n/\delta)}{N} + (p^*_{\mathsf{max}}\,n^2)^{1/3}\Big(\log\big(L(p^*)\big)+1\Big)^{2/3}\Big(\frac{\log(n/\delta)\log(n)}{N}\Big)^{2/3}$$

with probability at least $1-4\delta$. Moreover, recall definition (2.8) and note that

$$p^*_i \le \frac{\rho_{\max}}{n^2} \quad \text{and} \quad L(p^*) \le \frac{\rho^2_{\max}}{\rho^2_{\min}}.$$

It then follows that

$$\mathsf{h}^2(\hat\rho,\bar\rho) \lesssim \frac{n\log(n/\delta)}{N} + (\rho_{\max})^{1/3}\Big(\log\Big(\frac{\rho_{\max}}{\rho_{\min}}\Big)+1\Big)^{2/3}\Big(\frac{\log(n/\delta)\log(n)}{N}\Big)^{2/3}. \tag{6.25}$$

To bound the second term on the right-hand side of (6.24), note that by the mean value theorem for integrals and the continuity of ρ^*, for all $i,j\in[n]$, there exist $\zeta_{i,j}$ such that

$$\bar\rho(x) = n^2\int_{S_{i,j}}\rho^*(y)\,\mathrm{d}y = \rho^*(\zeta_{i,j}).$$

Note that for any $a,b>0$, it holds that

$$\big|\sqrt{a}-\sqrt{b}\big| \le \frac{|a-b|}{\sqrt{a}\vee\sqrt{b}}.$$

Moreover, assumptions (3.2) and (3.3) imply that

$$|\rho^*(x)-\rho^*(y)| \le R\|x-y\|_2^{\tilde\beta}, \quad x,y\in[0,1]^2,$$

where we recall $\tilde\beta = \beta\wedge 1$. Combining the above facts, we obtain

$$\begin{aligned}\mathsf{h}^2(\bar\rho,\rho^*) &= \sum_{i,j=1}^n\int_{S_{i,j}}\Big(\sqrt{\rho^*(x)}-\sqrt{\rho^*(\zeta_{i,j})}\Big)^2\,\mathrm{d}x\\ &\le \frac{1}{\rho_{\min}}\sum_{i,j=1}^n\int_{S_{i,j}}\Big(\rho^*(x)-\rho^*(\zeta_{i,j})\Big)^2\,\mathrm{d}x\end{aligned}$$

$$\leq \frac{R^2}{\rho_{\min}} \sum_{i,j=1}^{n} \int_{S_{i,j}} \mathrm{diam}(S_{i,j})^{2\tilde{\beta}} = \frac{2R^2}{\rho_{\min}} n^{-2\tilde{\beta}}. \tag{6.26}$$

Plugging inequalities (6.25) and (6.26) into (6.24), we conclude that for $N \geq \frac{12n^2\log(n^2/\delta)}{\rho_{\min}}$,

$$\begin{aligned} \mathsf{h}^2(\hat{\rho}, \rho^*) \lesssim{}& \frac{n\log(n/\delta)}{N} \\ &+ (\rho_{\max})^{1/3}\Big(\log\Big(\frac{\rho_{\max}}{\rho_{\min}}\Big)+1\Big)^{2/3}\Big(\frac{\log(n/\delta)\log(n)}{N}\Big)^{2/3} + \frac{R^2}{\rho_{\min}} n^{-2\tilde{\beta}} \end{aligned}$$

with probability $1-4\delta$. Setting

$$n = \Big\lfloor \Big(\frac{R^2 N}{\rho_{\min}}\Big)^{1/(2\tilde{\beta}+1)} \wedge \Big(\frac{\rho_{\min} N}{24\log(\frac{\rho_{\min}N}{12\delta})}\Big)^{1/2} \Big\rfloor$$

then leads to the bound

$$\begin{aligned} \mathsf{h}^2(\hat{\rho}, \rho^*) \lesssim{}& \frac{\log(RN/\delta)}{\rho_{\min}^{1/(2\tilde{\beta}+1)}} \frac{R^{2/(2\tilde{\beta}+1)}}{N^{2\tilde{\beta}/(2\tilde{\beta}+1)}} \\ &+ (\rho_{\max})^{1/3}\Big(\log\Big(\frac{\rho_{\max}}{\rho_{\min}}\Big)+1\Big)^{2/3}\Big(\frac{\log(RN/\delta)\log(RN)}{N}\Big)^{2/3}, \end{aligned} \tag{6.27}$$

for $\beta > 0.5$ and $N \gtrsim \big[\frac{R^4}{\rho_{\min}^{2\tilde{\beta}+3}} \log^{2\tilde{\beta}+1}(\frac{R}{\rho_{\min}\delta})\big]^{\frac{1}{2\tilde{\beta}-1}}$, and

$$\begin{aligned} &\mathsf{h}^2(\hat{\rho}, \rho^*) \lesssim \Big[\frac{\rho_{\min}\log(N/\delta)}{N}\Big]^{1/2} \\ &+ (\rho_{\max})^{1/3}\Big(\log\Big(\frac{\rho_{\max}}{\rho_{\min}}\Big)+1\Big)^{2/3}\Big(\frac{\log(N/\delta)\log(N)}{N}\Big)^{2/3} + \frac{R^2}{\rho_{\min}}\Big[\frac{\log(N/\delta)}{\rho_{\min}N}\Big]^{\tilde{\beta}}, \end{aligned} \tag{6.28}$$

for $0 < \beta \leq 0.5$, where the hidden constants depend on β. Choosing $\delta = 1/(4N^4)$ completes the proof.

6.3. Proof of Theorem 4

We prove the theorem by treating the two terms n_1/N and $1/N^{2/3}$ separately.

6.3.1. The first term n_1/N

Without loss of generality, we assume that 8 divides n_1. Let d_{H} denote the Hamming distance between two binary vectors. By the Gilbert-Varshamov bound (see, for example, [32, Lemma 4.10]), there exists a set $\{w^{(k)}\}_{k=1}^{M}$ of points in $\{0,1\}^{n_1}$ such that

- $d_{\mathsf{H}}(0, w^{(k)}) = n_1/4$,
- $d_{\mathsf{H}}(w^{(k)}, w^{(\ell)}) \geq n_1/8$ for all distinct $k, \ell \in [M]$, and
- $\log(M) \geq n_1/30$.

For $\delta \in [0,1]$ and each $k \in [M]$, let us define a density $p^{(k)}$ on $[n_1] \times [n_2]$ by

$$p_{i,j}^{(k)} = \frac{4(1+\delta w_i^{(k)})}{(4+\delta)n_1 n_2} \quad \text{for } (i,j) \in [n_1] \times [n_2].$$

Note that each $p^{(k)}$ is indeed a density because $\sum_i w_i^{(k)} = n_1/4$ and thus

$$\sum_{i,j} p_{i,j}^{(k)} = \sum_{i,j} \frac{4(1+\delta w_i^{(k)})}{(4+\delta)n_1 n_2} = \frac{4}{4+\delta} + \frac{4\delta}{(4+\delta)n_1} \sum_i w_i^{(k)} = 1.$$

Also, each $p^{(k)}$ is totally positive because it has constant rows and therefore $p_{i,j}^{(k)} p_{i+1,j+1}^{(k)} = p_{i,j+1}^{(k)} p_{i+1,j}^{(k)}$.

Furthermore, since $\delta \in [0,1]$, we see that $\frac{4}{5n_1n_2} \leq p_{i,j}^{(k)} \leq \frac{8}{5n_1n_2}$. The relation (6.1) yields

$$\mathsf{KL}(p^{(k)}, p^{(\ell)}) \leq 5\, \mathsf{h}^2(p^{(k)}, p^{(\ell)}). \tag{6.29}$$

Note that $p_{i,j}^{(k)}$ can only take two possible values $\frac{4}{(4+\delta)n_1n_2}$ or $\frac{4(1+\delta)}{(4+\delta)n_1n_2}$, so $\left(\sqrt{p_{i,j}^{(k)}} - \sqrt{p_{i,j}^{(\ell)}}\right)^2$ is either 0 or $\frac{4(\sqrt{1+\delta}-1)^2}{(4+\delta)n_1n_2}$. Therefore, we have

$$\mathsf{h}^2(p^{(k)}, p^{(\ell)}) = \sum_{i \in [n_1],\, j \in [n_2]} \left(\sqrt{p_{i,j}^{(k)}} - \sqrt{p_{i,j}^{(\ell)}}\right)^2 = d_{\mathsf{H}}(w^{(k)}, w^{(\ell)}) \frac{4(\sqrt{1+\delta}-1)^2}{(4+\delta)n_1}. \tag{6.30}$$

Together with the condition $n_1/8 \leq d_{\mathsf{H}}(w^{(k)}, w^{(\ell)}) \leq n_1$ for $k \neq \ell$, relations (6.29) and (6.30) yield

$$\mathsf{h}^2(p^{(k)}, p^{(\ell)}) \geq \frac{(\sqrt{1+\delta}-1)^2}{2(4+\delta)} \quad \text{and} \quad \mathsf{KL}(p^{(k)}, p^{(\ell)}) \leq \frac{20(\sqrt{1+\delta}-1)^2}{(4+\delta)}. \tag{6.31}$$

Additionally, since the KL divergence tensorizes, if we let $p^{\otimes N}$ denote the distribution of N independent observations sampled according to the density p on $[n_1] \times [n_2]$, then

$$\mathsf{KL}\left((p^{(k)})^{\otimes N}, (p^{(\ell)})^{\otimes N}\right) = N\ \mathsf{KL}(p^{(k)}, p^{(\ell)}) \leq N \frac{20(\sqrt{1+\delta}-1)^2}{(4+\delta)}. \tag{6.32}$$

For a sufficiently small positive constant c_1, we choose $\delta \in [0,1]$ so that $N\frac{20(\sqrt{1+\delta}-1)^2}{(4+\delta)} = c_1 n_1 \leq 0.1 \log(M)$. We can apply [46, Theorem 2.5] together with (6.31) and (6.32) to obtain that

$$\inf_{\hat{p}} \sup_{p^* \ \mathrm{MTP}_2} \mathbb{P}_{(p^*)^{\otimes N}} \left\{ \mathsf{h}^2(\hat{p}, p^*) \geq c_2 \frac{n_1}{N} \right\} \geq 0.1.$$

6.3.2. *The second term* $1/N^{2/3}$

We turn to the second term in the lower bound. Consider positive integers $k_1 \le n_1$ and $k_2 \le n_2$ such that 4 divides k_1k_2, and k_i divides n_i for $i = 1, 2$, without loss of generality. If k_i does not divide n_i, with minor revision, the proof works on a sub-grid $[n_1'] \times [n_2']$ where k_i divides n_i' and $n_i/2 \le n_i' \le n_i$. Thus we make these mild assumptions to ease the notation.

The strategy of proving the lower bound is based on constructing an appropriate packing of supermodular log-densities, which correspond to totally positive densities. By the Gilbert-Varshamov bound [32, Lemma 4.10] again, we obtain a set $\{\tau^{(\ell)}\}_{\ell=1}^{M}$ of matrices in $\{0,1\}^{k_1 \times k_2}$ such that

- $d_{\mathsf{H}}(0, \tau^{(\ell)}) = k_1k_2/4$,
- $d_{\mathsf{H}}(\tau^{(\ell)}, \tau^{(r)}) \ge k_1k_2/8$ for all distinct $\ell, r \in [M]$, and
- $\log(M) \ge k_1k_2/30$.

For each $\tau^{(\ell)}$, we need to carefully define a log-density $\theta^{(\ell)} \in \mathbb{R}^{n_1 \times n_2}$ that is supermodular and amenable to distance calculation. To that end, we have the following construction which simplifies computation later. For $i \in [n_1]$ and $j \in [n_2]$, define $u_i := \lceil ik_1/n_1 \rceil \in [k_1]$ and $v_j := \lceil jk_2/n_2 \rceil \in [k_2]$. Moreover, for any $\delta \in [0, 1/6]$ and $(i, j) \in [n_1] \times [n_2]$, we define $\tilde{\delta}_{u_i,v_j} \in [\delta, 3\delta] \subset [0, 1/2]$ so that

$$\exp\left(\frac{u_iv_j}{k_1k_2} + \frac{\tilde{\delta}_{u_i,v_j}}{k_1k_2}\right) - \exp\left(\frac{u_iv_j}{k_1k_2}\right) = \exp\left(1 + \frac{\delta}{k_1k_2}\right) - \exp(1). \tag{6.33}$$

To see why $\tilde{\delta}_{u_i,v_j}$ is properly defined in the range $[\delta, 3\delta]$, first note that the quantity $\tilde{\delta}_{u_i,v_j}$ is larger for smaller $u_iv_j \in [k_1k_2]$. Hence it suffices to check that there exists $\delta' \in [\delta, 3\delta]$ such that

$$\exp\left(\frac{\delta'}{k_1k_2}\right) - \exp(0) = \exp\left(1 + \frac{\delta}{k_1k_2}\right) - \exp(1).$$

This follows from that for $x \in [0, 1/6]$,

$$\exp(x) - \exp(0) \le \exp(1 + x) - \exp(1) \le \exp(3x) - \exp(0).$$

With $\tau^{(\ell)}_{u_i,v_j}$ chosen earlier and $\tilde{\delta}_{u_i,v_j}$ defined in (6.33), we consider the quantity

$$\tilde{\theta}^{(\ell)}_{i,j} := \frac{u_iv_j}{k_1k_2} + \tau^{(\ell)}_{u_i,v_j}\frac{\tilde{\delta}_{u_i,v_j}}{k_1k_2}, \tag{6.34}$$

and further define the log-density

$$\begin{aligned}\theta^{(\ell)}_{i,j} &:= \tilde{\theta}^{(\ell)}_{i,j} - \log \sum_{s\in[n_1],\, t\in[n_2]} \exp\left(\tilde{\theta}^{(\ell)}_{i,j}\right) \\ &= \frac{u_iv_j}{k_1k_2} + \tau^{(\ell)}_{u_i,v_j}\frac{\tilde{\delta}_{u_i,v_j}}{k_1k_2} - \log \underbrace{\sum_{s\in[n_1],\, t\in[n_2]} \exp\left(\frac{u_sv_t}{k_1k_2} + \tau^{(\ell)}_{u_s,v_t}\frac{\tilde{\delta}_{u_s,v_t}}{k_1k_2}\right)}_{=:\,\mathfrak{N}}. \end{aligned} \tag{6.35}$$

Finally, the density $p^{(\ell)}$ is defined by $p^{(\ell)}_{i,j} := \exp(\theta^{(\ell)}_{i,j})$.

The normalization factor $\mathfrak{N}$ in (6.35) guarantees that $\sum_{i,j}\exp(\theta^{(\ell)}_{i,j})=1$, so $p^{(\ell)}$ is indeed a density. Moreover, crucial to our computation later, the normalization factor in fact does not depend on $\ell\in[M]$ thanks to the definition of $\tilde{\delta}_{u_i,v_j}$ in (6.33); namely,

$$
\begin{aligned}
\mathfrak{N} &= \sum_{s\in[n_1],\, t\in[n_2]} \exp\left(\frac{u_s v_t}{k_1k_2} + \tau^{(\ell)}_{u_s,v_t}\frac{\tilde{\delta}_{u_s,v_t}}{k_1k_2}\right)\\
&= \sum_{s\in[n_1],\, t\in[n_2]} \exp\left(\frac{u_s v_t}{k_1k_2}\right)\\
&\quad + \sum_{(s,t):\, \tau^{(\ell)}_{u_s,v_t}=1} \left[\exp\left(\frac{u_s v_t}{k_1k_2}+\frac{\tilde{\delta}_{u_s,v_t}}{k_1k_2}\right) - \exp\left(\frac{u_s v_t}{k_1k_2}\right)\right]\\
&= \sum_{s\in[n_1],\, t\in[n_2]} \exp\left(\frac{u_s v_t}{k_1k_2}\right) + \left|\left\{(s,t):\, \tau^{(\ell)}_{u_s,v_t}=1\right\}\right|\cdot\left[\exp\left(1+\frac{\delta}{k_1k_2}\right)-e\right]\\
&= \sum_{s\in[n_1],\, t\in[n_2]} \exp\left(\frac{u_s v_t}{k_1k_2}\right) + \frac{n_1n_2}{4}\left[\exp\left(1+\frac{\delta}{k_1k_2}\right)-e\right]
\end{aligned}
$$

where the last equality follows from that $d_{\mathsf{H}}(0,\tau^{(\ell)})=k_1k_2/4$ and the definitions of u_s and v_t.

Next, we check that $\theta^{(\ell)}$ is supermodular, so that $p^{(\ell)}$ is totally positive. Since $\tilde{\theta}^{(\ell)}\in\mathbb{R}^{n_1\times n_2}$ defined in (6.34) is equal to $\theta^{(\ell)}$ plus a common constant on each entry, it suffices to check that $\tilde{\theta}^{(\ell)}_{i,j}+\tilde{\theta}^{(\ell)}_{i+1,j+1}-\tilde{\theta}^{(\ell)}_{i,j+1}-\tilde{\theta}^{(\ell)}_{i+1,j}\ge 0$. There are two cases:

1. If $u_i=u_{i+1}$ or $v_j=v_{j+1}$, then we have, respectively, either $\tilde{\theta}^{(\ell)}_{i,j}=\tilde{\theta}^{(\ell)}_{i+1,j}$, $\tilde{\theta}^{(\ell)}_{i,j+1}=\tilde{\theta}^{(\ell)}_{i+1,j+1}$ or $\tilde{\theta}^{(\ell)}_{i,j}=\tilde{\theta}^{(\ell)}_{i,j+1}$, $\tilde{\theta}^{(\ell)}_{i+1,j}=\tilde{\theta}^{(\ell)}_{i+1,j+1}$. In both cases, the difference above is 0.
2. Otherwise, we have $u_{i+1}=u_i+1$ and $v_{j+1}=v_j+1$. Then it holds

$$
\begin{aligned}
&\tilde{\theta}^{(\ell)}_{i,j}+\tilde{\theta}^{(\ell)}_{i+1,j+1}-\tilde{\theta}^{(\ell)}_{i,j+1}-\tilde{\theta}^{(\ell)}_{i+1,j}\\
&= \frac{u_iv_j+(u_i+1)(v_j+1)-u_i(v_j+1)-(u_i+1)v_j}{k_1k_2}\\
&\quad + \frac{\tau^{(\ell)}_{u_i,v_j}\tilde{\delta}_{u_i,v_j}+\tau^{(\ell)}_{u_{i+1},v_{j+1}}\tilde{\delta}_{u_{i+1},v_{j+1}}-\tau^{(\ell)}_{u_i,v_{j+1}}\tilde{\delta}_{u_i,v_{j+1}}-\tau^{(\ell)}_{u_{i+1},v_j}\tilde{\delta}_{u_{i+1},v_j}}{k_1k_2}\\
&\ge \frac{1}{k_1k_2}-\frac{6\delta}{k_1k_2}\ge 0,
\end{aligned}
$$

since $\tilde{\delta}_{u_i,v_j}\le 3\delta\le 1/2$.

Having verified that each $\theta^{(\ell)}$ is a supermodular log-density, we proceed to study $\mathsf{h}^2(\theta^{(\ell)}, \theta^{(r)})$ for distinct $\ell, r \in [M]$. Since the normalization term $\mathfrak{N}$ does not depend on the index ℓ, definition (6.35) yields

$$|\theta^{(\ell)}_{i,j} - \theta^{(r)}_{i,j}| = \big|\tau^{(\ell)}_{u_i,v_j} - \tau^{(r)}_{u_i,v_j}\big| \frac{\tilde{\delta}_{u_i,v_j}}{k_1 k_2} \le \frac{1}{2}.$$

By the definition of the Hellinger distance, it holds that

$$\begin{aligned}
\mathsf{h}^2(p^{(\ell)}, p^{(r)}) &= \sum_{i\in[n_1],\, j\in[n_2]} \Big(\sqrt{p^{(\ell)}_{i,j}} - \sqrt{p^{(r)}_{i,j}}\Big)^2 \\
&= \sum_{i\in[n_1],\, j\in[n_2]} \Big(\exp(\theta^{(\ell)}_{i,j}/2) - \exp(\theta^{(r)}_{i,j}/2)\Big)^2 \\
&= \sum_{i\in[n_1],\, j\in[n_2]} p^{(\ell)}_{i,j}\Big(1 - \exp\big((\theta^{(r)}_{i,j} - \theta^{(\ell)}_{i,j})/2\big)\Big)^2.
\end{aligned}$$

Using the approximation $x^2/2 \le (1-e^x)^2 \le 2x^2$ for $|x| \le 1/4$, we obtain

$$\frac{1}{8} \sum_{i\in[n_1],\, j\in[n_2]} p^{(\ell)}_{i,j}\big(\theta^{(r)}_{i,j} - \theta^{(\ell)}_{i,j}\big)^2 \le \mathsf{h}^2(p^{(\ell)}, p^{(r)}) \le \frac{1}{2} \sum_{i\in[n_1],\, j\in[n_2]} p^{(\ell)}_{i,j}\big(\theta^{(r)}_{i,j} - \theta^{(\ell)}_{i,j}\big)^2.$$

Furthermore, it is easily seen from (6.35) that $|\theta^{(\ell)}_{i,j} - \theta^{(\ell)}_{i,'j'}| \le 1.5$, so that $1/5 \le p^{(\ell)}_{i,j}/p^{(\ell)}_{i,'j'} \le 5$ for any $(i,j), (i',j') \in [n_1] \times [n_2]$. As a result, we obtain $\frac{1}{5n_1n_2} \le p^{(\ell)}_{i,j} \le \frac{5}{n_1n_2}$ and thus

$$\frac{1}{40 n_1 n_2}\big\|\theta^{(\ell)} - \theta^{(r)}\big\|_2^2 \le \mathsf{h}^2(p^{(\ell)}, p^{(r)}) \le \frac{5}{2n_1n_2}\big\|\theta^{(\ell)} - \theta^{(r)}\big\|_2^2. \tag{6.36}$$

In addition, it follows from (6.1) that

$$\mathsf{KL}(p^{(\ell)}, p^{(r)}) \le 11\ \mathsf{h}^2(p^{(\ell)}, p^{(r)}). \tag{6.37}$$

It remains to study $\|\theta^{(\ell)} - \theta^{(r)}\|_2^2$. To this end, we obtain from (6.35) that

$$\begin{aligned}
\sum_{i\in[n_1], j\in[n_2]} (\theta^{(\ell)}_{i,j} - \theta^{(r)}_{i,j})^2 &= \sum_{i\in[n_1], j\in[n_2]} \big(\tau^{(\ell)}_{u_i,v_j} - \tau^{(r)}_{u_i,v_j}\big)^2 \Big(\frac{\tilde{\delta}_{u_i,v_j}}{k_1k_2}\Big)^2 \\
&= \sum_{u\in[k_1], v\in[k_2]} \frac{n_1n_2}{k_1k_2}\big(\tau^{(\ell)}_{u,v} - \tau^{(r)}_{u,v}\big)^2 \frac{\tilde{\delta}^2_{u,v}}{k_1^2k_2^2}.
\end{aligned}$$

Since $\tilde{\delta}_{u,v} \in [\delta, 3\delta]$, we have the bounds

$$\frac{\delta^2 n_1 n_2}{k_1^3 k_2^3} \sum_{u\in[k_1], v\in[k_2]} \big(\tau^{(\ell)}_{u,v} - \tau^{(r)}_{u,v}\big)^2 \le \|\theta^{(\ell)} - \theta^{(r)}\|_2^2$$

$$\leq \frac{9\delta^2 n_1 n_2}{k_1^3 k_2^3} \sum_{u\in[k_1], v\in[k_2]} \left(\tau_{u,v}^{(\ell)} - \tau_{u,v}^{(r)}\right)^2.$$

By the construction of the packing $\{\tau^{(\ell)}\}_{\ell\in[M]}$,

$$\frac{k_1k_2}{8} \leq \sum_{u\in[k_1], v\in[k_2]} \left(\tau_{u,v}^{(\ell)} - \tau_{u,v}^{(r)}\right)^2 = d_{\mathsf{H}}(\tau^{(\ell)}, \tau^{(r)}) \leq k_1k_2.$$

Therefore, combining the above bounds yields that

$$\frac{\delta^2 n_1 n_2}{8k_1^2k_2^2} \leq \|\theta^{(\ell)} - \theta^{(r)}\|_2^2 \leq \frac{9\delta^2 n_1 n_2}{k_1^2 k_2^2}.$$

This together with (6.36) implies that

$$\frac{\delta^2}{320k_1^2k_2^2} \leq \mathsf{h}^2(p^{(\ell)}, p^{(r)}) \leq \frac{45\delta^2}{2k_1^2k_2^2}. \tag{6.38}$$

To complete the proof, we may choose $\delta = c_1\big(\frac{k_1^3k_2^3}{N}\big)^{1/2} \in [0, 1/6]$ for a sufficiently small constant $c_1 > 0$, provided that $k_1^3k_2^3 \lesssim N$. Then the bounds (6.38) and (6.37) combined imply that

$$\mathsf{KL}\left((p^{(k)})^{\otimes N}, (p^{(\ell)})^{\otimes N}\right) = N\ \mathsf{KL}(p^{(\ell)}, p^{(r)}) \leq c_2k_1k_2 \leq 0.1\log(M).$$

Thus, we can apply [46, Theorem 2.5] together with the lower bound in (6.38) to see that

$$\inf_{\hat{p}} \sup_{p^*\ \mathrm{MTP}_2} \mathbb{P}_{(p^*)^{\otimes N}}\left\{ \mathsf{h}^2(\hat{p}, p^*) \geq c_2 \frac{k_1k_2}{N}\right\} \geq 0.1,$$

where we continue to use the notation $\hat{\theta}_{i,j} = \log \hat{p}_{i,j}$ and $\theta^*_{i,j} = \log p^*_{i,j}$.

Note that k_1k_2 needs to be chosen so that $\delta = c_1\big(\frac{k_1^3k_2^3}{N}\big)^{1/2} \leq 1/6$. Hence if $N \lesssim n_1^3n_2^3$, then we choose $k_1k_2 \asymp N^{1/3}$ to obtain the lower bound of order $N^{-2/3}$. If $N \gtrsim n_1^3n_2^3$, then we choose $k_1 = n_1$ and $k_2 = n_2$ to obtain the lower bound of order $\frac{n_1n_2}{N}$.

6.4. Proof of Theorem 7

We first set up the proof for smooth densities, and then prove the theorem for each regime of β.

6.4.1. Differential characterization and setup

We begin by stating a short lemma that yields a condition for total positivity in terms of the derivatives of a density.

Lemma 13. *A function $f \in C^2([0,1]^2)$ fulfills*

$$f(w,z) + f(x,y) \geq f(x,z) + f(w,y), \quad \textit{for all } 0 \leq x \leq w \leq 1,\ 0 \leq y \leq z \leq 1 \tag{6.39}$$

if and only if

$$\partial_1 \partial_2 f(x,y) \geq 0, \quad \textit{for all } x, y \in [0,1]. \tag{6.40}$$

Moreover, if $\rho > 0$ is a probability density in $C^2([0,1]^2)$, then ρ is totally positive if and only if $\log \rho$ fulfills (6.40), if and only if

$$-\frac{1}{\rho(x,y)^2} \partial_1 \rho(x,y) \partial_2 \rho(x,y) + \frac{1}{\rho(x,y)} \partial_1 \partial_2 \rho(x,y) \geq 0, \quad \textit{for all } x, y \in [0,1]. \tag{6.41}$$

Proof. The first claim follows easily from the fundamental theorem of calculus and the continuity of f. To obtain the second claim, note that because ρ is bounded away from zero, we can take logarithms, and the MTP_2 condition (3.1) is equivalent to (6.39) with $f = \log \rho$. Computing the derivative of $\log \rho$ by applying the chain rule finally yields the condition (6.41). □

To prove Theorem 7, we distinguish three cases, $\beta \leq 1$, $\beta \geq 2$, and $1 < \beta < 2$. In the first case where $\beta \leq 1$, it suffices to consider densities that only depend on one variable, which fulfill the MTP_2 constraint automatically, leading to the rate $N^{-2\beta/(2\beta+1)}$ for the estimation of a one-dimensional Hölder function. On the other hand, in the case $\beta \geq 2$, we appeal to two-dimensional constructions in density estimation. The MTP_2 condition is a second-order constraint and hence can be satisfied by a carefully chosen set of Hölder functions for $\beta \geq 2$, leading to the rate $N^{-2\beta/(2\beta+2)}$. Finally, in the remaining regime $1 \leq \beta \leq 2$, we in fact use the construction for $\beta = 2$, yielding the rate $N^{-2/3}$.

6.4.2. Case $\beta \leq 1$

For $\beta \leq 1$, we apply an argument based on Fano's inequality, [46, Theorem 2.5]. The following construction is standard in proving lower bounds for nonparametric estimation; see [46, Section 2.6], for example.

Fix a nonzero function $g \in C^\infty(\mathbb{R})$ supported in $[0,1]$ such that $\int_{\mathbb{R}} g(x)\,\mathrm{d}x = 0$ and g is $1/2$-Lipschitz. Let $k \in \mathbb{N}$ and assume without loss of generality that it is divisible by 4. Denote by x_j, $j = 1, \ldots, k$, the left endpoints of an equidistant subdivision of the interval $[0,1]$, that is,

$$x_j = \frac{j-1}{k}, \quad j = 1, \ldots, k,$$

leading to the subdivision

$$S_j = (x_j, 0) + [0, \frac{1}{k}] \times [0,1], \quad j = 1, \ldots, k,$$

of the square $[0,1]^2$. With this, define the functions $g_j \in C^\infty([0,1]^2)$ as

$$g_j(x,y) = \frac{1}{k^\beta} g(k(x-x_j)), \quad (x,y) \in [0,1]^2,\ j = 1,\dots,k.$$

Note that each g_j is supported in S_j. Moreover, we have $g_j \in \mathcal{D}(\beta, 1/2)$ for $\beta \le 1$: for $(x,y),(x',y') \in S_j$, by the 1/2-Lipschitzness of g,

$$\begin{aligned}|g_j(x,y) - g_j(x',y')| &= \frac{1}{k^\beta}|g(k(x-x_j)) - g(k(x'-x_j))| \\ &\le \frac{1}{2k^\beta}|k(x-x')| \le \frac{1}{2k^\beta}|k(x-x')|^\beta = \frac{1}{2}|x-x'|^\beta.\end{aligned}$$

For $(x,y) \in S_j$ and $(x',y') \in S_{j'}$ in the case $j \neq j'$, we obtain the same estimate by applying the above to segments of the line connecting the two points. This also shows that $\|g_j\|_\infty \le 1/2$ as we can take (x,y) at the boundary of S_j so that $g_j(x,y) = 0$.

By the Gilbert-Varshamov bound [32, Lemma 4.10], there is a set $\{\tau^{(\ell)}\}_{\ell=1}^M$ with $\tau^{(\ell)} \in \{0,1\}^k$ such that

- $d_{\mathsf{H}}(0, \tau^{(\ell)}) = k/4$,
- $d_{\mathsf{H}}(\tau^{(\ell)}, \tau^{(r)}) \ge k/8$ for all distinct $\ell, r \in [M]$, and
- $\log(M) \ge k/30$.

For each $\ell \in [M]$, set

$$\rho^{(\ell)}(x,y) = 1 + \sum_{j=1}^k \tau_j^{(\ell)} g_j(x,y).$$

As $\|g_j\|_\infty \le 1/2$ for each j, all $\rho^{(\ell)}$ are bounded within $[1/2, 3/2]$. Since g is mean-zero, $\rho^{(\ell)}$ is a density. We also have that $\rho^{(\ell)} \in \mathcal{D}(\beta, 1/2)$ by definition because $g_j \in \mathcal{D}(\beta, 1/2)$. Moreover, checking condition (6.41) in Lemma 13 yields that all densities are MTP_2, since they only depend on x.

To check the conditions of [46, Theorem 2.5], we first apply (6.1) to obtain that

$$\mathsf{KL}(\rho^{(\ell)}, \rho^{(r)}) \lesssim h^2(\rho^{(\ell)}, \rho^{(r)}), \quad \ell, k \in [M],$$

since all densities are bounded from above and below. Next, the boundedness of the densities and the mean value theorem together imply that

$$h^2(\rho^{(\ell)}, \rho^{(r)}) \asymp \int_{[0,1]^2} (\rho^{(\ell)}(z) - \rho^{(r)}(z))^2 \,\mathrm{d}z = \|\rho^{(\ell)} - \rho^{(r)}\|_{L^2([0,1]^2)},$$

which can be estimated as

$$\begin{aligned}\|\rho^{(\ell)} - \rho^{(r)}\|_{L^2([0,1]^2)} &= \sum_{j=1}^k \int_{S_j} \left(\rho^{(\ell)}(z) - \rho^{(r)}(z)\right)^2 \,\mathrm{d}z \\ &= \sum_{j:\tau_j^{(\ell)} \neq \tau_j^{(r)}} \int_{S_j} (g_j(z))^2 \,\mathrm{d}z\end{aligned}$$

$$= \frac{1}{k^{2\beta+1}} \sum_{j:\tau_j^{(\ell)} \neq \tau_j^{(r)}} \int_0^1 (g(x))^2 \, \mathrm{d}x$$
$$\asymp d_{\mathsf{H}}(\tau^{(\ell)}, \tau^{(r)}) k^{-2\beta-1} \asymp k^{-2\beta},$$

where the hidden constants only depend on the choice of g, which can be made absolute.

That means on the one hand that

$$\mathsf{KL}((\rho^{(\ell)})^{\otimes N}, (\rho^{(r)})^{\otimes N}) \lesssim N k^{-2\beta},$$

and on the other hand that

$$\mathsf{h}^2(\rho^{(\ell)}, \rho^{(r)}) \gtrsim k^{-2\beta}.$$

Thus, if we pick

$$k = C \lceil N^{1/(2\beta+1)} \rceil$$

for a sufficiently large constant $C > 0$, we can ensure that

$$\mathsf{KL}((\rho^{(\ell)})^{\otimes N}, (\rho^{(r)})^{\otimes N}) \leq 0.1 \log(M),$$

and we conclude by [46, Theorem 2.5] that

$$\inf_{\tilde{\rho}} \sup_{\rho^* \in \mathcal{D}(\beta, R)} \mathbb{P}_{(P^*)^{\otimes N}} \left(\mathsf{h}^2(\tilde{\rho}, \rho^*) \geq c_2 N^{\frac{-2\beta}{2\beta+1}} \right) \geq \frac{1}{3},$$

for a constant $c_2 > 0$.

6.4.3. Case $\beta \geq 2$

For $\beta \geq 2$, we need to construct hypotheses that depend on both variables x and y. Let $k \in \mathbb{N}$ to be determined later, and without loss of generality, assume that k is divisible by 4. Fix a non-zero, non-negative function $f \in C^\infty(\mathbb{R})$ with support in $[0, 1]$, and set

$$g(x, y) = f(x) f(y).$$

Moreover, for $i, j \in [k]$, define $w_{i,j}$ as the corners of an equidistant partition of $[0, 1]^2$ denoted by $S_{i,j}$, that is,

$$w_{i,j} = \left(\frac{i-1}{k}, \frac{j-1}{k} \right)^\top \quad \text{and} \quad S_{i,j} = w_{i,j} + \left[0, \frac{1}{k}\right]^2, \quad i, j \in [k].$$

In addition, we let

$$g_{i,j}(z) = \frac{1}{k^\beta} g(k(z - w_{i,j})), \quad z \in [0, 1]^2,$$

which is supported in $S_{i,j}$. Note that for any $r \in \mathbb{N}$ and $\alpha \in \{1,2\}^r$, we have

$$\partial^\alpha g_{i,j}(z) = \frac{k^r}{k^\beta} \partial^\alpha g(k(z - w_{i,j})).$$

Since $g \in C^\infty(\mathbb{R}^2)$, it is easily verified that by definition, $g_j \in \mathcal{D}(\beta, C_1)$ for some constant $C_1 > 0$ that only depends on g.

By the Gilbert-Varshamov bound [32, Lemma 4.10], there is a set $\{\tau^{(\ell)}\}_{\ell=1}^M$ such that $\tau^{(\ell)} \in \{0,1\}^{k\times k}$ and

- $d_{\mathsf{H}}(0, \tau^{(\ell)}) = k^2/4$,
- $d_{\mathsf{H}}(\tau^{(\ell)}, \tau^{(r)}) \geq k^2/8$ for all distinct $\ell, r \in [M]$, and
- $\log(M) \geq k^2/30$.

Next, we associate a density to each $\tau^{(\ell)}$ such that we can control the pairwise distances between these densities. Similar to the proof of Theorem 4, we claim that there exists a useful choice of scaling constants that ensures that the normalization factor of the log-densities stays the same among all ℓ.

For a fixed $\delta \in [0,1]$, we claim that there exists a constant C_3 such that for every $(i,j) \in [k]^2$, there exists $\tilde{\delta}_{i,j} \in [\delta, C_3\delta]$ with

$$\begin{aligned} &\underbrace{\int_{S_{i,j}} \left[\exp\left(z_1 z_2 + \tilde{\delta}_{i,j} g_{i,j}(z)\right) - \exp\left(z_1 z_2\right)\right] \mathrm{d}z}_{=:h_{i,j}(\tilde{\delta}_{i,j})} \\ &= \underbrace{\int_{S_{1,1}} \left(\exp\left(1 + \delta g_{1,1}(z)\right) - e\right) \mathrm{d}z}_{=:H}. \end{aligned} \tag{6.42}$$

To see why this is true, denote the left-hand side of (6.42) by $h_{i,j}(\tilde{\delta}_{i,j})$ and the right-hand side by H as above. Observe that $h_{i,j}(\tilde{\delta}_{i,j})$ is a continuous function of $\tilde{\delta}_{i,j}$ as a consequence of the bounded convergence theorem. Hence, the intermediate value theorem allows us to conclude (6.42) if we can show that $h_{i,j}(\delta) \leq H$ and $h_{i,j}(C_3\delta) \geq H$. The first inequality $h_{i,j}(\delta) \leq H$ follows from the fact that

$$\exp\left(z_1 z_2 + \tilde{\delta}_{i,j} g_{i,j}(z)\right) - \exp(z_1 z_2) \leq \exp\left(1 + \tilde{\delta}_{i,j} g_{i,j}(z)\right) - \exp(1)$$

and changing the limits of the integral. The second inequality $h_{i,j}(C_3\delta) \geq H$ follows from the following estimates. For $h_{i,j}(C_3\delta)$, we have by the fundamental theorem of calculus and the fact that $\exp(t) \geq 1$ for $t \geq 0$, that

$$\begin{aligned} h_{i,j}(C_3\delta) &= \int_{S_{i,j}} \int_{z_1 z_2}^{z_1 z_2 + C_3 \delta g_{i,j}(z)} \exp(t)\, \mathrm{d}t\, \mathrm{d}z \\ &\geq \int_{S_{i,j}} \int_0^{C_3 \delta g_{i,j}(z)} \exp(t)\, \mathrm{d}t\, \mathrm{d}z \\ &= \int_{S_{1,1}} \int_0^{C_3 \delta g_{1,1}(z)} \exp(t)\, \mathrm{d}t\, \mathrm{d}z \end{aligned}$$

$$\geq C_3\delta \int_{S_{1,1}} g_{1,1}(z)\,\mathrm{d}z.$$

On the other hand, for H, we similarly have

$$\begin{aligned}
H &= \int_{S_{1,1}} \int_0^{\delta g_{1,1}(z)} \exp(1+t)\,\mathrm{d}t\,\mathrm{d}z \\
&= e \int_{S_{1,1}} \int_0^{\delta g_{1,1}(z)} \exp(t)\,\mathrm{d}t\,\mathrm{d}z \\
&\leq \delta e \sup_{z\in S_{1,1}} \exp(\delta g_{1,1}(z)) \int_{S_{1,1}} g_{1,1}(z)\,\mathrm{d}z \\
&\leq e \sup_{z\in S_{1,1}} \exp(g_{1,1}(z)) \int_{S_{1,1}} g_{1,1}(z)\,\mathrm{d}z.
\end{aligned}$$

By definition, it holds that $\|g_{1,1}\|_\infty \leq \|g\|_\infty$. Therefore, with the above estimates combined, we see that if $C_3 \geq e\cdot\exp(\|g\|_\infty)$, then $h_{i,j}(C_3\delta) \geq H$, and thus (6.42) is proved.

For each $\tau^{(\ell)}$, let us define

$$\tilde{\eta}^{(\ell)}(z) := z_1 z_2 + \sum_{i,j\in[k]} \tau_{i,j}^{(\ell)} \tilde{\delta}_{i,j} g_{i,j}(z),$$

which after normalization leads to the log-densities

$$\eta^{(\ell)}(z) := \tilde{\eta}^{(\ell)}(z) - \log \underbrace{\int_{[0,1]^2} \exp(\tilde{\eta}^{(\ell)}(w))\,\mathrm{d}w}_{=:\mathfrak{N}},$$

and the densities $\rho^{(\ell)}(z) := \exp(\eta^{(\ell)}(z))$.

As in the proof of Theorem 4, $\mathfrak{N}$ does not depend on ℓ, since by the fact that $g_{i,j}$ is supported on $S_{i,j}$, then (6.42) and that $d_{\mathsf{H}}(0,\tau^{(\ell)}) = k^2/4$, we have

$$\begin{aligned}
\mathfrak{N} &= \int_{[0,1]^2} \exp\left(z_1 z_2 + \sum_{i,j\in[k]} \tau_{i,j}^{(\ell)} \tilde{\delta}_{i,j} g_{i,j}(z)\right) \mathrm{d}z \\
&= \sum_{i,j\in[k]} \int_{S_{i,j}} \exp\left(z_1 z_2 + \tau_{i,j}^{(\ell)} \tilde{\delta}_{i,j} g_{i,j}(z)\right) \mathrm{d}z \\
&= \sum_{i,j\in[k]} \int_{S_{i,j}} \exp\left(z_1 z_2\right)\,\mathrm{d}z \\
&\quad + \sum_{(i,j):\tau_{i,j}^{(\ell)}=1} \left[\int_{S_{i,j}} \exp\left(z_1 z_2 + \tilde{\delta}_{i,j} g_{i,j}(z)\right)\,\mathrm{d}z - \int_{S_{i,j}} \exp\left(z_1 z_2\right)\,\mathrm{d}z\right] \\
&= \int_{[0,1]^2} \exp\left(z_1 z_2\right)\,\mathrm{d}z + d_{\mathsf{H}}(0,\tau^{(\ell)}) \int_{S_{1,1}} \left(\exp\left(1 + \delta g_{1,1}(z)\right) - e\right) \mathrm{d}z
\end{aligned}$$

$$= \int_{[0,1]^2} \exp(z_1 z_2)\, \mathrm{d}z + \frac{k^2}{4} \int_{S_{1,1}} (\exp(1 + \delta g_{1,1}(z)) - e)\, \mathrm{d}z$$

$$= \int_{[0,1]^2} \exp(z_1 z_2)\, \mathrm{d}z + \frac{1}{4} \int_{[0,1]^2} \left(\exp\left(1 + \frac{\delta}{k^\beta} g(z)\right) - e \right) \mathrm{d}z. \tag{6.43}$$

Additionally, from the last line (6.43) of the above calculation, we can also conclude that $\mathfrak{N}$ can be bounded from above and below by positive constants independent from k.

Moreover, for all ℓ and $z \in [0,1]^2$, it holds that

$$\begin{aligned} \partial_1 \partial_2 \eta^{(\ell)}(z) &= \partial_1 \partial_2 \tilde{\eta}^{(\ell)}(z) \\ &= 1 + \sum_{i,j \in [k]} \tau_{i,j}^{(\ell)} \tilde{\delta}_{i,j} \partial_1 \partial_2 g_{i,j}(z) \\ &= 1 + \sum_{i,j \in [k]} \tau_{i,j}^{(\ell)} \tilde{\delta}_{i,j} \frac{k^2}{k^\beta} \partial_1 \partial_2 g(k(z - w_{i,j})) \\ &\geq 1 - C_3 \delta \frac{k^2}{k^\beta} \sup_z |\partial_1 \partial_2 g(z)|, \end{aligned}$$

which, in view of $\beta \geq 2$, can be made positive if δ is chosen to be a sufficiently small constant. Lemma 13 then implies that all $\rho^{(\ell)}$ are MTP_2 densities.

In addition, for any $r \in \mathbb{N}$ and $\alpha \in \{1,2\}^r$, we have by definition

$$\begin{aligned} \partial^\alpha \rho^{(\ell)}(z) &= \rho^{(\ell)}(z) \cdot \partial^\alpha \eta^{(\ell)}(z) \\ &= \exp\left(z_1 z_2 + \sum_{i,j \in [k]} \tau_{i,j}^{(\ell)} \tilde{\delta}_{i,j} g_{i,j}(z) - \log \mathfrak{N} \right) \\ &\quad \cdot \left(\partial^\alpha (z_1 z_2) + \sum_{i,j \in [k]} \tau_{i,j}^{(\ell)} \tilde{\delta}_{i,j} \partial^\alpha g_{i,j}(z) \right). \end{aligned}$$

Since $\mathfrak{N}$ is bounded from above and below by positive constants, the first factor (that is, $\rho^{(\ell)}(z)$) is bounded from above and below. Also, we have that $g_{i,j} \in \mathcal{D}(\beta, C_1)$ and $\tilde{\delta}_{i,j} \leq C_3 \delta$, so it is easily seen that if δ is chosen to be a sufficiently small constant, then $\rho^{(\ell)} \in \mathcal{D}(\beta, 1)$ by definition.

Finally, we bound $\mathsf{KL}(\rho^{(\ell)}, \rho^{(r)})$ and $\mathsf{h}(\rho^{(\ell)}, \rho^{(r)})$. We have seen above that $\rho^{(\ell)}$ can be bounded from above and below, that is, $C_4^{-1} \leq \rho^{(\ell)} \leq C_4$ for a constant $C_4 > 0$. Moreover, we can choose C_4 so that $-C_4 \leq \eta^{(\ell)} \leq C_4$. For the Hellinger distance, we write

$$\begin{aligned} \mathsf{h}^2(\rho^{(\ell)}, \rho^{(r)}) &= \int_{[0,1]^2} \left(\sqrt{\rho^{(\ell)}(z)} - \sqrt{\rho^{(r)}(z)} \right)^2 \mathrm{d}z \\ &= \int_{[0,1]^2} \left(\exp(\eta^{(\ell)}(z)/2) - \exp(\eta^{(r)}(z)/2) \right)^2 \mathrm{d}z \\ &= \int_{[0,1]^2} \rho^{(\ell)}(z) \left(1 - \exp(\eta^{(r)}(z)/2 - \eta^{(\ell)}(z)/2) \right)^2 \mathrm{d}z. \end{aligned}$$

By the Taylor expansion, we can obtain a quadratic control of the exponential term of the form

$$x^2 \le (1 - \exp(x))^2 \le C_5 x^2, \quad x \in [-C_4, C_4],$$

where $C_5 > 0$. This allows us to bound

$$\begin{aligned}\frac{1}{4C_4} \int_{[0,1]^2} (\tilde{\eta}^{(\ell)}(z) - \tilde{\eta}^{(r)}(z))^2 \,\mathrm{d}z &\le \mathsf{h}^2(\rho^{(\ell)}, \rho^{(r)}) \\ &\le \frac{C_5 C_4}{4} \int_{[0,1]^2} (\tilde{\eta}^{(\ell)}(z) - \tilde{\eta}^{(r)}(z))^2 \,\mathrm{d}z.\end{aligned}$$

Taking into account (6.1), it remains to bound the L^2 distance between $\tilde{\eta}^{(\ell)}$ and $\tilde{\eta}^{(r)}$. Using again that the support of $g_{i,j}$ is in $S_{i,j}$, we have

$$\begin{aligned}\int_{[0,1]^2} (\tilde{\eta}^{(\ell)}(z) - \tilde{\eta}^{(r)}(z))^2 \,\mathrm{d}z &= \int_{[0,1]^2} \left(\sum_{i,j \in [k]} (\tau_{i,j}^{(\ell)} - \tau_{i,j}^{(r)}) \tilde{\delta}_{i,j} g_{i,j}(z) \right)^2 \mathrm{d}z \\ &= \sum_{i,j \in [k]} (\tau_{i,j}^{(\ell)} - \tau_{i,j}^{(r)})^2 \tilde{\delta}_{i,j}^2 \int_{S_{i,j}} g_{i,j}(z)^2 \,\mathrm{d}z \\ &\le \sum_{i,j \in [k]} (\tau_{i,j}^{(\ell)} - \tau_{i,j}^{(r)})^2 C_3^2 \delta^2 \int_{S_{1,1}} g_{1,1}(z)^2 \,\mathrm{d}z \\ &= d_{\mathsf{H}}(\tau^{(\ell)}, \tau^{(r)}) C_3^2 \delta^2 \frac{1}{k^{2\beta+2}} \int_{[0,1]^2} g(z)^2 \,\mathrm{d}z \\ &\le \frac{C_3^2 \delta^2}{k^{2\beta}} \int_{[0,1]^2} g(z)^2 \,\mathrm{d}z,\end{aligned}$$

where we changed the limits of integration by substitution and used the bound $d_{\mathsf{H}}(\tau^{(\ell)}, \tau^{(r)}) \le k^2$. Similarly, we can derive a lower bound of the same order, using that $\tilde{\delta}_{i,j} \ge \delta$ and $d_{\mathsf{H}}(\tau^{(\ell)}, \tau^{(r)}) \ge k^2/4$. In conclusion, there exists a constant $C_6 > 0$ such that

$$\frac{1}{C_6} \frac{\delta^2}{k^{2\beta}} \le \mathsf{h}^2(\rho^{(\ell)}, \rho^{(r)}) \le C_6 \frac{\delta^2}{k^{2\beta}}.$$

To finish, we note that for the KL condition of [46, Theorem 2.5] to hold, we need

$$\begin{aligned}\mathsf{KL}((\rho^{(\ell)})^{\otimes N}, (\rho^{(r)})^{\otimes N}) &= N \,\mathsf{KL}(\rho^{(\ell)}, \rho^{(r)}) \\ &\le C_7 N \,\mathsf{h}^2(\rho^{(\ell)}, \rho^{(r)}) \le N C_6 C_7 \frac{\delta^2}{k^{2\beta}} \le 0.1 \log(M),\end{aligned}$$

which, in view of the bound $\log(M) \ge k^2/30$, can be fulfilled by choosing δ to be a sufficiently small constant and $k = \lceil N^{\frac{1}{2\beta+2}} \rceil$. This then leads to a separation of the hypotheses of

$$\mathsf{h}(\rho^{(\ell)}, \rho^{(r)}) \ge c_8 N^{\frac{-2\beta}{2\beta+2}},$$

so [46, Theorem 2.5] yields

$$\inf_{\tilde{\rho}} \sup_{\rho^* \in \mathcal{D}(\beta,R)} \mathbb{P}_{(P^*)^{\otimes N}} \left(\mathsf{h}^2(\tilde{\rho}, \rho^*) \geq c_8 N^{\frac{-2\beta}{2\beta+2}} \right) \geq \frac{1}{3}.$$

6.4.4. *Case* $1 < \beta < 2$

For $1 < \beta < 2$, note that $\mathcal{D}(2,R) \subseteq \mathcal{D}(\beta,R)$, so the above construction in the case $\beta = 2$ still remains valid, which we can use to conclude

$$\inf_{\tilde{\rho}} \sup_{\rho^* \in \mathcal{D}(\beta,R)} \mathbb{P}_{(P^*)^{\otimes N}} \left(\mathsf{h}^2(\tilde{\rho}, \rho^*) \geq c_2 N^{-\frac{2}{3}} \right) \geq \frac{1}{3},$$

which finishes the proof.

7. Conclusion and discussion

In this work, we studied minimax estimation of discrete and continuous two-dimensional totally positive distributions. In particular, for estimation of β-Hölder smooth distributions, we established the minimax rates of estimation in the squared Hellinger distance up to polylogarithmic factors, for any $\beta \geq 0.5$. In addition, we proposed and implemented efficient algorithms to compute our estimators. The numerical experiments supported our theoretical findings.

Several questions are left open for future research. First, for $\beta \in (0, 0.5)$, the upper bound for our estimator does not match the minimax lower bound. Moreover, our bounds do not capture the optimal dependency on the pointwise infimum or supremum of the ground-truth density. These are possibly artifacts of our estimation procedure or proofs. Second, we studied a variant of the MLE with an extra box constraint. While this box-constrained MLE has almost the same computational cost and empirical performance as the original MLE, it is theoretically more desirable to establish the same guarantees for the original MLE. Third, it is of significant interest to study estimation of totally positive distributions in general dimensions. However, our current proof techniques do not generalize to higher dimensions straightforwardly, and we leave this to future research.

Appendix A: Nonexistence of MLE under MTP_2 constraint alone

In this section, we show that without further regularity assumptions on the underlying densities, the MLE under the MTP_2 constraint does not exist.

Lemma 14. *Let* ρ^* *be an* MTP_2 *density on* $[0,1]^2$ *with respect to the Lebesgue measure. Let* $X_1, \ldots, X_N$ *be* N *i.i.d. observations from the corresponding probability distribution. Then, the optimization problem*

$$\max \sum_{i=1}^{N} \log \rho(X_i) \quad \textit{s.t. } \rho \textit{ is an } \mathrm{MTP}_2 \textit{ density w.r.t. the Lebesgue measure}$$

is almost surely unbounded. Consequently, the MLE under the MTP_2 *constraint does not exist.*

Proof. Denote by $\mathbb{P}_{\rho^*}$ the probability distribution corresponding to ρ^* and by $\mathbb{P}_{\rho^*}^{\otimes N}$ the probability distribution of N i.i.d. observations from $\mathbb{P}_{\rho^*}$. Let

$$\mathcal{A} = \{(X_i)_1 \neq (X_j)_1 \text{ for all } i \neq j\}.$$

Then,

$$\begin{aligned}\mathbb{P}_{\rho^*}^{\otimes N}(\mathcal{A}) &= 1 - \mathbb{P}_{\rho^*}^{\otimes N}\Big(\bigcup_{i,j \in [N],\, i \neq j} \{(X_i)_1 = (X_j)_1\} \Big) \\ &\geq 1 - \sum_{i,j\in[N],\, i\neq j} \int_{[0,1]^4} \mathbb{1}\{x_1 = x_2\} \rho^*(x_1,y_1)\rho^*(x_2,y_2) dx_1\, dx_2\, dy_1\, dy_2 \qquad \text{(A.1)} \\ &= 1, \qquad \text{(A.2)}\end{aligned}$$

where the second line (A.1) follows from the sub-additivity of the probability measure and the definition of $\mathcal{A}$, and the third line (A.2) follows from the fact that the integrand in (A.1) is only non-zero on a lower dimensional subset of $[0,1]^4$ and hence is zero almost everywhere with respect to the Lebesgue measure.

Similarly, if $\mathcal{B} = \{X_i \notin \{0,1\} \text{ for all } i\}$, then $(P^*)^{\otimes N}(\mathcal{B}) = 1$.

For the rest of the proof, assume that the event $\mathcal{A} \cap \mathcal{B}$ occurred. Because of the definitions of $\mathcal{A}$ and $\mathcal{B}$, and the fact that N is finite, the minimum distance between the first coordinates is positive, as is the minimum distance to any of the interval boundaries, that is,

$$\epsilon_0 = \Big(\min_{i,j\in[N],\, i\neq j} |(X_i)_1 - (X_j)_1| \Big) \wedge \Big(\min_{i\in[N]} (X_i)_1 \Big) \wedge \Big(\min_{i\in[N]} \big(1 - (X_i)_1\big) \Big) > 0.$$

Let $f \in C^\infty(\mathbb{R})$ be a non-negative bump function supported in $[-1,1]$ such that $\int_{\mathbb{R}} f(x)dx = 1$ and $f(0) = f_0 > 0$. For $0 < \epsilon < \epsilon_0/2$, set

$$\rho_\epsilon(x,y) = \frac{1}{N}\sum_{i=1}^N \frac{1}{\epsilon} f\left(\frac{x - (X_i)_1}{\epsilon} \right).$$

Then,

$$\begin{aligned}\int_{[0,1]^2} \rho_\epsilon(x,y)\, dx\, dy &= \int_{[0,1]} \frac{1}{N}\sum_{i=1}^N \frac{1}{\epsilon} f\left(\frac{x - (X_i)_1}{\epsilon} \right) dx \\ &= \frac{1}{N}\sum_{i=1}^N \int_{\mathbb{R}} f(\xi)\, d\xi = 1,\end{aligned}$$

so ρ_ϵ is again a probability distribution on $[0,1]^2$. Moreover, because ρ_ϵ does not depend on y, by Lemma 13, ρ_ϵ is MTP_2.

Finally, for the log-likelihood, we obtain

$$\begin{aligned}\sum_{i=1}^{N} \log(\rho_\epsilon(X_i)) &= \sum_{i=1}^{N} \log\left(\frac{1}{N}\sum_{j=1}^{N}\frac{1}{\epsilon} f\left(\frac{(X_i)_1-(X_j)_1}{\epsilon}\right)\right) \\ &= \sum_{i=1}^{N}\log\left(\frac{f_0}{\epsilon}\right) \qquad \text{(A.3)} \\ &= N\log\left(\frac{f_0}{\epsilon}\right)\to\infty, \quad \text{for } \epsilon\to 0,\end{aligned}$$

where (A.3) follows because by the definition of $\mathcal{A}$ and of ρ_ϵ, the individual bumps centered at the observations X_i do not intersect. Combined, by choosing ϵ arbitrarily small, we can obtain an arbitrarily large log-likelihood. In turn, the MLE does not exist. □

Remark 15. *Even if the MLE is not defined, there could potentially exist a different estimator over the whole class of* MTP_2 *with good estimation properties. However, the estimation problem over the whole* MTP_2 *class bears other signs of ill-posedness: Since*

$$\bigcup_{\beta\in(0,1)} \mathcal{D}(\beta, R) \subseteq \{\rho : \rho \text{ is } \mathrm{MTP}_2\},$$

the lower bound in Theorem 7 suggests that no estimator $\hat\rho$ *can attain a polynomial estimation rate of*

$$\mathsf{h}^2(\hat\rho,\rho^*) \lesssim N^{-\alpha},$$

for any $\alpha > 0$ *over the whole* MTP_2 *class. While this does not explicitly exclude possibly slower rates of convergence such as* $\log(N)^{-1}$*, this still serves to show that the estimation problem without further regularity assumptions is ill-posed in the sense of not admitting polynomially fast rates.*

Appendix B: Existing results

We state and prove some results that are known or follow easily from existing ones.

B.1. Concentration of multinomial random variables

The following is a standard tail bound for a binomial random variable.

Lemma 16. *Suppose that* Y *has the binomial distribution* $\mathsf{Bin}(N, x)$*, where* N *is a positive integer and* $x \in (0, 1)$*. Then for* $y \in [0, 1]$*, we have* $|Y - Nx| \le Ny$ *with probability at least* $1 - 2\exp\big(-N\frac{y^2}{2(x+y)}\big)$*.*

Proof. This follows immediately from Lemma 6 of [31] by taking $r = (x - y) \vee 0$ and $s = (x + y) \wedge 1$. □

Next, we present a lemma that follows from Bernstein's inequality. Recall that for a vector $a \in \mathbb{R}^m$ and an entrywise positive vector $b \in \mathbb{R}^m$, we denote the b-weighted ℓ_2-norm of a by $\|a\|_b = (\sum_{i=1}^m b_i a_i^2)^{1/2}$.

Lemma 17. *Suppose that Y is a random vector in $\mathbb{R}^m$ having the multinomial distribution $\mathsf{Multi}(N, p)$, where N is a positive integer and $p = (p_1, \dots, p_m)^\top$ is a vector in $(0,1)^m$ with $\sum_{i=1}^m p_i = 1$. Then, for any vector $a \in \mathbb{R}^m$,*

$$\mathbb{P}\Big\{\big|\langle Y - Np, a\rangle\big| \geq t\Big\} \leq 2\exp\Big(\frac{-3t^2}{6N\|a\|_p^2 + 4\|a\|_\infty t}\Big).$$

Proof. Let $I_1, \dots, I_N$ be i.i.d. $\mathsf{Multi}(1, p)$ random variables. That is, we have $I_j = i$ with probability p_i for each $i \in [m]$ and $j \in [N]$. Then we have $Y_i = \sum_{j=1}^N \mathbb{1}\{I_j = i\}$, and thus

$$\begin{aligned}\langle Y - Np, a\rangle &= \sum_{i=1}^m (Y_i - Np_i)a_i = \sum_{i=1}^m \Big(\sum_{j=1}^N \mathbb{1}\{I_j = i\} - Np_i\Big)a_i \\ &= \sum_{j=1}^N \sum_{i=1}^m (\mathbb{1}\{I_j = i\} - p_i)a_i = \sum_{j=1}^N \Big(a_{I_j} - \sum_{i=1}^m p_i a_i\Big) = \sum_{j=1}^N \Big(a_{I_j} - \mathbb{E}[a_{I_j}]\Big).\end{aligned}$$

Since this is a sum of i.i.d. zero-mean random variables with absolute values bounded by $2\|a\|_\infty$, Bernstein's inequality (Theorem 2.8.4 of [47]) implies that

$$\mathbb{P}\Big\{\big|\langle Y - Np, a\rangle\big| \geq t\Big\} \leq 2\exp\Big(\frac{-t^2/2}{\sigma^2 + 2\|a\|_\infty t/3}\Big),$$

where $\sigma^2 = N\mathbb{E}(a_{I_j} - \mathbb{E}[a_{I_j}])^2 \leq N\mathbb{E}[a_{I_j}^2] = N\sum_{i=1}^m p_i a_i^2 = N\|a\|_p^2$. $\square$

The following lemma is concerned with projections of a multinomial random vector.

Lemma 18. *Suppose that Y is a random vector in $\mathbb{R}^m$ having the multinomial distribution $\mathsf{Multi}(N, p)$, where N is a positive integer and $p = (p_1, \dots, p_m)^\top$ is a vector in $(0,1)^m$ with $\sum_{i=1}^m p_i = 1$. Given vectors $v_1, \dots, v_\ell \in \mathbb{R}^m$, for any $\delta \in (0,1]$, it holds with probability at least $1 - \delta$ that*

$$\max_{j\in[\ell]} \big|\langle Y - Np, v_j\rangle\big| \lesssim \Big(\max_{j\in[\ell]} \|v_j\|_p\Big)\sqrt{N\log(\ell/\delta)} + \Big(\max_{j\in[\ell]} \|v_j\|_\infty\Big)\log(\ell/\delta).$$

Proof. The result follows from Lemma 17 and a union bound, with the choice of t equal to a constant times the right-hand side of the above inequality. $\square$

B.2. MLE for $\mathbf{MTP_2}$ *distributions on a grid*

Given the observation Y defined by (2.2), it is well known that the MLE (2.4) can be equivalently defined using the following convex program, which can be solved efficiently:

$$\hat{\theta}^{\mathsf{MLE}} := \operatorname*{argmax}_{D\theta\tilde{D}\geq 0} \frac{1}{N}\langle Y, \theta\rangle - \sum_{i,j} e^{\theta_{i,j}}. \tag{B.1}$$

Lemma 19. *The two definitions* (2.4) *and* (B.1) *of the MLE* $\hat\theta = \hat\theta^{\mathsf{MLE}}$ *are equivalent.*

Proof. It suffices to verify that $\hat\theta$ given by program (B.1) always satisfies that $\sum_{i,j} e^{\hat\theta_{i,j}} = 1$. Suppose this is not the case, and define $\tilde\theta \in \mathbb{R}^{n_1\times n_2}$ by $\tilde\theta_{i,j} = \hat\theta_{i,j} - \log\sum_{k,\ell} e^{\hat\theta_{k,\ell}}$ so that $\sum_{i,j} e^{\tilde\theta_{i,j}} = 1$. Then we have

$$\begin{aligned}\frac{1}{N}\langle Y,\tilde\theta\rangle - \sum_{i,j} e^{\tilde\theta_{i,j}} &= \frac{1}{N}\langle Y,\hat\theta\rangle - \frac{1}{N}\Big(\sum_{i,j} Y_{i,j}\Big)\log\sum_{i,j} e^{\hat\theta_{i,j}} - 1\\ &> \frac{1}{N}\langle Y,\hat\theta\rangle - \sum_{i,j} e^{\hat\theta_{i,j}},\end{aligned}$$

since $\sum_{i,j} Y_{i,j} = N$ and $\log(x)+1 < x$ for any $x \neq 1$. However, this gives a contradiction. □

B.3. Rate of convergence of the empirical frequency matrix

Let us consider the empirical frequency matrix Y/N in the discrete setting, where Y is defined by (2.2). Without leveraging the MTP_2 constraint, it achieves the following trivial rate of estimation.

Lemma 20. *In the setting of Section 2, the empirical frequency matrix* Y/N *satisfies*

$$\mathbb{E}\big[\,\mathsf{h}^2(p^*, Y/N)\big] \le \frac{n_1 n_2}{N}.$$

Proof. For any $i \in [n_1]$ and $j \in [n_2]$, we have $Y_{i,j} \sim \mathsf{Bin}(N, p^*_{i,j})$ marginally. Thus we have

$$\begin{aligned}\mathbb{E}\big[\,\mathsf{h}^2(p^*, Y/N)\big] &= \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\mathbb{E}\Big(\sqrt{p^*_{i,j}} - \sqrt{Y_{i,j}/N}\Big)^2\\ &= \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\mathbb{E}\Big(\frac{p^*_{i,j} - Y_{i,j}/N}{\sqrt{p^*_{i,j}} + \sqrt{Y_{i,j}/N}}\Big)^2\\ &\le \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\mathbb{E}\frac{(p^*_{i,j} - Y_{i,j}/N)^2}{p^*_{i,j}}\\ &= \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\frac{p^*_{i,j}(1-p^*_{i,j})}{p^*_{i,j}N} \le \frac{n_1 n_2}{N}.\end{aligned}$$

□

Appendix C: Further details on numerical experiments

C.1. Implementation details

All simulations are run with Julia 1.4.1 [5], where, besides the standard library, we use the libraries Cubature (version 1.5.1), Distributions [30, 4] (version 0.23.2), StatsBase (version 0.33.0), PyPlot (version 2.9.0), and GLM [3] (version 1.3.9).

Algorithm 2 is stopped at a relative distance in the Frobenius norm between two consecutive iterates of less than 10^{-6} or 400,000 iterations, whichever comes first. Similarly, Algorithm 3 is stopped at a relative distance of 10^{-5} or 100 iterations. The distribution p^* in is sampled as a multinomial distribution via the Distributions package, while the distribution corresponding to the density ρ^* in is sampled via rejection sampling from the corresponding Gaussian distribution. For the calculation of Hellinger distances in the continuous case, we use numerical integration with the Cubature package.

For the calculation of the oracle estimator in Figure 4a, we computed the corresponding estimators for $n \in \{4, 7, 10, 15, 23, 36, 55, 84, 130, 201\}$ and picked the n achieving the best squared Hellinger distance to the ground truth in each case.

C.2. Numerical instability of the MLE for small values of N

As observed in Section 5, for small values of N, Algorithm 3 can become unstable and values in the iterate θ can underflow due to a large number of zeros in the empirical frequency matrix. To illustrate this, we perform the same experiment as in Figure 2a with $N = 100$, which leads to a large error for the unconstrained MLE, which we plot in Figure 6.

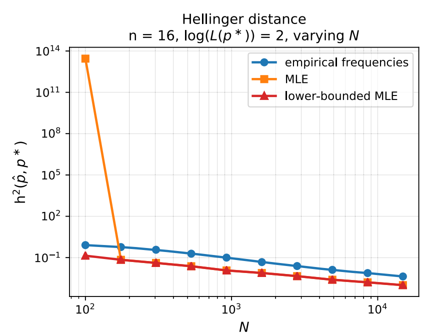


FIG 6. *Instability for small sample sizes in the Hellinger distance for varying N and* $\log(L(p^*)) = 2$

However, this behavior can be remedied by introducing an additional constraint of

$$\theta \in \tilde{\mathcal{C}} := \{\theta : \exp(\theta_{i,j}) \geq \epsilon, \quad i, j \in [n_1] \times [n_2]\}. \tag{C.1}$$

in the calculation of the MLE, where ϵ is small. For example, in this experiment, we specify $\epsilon = e^{-30}$. This leads to the estimator

$$\tilde{\theta}^{\mathrm{lb}} := \underset{\substack{D\theta\tilde{D}^\top \geq 0 \\ \theta \in \tilde{\mathcal{C}}}}{\operatorname{argmax}} \frac{1}{N}\langle Y, \theta\rangle - \sum_{i\in[n_1],\, j\in[n_2]} e^{\theta_{i,j}},$$

$$\hat{\theta}^{\mathrm{lb}}_{i,j} := \tilde{\theta}^{\mathrm{lb}}_{i,j} - \log \sum_{r\in[n_1],\, s\in[n_2]} e^{\tilde{\theta}^{\mathrm{lb}}_{r,s}} \quad \text{for } i \in [n_1],\, j \in [n_2].$$

The constraint (C.1) can be incorporated into Algorithm 2 in the same way as the constraint $\theta \in \mathcal{C}(Y)$ by iterative projection of each component $\theta_{i,j}$ onto the corresponding interval $[\log(\epsilon), \infty)$. As can be seen in Figure 6, this modification ("lower-bounded MLE") is sufficient to overcome the problem of numerical instability when facing a small sample size.

Acknowledgments

PR was supported in part by NSF awards IIS-1838071, DMS-1712596 and DMS-TRIPODS-1740751; ONR grant N00014-17-1-2147 and grant 2018-182642 from the Chan Zuckerberg Initiative DAF. ER was supported in part by an NSF MSPRF DMS-1703821. We thank the anonymous reviewers for their constructive comments.

References

[1] R. Agrawal, U. Roy, and C. Uhler. Covariance matrix estimation under total positivity for portfolio selection, *preprint* arXiv:1909.04222, 2019.
[2] F. Bartolucci and A. Forcina. A likelihood ratio test for *mtp_2* within binary variables. *The Annals of Statistics*, 28(4):1206–1218, 2000. MR1811325
[3] D. Bates, S. Kornblith, A. Noack, M. Bouchet-Valat, M. K. Borregaard, J. M. White, A. Arslan, D. Kleinschmidt, G. Lynch, S. Lendle, P. K. Mogensen, I. Dunning, J. B. S. Calderón, D. Aluthge, pdeffebach, P. Bastide, J. Quinn, C. DuBois, B. Setzler, B. Born, R. Herikstad, M. Grechkin, L. Hein, J. TagBot, J. Adenbaum, H. Q. Ngo, D. Lin, and C. Caine. Juliastats/glm.jl: v1.3.9, Apr. 2020.
[4] M. Besançon, D. Anthoff, A. Arslan, S. Byrne, D. Lin, T. Papamarkou, and J. Pearson. Distributions.jl: Definition and modeling of probability distributions in the juliastats ecosystem. *arXiv e-prints*, arXiv:1907.08611, Jul 2019. MR4116860
[5] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–98, 2017. MR3605826
[6] L. Birgé and P. Massart. Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, 97(1-2):113–150, 1993. MR1240719
[7] Y. Chen. A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology*, 1(1):161–187, 2017.
[8] A. Colangelo, M. Scarsini, and M. Shaked. Some notions of multivariate positive dependence. *Insurance: Mathematics and Economics*, 37(1):13–26, 2005. MR2156593

[9] M. Cule, R. Samworth, and M. Stewart. Maximum likelihood estimation of a multi-dimensional log-concave density. *Journal of the Royal Statistical Society, Series B*, 72:545–607, 2010. MR2758237

[10] A. Domahidi, E. Chu, and S. Boyd. ECOS: An SOCP solver for embedded systems. In *European Control Conference (ECC)*, pages 3071–3076, 2013.

[11] L. Dümbgen and K. Rufibach. Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency. *Bernoulli*, 15:40–68, 2009. MR2546798

[12] B. Fang, A. Guntuboyina, and B. Sen. Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and Hardy-Krause variation. *arXiv preprint* arXiv:1903.01395, 2019. MR2625527

[13] J. Felsenstein. Maximum-likelihood estimation of evolutionary trees from continuous characters. *American Journal of Human Genetics*, 25(471-492), 1973.

[14] C. Fortuin, P. Kasteleyn, and J. Ginibre. Correlation inequalities on some partially ordered sets. *Communications in Mathematical Physics*, 22(2):89–103, 1971. MR0309498

[15] U. Grenander. On the theory of mortality measurement II. *Skandinavisk Aktuarietidskrift*, 39:125–153, 1956. MR0093415

[16] P. Groeneboom and G. Jongbloed. *Nonparametric Estimation under Shape Constraints*. Cambridge University Press, Cambridge, 2014. MR3445293

[17] P. Groeneboom, G. Jongbloed, and J. A. Wellner. Estimation of a convex function: Characterizations and asymptotic theory. *Annals of Statistics*, 29:1653–1698, 2001. MR1891742

[18] A. J. Hoffman. On simple linear programming problems. In *Proceedings of Symposia in Pure Mathematics*, volume 7, pages 317–327, 1963. MR0157778

[19] J. Hütter, C. Mao, P. Rigollet, and E. Robeva. Estimation of Monge matrices. *arXiv preprint* arXiv:1904.03136, 2019.

[20] A. J. Izenman. Review papers: Recent developments in nonparametric density estimation. *Journal of the American Statistical Association*, 86(413):205–224, 1991. MR1137112

[21] A. Juditsky, P. Rigollet, and A. B. Tsybakov. Learning by mirror averaging. *The Annals of Statistics*, 36(5):2183–2206, 2008. MR2458184

[22] S. Karlin. *Total Positivity*, volume 1. Stanford University Press, 1968. MR0230102

[23] S. Karlin and J. McGregor. Coincidence probabilities. *Pacific Journal of Mathematics*, 9(4):1141–1164, 1959. MR0114248

[24] S. Karlin and Y. Rinott. Classes of orderings of measures and related correlation inequalities. I. Multivariate totally positive distributions. *Journal of Multivariate Analysis*, 10(4):467–498, 1980. MR0599685

[25] S. Karlin and Y. Rinott. Total positivity properties of absolute value multinormal variables with applications to confidence interval estimates and related probabilistic inequalities. *Annals of Statistics*, 9:1035–1049, 1981. MR0628759

[26] S. Karlin and Y. Rinott. M-matrices as covariance matrices of multinormal distributions. *Linear Algebra and its Applications*, 52:419–438, 1983.

MR0709364

[27] S. Lauritzen, C. Uhler, and P. Zwiernik. Maximum likelihood estimation in gaussian models under total positivity. arXiv:1702.04031, 2017. MR3953437

[28] J. Lebowitz. Bounds on the correlations and analyticity properties of ferromagnetic ising spin system. *Commun. Math. Phys*, 28(313-321), 1972. MR0323271

[29] J. D. Lee, Y. Sun, and M. A. Saunders. Proximal Newton-type methods for minimizing composite functions. *SIAM Journal on Optimization*, 24(3):1420–1443, 2014. MR3252813

[30] D. Lin, J. M. White, S. Byrne, D. Bates, A. Noack, J. Pearson, A. Arslan, K. Squire, D. Anthoff, T. Papamarkou, M. Besançon, J. Drugowitsch, M. Schauer, and other contributors. JuliaStats/Distributions.jl: a Julia package for probability distributions and associated functions, 2019.

[31] C. Mao, J. Weed, and P. Rigollet. Minimax rates and efficient algorithms for noisy sorting. In *Proceedings of the 28th International Conference on Algorithmic Learning Theory*, 2018. MR3857331

[32] P. Massart. *Concentration inequalities and model selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII - 2003*. Number no. 1896 in Ecole d'Eté de Probabilités de Saint-Flour. Springer-Verlag, 2007.

[33] A. Nemirovski. Topics in non-parametric statistics. *Ecole d'Eté de Probabilités de Saint-Flour*, 28:85, 2000. MR1775640

[34] B. O'Donoghue, E. Chu, N. Parikh, and S. Boyd. Conic optimization via operator splitting and homogeneous self-dual embedding. *Journal of Optimization Theory and Applications*, 169(3):1042–1068, June 2016. MR3501397

[35] B. O'Donoghue, E. Chu, N. Parikh, and S. Boyd. SCS: Splitting conic solver, version 2.0.2. `https://github.com/cvxgrp/scs`, Nov. 2017.

[36] W. Polonik. The silhouette, concentration functions and ML-density estimation under order restrictions. *Annals of Statistics*, 26:1857–1877, 1998. MR1673281

[37] M. Queyranne, F. Spieksma, and F. Tardella. A general class of greedily solvable linear programs. *Mathematics of Operations Research*, 23(4):892–908, 1998. MR1662430

[38] E. Robeva, B. Sturmfels, N. Tran, and C. Uhler. Maximum likelihood estimation for totally positive log-concave densities. *arXiv preprint* arXiv:1806.10120, 2018.

[39] E. Robeva, B. Sturmfels, and C. Uhler. Geometry of log-concave density estimation. arXiv:1704.01910, 2017. MR3925548

[40] R. J. Samworth. Recent progress in log-concave density estimation. *Statist. Sci.*, 33(4):493–509, 11 2018. MR3881205

[41] D. W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, 2015. MR3329609

[42] A. Seregin and J. A. Wellner. Nonparametric estimation of convex-transformed densities. *Annals of Statistics*, 38:3751–3781, 2010. MR2766867

[43] B. W. Silverman. *Density Estimation for Statistics and Data Analysis.*

Routledge, 2018. MR0848134

[44] M. Slawski and M. Hein. Estimation of positive definite M-matrices and structure learning for attractive Gaussian Markov random field. *Linear Algebra and its Applications*, 473:145–179, 2015. MR3338330

[45] G. Strang. *Computational Science and Engineering*, volume 791. Wellesley-Cambridge Press, Wellesley, 2007. MR2742791

[46] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009. MR2724359

[47] R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press, 2018. MR3837109

[48] G. Walther. Detecting the presence of mixing with multiscale maximum likelihood. *Journal of the American Statistical Association*, 97:508–513, 2002. MR1941467

[49] L. Wasserman. Topological data analysis. *Annual Review of Statistics and Its Application*, 5:501–532, 2016. MR3774757