Journal of Machine Learning Research 25 (2024) 1-57 Submitted 1/22; Published 5/24

Permuted and Unlinked Monotone Regression in $\mathbb{R}^d$: an approach based on mixture modeling and optimal transport

Martin Slawski MSLAWSK3@GMU.EDU
Department of Statistics
George Mason University
Fairfax, VA 22030-4444, USA

Bodhisattva Sen BODHI@STAT.COLUMBIA.EDU
Department of Statistics
Columbia University
New York, NY 10027, USA

Editor: Marco Cuturi

Abstract

Suppose that we have a regression problem with response variable $Y \in \mathbb{R}^d$ and predictor $X \in \mathbb{R}^d$, for $d \geq 1$. In *permuted or unlinked regression* we have access to *separate unordered* data on X and Y, as opposed to data on (X, Y)-pairs in usual regression. So far in the literature the case $d = 1$ has received attention, see e.g., the recent papers by Rigollet and Weed [*Information & Inference*, **8**, 619–717] and Balabdaoui *et al.* [*J. Mach. Learn. Res.*, **22**(172), 1–60]. In this paper, we consider the general multivariate setting with $d \geq 1$. We show that the notion of *cyclical monotonicity* of the regression function is sufficient for identification and estimation in the permuted/unlinked regression model. We study permutation recovery in the permuted regression setting and develop a computationally efficient and easy-to-use algorithm for denoising based on the Kiefer-Wolfowitz [*Ann. Math. Statist.*, **27**, 887–906] nonparametric maximum likelihood estimator and techniques from the theory of optimal transport. We provide explicit upper bounds on the associated mean squared denoising error for Gaussian noise. As in previous work on the case $d = 1$, the permuted/unlinked setting involves slow (logarithmic) rates of convergence rooted in the underlying deconvolution problem. We also provide an extension to a certain class of elliptic noise distributions that includes a multivariate generalization of the Laplace distribution, for which polynomial rates can be obtained. Numerical studies complement our theoretical analysis and show that the proposed approach performs at least on par with the methods in the aforementioned prior work in the case $d = 1$ while achieving substantial reductions in terms of computational complexity.

Keywords: cyclical monotonicity; deconvolution; estimation of transport maps; isotonic regression; mixture estimation; multivariate Laplace distribution; nonparametric maximum likelihood; permutation recovery.

Acknowledgments

We would like to acknowledge support for this project from the National Science Foundation (NSF grants CCF-1849876 and DMS-2015376).

©2024 Martin Slawski and Bodhisattva Sen.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v25/22-0058.html.

1. Introduction

In their seminal paper DeGroot et al. (1971) considered the following problem: given photographs of n film stars and another set of photographs of the same film stars taken at a younger age, can we identify corresponding pairs of photographs (i.e., belonging to the same film star) based on, e.g., d biometric measurements extracted from each photograph? A specific variant of this problem (illustrated in Figure 1) is studied in the present paper. Let $\mathcal{X}_n = \{X_i\}_{i=1}^n$ and $\mathcal{Y}_n = \{Y_i\}_{i=1}^n$ be given $\mathbb{R}^d$-valued ($d \geq 1$) samples of data (e.g., $\mathcal{X}_n$ denoting past photographs and $\mathcal{Y}_n$ recent photographs) pertaining to a common set of n entities, and suppose that there is a function $f^* : \mathbb{R}^d \to \mathbb{R}^d$ transforming data in $\mathcal{X}_n$ to their matching counterparts in $\mathcal{Y}_n$, modulo additive noise, i.e., for some *unknown* permutation π^* of $\{1, \ldots, n\}$, we have that

$$Y_i = f^*(X_{\pi^*(i)}) + \epsilon_i, \quad 1 \leq i \leq n, \tag{1}$$

where the $\{\epsilon_i\}_{i=1}^n$ represent i.i.d. zero-mean additive noise. Note that if π^* was known, the problem boils down to a standard regression / (non-parametric) function estimation setup. On the other hand, if f^* was known, the problem boils down to a standard matching problem (Burkard et al., 2009; Collier and Dalalyan, 2016). In this paper, both f^* and π^* are assumed to be unknown, and the following tasks are considered:

(**T1**): (Exact) *Permutation recovery*, i.e., inferring the permutation π^* without error,

(**T2**): *Denoising*, i.e., the construction of estimators $\{\widehat{f}(X_i)\}_{i=1}^n$ for $\{\theta_i^* \equiv f^*(X_i)\}_{i=1}^n$.

Task (**T2**) will also be studied in a slightly more general setup in which samples of different size, say, $\mathcal{X}_n$ and $\mathcal{Y}_m$ are observed such that samples in the latter are i.i.d. copies of $Y \stackrel{\mathcal{D}}{=} f^*(X) + \epsilon$ and samples in $\mathcal{X}_n$ are i.i.d. copies of $X \sim \mu$ for some suitable probability measure μ on $\mathbb{R}^d$, with $\stackrel{\mathcal{D}}{=}$ denoting equality in distribution. Adopting the terminology in Balabdaoui et al. (2021), this generalized setup will be referred to as *unlinked regression*, whereas the basic setup (1) will be referred to as *permuted regression*. In the latter case, $\{X_i\}_{i=1}^n$ will be considered as fixed, unless stated otherwise.

Applications. The problem outlined above arises in a series of applications in various domains. In computer vision, a common task is to identify corresponding pairs of images, with one image arising as a distorted image of the other (Hartley and Zisserman, 2004); in this context, the function f^* may represent a specific combination of distortions (e.g., scaling, rotations, blur, etc.). Specific instances of (1) that have received considerable attention lately are *unlabeled sensing* or *linear regression with unknown permutation*, (e.g., Unnikrishnan et al., 2018; Pananjady et al., 2018, 2017; Abid et al., 2017; Hsu et al., 2017; Slawski and Ben-David, 2019; Tsakiris et al., 2020; Tsakiris and Peng, 2019; Slawski et al., 2020, 2021; Zhang et al., 2022) in which case f^* is an affine transformation (albeit not necessarily from $\mathbb{R}^d$ to $\mathbb{R}^d$). Among these works, Slawski and Ben-David (2019); Slawski et al. (2021) discuss applications in record linkage (Herzog et al., 2007; Christen, 2012; Winkler, 2014), specifically post-linkage data analysis (Scheuren and Winkler, 1993, 1997; Lahiri and Larsen, 2005). Grave et al. (2019); Shi et al. (2021) consider the case in which $\mathcal{X}_n$ and $\mathcal{Y}_n$ are points on the unit sphere in $\mathbb{R}^d$ and f^* is a unitary map with applications in automated translation between different word embeddings. As elaborated in more detail in §3.1 below,

Figure 1: Illustration of the film stars correspondence problem described in DeGroot et al. (1971). In terms of model (1), one can potentially think of $\{Y_1, \dots, Y_n\}$ as the image of $\{X_1, \dots, X_n\}$ under some "morphing" function f^*, modulo unstructured noise.

the setup (1) also arises in matrix estimation problems, in which a noisy row-permuted version of a matrix, whose columns exhibit the same ordering pattern (decreasing or increasing), is observed. Flammarion et al. (2019); Ma et al. (2020, 2021) discuss applications in statistical seriation (Liiv, 2010) and microbiome data analysis. Finally, model (1) bears a relation to linkage attacks in the literature on data privacy (Sweeney, 2001; Narayanan and Shmatikov, 2008): here, $\mathcal{Y}_n$ may represent (anonymized) sensitive data while an adversary holds auxiliary data $\mathcal{X}_n$ along with identifiers (e.g., individuals' names) and tries to leverage the functional relationship between the two data sets to guess the values of the sensitive attributes contained in $\mathcal{Y}_n$ for each or a subset of the identifiers.

Summary of contributions and related work. In a nutshell, the current paper can be seen as an extension of the setup in Carpentier and Schlüter (2016); Rigollet and Weed (2019); Balabdaoui et al. (2021) which consider (variants of) (1) with $d = 1$ and f^* monotone with known direction of monotonicity (say, non-decreasing). A fundamental question associated with (1) asks for what class of functions f^* it is possible to perform tasks (**T1**) and (**T2**) in a statistically consistent manner. In fact, even in the absence of noise and the additional requirement that f^* be smooth, (**T1**) is generally hopeless already for $d = 1$ as can be seen from a simple example (cf. Remark 1 in §2).

In this paper, we establish that (**T1**) and (**T2**) can be accomplished if

$$f^* = \nabla \psi_{f^*}$$

where $\psi_{f^*} : \mathbb{R}^d \to \mathbb{R}$ is a strictly convex function and by $\nabla \psi_{f^*}$ we mean the gradient of the function ψ_{f^*}. Such functions f^* provide a natural generalization of increasing functions for $d = 1$ in view of the property that

$$\langle \nabla \psi_{f^*}(y) - \nabla \psi_{f^*}(x), y - x \rangle > 0 \quad \text{for all } x, y \in \mathbb{R}^d.$$

Note that in particular, functions of the form $f^* = (f_1^*, \dots, f_d^*)$ with f_j^* increasing on $\mathbb{R}$, $1 \le j \le d$, as studied in Flammarion et al. (2019); Ma et al. (2020, 2021) are included, corresponding to component-wise separable additive (strictly) convex functions of the form

$$\psi_{f^*}(x_1, \dots, x_d) = \textstyle\sum_{j=1}^d \psi_{f_j^*}(x_j).$$

Permutation recovery in the presence of noise based on the solution of a linear assignment problem associated with $\mathcal{X}_n$ and $\mathcal{Y}_n$ is shown to succeed (see Proposition 4) if a certain

minimum signal condition similar to conditions in related papers (Flammarion et al., 2019; Ma et al., 2020, 2021; Zhang et al., 2022) is met. As a byproduct, the result on permutation recovery herein yields the novel insight that the unlabeled sensing problem in Zhang et al. (2022) can be solved efficiently whenever the unknown linear transformation is positive (semi)-definite.

Regarding the task (**T2**) of denoising, we leverage a connection to the Brenier theorem in optimal transportation (see e.g., Peyré and Cuturi, 2019; Villani, 2009, 2003; Santambrogio, 2015). According to this connection, the sample $\mathcal{Y}_n$ is thought of as the image of $\mathcal{X}_n$ under an optimal transport map $f^* = \nabla\psi_{f^*}$, contaminated by additive noise. Denoising is achieved via deconvolution of the measure $\frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$ and subsequent computation of an optimal coupling $\widehat{\gamma}$ between the deconvolution estimate and the measure $\frac{1}{n}\sum_{i=1}^n \delta_{X_i}$; finally, we take $\{\widehat{f}(X_i)\}_{i=1}^n$ as the so-called *barycentric projection* of $\widehat{\gamma}$. Deconvolution is based on the Kiefer-Wolfowitz NPMLE for location mixtures (Kiefer and Wolfowitz, 1956; Koenker and Mizera, 2014) and requires knowledge of the noise distribution. The approach developed herein (cf. Algorithm 1) is free of tuning parameters, and directly generalizes to the unlinked regression setting with samples $\mathcal{X}_n$ and $\mathcal{Y}_m$ of different sizes described above at the end of the first paragraph. We provide upper bounds on the mean-square denoising error $\frac{1}{n}\sum_{i=1}^n \|f^*(X_i) - \widehat{f}(X_i)\|_2^2$ in terms of the Hellinger distance of the Kiefer-Wolfowitz NPMLE to the underlying location mixture generating $\mathcal{Y}_n$ and the rate of decay of the characteristic function of the noise distribution, the latter being a common ingredient in deconvolution problems.

Specificially, our results cover (i) Gaussian errors and (ii) errors from a certain class of elliptic distributions that is characterized by light tails and a polynomial rate of decay of the characteristic function; in particular, this class includes the *generalized multivariate Laplace distribution* (Kozubowski et al., 2013) that generalizes the well-known Laplace distribution to higher dimensions. Regarding (i), we obtain logarithmic rates of convergence (see Theorems 6 & 7) in alignment with prior work (Carpentier and Schlüter, 2016; Rigollet and Weed, 2019; Balabdaoui et al., 2021) on the case $d = 1$. For (ii), polynomial rates can be shown (see Theorem 11), in agreement with results in Balabdaoui et al. (2021) on the case $d = 1$ concerning distributions whose characteristic functions exhibit polynomial decay. As an important intermediate step towards obtaining (ii), we derive rates of convergence of the associated Kiefer-Wolfowitz NPMLE in Hellinger distance (cf. Theorem 10). This result is considered to be of independent interest.

The main innovations of the present work over Carpentier and Schlüter (2016); Rigollet and Weed (2019); Balabdaoui et al. (2021) is the generalization to arbitrary dimension d, whereas the previous papers only consider $d = 1$. All three works are based on deconvolution, and a connection to optimal transportation, albeit for $d = 1$, is already made in Rigollet and Weed (2019). However, even for $d = 1$, we argue that the approach developed in this paper is computationally more appealing than those in Carpentier and Schlüter (2016); Rigollet and Weed (2019); Balabdaoui et al. (2021). The method in Carpentier and Schlüter (2016) is based on the truncated characteristic function estimator originating in the deconvolution literature and hence entails a tuning parameter. The method in Rigollet and Weed (2019) is tuning-free and based on convex optimization; however, their deconvolution procedure involves Wasserstein distance minimization and in turn a non-smooth optimiza-

tion problem that is less straightforward to solve than the Kiefer-Wolfowitz NPMLE. The method in Balabdaoui et al. (2021) is based on a non-convex optimization problem.

The theoretical results presented in Carpentier and Schlüter (2016); Rigollet and Weed (2019); Balabdaoui et al. (2021) are of different flavors, and hence not directly comparable. Carpentier and Schlüter (2016) do not provide explicit rates of convergence. The paper by Balabdaoui et al. (2021) is different from Rigollet and Weed (2019) in the sense that the former emphasizes on the unlinked regression setting and provides rates for function estimation in the L_1-distance, whereas Rigollet and Weed (2019) study the mean-squared denoising error in the permuted regression setting (1). For $d = 1$, the denoising performance metric in Rigollet and Weed (2019) (mean squared error at the $\{X_i\}_{i=1}^n$) coincides with what is considered in the present paper. The rate herein is slightly slower than the minimax rate shown in Rigollet and Weed (2019), but given that both rates decrease only logarithmically in n, the gap is not that pronounced. More detailed comparisons are postponed to later sections in this paper.

The paper by Meis and Mammen (2021) studies the setting in Rigollet and Weed (2019) under discrete errors. After submitting a first version of our paper, we became aware of Azadkia and Balabdaoui (2023) that adopts the deconvolution perspective on unlinked regression for linear regression setups. Parametric rates are obtained in both Meis and Mammen (2021) and Azadkia and Balabdaoui (2023).

The approach taken in this paper and the techniques used for its analysis bear various connections to recent developments in the literature on optimal transport, e.g., on the estimation of (smooth) optimal transport maps (Ghosal and Sen, 2022; Hütter and Rigollet, 2021; Deb et al., 2021; Manole et al., 2024; Chizat et al., 2020). Key steps in our proofs are based on adaptations of parts of the analysis in Chizat et al. (2020); Deb et al. (2021); Manole et al. (2024). At a technical level, the main distinction of the present work compared to these earlier works is the convolution setting considered herein.

Paper outline. This paper is organized as follows. Section §2 provides a more detailed discussion of the problem sketched in the introduction, and presents an overview of the technical approach taken. The theoretical properties of that approach are studied in §3 and corroborated with numerical results in §4. A conclusion is provided in §5. Proofs of our results and additional technical details can be found in the Appendix.

Notation. For the convenience of the reader, notation that is used frequently in this paper is summarized in the following table. In addition, we would like to recall model (1) and that $\{\theta_i^*\}_{i=1}^n = \{f^*(X_i)\}_{i=1}^n$, where f^* is the function of interest and the $\{X_i\}_{i=1}^n$ are the design points.

ν_n^*	measure $\frac{1}{n}\sum_{i=1}^n \delta_{\theta_i^*}$	$\widehat{\nu}$	mixing measure associated with $\widehat{\mathsf{f}}_n$
ν_n	measure $\frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$	ψ_{f^*}	convex function associated with f^*
μ_n	measure $\frac{1}{n}\sum_{i=1}^n \delta_{X_i}$	$\psi_{f^*}^\star$	conjugate of ψ_{f^*}
φ	PDF of ϵ/σ, $\mathrm{Cov}(\epsilon) = \sigma^2 I_d$	π^*	ground truth permutation of $\{1,\dots,n\}$
φ_σ	PDF of ϵ	Π^*	permutation matrix corresponding to π^*
$\star$	convolution	$\mathcal{P}(n)$	set of permutation matrices of order n
$\mathsf{f}_n = \varphi_\sigma \star \nu_n^*$	average location mixture PDF	π, Π	generic elements of $\mathcal{P}(n)$
$\widehat{\mathsf{f}}_n$	NPMLE of f_n	$\mathbf{F}$	Fourier transformation

We often refer to a permutation via the underlying map π and the corresponding matrix Π in an interchangeable fashion, and accordingly $\mathcal{P}(n)$ may refer to both maps and matrices.

2. Estimation strategy

In this section we describe our estimation procedure for both tasks — (**T1**) and (**T2**). We start with a simple example that illustrates the non-identifiability of f^* and π^* in (1) without further assumptions on the structure of f^*; see Remark 1 below. It turns out that if f^* is *cyclically monotone* (see §2.1 where we formally define this notion along with other related concepts) then model (1) is identifiable and consistent estimation can be successfully carried out. To solve the denoising problem (**T2**) we leverage ideas from the theory of optimal transport and the Kiefer-Wolfowitz NPMLE for location mixtures which is discussed in detail in §2.2. We also give our main algorithm (see Algorithm 1) and discuss the computational approach in §2.2.

Remark 1 (A negative example) *To gain some insights into the feasibility of tasks (**T1**) and (**T2**) given the observation model* (1)*, let us first consider a simple example which shows that recovery of f^* or π^* is generally hopeless even in seemingly benign settings ($d = 1$, no noise, f^* smooth). Specifically, suppose that $X_i = Y_i = i/M$, $1 \le i \le n = M-1$ for $M \ge 2$. Then both pairs $f_1^*(x) = x$ with $\pi_1^*(i) = i$, $1 \le i \le n$, and $f_2^*(x) = 1 - x$ with $\pi_2^*(i) = n - i$, $1 \le i \le n$, satisfy* (1)*. Clearly, additionally requiring that f^* be increasing rules out this ambiguity. In fact, estimation of monotone f^* with known direction of monotonicity under the permuted regression setup* (1) *has been shown to be feasible even in the presence of noise (Carpentier and Schlüter, 2016; Rigollet and Weed, 2019; Balabdaoui et al., 2021). At the same time, estimation of the direction of monotonicity itself is generally not possible even if f^* is linear (DeGroot and Goel, 1980; Bai and Hsing, 2005).*

2.1 Monotone operators and linear assignment problems

The example above for $d = 1$ (and in the absence of noise) provides some useful clues regarding the generalization to arbitrary dimension $d \ge 1$. If f^* is known to be increasing, the underlying permutation π^* is immediately determined by the requirement that Y_i must match the corresponding order statistic in $\mathcal{X}_n$, i.e., $X_{\pi^*(i)} = X_{\mathrm{rank}(i)}$, where $\mathrm{rank}(i)$ denotes the rank of Y_i among $\mathcal{Y}_n$, $1 \le i \le n$. It can also be shown that π^* minimizes the optimization problem

$$\min_{\pi} \frac{1}{2} \sum_{i=1}^{n} |Y_i - X_{\pi(i)}|_2^2 = -\max_{\pi} \sum_{i=1}^{n} X_{\pi(i)} Y_i + c, \qquad c := \frac{1}{2} \sum_{i=1}^{n} (X_i^2 + Y_i^2) \tag{2}$$

over all permutations π of $\{1, \ldots, n\}$. The above problem is a specifically simple instance of the class of *linear assignment problems* (LAPs) that are of the form

$$\min_{\Pi \in \mathcal{P}(n)} \sum_{i=1}^{n} \sum_{j=1}^{n} \Pi_{ij} C_{ij} = \min_{\Pi \in \mathcal{P}(n)} \mathrm{tr}(C^\top \Pi), \tag{3}$$

where $\mathcal{P}(n) = \{\Pi \in \{0,1\}^{n \times n} : \sum_{i=1}^{n} \Pi_{ij} = 1,\ 1 \le j \le n,\ \sum_{j=1}^{n} \Pi_{ij} = 1,\ 1 \le i \le n\}$ denotes the set of permutation matrices of dimension n and $C = (c_{ij})$ is a cost matrix with

entry (i, j) representing the cost associated with the pairing (i, j), $1 \leq i, j \leq n$. LAPs (3) constitute a well-studied class of optimization problems that are known as bipartite matching problems in the literature on combinatorial optimization (Burkard et al., 2009). In light of the celebrated Birkhoff-von Neumann theorem (Ziegler, 1995), (3) can be solved efficiently via linear programming. Tailored algorithms such as the Hungarian Algorithm (Bertsekas and Castanon, 1992) and the Auction Algorithm (Kuhn, 1955) have runtime complexity $O(n^3)$, and approximate solutions can be obtained via Sinkhorn iterations in time $O(n^2 \log n)$ (Cuturi et al., 2019); especially simple instances such as (2) in which C has rank one reduce to sorting (e.g., Burkard and Cela, 1999, Section 2).

The crucial insight here is that knowing f^* is monotone increasing immediately allows us to recover π^* in the absence of noise via the optimization problem (2). This observation prompts the following generalization. Let $\mathbb{X}^n \subset \mathbb{R}^d \times \ldots \times \mathbb{R}^d$ be a domain containing all possible samples $\mathcal{X}_n$. We require that for all $\mathcal{X}_n \subset \mathbb{X}^n$ and all $n \geq 1$, f^* has the property that

$$\min_{\pi} \frac{1}{2} \sum_{i=1}^{n} \|Y_i - X_{\pi(i)}\|_2^2, \tag{4}$$

is (uniquely) minimized by $\pi = \pi^*$, where $Y_i = f^*(X_{\pi^*(i)})$, $1 \leq i \leq n$. This requirement can be expressed more succinctly via the notion of (strict) *cyclical monotonicity*, a notion that arises in the study of monotone operators in convex analysis (Bauschke and Combettes, 2011) as well as in optimal transportation (e.g., McCann and Guillen, 2011, Definition 2.1), a connection that plays a fundamental role in the developments further below.

Proposition 2 *Suppose $Y_i = f^*(X_{\pi^*(i)})$, $1 \leq i \leq n$, for some function $f^* : \mathbb{R}^d \to \mathbb{R}^d$. Without loss of generality, suppose that π^* equals the identity* id *permutation. The optimization problem* (4) *is uniquely minimized by $\pi = \mathsf{id}$ iff $\Gamma_{f^*} := \{(x, f^*(x)) : x \in \mathbb{X}\} \subset \mathbb{R}^d \times \mathbb{R}^d$ is a (strictly) cyclically monotone set (with respect to the Euclidean norm), i.e., if for all $k \geq 1$ and all $\{(x_i, y_i)\}_{i=1}^k \subset \Gamma_{f^*}$, it holds that*

$$-\sum_{i=1}^{k} \langle x_i, y_i \rangle < \sum_{i=1}^{k} -\langle x_{i+1}, y_i \rangle, \qquad x_{k+1} := x_1. \tag{5}$$

Proposition 2 is obtained by omitting the square terms in the objective as in (2) and decomposing permutations into their disjoint cycles; a formal proof is omitted for the sake of brevity. The following result, due to Rockafellar, precisely characterizes the class of functions $f : \mathbb{R}^d \to \mathbb{R}^d$ whose graphs are cyclically monotone.

Theorem 3 (Rockafellar, 1966) *The graph of the sub-differential $\partial\psi$ of a convex function $\psi : \mathbb{R}^d \to \mathbb{R}^d$, i.e., $\Gamma_{\partial\psi} := \{(x, y) \in \mathbb{R}^d \times \mathbb{R}^d : \psi(z) \geq \psi(x) + \langle z - x, y \rangle \ \forall z \in \mathbb{R}^d\}$ is a cyclically monotone subset of $\mathbb{R}^d \times \mathbb{R}^d$. Moreover, any cyclically monotone subset of $\mathbb{R}^d \times \mathbb{R}^d$ is contained in such a set.*

The subdifferential of a convex function ψ is a *monotone operator* in the sense that the relation $\{(x, \partial\psi(x)) : x \in \mathbb{R}^d\}$ has the property that $\langle x - z, g_x - g_z \rangle \geq 0$ for all $x, z \in \mathbb{R}^d$ and all $g_x \in \partial\psi(x)$, $g_z \in \partial\psi(z)$, which is in analogy to the fact monotone functions on the real line arise as derivatives of convex functions.

In combination, Proposition 2 and Rockafellar's theorem above prompt the requirement

$$f^* = \nabla \psi_{f^*}$$

for a convex function $\psi_{f^*} : \mathbb{R}^d \to \mathbb{R}^d$. Working with gradients instead of subdifferentials is needed in order to ensure that f^* is actually a map from $\mathbb{R}^d$ to $\mathbb{R}^d$ even though the distinction is somewhat minor in light of the fact that convex functions are differentiable (Lebesgue) almost everywhere.

For the purpose of permutation recovery (**T1**) and in turn for *strict* cyclical monotonicity to hold, we need to impose the additional requirement that ψ_{f^*} be *strictly* convex, i.e., the strengthened first-order convexity condition

$$\psi_{f^*}(z) > \psi_{f^*}(x) + \langle \nabla \psi_{f^*}(x), z - x \rangle \quad \forall\, x, z \in \mathbb{R}^d,\ x \neq z. \tag{6}$$

Note that $\nabla \psi_{f^*} : \mathbb{R}^d \to \mathbb{R}^d$ is injective if and only if (6) holds. In the presence of noise, strict convexity will further be strengthened to strong convexity (cf. Proposition 4 in §3 below).

2.2 A path towards denoising (T2) via optimal transportation

Gradients of convex functions are also known as Brenier maps in the field of optimal (measure) transportation (e.g., Villani, 2009, 2003; Santambrogio, 2015). Specifically, for random variables $U \sim \rho$ and $V \sim \tau$ with ρ and τ absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$ such that $\mathbf{E}_{U\sim\rho}[\|U\|_2^2], \mathbf{E}_{V\sim\tau}[\|V\|_2^2]$ are both finite, Brenier's theorem (in short) states that the minimization problem

$$\inf_T \frac{1}{2}\, \mathbf{E}_{U\sim\rho}[\|U - T(U)\|_2^2],$$

over all measurable functions T such that $T(U) \sim \tau$ has a solution $T^* = \nabla \psi_{T^*}$ for a convex function ψ_{T^*} with T^* being uniquely determined almost everywhere. Moreover, the solution of the reverse problem in which τ is optimally transported to ρ in the above sense is the optimal transport map given by $\nabla \psi^\star_{T^*}$ with $\psi^\star_{T^*}$ denoting the Legendre-Fenchel conjugate of ψ_{T^*}; we refer to Appendix H for a more detailed background and references.

Linear assignment problems of the form (2) and (4) can be interpreted as specific discrete optimal transport problems between the atomic measures

$$\mu_n := \tfrac{1}{n} \textstyle\sum_{i=1}^n \delta_{X_i}, \qquad \text{and} \qquad \nu_n := \tfrac{1}{n} \textstyle\sum_{i=1}^n \delta_{Y_i}.$$

The requirement that $\mu_n(T^{-1}(Y_i)) = 1/n$, $1 \leq i \leq n$, immediately implies that the resulting optimal transport problem seeks for an optimal pairing $\{(X_{\pi(i)}, Y_i)\}_{i=1}^n$ over all permutations π of $\{1, \ldots, n\}$.

The connection to optimal transportation turns out to be fruitful since it suggests a natural approach for the task of denoising (**T2**), i.e., the construction of estimators $\{\widehat{f}(X_i)\}_{i=1}^n$ of $\{f^*(X_i)\}_{i=1}^n$ under the permuted regression model (1). Note that solving the linear assignment problem (4) to find an optimal collection of (X, Y)-pairs is not suitable for this task in general since all noise inherent in the $\{Y_i\}_{i=1}^n$ is retained (cf. Figure 2). In fact, we are interested in the pairings $\{(X_i, \theta_i^*)\}_{i=1}^n$ with $\theta_i^* = f(X_i^*)$, $1 \leq i \leq n$, which corresponds

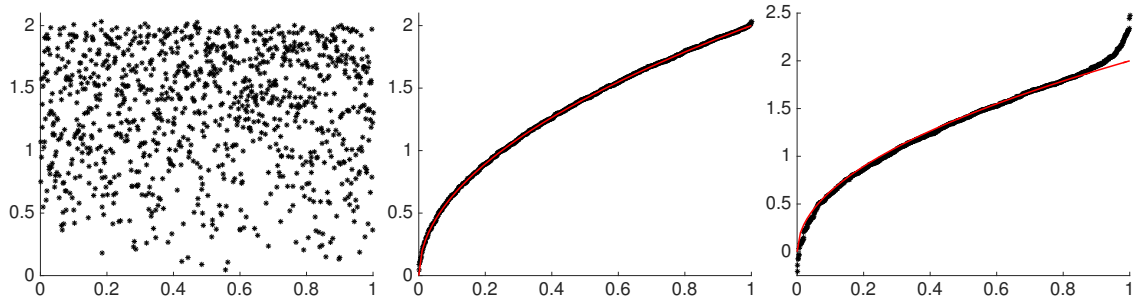


Figure 2: Left: Shuffled Data $\{(X_i, Y_i)\}_{i=1}^n$. Middle: Sorted Data $(X_{(i)}, Y_{(i)})_{i=1}^n$ in case of negligible noise; underlying function $x \mapsto f^*(x) := 2\sqrt{x}$ in red. Right: Sorted Data $(X_{(i)}, Y_{(i)})_{i=1}^n$ in case of substantial noise. The results indicates a serious amount of bias, particularly near the boundaries.

to the optimal transportation problem between μ_n and $\nu_n^* := \frac{1}{n}\sum_{i=1}^n \delta_{\theta_i^*}$. The fact that the latter is not given — in fact, it corresponds to the target to be recovered — suggests the need for its estimation. Below, we shall present an atomic estimator $\widehat{\nu}$ of ν_n^* of the form

$$\widehat{\nu} := \textstyle\sum_{j=1}^p \widehat{\alpha}_j \delta_{\widehat{\theta}_j}$$

with atoms $\{\widehat{\theta}_j\}_{j=1}^p \subset \mathbb{R}^d$ and masses (i.e., positive numbers summing to one) $\{\widehat{\alpha}_j\}_{j=1}^p$. Since $p \neq n$ in general, there does not exist a transport map between μ_n and $\widehat{\nu}$[1]. However, the *Kantorovich problem*, a relaxation of the optimal transportation problem (cf. Appendix H), can be used to obtain a proxy as follows. The Kantorovich problem is given by the optimization problem

$$\min_{\gamma \in \Pi(\mu_n, \widehat{\nu})} \int \int \frac{1}{2}\|x - \theta\|_2^2 \, d\gamma(x, \theta), \tag{7}$$

where the minimum is over all couplings γ of μ_n and $\widehat{\nu}$, i.e., all probability measures on the set $\{X_i\}_{i=1}^n \times \{\widehat{\theta}_j\}_{j=1}^p$ whose marginal distributions are given by μ_n and $\widehat{\nu}$, respectively.

Let $\widehat{\gamma}$ denote a minimizer of (7). We then use the estimator

$$\widehat{f}(X_i) := \mathbf{E}_{(\theta,X)\sim\widehat{\gamma}}[\theta | X = X_i] = \frac{\int_\theta \theta \, d\widehat{\gamma}(\theta, X_i)}{\int_\theta \, d\widehat{\gamma}(\theta, X_i)} = \frac{\int_\theta \theta \, d\widehat{\gamma}(\theta, X_i)}{\mu_n(\{X_i\})}, \quad 1 \leq i \leq n, \tag{8}$$

i.e., the conditional expectation of θ given $X = X_i$, $1 \leq i \leq n$, resulting from the optimal coupling $\widehat{\gamma}$. The map $x \mapsto \mathbf{E}_{(X,\theta)\sim\widehat{\gamma}}[\theta | X = x]$, $x \in \mathcal{X}_n$, is usually referred to as the *barycentric projection* of $\widehat{\gamma}$ in the optimal transport literature (Paty et al., 2020, Defn. 2).

In order to finalize the outline of our approach for task (**T2**), which is summarized in Algorithm 1, it remains to present a specific estimator $\widehat{\nu}$ of ν_n^*. Let φ denote the density of the i.i.d. standardized noise terms $\{\epsilon_i/\sigma\}_{i=1}^n$, where we assume that $\mathbf{E}[\epsilon_1] = 0$ and $\mathrm{Cov}(\epsilon_1) = \sigma^2 I_d$, for $\sigma > 0$. Then the average density of the $\{Y_i\}_{i=1}^n$ is given by the location mixture density $\mathsf{f}_n := \varphi_\sigma \star \nu_n^*$ with $\star$ denoting convolution and $\varphi_\sigma(\cdot) := \sigma^{-d}\varphi(\cdot/\sigma)$, i.e.,

$$\mathsf{f}_n(y) := \int \varphi_\sigma(y - \theta) \, d\nu_n^*(\theta) = \frac{1}{n}\sum_{i=1}^n \varphi_\sigma(y - \theta_i^*), \quad y \in \mathbb{R}^d.$$

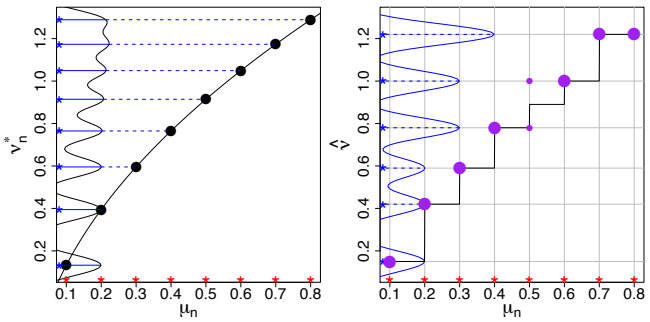
1. The measure preservation property $\mu_n(T^{-1}(\widehat{\theta}_j)) = 1/n$, $1 \leq j \leq p$, cannot hold since in general $n \neq p$.

Algorithm 1 Denoising for Permuted or Unlinked Regression

Inputs: $\mathcal{X}_n$, $\mathcal{Y}_m$, $\varphi_\sigma, \mathbb{G}$.

1. Solve problem (10):
$\rightsquigarrow \widehat{\nu} = \sum_{j=1}^p \widehat{\alpha}_j \delta_{\widehat{\theta}_j}$
2. Compute an optimal coupling between μ_n and $\widehat{\nu}$ via the linear program (12).
$\rightsquigarrow \widehat{\Gamma} \in \mathbb{R}_+^{n \times p}$.

Return $\widehat{f}(X_i) = n \sum_j \widehat{\Gamma}_{ij} \widehat{\theta}_j$, $1 \le i \le n$.



Left figure: $\nu_n^* = \frac{1}{n}\sum_{i=1}^n \delta_{\theta_i^*}$ with $\theta_i^* = 2\sqrt{X}_i - 0.5$, $1 \le i \le n$, with $\mathcal{X}_n = \{0.1, 0.2, \ldots, 0.8\}$. The solid black line drawn over the vertical axis represents the mixture density $\mathsf{f}_n = \nu_n^* \star \varphi_\sigma$. Right figure: Estimated mixture density $\widehat{\mathsf{f}}_n$ and mixing measure $\widehat{\nu}$ (blue). The resulting optimal coupling $\widehat{\Gamma}$ between μ_n and $\widehat{\nu}$ is represented by purple dots (with sizes proportional to the corresponding entry of $\widehat{\Gamma}$. Solid black line: Function estimate $\widehat{f}$ obtained by constant interpolation based on $\{(X_i, \widehat{f}(X_i))\}_{i=1}^n$.

We propose to estimate f_n via the Kiefer-Wolfowitz nonparametric maximum likelihood estimator[2] (NPMLE, cf., Kiefer and Wolfowitz, 1956; Koenker and Mizera, 2014) given by

$$\inf_{\mathsf{f} \in \mathcal{F}_{\varphi,\sigma}} -\sum_{i=1}^n \log \mathsf{f}(Y_i), \quad \mathcal{F}_{\varphi,\sigma} := \left\{ \mathsf{f} = \int \varphi_\sigma(y-\theta) d\nu(\theta) : \ \nu \text{ distribution on } \mathbb{R}^d \right\}. \tag{9}$$

Even though the optimization problem (9) is infinite-dimensional, it can be shown to be convex and that a solution $\widehat{\mathsf{f}}_n$ exists, and that the associated mixing measure $\widehat{\nu}$ is atomic with a finite number of atoms (Koenker and Mizera, 2014; Lindsay, 1983). We shall use $\widehat{\nu}$ as an estimator of ν_n^* that is then plugged into the Kantorovich problem (7). Note that the Kiefer-Wolfowitz problem assumes knowledge of the density φ_σ, i.e., the noise distribution. This assumption is common in deconvolution problems (Meister, 2009) as encountered here; in fact, without any knowledge about the noise distribution, deconvolution problems are generally ill-defined. The assumption of known σ can potentially be relaxed (cf. §5).

Unlinked Regression. The estimator (8) remains applicable in the unlinked regression setting in which $\mathcal{X}_n$ and $\mathcal{Y}_m$ are of different sizes $n \neq m$ as described in the Introduction with the elements of $\mathcal{Y}_m$ being i.i.d. as $Y \stackrel{\mathcal{D}}{=} f^*(X) + \epsilon$ with $X \sim \mu$ for some absolutely continuous probability measure μ supported on a compact subset of $\mathbb{R}^d$ and $f^* = \nabla \psi_{f^*}$ for ψ_{f^*} convex. In fact, $\mathcal{Y}_m$ can be used to obtain an estimator $\widehat{\nu}$ of $\frac{1}{n}\sum_{i=1}^n \delta_{f^*(X_i)}$ as before via (9), and all subsequent steps in Algorithm 1 can be executed. The rates of convergence for the denoising error are almost identical to the permuted regression setting with $n = m$ as long as $n \asymp m$, cf. §3.2.

2. Terminology varies in the literature; Zhang (2009) uses the term "generalized MLE".

Computation. Algorithm 1 requires computation of the Kiefer-Wolfowitz NPMLE, the Kantorovich problem (7), and finally the barycentric projections (8). The Kiefer-Wolfowitz problem can be reformulated as a (non-convex) finite mixture likelihood optimization problem, and then solved via the EM algorithm (Jiang and Zhang, 2009). Instead, in order to preserve convexity, we approximate the solution of (9) via the finite-dimensional optimization problem

$$\inf_{\mathsf{f}\in\mathcal{F}^{\mathbb{G}}_{\varphi,\sigma}} -\sum_{i=1}^{n}\log \mathsf{f}(Y_i), \quad \mathcal{F}^{\mathbb{G}}_{\varphi,\sigma} := \left\{\mathsf{f}=\int \varphi_\sigma(y-\theta)d\nu(\theta):\ \nu \text{ distribution on } \mathbb{G}\right\}, \tag{10}$$

where $\mathbb{G}$ is a finite set of points in $\mathbb{R}^d$. Problem (10) can be rewritten as

$$\inf_{\alpha\in\Delta^{|\mathbb{G}|}} -\sum_{i=1}^{n}\log\left(\sum_{j=1}^{|\mathbb{G}|}\alpha_j\varphi_\sigma(Y_i-\theta_j)\right), \tag{11}$$

where $\Delta^r := \{x\in\mathbb{R}^r_+ : \sum_{j=1}^r x_j = 1\}$ denotes the probability simplex in $\mathbb{R}^r$, $r\geq 1$. There is a variety of convex optimization algorithms that can be used to solve (11). Our experiments are based on a primal-dual interior point method (Boyd and Vandenberghe, 2004) that yields fast and highly accurate results even if $|\mathbb{G}|$ includes several thousand points. Regarding $\mathbb{G}$, our default choice is $\mathbb{G}=\mathcal{Y}_n$ for $d\geq 2$ and $\mathbb{G}$ being a set of g_n linearly spaced points in the interval $[\min_i Y_i, \max_i Y_i]$ with $g_n = n$ or $g_n = n^{1/2}$ for $d=1$. In Dicker and Zhao (2016) it is shown that the latter choice suffices to ensure comparable statistical performance to the solution of the infinite-dimensional problem (9).

Solving (10) yields the estimator $\widehat{\nu}=\sum_{j=1}^p \widehat{\alpha}_j\delta_{\widehat{\theta}_j}$, where $\{\widehat{\alpha}_j\}_{j=1}^p$ represent the non-zero entries of the resulting minimizer of (11) and $\{\widehat{\theta}_j\}_{j=1}^p\subseteq\mathbb{G}$ represent the corresponding atoms. Computing an optimal coupling between the two finitely supported measures μ_n and $\widehat{\nu}$ according to problem (7) amounts to solving the linear program

$$\min_{\Gamma\in\mathbb{R}^{n\times p}_+}\operatorname{tr}(C^\top\Gamma) \quad \text{subject to } \sum_{i=1}^n\Gamma_{ij}=\widehat{\alpha}_j,\ 1\leq j\leq p,\ \ \sum_{j=1}^p\Gamma_{ij}=\frac{1}{n},\ 1\leq i\leq n, \tag{12}$$

where $C=(\|X_i-\widehat{\theta}_j\|_2^2/2)_{1\leq i\leq n,1\leq j\leq p}$, and the row and column sum constraints represent the requirements on the two marginal distributions. Solving (12) exhibits similar computational complexity to the linear assignment problem (3). For the numerical examples presented in this paper, we used the routine `cplexlp` in CPLEX (IBM, 2016). Fast approximate solution can be obtained via Sinkhorn iterations (Cuturi et al., 2019). For $d=1$, problem (12) becomes considerably simpler due to the natural ordering of the real line, and can be solved in time $O(n+p)$ via the so-called "Northwest Corner Rule" (Peyré and Cuturi, 2019, §3.4.2) after sorting the $\{X_i\}_{i=1}^n$ and $\{\widehat{\theta}_j\}_{j=1}^p$.

Finally, given a minimizer $\widehat{\Gamma}$ of (12), the barycentric projections (8) can be computed as

$$\widehat{f}(X_i) := \sum_{j=1}^p\widehat{\Gamma}_{ij}\widehat{\theta}_j\Big/\sum_{j=1}^p\widehat{\Gamma}_{ij} = n\sum_{j=1}^p\widehat{\Gamma}_{ij}\widehat{\theta}_j, \quad 1\leq i\leq n.$$

3. Main results

In this section, we first analyze permutation recovery (**T1**) based on the linear assignment problem in (4) with the distinction that $\{Y_i\}_{i=1}^n$ may be contaminated by Gaussian additive noise, i.e., $Y_i = f(X_{\pi^*(i)}) + \epsilon_i$, $1 \leq i \leq n$, with $\{\epsilon_i\}_{i=1}^n$ being i.i.d. $N(0, \sigma^2 I_d)$-distributed random variables. The Gaussianity assumption is not essential; generalizations to the non-isotropic case or other noise distributions satisfying various tail conditions (sub-Gaussian, sub-Exponential, ...) appear rather straightforward, and are not pursued in this paper to simplify the exposition and to facilitate the comparison to related results in previous literature, specifically Ma et al. (2020); Zhang et al. (2022); Flammarion et al. (2019).

The main technical contribution of this paper is the analysis of Algorithm 1 for the purpose of denoising (**T2**), which is presented subsequently.

3.1 Permutation recovery

Consider the following linear assignment problem under the permuted regression setup (1):

$$\min_{\pi} \frac{1}{2} \sum_{i=1}^n \|Y_i - X_{\pi(i)}\|_2^2, \tag{13}$$

where the minimization is over all permutations π of $\{1, \ldots, n\}$. Let $\widehat{\pi}$ denote the minimizer of (13). Assuming i.i.d. Gaussian errors, the following result (Proposition 4) states sufficient conditions for exact permutation recovery, i.e., the event $\{\widehat{\pi} = \pi^*\}$, to occur with high probability. Comparison to existing results will indicate that the required conditions cannot substantially be relaxed.

The discussion below Theorem 3 in §2 has indicated the necessity of the requirement that ψ_{f^*} be strictly convex already in the absence of noise. A further strengthening to strong convexity, i.e.,

$$\psi_{f^*}(z) \geq \psi_{f^*}(x) + \langle \nabla \psi_{f^*}(x), z - x \rangle + \frac{\lambda}{2} \|x - z\|_2^2 \quad \forall x, z \in \mathbb{R}^d \tag{14}$$

becomes necessary to counteract noise[3]. Equipped with strong convexity, we are in position to state the following result (proved in Appendix A).

Proposition 4 *Suppose that $Y_i = f^*(X_{\pi^*(i)}) + \epsilon_i$, $1 \leq i \leq n$ with $f^* = \nabla \psi_{f^*}$ being the gradient of a λ-strongly convex function ψ_{f^*}, for fixed vectors $\{X_i\}_{i=1}^n \subset \mathbb{R}^d$ and i.i.d. errors $\{\epsilon_i\}_{i=1}^n \sim N(0, \sigma^2 I_d)$. Let $\widehat{\pi}$ denote the minimizer of the optimization problem* (13)*. If $\min_{i<j} \|X_i - X_j\|_2 > \lambda^{-1} \sigma \sqrt{6 \log n}$, it holds with probability at least $1 - 1/n$ that $\widehat{\pi} = \pi^*$.*

Discussion. Comparison to previous work indicates that the separation condition

$$\min_{i<j} \|X_i - X_j\|_2 \geq \lambda^{-1} \sigma \sqrt{6 \log n} \tag{15}$$

3. To obtain more intuition, note that (14) is equivalent to $\langle \nabla \psi_{f^*}(z) - \nabla \psi_{f^*}(x), z - x \rangle \geq \lambda \|x - z\|_2^2$; the left hand side of this expression is obtained when evaluating the difference of the objectives of the LAP (13) in the absence of noise for $n = 2$ at $\pi_1 = \mathsf{id}$ and $\pi_2 = (2\ 1)$, respectively, with (X_1, Y_1) corresponding $(x, \nabla \psi_{f^*}(x))$ and with (X_2, Y_2) corresponding to $(z, \nabla \psi_{f^*}(z))$.

cannot be substantially relaxed. Zhang et al. (2022) consider the case in which $f^*(x) = B^*x$ is a linear transformation, which corresponds to $\psi_{f^*}(x) = \frac{1}{2}x^\top B^* x$ (up to an additive constant). Under the assumption of Gaussian noise as in Proposition 4 and Gaussian design, i.e., $\{X_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} N(0, I_d)$, it is shown that permutation recovery *fails* for *any estimator* with probability at least $1/2$ whenever

$$\sum_{j=1}^{d} \log\left(1 + \frac{\lambda_j^2}{\sigma^2}\right) \leq \log n, \tag{16}$$

where $\{\lambda_j\}_{j=1}^d$ are the singular values of B^*. In the setting of this paper, B^* is required to be symmetric positive semidefinite. Suppose that B^* has bounded condition number, i.e., $\lambda I_d \preceq B^* \preceq C\,\lambda I_d$ for some constant $C \geq 1$. In this case, the left hand side of (16) becomes proportional to $d\log(1 + \frac{\lambda^2}{\sigma^2}) \leq d\frac{\lambda^2}{\sigma^2}$, and hence in summary, permutation recovery cannot succeed if $d \lesssim \lambda^{-2}\sigma^2 \log n$. On the other hand, for $\{X_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} N(0, I_d)$, concentration results for Gaussian random vectors and the union bound yields that $\min_{i<j}\|X_i - X_j\|_2 \gtrsim \sqrt{d} - \sqrt{\log n} \gtrsim \sqrt{d}$ for $d \gtrsim \log n$ with high probability, which, when substituted into (15), implies that the condition $d \gtrsim \lambda^{-2}\sigma^2 \log n$ suffices for permutation recovery to succeed.

The above example shows that the condition in Proposition 4 is generally sharp, up to a constant factor. Moreover, the example reveals a "blessing of dimensionality" phenomenon in the sense that permutation recovery can typically (only) be hoped for in the regime $d \gtrsim \log n$. Indeed, for sub-Gaussian random designs, in that regime the scaling of the minimum separation $\min_{i<j}\|X_i - X_j\|_2$ begins to outweigh the $\sqrt{\log n}$ factor on the right hand side of the sufficient condition (15), (cf., Slawski et al., 2020, Lemma B.1). By contrast, for $d = O(1)$, $\min_{i<j}\|X_i - X_j\|_2$ may exhibit polynomial decay in n (Slawski et al., 2020, Lemma 2).

Finally, the specialization of Proposition 4 to a linear map shows that so-called unlabeled sensing problems (Unnikrishnan et al., 2018) (i.e., permuted regression problems with f^* linear) can be solved *efficiently* via the linear assignment problem (13) if the underlying linear map is *positive definite* and the noise level is small enough. So far, no computationally efficient approach to unlabeled sensing problems with provable recovery guarantees was known except for the case of "sparse shuffling" in which π^* is known to permute only a somewhat small fraction of $\{1, \ldots, n\}$ (Slawski et al., 2020, 2019; Zhang et al., 2022; Zhang and Li, 2020; Peng et al., 2021).

Connection to recovery results in the "permuted monotone matrix model". Ma et al. (2020) consider the model

$$\mathbf{Y} = \Pi^*\Theta^* + \mathbf{Z}, \tag{17}$$

where Π^* and Θ^* are unknown permutation and "signal" matrices of dimension n-by-n and n-by-d, respectively, and the entries of the noise matrix $\mathbf{Z}$ are i.i.d. $N(0, \sigma^2)$-distributed. Moreover, the entries of each of the columns of Θ^* are arranged in increasing order, i.e., for all $1 \leq j \leq d$, it holds that $\Theta^*_{ij} < \Theta^*_{(i+1)j}$, for $1 \leq i \leq n-1$.

Ma et al. (2020) study the problem of recovering Π^* from $\mathbf{Y}$. One can think of the entries (Θ^*_{ij}) as evaluations of monotone increasing functions $\{f_j^*\}_{j=1}^d$ at (unknown) design points X_{ij}, i.e., $\Theta^*_{ij} = f_j^*(X_{ij})$, $1 \leq i \leq n$, $1 \leq j \leq d$. Observe that functions of the form

$f^*(x) \equiv f^*(x_1, \ldots, x_d) = (f_1^*(x_1), \ldots, f_d^*(x_d))$ with $\{f_j^*\}_{j=1}^d$ monotone increasing equal the gradient of a sum of *univariate* convex functions, i.e., $f^*(x) = \nabla(\psi_{f_1^*}(x_1) + \ldots + \psi_{f_d^*}(x_d))$ with $\{\psi_{f_j^*}\}_{j=1}^d$ convex, which constitutes an important special case of the class of functions that are gradients of convex functions. As opposed to the setup under consideration in this paper, the setting in Ma et al. (2020) does not involve any design points $\{X_i\}_{i=1}^n$. However, specific (user-designed) choices of those points in conjunction with the linear assignment problem (13) with Y_i (the i-th row of $\mathbf{Y}$), $1 \le i \le n$, can lead to specific approaches for recovering Π^*. Perhaps the most straightforward choice is given by $X_i = x_i \mathbf{1}_d$, $1 \le i \le n$, for any increasing sequence of scalars $\{x_i\}_{i=1}^n \subset \mathbb{R}$; in this case, the LAP (13) reduces to sorting the rows of $\mathbf{Y}$ according to their row sums, which is also a rather intuitive strategy. In Ma et al. (2020) the leading right singular vector of $\mathbf{Y}$ is used instead of $\mathbf{1}_d$, which yields improved recovery results.

Remark 5 (Comparison to results in Flammarion et al. (2019); Ma et al. (2020)) *The conditions for permutation recovery in Ma et al. (2020) very much align with our condition* (15)*. The agreement can be seen best if $\{f_j^*\}_{j=1}^d$ are linear functions with non-negative slopes $\{\eta_j\}_{j=1}^d$ and $\Theta_{ij}^* \equiv f_j^*(X_{ij}) = f_j^*(x_i)$, $1 \le i \le n$, $1 \le j \le d$, for scalars $x_1 < \ldots < x_n$, in which case $\Theta^* = \mathbf{x}\boldsymbol{\eta}^\top$ with $\mathbf{x} = (x_1, \ldots, x_n)^\top$ and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_d)^\top$. It is shown in Ma et al. (2020) that the condition $\|\boldsymbol{\eta}\|_2 \gtrsim \sigma\sqrt{\log n}$ is necessary (in a minimax sense) for exact permutation recovery. Observe that $\|\boldsymbol{\eta}\|_2 \asymp \sqrt{d}\min_{1\le j\le d}\eta_j$ as long as the slopes are of the same order, which agrees with the recovery condition* (15) *up to constant factors noting that here $\lambda = \min_{1\le j\le d}\eta_j$ and assuming the scaling $\min_{i<j}\|X_i - X_j\|_2 \asymp \sqrt{d}$ as explained above. In particular, the requirement $d \gtrsim \log n$ becomes manifest once more, and also appears as a crucial condition in Flammarion et al. (2019) even though the latter paper studies model* (17) *with the goal of estimating the signal Θ^* rather than the permutation Π^*. Flammarion et al. (2019) show that the excess error in estimating Θ^* relative to an oracle that is equipped with knowledge of Π^* is proportional to $\log(n)/d$.*

3.2 Denoising

In this subsection, we present our main results on the denoising task **(T2)** based on Algorithm 1. In particular, we provide upper bounds on the mean squared error that indicate that this task can indeed be accomplished, albeit at slow rates.

The subsection is organized as follows: (i) we first present a result under the assumption of Gaussian errors for the permuted regression setting (1), which is readily extended to (ii) the *unlinked regression* setting with samples $\mathcal{X}_n$ and $\mathcal{Y}_m$ of different size; (iii) the univariate case $d = 1$ admits relaxed assumptions and a considerably simpler proof.

The following theorem addresses item (i). We first list the key assumptions on $f^* = \nabla\psi_{f^*}$.

(A1) The function ψ_{f^*} is λ-strongly convex, i.e., (14) holds.

(A2) The function ψ_{f^*} is L-smooth, i.e.,

$$\psi_{f^*}(z) \le \psi_{f^*}(x) + \langle \nabla\psi_{f^*}(x), z - x\rangle + \frac{L}{2}\|x - z\|_2^2 \quad \forall x, z \in \mathbb{R}^d. \tag{18}$$

(A3) Boundedness, i.e., $\mathbf{P}_{X\sim\mu}(\|f^*(X)\|_2 \le B) = 1$ with μ denoting the distribution generating the $\{X_i\}_{i=1}^n$.

Theorem 6 *Consider the permuted regression problem* (1) *with* $\{X_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mu$ *with* μ *being an absolutely continuous distribution on* $\mathbb{R}^d$*. Moreover, suppose that* $\{\epsilon_i\}_{i=1}^n$ *are i.i.d. Gaussian errors with zero mean and covariance* $\sigma^2 I_d$ *independent of* $\{X_i\}_{i=1}^n$*. Let* $\{\widehat{f}(X_i)\}_{i=1}^n$ *be the output of Algorithm 1. Suppose further that assumptions* **(A1)**, **(A2)** *and* **(A3)** *hold. Then if* $n \ge C_0(d, B)$*, with probability at least* $1 - 5/n$*, it holds that*

$$\frac{1}{n}\sum_{i=1}^n \|\widehat{f}(X_i) - f^*(X_i)\|_2^2 \lesssim_{\sigma,d,B} \frac{L}{\lambda}\frac{1}{\log n},$$

where $\lesssim_{[\ldots]}$ *indicates the presence of a positive multiplicative constant depending only on the quantities* $[\ldots]$ *given in the subscripts, and* $C_0(\ldots)$ *is a positive constant depending only on the quantities in the parentheses.*

The above theorem (proved in Appendix B) indicates a rather slow rate of convergence proportional to $1/\log n$. For ease of exposition, we refrain from elaborating on the constants in terms of σ, d, and B; details can be found in the Appendix containing the proofs.

Even though this paper does not present a (minimax) lower bound, rates faster than logarithmic decay generally appear implausible in view of results in the deconvolution literature (e.g., Hall and Lahiri, 2008; Fan, 1991; Dattner et al., 2011). Our simulation results in §4 in part corroborate the rate in Theorem 6.

In the following Theorem 7, we obtain a result (albeit a bit weaker one) similar to Theorem 6 without assumption **(A1)**, i.e., without requiring strong convexity of ψ_{f^*}. Note that there are several popular examples of convex functions that are *not* strongly convex, e.g., the maps $x \mapsto \|x\|_2$ (Euclidean norm) and $x \mapsto \log\big(\sum_{j=1}^d \exp(x_j)\big)$ ("log-sum-exp", the conjugate of the negative entropy). We observe that both these examples satisfy assumption **(A2)** with $L = 1$ each.

Theorem 7 *Consider the permuted regression problem* (1) *with* $\{X_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mu$ *with* μ *being an absolutely continuous distribution on* $\mathbb{R}^d$*. Further suppose that* $\{\epsilon_i\}_{i=1}^n$ *are i.i.d. Gaussian errors with zero mean and covariance* $\sigma^2 I_d$ *independent of* $\{X_i\}_{i=1}^n$*. Let* $\{\widehat{f}(X_i)\}_{i=1}^n$ *be the output of Algorithm 1. Suppose that assumptions* **(A2)** *and* **(A3)** *hold. Moreover, assume that there exists a* $B_{\mathcal{X}} > 0$ *so that the support of* μ *is contained in the Euclidean ball of radius* $B_{\mathcal{X}}$ *centered at zero. Then if* $n \ge C_0(d, B)$*, with probability at least* $1 - 5/n$*, it holds that*

$$\frac{1}{n}\sum_{i=1}^n \|\widehat{f}(X_i) - f^*(X_i)\|_2^2 \lesssim_{\sigma,d,B,B_{\mathcal{X}},L} \frac{1}{\sqrt{\log n}}.$$

Note that the upper bound in the above display is of the order $1/\sqrt{\log n}$, i.e., the rate is slower compared to Theorem 6. The slower rate is a consequence of employing deconvolution rates in Wasserstein-1 distance instead of the squared Wasserstein-2 distance. To avoid cluttering, all subsequent results (with the exception of Proposition 9) will require assumptions **(A1)** through **(A3)**.

Unlinked Regression. Our next result (proved in Appendix B) constitutes a counterpart to Theorem 6 in the unlinked regression setting.

Theorem 8 *Consider random variables $X \sim \mu$ with μ being an absolutely continuous distribution on $\mathbb{R}^d$ and $Y \stackrel{\mathcal{D}}{=} f^*(X) + \epsilon$ with $f^* = \nabla \psi_{f^*}$ such that* **(A1)**, **(A2)** *and* **(A3)** *hold true, and $\epsilon \sim N(0, \sigma^2 I_d)$ independent of X. Let $\{\widehat{f}(X_i)\}_{i=1}^n$ denote the output of Algorithm 1 given samples $\mathcal{X}_n = \{X_i\}_{i=1}^n$ and $\mathcal{Y}_m = \{Y_i\}_{i=1}^m$ consisting of i.i.d. copies of X and Y, respectively. Then if $m \geq C_1(d, B)$, with probability at least $1 - 5/m - C(n^{-c} + m^{-c})$, it holds that*

$$\frac{1}{n}\sum_{i=1}^n \|\widehat{f}(X_i) - f^*(X_i)\|_2^2 \leq \frac{2L}{\lambda}\left(\frac{C_2(\sigma, d, B)}{\log m} + 2\sqrt{\frac{\log n}{n}} \vee \Big(\frac{\log n}{n}\Big)^{2/d} + 2\sqrt{\frac{\log m}{m}} \vee \Big(\frac{\log m}{m}\Big)^{2/d}\right)$$

for absolute constants $C, c > 0$ and constants $C_1 > 0$ and $C_2 > 0$ depending only on the quantities in the parentheses.

The above statement indicates that the unlinked regression case does not behave fundamentally differently from the permuted regression setting. Specifically, as long as $n \asymp m$ the extra terms in Theorem 8 incurred in distinction to Theorem 6 are lower order terms; they reflect the Wasserstein distance between the two measures $\frac{1}{n}\sum_{i=1}^n \delta_{f^*(X_i)}$ and $\frac{1}{m}\sum_{i=1}^m \delta_{\theta_i^*}$ with $\theta_i^* \stackrel{\mathcal{D}}{=} f^*(X_1)$, $1 \leq i \leq n$. This Wasserstein distance decays more rapidly than the Wasserstein deconvolution rate of the NPMLE, which reflects the error incurred in step 1 in Algorithm 1.

We now state a separate result for the case $d = 1$; see Appendix B.7 for a proof. Even though the rates remain unchanged, it is noteworthy that assumptions **(A1)** and **(A2)** are no longer required.

Proposition 9 *Suppose that $d = 1$. Then in the situation of Theorem 6 (without requiring* **(A1)** *and* **(A2)** *to hold),*

$$\frac{1}{n}\sum_{i=1}^n |\widehat{f}(X_i) - f^*(X_i)|^2 \lesssim_{\sigma, B} \frac{1}{\log n}.$$

Furthermore, in the situation of Theorem 8 (without requiring **(A1)** *and* **(A2)** *to hold),*

$$\frac{1}{n}\sum_{i=1}^n |\widehat{f}(X_i) - f^*(X_i)|^2 \leq C(\sigma, B)\frac{1}{\log m} + 4\left(\sqrt{\frac{\log n}{n}} + \sqrt{\frac{\log m}{m}}\right)$$

where $C(\sigma, B)$ is a constant depending only on σ and B.

At this point, it is worth comparing the rates in Proposition 9, in the case $d = 1$, to previous results in the literature. Regarding the permuted regression setting, the rate in Proposition 9 falls slightly short of the minimax rate $\{\log\log n/\log n\}^2$ in Rigollet and Weed (2019). At the same time, the approach taken herein yields slightly faster rates in the unlinked regression setting than Balabdaoui et al. (2021) who bound the mean absolute error rather than the mean squared error; for Gaussian errors, they obtain the rate $1/(\log n)^{1/4}$, whereas a minor adaptation of the proof of Proposition 9 yields the rate $1/(\log n)^{1/2}$ for the mean absolute error for the proposed estimator.

3.3 Extension to other noise distributions

Note that the statements in the preceding subsection were based on the assumption of Gaussian noise. It is of interest whether comparable results can be established for other noise distributions. In this subsection, we develop results for a specific class of elliptic distributions that are characterized by a polynomial decay of the associated characteristic functions and an exponential-type tail condition. In particular, this class contains a subfamily of the *generalized multivariate Laplace distribution*, a generalization of the Laplace distribution for $d \geq 1$ (Kozubowski et al., 2013). The main effort in deriving results similar to those above goes into the analysis of the Kiefer-Wolfowitz NPMLE for the class of distributions under consideration. To our knowledge, the associated result is novel and of independent interest.

Consider model (1) where the scaled noise variables $\{\epsilon_i/\sigma\}_{i=1}^n$ are now assumed to have a density φ satisfying the following conditions:

(**D1**) $\varphi(z) = \psi(\|z\|_2)$, $z \in \mathbb{R}^d$, where $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ is decreasing, bounded at the origin, and Lipschitz continuous.

(**D2**) *Decay in the Fourier domain*: there exists $\alpha \geq d+1$ such that

$$\mathbf{F}[\varphi](\omega) := \mathbf{E}_{Z\sim\varphi}[\exp(i\omega^\top Z)] \lesssim \|\omega\|_2^{-\alpha}, \quad \omega \in \mathbb{R}^d.$$

(**D3**) *Tail behavior*: there exist constants $\beta > 0$ and $u_* > 0$ such that for all $u \geq u_*$

$$\mathbf{P}_{Z\sim\varphi}(\|Z\|_2 \geq C_{d,\alpha}\, u) \lesssim \exp(-cu^\beta),$$

where $c > 0$ is a universal constant and $C_{d,\alpha} > 0$ is a constant that may depend on d and α.

Conditions (**D1**) through (**D3**) are satisfied, for example, by the generalized multivariate Laplace distribution with parameter $\kappa \geq \frac{d+1}{2}$ whose Fourier transform is given by

$$\mathbf{F}[\varphi](\omega) = \left(\frac{1}{1+\frac{1}{2}\|\omega\|_2^2}\right)^\kappa, \quad \omega \in \mathbb{R}^d. \tag{19}$$

Note that for $\kappa = 1$ and $d = 1$, this expression equals the characteristic function of the (usual) Laplace distribution. In light of the above requirement on κ, property (**D2**) immediately follows. Properties (**D1**) and (**D3**) are verified in Appendix J.

As an intermediate result that we consider of independent interest, we provide an upper bound on the rate of convergence of the Kiefer-Wolfowitz NPMLE (9) in Hellinger distance (denoted by the symbol H), i.e., for two densities g_1, g_2 on $\mathbb{R}^d$, $\mathsf{H}^2(g_1, g_2) := \int(\sqrt{g_1} - \sqrt{g_2})^2$.

Theorem 10 *Let $\{(Y_i, \zeta_i)\}_{i=1}^n$ be independent pairs of random vectors such that the $\{\zeta_i\}_{i=1}^n$ are distributed according to a probability measure P_n supported in the Euclidean ball of radius R around the origin, and $Y_i|\zeta_i \sim \varphi_\sigma(\cdot - \zeta_i)$, $1 \leq i \leq n$, with φ satisfying properties* (**D1**) *through* (**D3**). *Let $\widehat{\mathsf{f}}_n$ denote the Kiefer-Wolfowitz NPMLE of $\mathsf{f}_n = \varphi_\sigma \star P_n$ given data $\{Y_i\}_{i=1}^n$, and let further $r_n := (\log n)^{(d/\beta+1)/2}\, n^{-\frac{1}{6}\frac{\alpha-d}{\alpha-1}}$. Then there exists a constant $t_* = t_*(R, d, \alpha, \beta)$ such that, for all $t \geq t_*$,*

$$\mathbf{P}(\mathsf{H}(\mathsf{f}_n, \widehat{\mathsf{f}}_n) \geq 2t \cdot r_n) \leq 2\exp\left(-\frac{nt^2 r_n^2}{20}\right) + \frac{1}{n}.$$

Compared with the rate of the NPMLE in the Gaussian case in Saha and Guntuboyina (2020) (cf. Appendix D) the rate in Theorem 10 is substantially slower, dropping from $n^{-1/2}$ to a rate slower than $n^{-1/6}$ (modulo logarithmic factors).

Theorem 10 paves the way for establishing results similar to Theorems 6 through 8. We here only state a counterpart to Theorem 6 and note that counterparts to Theorems 7 and 8 can be shown similarly.

Theorem 11 *Consider the permuted regression problem* (1) *with* $\{X_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mu$ *with* μ *being an absolutely continuous distribution on* $\mathbb{R}^d$. *Moreover, suppose that* $\{\epsilon_i\}_{i=1}^n$ *are i.i.d. errors independent of* $\{X_i\}_{i=1}^n$ *with density* φ_σ *so that* φ *satisfies* (**D1**) *through* (**D3**) *with associated constants* $\alpha \geq d+1, \beta > 0$. *Let* $\{\widehat{f}(X_i)\}_{i=1}^n$ *be the output of Algorithm 1. Suppose further that assumptions* **(A1)**, **(A2)** *and* **(A3)** *hold. Then if* $n \geq C_0(d, \alpha, \beta, B)$, *with probability at least* $1 - 7/n$, *it holds that*

$$\frac{1}{n}\sum_{i=1}^n \|\widehat{f}(X_i) - f^*(X_i)\|_2^2 \lesssim_{\alpha,\beta,\sigma,d,B} \frac{L}{\lambda}(\log n)^{\frac{d}{\beta(d+6)(2\alpha+d)}} n^{-\frac{1}{3(d+6)(2\alpha+d)}\frac{\alpha-d}{\alpha-1}},$$

where $\lesssim_{[\ldots]}$ *indicates the presence of a positive multiplicative constants depending only on the quantities* $[\ldots]$ *given in the subscripts and* $C_0(\ldots)$ *is a positive constant depending only on the quantities in the parentheses.*

The rate in Theorem 11 is *faster* than in the Gaussian case, with a polynomial (modulo log factors) rather than a logarithmic decay (as in Theorem 6). This is unsurprising since it is well-known (Nguyen, 2013; Gao and van der Vaart, 2016) that deconvolution rates are faster if the characteristic function of the errors exhibits polynomial decay according to (**D2**).

4. Numerical Results

In this section, we study key aspects of our rationale and analysis in the preceding sections via numerical examples. The empirical performance of the proposed approach with regard to denoising (**T2**) will also be investigated in detail, and compared to two competing methods (Balabdaoui et al., 2021; Rigollet and Weed, 2019) proposed previously for the case $d = 1$.

4.1 Permutation Recovery

This subsection is intended as an illustration of Proposition 4 concerning task (**T1**), i.e., exact permutation recovery. Three different settings are considered:

`psd`: $f^*(x) = Bx$, where B is a symmetric positive definite matrix, corresponding to the gradient of the convex function $x \mapsto \frac{1}{2}x^\top Bx$. In each replication, we generate $B \sim \mathrm{df}^{-1}\mathrm{Wishart}(I_d, \mathrm{df} = 2 \cdot d)$, where "df" is short for "degrees of freedom", $\{X_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} N(0, I_d)$, and finally $Y_i = f^*(X_i) + \sqrt{3/2}\,\epsilon_i$, $1 \leq i \leq n$.

`sep`: $f^*(x) = (3/2) \cdot (\sqrt{x_1}, \ldots, \sqrt{x_d})$, corresponding to the gradient of the separable convex function $x \equiv (x_1, \ldots, x_d) \mapsto \sum_{j=1}^d x_j^{3/2}$ on $\mathbb{R}_+^d$. In each replication, we generate $\{X_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathsf{U}([0,1]^d)$ and $Y_i = f^*(X_i) + \sqrt{2/7}\,\epsilon_i$, $1 \leq i \leq n$, where $\mathsf{U}(\ldots)$ denotes the uniform distribution.

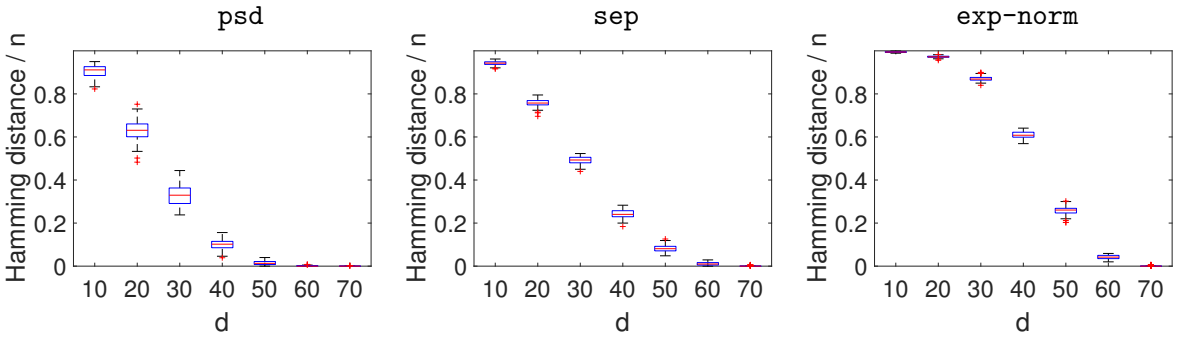


Figure 3: Boxplots of the scaled Hamming distance $\frac{1}{n}\sum_{i=1}^n \mathbb{I}(\widehat{\pi}(i) \neq \pi^*(i))$ between $\widehat{\pi}$ (from (13)) and the ground truth π^* based on 100 replications for each setting (plot) and each value of d (horizontal axis).

`exp-norm`: $f^*(x) = \frac{1}{2}\frac{x}{\|x\|_2} \cdot \exp(\|x\|_2/2)$, corresponding to the gradient of the convex function $x \mapsto \exp(\|x\|_2/2)$; convexity follows from the composition rules given in Boyd and Vandenberghe (2004), §3.2.4. In each replication, we generate $\{X_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} N(0, I_d)$, and $Y_i = f^*(X_i) + 4\epsilon_i$, $1 \leq i \leq n$.

In all three settings, we fix $n = 1,000$ and the noise terms $\{\epsilon_i\}_{i=1}^n$ are drawn i.i.d. from the $N(0, I_d)$-distribution. The noise variance is chosen specifically for each setting, to ensure comparable signal-to-noise ratios[4] across the three settings. The dimension d is varied between 10 and 70 in steps of 10. For each setting and each value of d, we perform 100 independent replications. In each replication, we solve the linear assignment problem (13), and obtain the scaled Hamming distance $\frac{1}{n}\sum_{i=1}^n \mathbb{I}(\widehat{\pi}(i) \neq i)$ (note that here $\pi^*(i) = i$, $1 \leq i \leq n$). The results are shown in Figure 3, and confirm the central insight that results from Proposition 4, namely that permutation recovery becomes considerably easier as the dimension d increases in view of the scaling of $\min_{i<j}\|X_i - X_j\|_2$. Ultimately, for d large enough, permutation recovery succeeds in all replications for all three settings.

4.2 Denoising, $d = 1$

In this subsection, we compare the performance of the proposed approach with regard to denoising **(T2)** to two methods proposed in earlier work (Rigollet and Weed, 2019; Balabdaoui et al., 2021). These two competing methods only discuss the case $d = 1$, hence our comparison is confined to this case. For our comparison, we adopt the five settings for the function f^* considered in Balabdaoui et al. (2021) and depicted in the left panel of Figure 4. Specifically, these five setting are given by

1. `linear`: $f^*(x) = x$, $x \in [0, 10]$,
2. `constant`: $f^*(x) = 0$, $x \in [0, 10]$,
3. `step2`: $f^*(x) = 2\mathbb{I}(x \in [0, 5)) + 8\mathbb{I}(x \in [0, 10])$,
4. `step3`: $f^*(x) = 5\mathbb{I}(x \in [10/3, 20/3)) + 10\mathbb{I}(x \in [20/3, 10])$,
5. `power`: $f^*(x) = -(x-5)^4\mathbb{I}(x \in (0, 5]) + (x-5)^4\mathbb{I}(x \in [5, 10])$.

4. Defined as the ratio of the left and right hand side in the recovery condition of Proposition 4.

The design points $\{X_i\}_{i=1}^n$ are sampled i.i.d. uniformly from the interval $[0, 10]$, and $Y_i = f^*(X_i) + \epsilon_i$, $1 \leq i \leq n$ (without loss of generality, we choose the permutation π^* as the identity). For the errors, we consider both Gaussian noise with zero mean and unit variance as well as Laplacian noise with zero mean and scale parameter equal to one. We consider $n = 100$ and $n = 1000$; comparison for larger n were not considered since the approach in Balabdaoui et al. (2021) does not scale favorably with n, incurring a runtime complexity of $O(n^2)$ per gradient iteration. Hundred independent replications are performed for each configuration in terms of the setting for f^*, noise distribution, and sample size.

The proposed approach (Slawski & Sen, short SS) is run by solving the approximate NPMLE problem (10) with $\mathbb{G}$ chosen as a linearly spaced grid of size $2\lceil n^{1/2} \rceil$ between $\min_i Y_i$ and $\max_i Y_i$, and the resulting deconvolution estimate $\widehat{\nu} = \sum_{j=1}^p \widehat{\alpha}_j \delta_{\widehat{\theta}_j}$ is used for the Kantorovich problem (12). The competitor BDD (initials of the last names of the authors of Balabdaoui et al. (2021)) is run based on an in-house implementation of the gradient descent method in that paper, using the starting values $\widehat{f}(X_{(i)}) = Y_{(i)}$, $1 \leq i \leq n$. Gradient descent is performed with constant step size; for the sake of fair comparison, six different values for the step size between 0.01 and 0.5 are considered, and for each configuration we report the result of the specific step size achieving minimum average error over the respective replications. The competitor RW (Rigollet and Weed, 2019) is run based on an in-house implementation of a subgradient descent method to solve the (discretized) Wasserstein deconvolution problem considered in that paper (cf. §2.2 therein). The size of the quantization alphabet is taken as $\lceil 2\sqrt{n} \rceil$, linearly spaced between $\min_i Y_i$ and $\max_i Y_i$. Each optimal transport problem required for subgradient computation is approximated via Sinkhorn's algorithm (Peyré and Cuturi, 2019, §4) with regularization parameter $\varepsilon = 0.1$. As for BDD, we consider six different values for the step size between $5 \cdot 10^{-5}$ and $2 \cdot 10^{-3}$, and select the results achieving minimum average error over these six choices.

Results. The results of our comparison are visualized in Figure 4 via boxplots showing the mean squared denoising errors over 100 replications. The general picture is that BDD achieves the best empirical performance (with optimized step size), while the performance of the proposed approach SS is often on par with BDD. In our comparison, the relative performance of SS is worse for "smooth" f^* (settings `linear` and `power`). By contrast, RW performs rather poorly for the settings `constant`, `step2`, and `step3`. Somewhat surprisingly, `RW` does not exhibit any noticeable decrease in error as the sample size is increased from 100 to 1000 with the exception of setting `linear` and Gaussian errors. Despite careful monitoring of convergence and inspection of potential computational issues, it is quite well possible that the performance of RW can be improved substantially with a refined implementation[5] since in fact all three approaches compared herein follow rather similar rationales, and gaps in performance are thus not expected.

4.3 Denoising, $d > 1$

This subsection is intended to corroborate and complement aspects of our theoretical results in §3.2 regarding task **(T2)** for general dimension. The competitors in the preceding section were developed for the case $d = 1$, hence we confine ourselves to the proposed method.

5. The authors of that method did not publish their implementation/code.

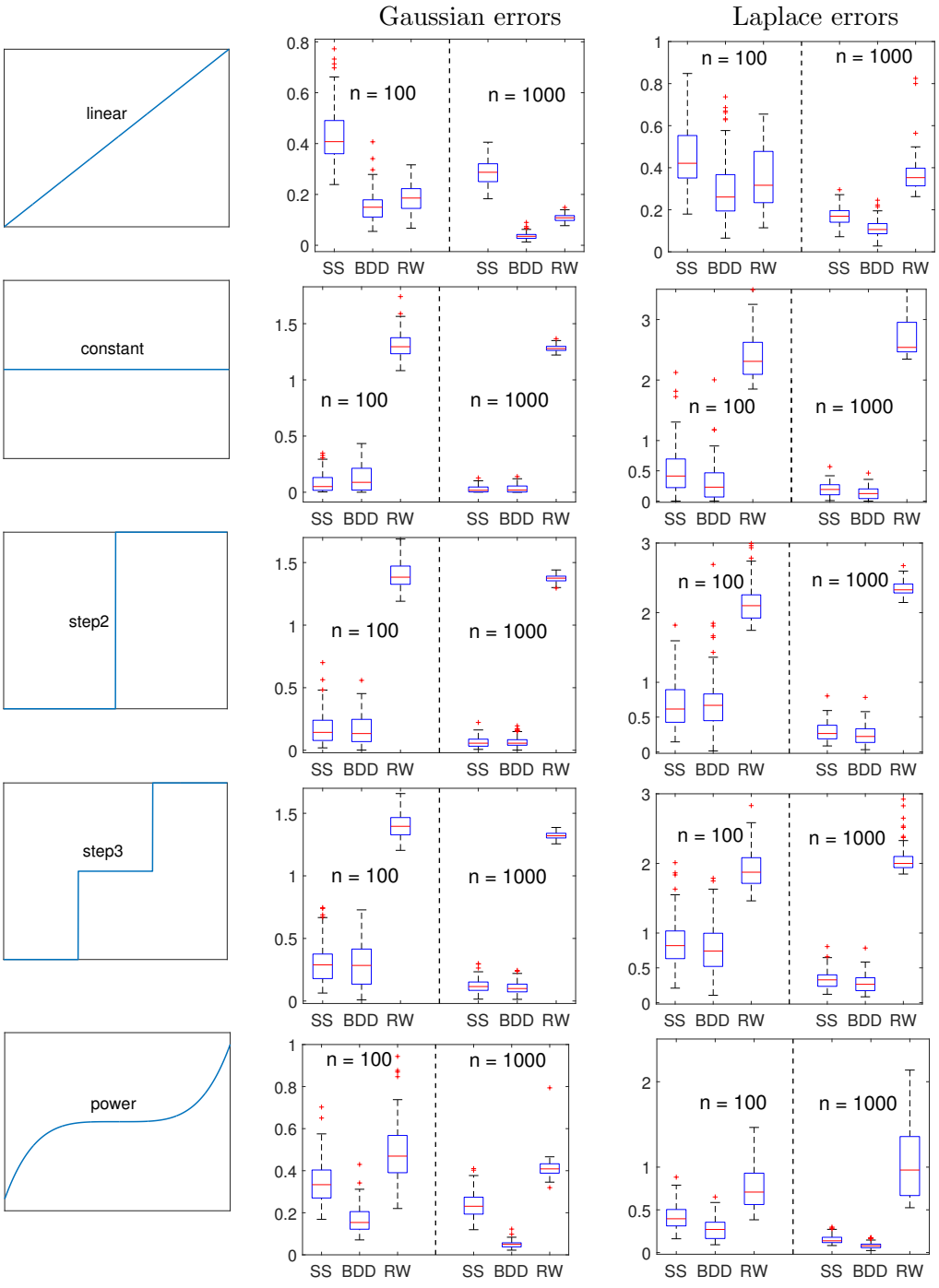


Figure 4: Results of the comparison of the three approaches under consideration for the denoising problem **(T2)**. Left: underlying function f^*. Middle and right: Boxplots of mean squared errors for Gaussian and Laplace errors, respectively.

	`cluster`	`linear`	`separable`	`sphere`	`radial`
$\psi_{f^*}(x)$	$\max_{1\le j\le k}\langle a_j, x\rangle$	$\frac{1}{2}x^\top \sum_{j=1}^k v_j v_j^\top x$	$\frac{2}{3}\sum_{j=1}^d (x_j+1)^{3/2}$	$\|x\|_2$	$\exp(\|x\|_2^2/2)$
$f^*(x)$	a_x	$\sum_{j=1}^k \langle x, v_j\rangle v_j$	$\sum_{j=1}^d (x_j+1)^{1/2}$	$\frac{x}{\|x\|_2}$	$x\cdot\exp(\|x\|_2^2/2)$

Table 1: Summary of the simulation settings considered in §4.3. In the 2nd column from the left, a_x is short for $\{a_j : \langle x, a_j\rangle = \psi_{f^*}(x)\}$.

We generate data following the permuted regression setup (1). The sample $\{X_i\}_{i=1}^n$ is sampled uniformly from the unit Euclidean ball in $\mathbb{R}^d$ ($d = 2, 4$), and subsequently we generate $Y_i = f^*(X_i)+\sigma\epsilon_i$, $1 \le i \le n$, where $\sigma = 1/16$ and the $\{\epsilon_i\}_{i=1}^n$ are sampled i.i.d. from the $N(0, I_d)$-distribution and alternatively from the multivariate Laplace distribution[6]. The settings considered for f^* are summarized in Table 1. For the sample size, we consider $n = 2^8, 2^9, \ldots, 2^{12} = 4096$ (and in some selected settings 2^{13}) in anticipation of slow rates as indicated by the results in §3.2.

The proposed approach is run as follows: we solve the approximate NPMLE problem (11) with $\mathbb{G} = \{Y_i\}_{i=1}^n$ equipped with knowledge of φ_σ, and use the resulting deconvolution estimate $\widehat{\nu}$ in the Kantorovich problem (12) to obtain $\{\widehat{f}(X_i)\}_{i=1}^n$. We then report the normalized MSE $\frac{1}{n\sigma^2}\sum_{i=1}^n \|\widehat{f}(X_i) - f^*(X_i)\|_2^2$. The results depicted in Figure 5 represent averages over 100 independent replications, with bars indicating $\pm$ standard error.

Results. First, the results shown in Figure 5 confirm that the rates are indeed slow as expected in light of the results in §3.2, with an error decay that is linear on a log-log scale for some instances (corresponding to a polynomial rate in n) and noticeably sublinear for others. While the discussion at the end of §3.2 suggests that Laplacian errors will yield faster rates, this is not confirmed by our simulations; the observed denoising error is often comparable if not higher than for Gaussian errors. Moreover, while the analysis in §3.2 suggests faster rates given strong convexity of ψ_{f^*}, the empirical results for several of the settings considered here (`cluster`, `linear` with $k < d$ and `sphere`) do not indicate that the lack of strong convexity generally prompts a substantially different scaling of the denoising error. In fact, the setting `cluster` corresponds to a clustering problem with a finite number of clusters, i.e., the underlying problem is parametric rather than non-parametric, and one would hence intuitively expect even faster rates. This intuition is confirmed by our results and is further supported by a recent result in Soloff et al. (2024) (cf. Theorem 12 therein). In a similar vein, we also observe smaller errors if the "intrinsic dimension" of the problem is smaller than the ambient dimension: in the setting `linear`, the parameter k reflects the intrinsic dimension, and Figure 5 indeed shows that the denoising error drops as k is reduced. For several settings with $d = 4$ (in particular `sphere` and `radial`) the denoising error is essentially flat and starts decreasing only after n becomes rather large. This behavior is not understood at this point; one possible explanation for the setting `sphere` might be the lack of strong convexity in conjunction with the absence of additional structure such as in the setting `cluster`.

6. Specifically, we generate $\epsilon_i = g_i \cdot \xi_i$, where $g_i \sim N(0, I_d)$-distribution and $\xi_i \sim \mathrm{Exp}(1)$, $1 \le i \le n$, where $\mathrm{Exp}(1)$ denotes the exponential distribution with unit scale.

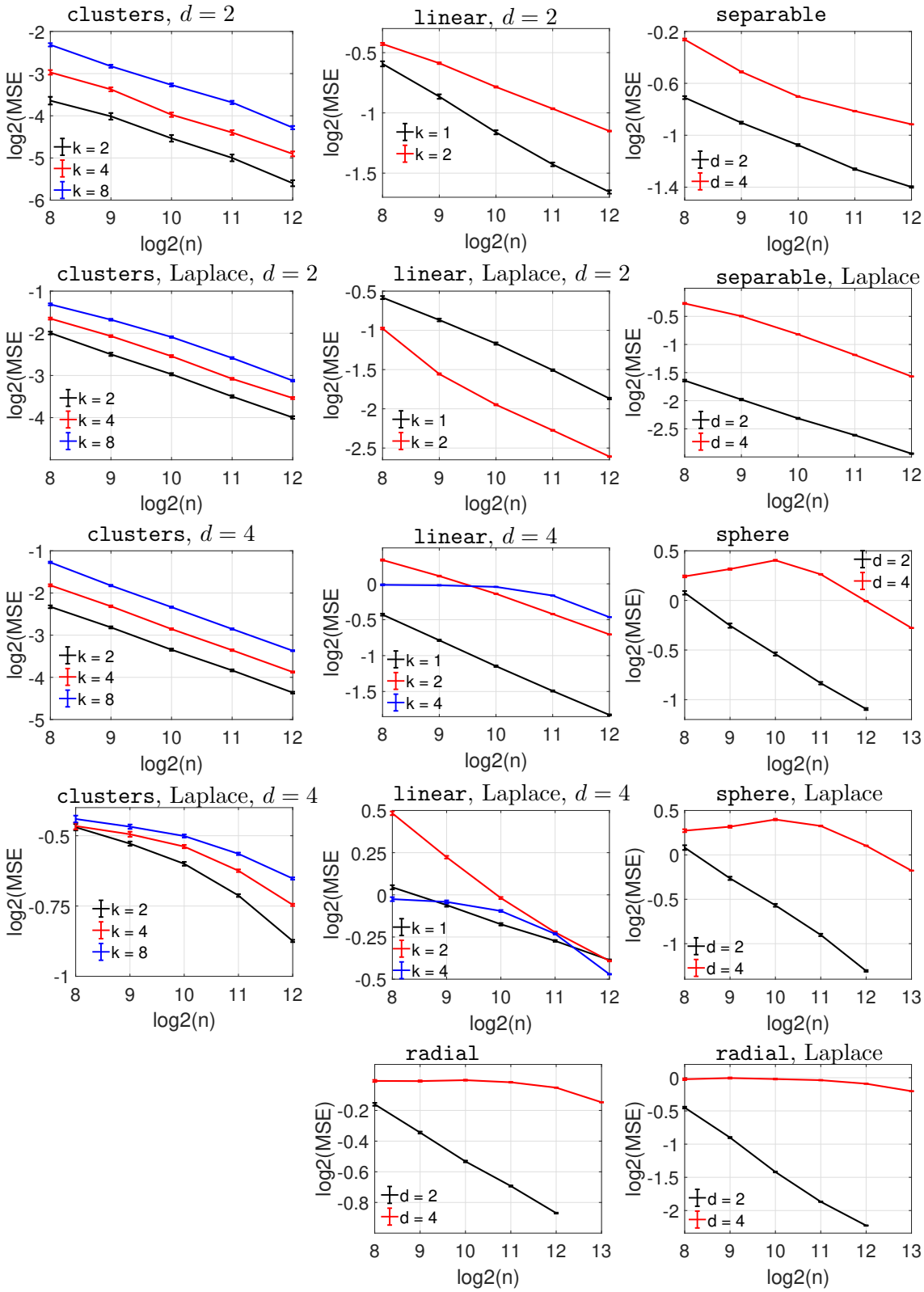


Figure 5: Denoising results for $d = 2, 4$; "MSE" refers to $\frac{1}{n\sigma^2}\sum_{i=1}^{n}\|\widehat{f}(X_i) - f^*(X_i)\|_2^2$. The results shown represent averages ± 1 standard error (the error bars are hardly visible for most instances) over 100 independent replications for the respective setting given in the plot captions (unless specified otherwise, the noise distribution is Gaussian).

5. Conclusion

In this paper, we have considered permuted and uncoupled regression for maps $f^* : \mathbb{R}^d \to \mathbb{R}^d$ that are gradients of convex functions, the multi-dimensional analog of monotone non-decreasing functions. This paper has studied exact permutation recovery and denoising, and has established several connections to several recent works involving related permuted data problems. The task of denoising is tackled via deconvolution based on the Kiefer-Wolfowitz NPMLE and optimal transport. The rich literature and the recently surging interest regarding the latter topic facilitates the analysis of the proposed approach. Compared to prior work on one-dimensional permuted regression problems, the implementation of our approach is particularly convenient since the underlying convex optimization problems are straightforward to solve and do not require careful tuning; currently, the only parameter to be specified is the grid for the approximate Kiefer-Wolfowitz problem, for which straightforward default options are available that yield reasonable results empirically (cf. §4). Note that the finer the grid the more accurate the result, so the approach is constrained only by the available computing power.

Despite the advances made in the current paper, there are several open problems and possible extensions from both practical and theoretical viewpoints as discussed below.

(I) *Towards deconvolution with unknown distribution of the errors.* Even though this objective appears not to be achievable in general, it is of great practical importance to relax the somewhat unrealistic assumption that the distribution of the error terms is fully known. As first steps, the following directions can be pursued: (i) the scale parameter σ is not known, and needs to be selected in a data-driven manner, and (ii) the model used for the errors is (mildly) misspecified.

(II) *Beyond denoising.* In this paper, we focus on denoising, i.e., the estimation of the values of the unknown function f^* at the sample points $\{X_i\}_{i=1}^n$. A next step is to develop an approach that provides a (smooth) estimate of f^* over, say, a compact domain.

(III) *Minimaxity and adaptation.* Concerning our results obtained for denoising, the minimax rate is yet unknown except for $d = 1$ (Rigollet and Weed, 2019). While slow rates appear inevitable in general, parts of our simulation results indicate that faster rates can be obtained for instances with additional structure such as piecewise affine functions and functions with low intrinsic dimensionality. In this context, it is of interest to study whether the proposed approach adapts to such underlying low-complexity structure.

(IV) *Wasserstein vs. maximum likelihood (ML) deconvolution.* The approach presented in this paper is based on the Kiefer-Wolfowitz problem and thus ML deconvolution. Our analysis, however, is based on bounding the distance to the underlying mixing measure in Wasserstein distance. This raises the question whether the use of the Wasserstein distance (as done in Rigollet and Weed (2019) for $d = 1$) instead of the Kullback-Leibler divergence is more suitable for the problem at hand. At the same time, ML deconvolution is considerably more convenient from a computational perspective. An interesting connection between ML deconvolution and entropic optimal transport is made in Rigollet and Weed (2018). It is of interest to study whether that connection can be leveraged to facilitate the analysis of the proposed approach.

(V) *Beyond equal dimensions.* The route taken in this paper requires f^* to be a map from $\mathbb{R}^d$ to $\mathbb{R}^d$. This requirement can be limiting in applications in which the two samples $\mathcal{X}_n$ and $\mathcal{Y}_n$ live in different dimensions.

References

A. Abid, A. Poon, and J. Zou. Linear Regression with Shuffled Labels. arXiv:1705.01342, 2017.

M. Azadkia and F. Balabdaoui. Linear regression with unmatched data: a deconvolution perspective. arXiv:2207.06320, September 2023.

Z. Bai and T. Hsing. The broken sample problem. *Probability Theory and Related Fields*, 131(4):528–552, 2005.

F. Balabdaoui, C.R. Doss, and C. Durot. Unlinked monotone regression. *Journal of Machine Learning Research*, 22(172):1–60, 2021.

H. Bauschke and P. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, 2011.

D. Bertsekas and D. Castanon. A forward/reverse auction algorihtm for asymmetric assignment problems. *Computational Optimization and Applications*, 1:277–297, 1992.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

R. Burkard and E. Cela. Linear assignment problems and extensions. In *Handbook of combinatorial optimization: Supplement volume A*, pages 75–149. Springer, 1999.

R. Burkard, M. Dell'Amico, and S. Martello. *Assignment Problems: Revised Reprint*. SIAM, 2009.

A. Carpentier and T. Schlüter. Learning relationships between data obtained independently. In *Proceedings of the International Conference on Artifical Intelligence and Statistics (AISTATS)*, pages 658–666, 2016.

Lenaic Chizat, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, and Gabriel Peyré. Faster Wasserstein distance estimation with the Sinkhorn divergence. arXiv:2006.08172, June 2020.

P. Christen. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer, 2012.

O. Collier and A. Dalalyan. Minimax Rates in Permutation Estimation for Feature Matching. *Journal of Machine Learning Research*, 17:1–31, 2016.

M. Cuturi, O. Teboul, and J.-P. Vert. Differentiable ranking and sorting using optimal transport. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Itai Dattner, Alexander Goldenshluger, and Anatoli Juditsky. On deconvolution of distribution functions. *The Annals of Statistics*, pages 2477–2501, 2011.

N. Deb, P. Ghosal, and B. Sen. Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections. In *Advances in Neural Information Processing Systems*, volume 34, pages 29736–29753, 2021.

M. DeGroot and P. Goel. Estimation of the correlation coefficient from a broken random sample. *The Annals of Statistics*, 8:264–278, 1980.

M. DeGroot, P. Feder, and P. Goel. Matchmaking. *The Annals of Mathematical Statistics*, 42:578–593, 1971.

L. Dicker and S. Zhao. High-dimensional classification via nonparametric empirical Bayes and maximum likelihood inference. *Biometrika*, 103(1):21–34, 2016.

J. Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, pages 1257–1272, 1991.

N. Flammarion, C. Mao, and P. Rigollet. Optimal Rates of Statistical Seriation. *Bernoulli*, 25:623–653, 2019.

N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, 2015.

F. Gao and A. van der Vaart. Posterior contraction rates for deconvolution of Dirichlet-Laplace mixtures. *Electronic Journal of Statistics*, 10(1):608–627, 2016.

P. Ghosal and B. Sen. Multivariate ranks and quantiles using optimal transport: Consistency, rates and nonparametric testing. *The Annals of Statistics*, 50(2):1012–1037, 2022.

Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised alignment of embeddings with wasserstein procrustes. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1880–1890, 2019.

Peter Hall and Soumendra N Lahiri. Estimation of distributions, moments and quantiles in deconvolution problems. *The Annals of Statistics*, 36(5):2110–2134, 2008.

R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2nd edition, 2004.

T. Herzog, F. Scheuren, and W. Winkler. *Data quality and record linkage techniques*. Springer, 2007.

D. Hsu, K. Shi, and X. Sun. Linear regression without correspondence. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1531–1540, 2017.

Jan-Christian Hütter and Philippe Rigollet. Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2):1166–1194, 2021.

IBM. ILOG CPLEX Optimization Studio. `http://www.ibm.com/us-en/marketplace/ibm-ilog-cplex`, 2016.

N. Ignatiadis and B. Sen. Empirical partially Bayes multiple testing and compound χ^2-decisions. arXiv:2303.02887, 2023.

W. Jiang and C.-H. Zhang. General maximum likelihood empirical Bayes estimation of normal means. *The Annals of Statistics*, 37:1647–1684, 2009.

S. Kakade, S. Shalev-Shwartz, and A. Tewari. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. `http://ttic.uchicago.edu/shai/papers/KakadeShalevTewari09.pdf`, 2009.

J. Kiefer and J. Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, pages 887–906, 1956.

R. Koenker and I. Mizera. Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *Journal of the American Statistical Association*, 109(506): 674–685, 2014.

T. Kozubowski, K. Podgórski, and I. Rychlik. Multivariate generalized Laplace distribution and related random fields. *Journal of Multivariate Analysis*, 113:59–72, 2013.

H. Kuhn. The Hungarian Method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955.

P. Lahiri and M. D. Larsen. Regression analysis with linked data. *Journal of the American Statistical Association*, 100(469):222–230, 2005.

I. Liiv. Seriation and matrix reordering methods: An historical overview. *Statistical Analysis and Data Mining*, 3:70–91, 2010.

Bruce G Lindsay. The geometry of mixture likelihoods: a general theory. *The Annals of Statistics*, 11:86–94, 1983.

R. Ma, T. Cai, and H. Li. Optimal permutation recovery in permuted monotone matrix model. *Journal of the American Statistical Association*, 116:1358–1372, 2020.

Rong Ma, T Tony Cai, and Hongzhe Li. Optimal estimation of bacterial growth rates based on a permuted monotone matrix. *Biometrika*, 108(3):693–708, 2021.

Tudor Manole, Sivaraman Balakrishnan, Jonathan Niles-Weed, and Larry Wasserman. Plug-in estimation of smooth optimal transport maps. arXiv:2107.12364; to appear in the Annals of Statistics, May 2024.

R. McCann and N. Guillen. *Five lectures on optimal transportation: geometry, regularity and applications*, pages 145 – 180. American Mathematical Society, 2011.

J. Meis and E. Mammen. Uncoupled isotonic regression with discrete errors. In *Advances in Contemporary Statistics and Econometrics*, pages 123–135. Springer, 2021.

A. Meister. *Deconvolution Problems in Nonparametric Statistics*. Springer, 2009.

A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *IEEE Symposium on Security and Privacy*, pages 111–125, 2008.

X.-L. Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *Annals of Statistics*, 41(1):370–400, 2013.

A. Pananjady, M. Wainwright, and T. Cortade. Denoising Linear Models with Permuted Data. arXiv:1704.07461, 2017.

A. Pananjady, M. Wainwright, and T. Cortade. Linear regression with shuffled data: Statistical and computational limits of permutation recovery. *IEEE Transactions on Information Theory*, 3826–3300, 2018.

François-Pierre Paty, Alexandre d'Aspremont, and Marco Cuturi. Regularity as regularization: Smooth and strongly convex brenier potentials in optimal transport. In *International Conference on Artificial Intelligence and Statistics*, pages 1222–1232, 2020.

L. Peng, B. Wang, and M. Tsakiris. Homomorphic sensing: Sparsity and noise. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8464–8475, 2021.

G. Peyré and M. Cuturi. Computational Optimal Transport: With Applications to Data Science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

P. Rigollet and J. Weed. Entropic optimal transport is maximum-likelihood deconvolution. *Comptes Rendus Mathematique*, 356(11-12):1228–1235, 2018.

P. Rigollet and J. Weed. Uncoupled isotonic regression via minimum Wasserstein deconvolution. *Information and Inference*, 8:691–717, 2019.

R. Rockafellar. Characterization of the subdifferentials of convex functions. *Pacific Journal of Mathematics*, 17(3):497–510, 1966.

S. Saha and A. Guntuboyina. On the nonparametric maximum likelihood estimator for Gaussian location mixture densities with application to Gaussian denoising. *Annals of Statistics*, 48(2):738–762, 2020.

F. Santambrogio. *Optimal Transport for Applied Mathematicians*. Birkäuser, NY, 2015.

F. Scheuren and W. Winkler. Regression analysis of data files that are computer matched I. *Survey Methodology*, 19:39–58, 1993.

F. Scheuren and W. Winkler. Regression analysis of data files that are computer matched II. *Survey Methodology*, 23:157–165, 12 1997.

X. Shi, X. Lu, and T. Cai. Spherical regression under mismatch corruption with application to automated knowledge translation. *Journal of the American Statistical Association*, 116 (536):1953–1964, 2021.

M. Slawski and E. Ben-David. Linear Regression with Sparsely Permuted Data. *Electronic Journal of Statistics*, 1:1–36, 2019.

M. Slawski, M. Rahmani, and P. Li. A Robust Subspace Recovery Approach to Linear Regression with Partially Shuffled Labels. In *Uncertainty in Artificial Intelligence (UAI)*, pages 38–48, 2019.

M. Slawski, E. Ben-David, and P. Li. A Two-Stage Approach to Multivariate Linear Regression with Sparsely Mismatched Data. *Journal of Machine Learning Research*, 21(204): 1–42, 2020.

M. Slawski, G. Diao, and E. Ben-David. A Pseudo-Likelihood Approach to Linear Regression with Partially Shuffled Data. *Journal of Computational and Graphical Statistics*, 30: 991–1003, 2021.

J.A. Soloff, A. Guntuboyina, and B. Sen. Multivariate, heteroscedastic empirical bayes via nonparametric maximum likelihood. arXiv:2109.03466; to appear in the Journal of the Royal Statistical Society Series B, 2024.

L. Sweeney. *Computational disclosure control: A primer on data privacy protection*. PhD thesis, Massachusetts Institute of Technology, 2001.

M. Tsakiris and L. Peng. Homomorphic sensing. In *International Conference on Machine Learning (ICML)*, pages 6335–6344, 2019.

M. Tsakiris, L. Peng, A. Conca, L. Kneip, Y. Shi, and H. Choi. An Algebraic-Geometric Approach to Shuffled Linear Regression. *IEEE Transactions on Information Theory*, 66: 5130–5144, 2020.

J. Unnikrishnan, S. Haghighatshoar, and M. Vetterli. Unlabeled sensing with random linear measurements. *IEEE Transactions on Information Theory*, 64:3237–3253, 2018.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018.

C. Villani. *Topics in Optimal Transportation*. American Mathematical Society, 2003.

C. Villani. *Optimal transport: old and new*. Springer, 2009.

W. E. Winkler. Matching and record linkage. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(5):313–325, 2014.

C.-H. Zhang. Generalized maximum likelihood estimation of normal mixture densities. *Statistica Sinica*, pages 1297–1318, 2009.

H. Zhang and P. Li. Optimal estimator for unlabeled linear regression. In *Proceedings of the 37th International Conference on Machine Learning*, pages 11153–11162, 2020.

H. Zhang, M. Slawski, and P. Li. The Benefits of Diversity: Permutation Recovery in Unlabeled Sensing from Multiple Measurement Vectors. *IEEE Transactions on Information Theory*, 68:2509–2529, 2022.

G. Ziegler. *Lectures on polytopes*. Graduate Texts in Mathematics. Springer, 1995. Updated 7th edition of first priting.

Organization of the proofs. This appendix contains proofs of our main results and additional technical background and discussion. The proofs of Theorems 6, 7, 8 and Proposition 9 are decomposed into several key pieces which are presented in dedicated sections. The specific constituents and their dependencies are outlined in Figure 6.

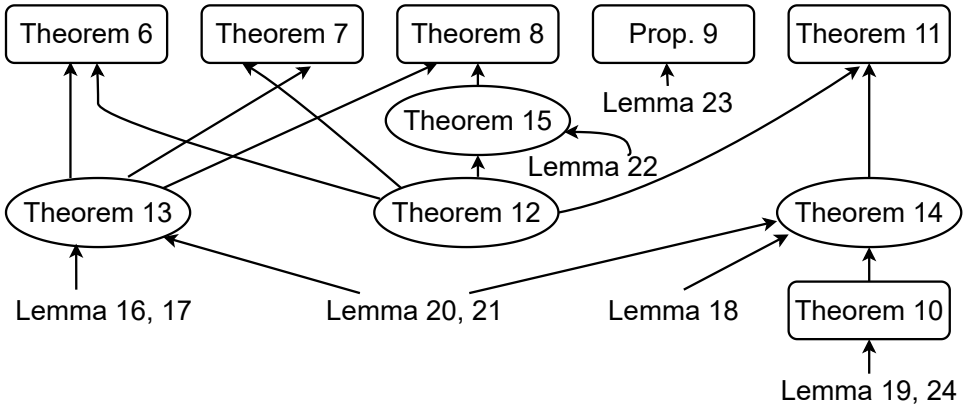


Figure 6: Chart summarizing the organization of the proofs of our main results on denoising **(T2)**. Boxed statements are provided in the body of the paper and circled statements are provided in the appendix.

Appendix A. Proof of Proposition 4

Without loss of generality, we may assume that π^* is the identity permutation, i.e., $\pi^*(i) = i$, $1 \leq i \leq n$. It then suffices to show that the set $\{(X_i, Y_i)\}_{i=1}^n$ is cyclically monotone with respect to the cost function $c_0(x, y) := \|x - y\|_2^2$, or equivalently the cost function $c(x, y) = -\langle x, y \rangle$, under the conditions stated in the proposition. For this purpose, we need to show that for any subset $\{(X_{i_j}, Y_{i_j})\}_{j=1}^k$ of $\{(X_i, Y_i)\}_{i=1}^n$ of size $k \geq 2$, it holds that

$$-\sum_{j=1}^{k} \langle X_{i_j}, Y_{i_j} \rangle \leq -\sum_{j=1}^{k} \langle X_{i_{j+1}}, Y_{i_j} \rangle, \quad i_{k+1} = i_1. \tag{20}$$

Expanding $Y_{i_j} = f^*(X_{i_j}) + \epsilon_{i_j}$, $1 \leq j \leq k+1$, the above inequality becomes

$$-\sum_{j=1}^{k} \langle X_{i_j}, f^*(X_{i_j}) + \epsilon_{i_j} \rangle \leq -\sum_{j=1}^{k} \langle X_{i_{j+1}}, f^*(X_{i_j}) + \epsilon_{i_j} \rangle.$$

Re-arranging in order to single out the contributions of the noise yields the following condition equivalent to (20)

$$\sum_{j=1}^{k} \langle X_{i_{j+1}} - X_{i_j}, f^*(X_{i_j}) \rangle + \sum_{j=1}^{k} \langle X_{i_{j+1}} - X_{i_j}, \epsilon_{i_j} \rangle \leq 0. \tag{21}$$

By λ-strong convexity of ψ_{f^*}, we have

$$\psi_{f^*}(X_{i_{j+1}}) - \psi_{f^*}(X_{i_j}) - \langle \underbrace{\nabla \psi_{f^*}(X_{i_j})}_{f^*(X_{i_j})}, X_{i_{j+1}} - X_{i_j} \rangle \geq \lambda \|X_{i_{j+1}} - X_{i_j}\|_2^2, \quad 1 \leq j \leq k.$$

Summation of the above inequality over j and using the cyclicity condition $i_{k+1} = i_1$ yields

$$\sum_{j=1}^{k} \left\langle X_{i_{j+1}} - X_{i_j}, f^*(X_{i_j}) \right\rangle \leq -\lambda \sum_{j=1}^{k} \|X_{i_{j+1}} - X_{i_j}\|_2^2. \tag{22}$$

We now upper bound the second term in (21). Conditional on the $\{X_i\}_{i=1}^n$ and using the independence of the errors, we have

$$\sum_{j=1}^{k} \left\langle X_{i_{j+1}} - X_{i_j}, \epsilon_{i_j} \right\rangle \sim N\Big(0, \sigma^2 \sum_{j=1}^{k} \|X_{i_{j+1}} - X_{i_j}\|_2^2\Big).$$

Now define the quantity

$$M_k := \max_{\{i_1,\ldots,i_k\} \subseteq \{1,\ldots,n\}} \frac{1}{\sqrt{\sum_{j=1}^{k} \|X_{i_{j+1}} - X_{i_j}\|_2^2}} \sum_{j=1}^{k} \left\langle X_{i_{j+1}} - X_{i_j}, \epsilon_{i_j} \right\rangle. \tag{23}$$

The standard Gaussian tail bound $\mathbf{P}(Z > z) \leq \exp(-z^2/2)$ for $z \geq 0$, $Z \sim N(0,1)$, combined with the union bound and the inequality $\binom{n}{k} \leq \left(\frac{en}{k}\right)^k$ yields

$$\mathbf{P}(M_k \geq t) \leq \exp\left(-\frac{t^2}{2\sigma^2} + k \log(en/k)\right), \quad t \geq 0.$$

Choosing $t = \sigma\sqrt{4 \log n + 2k \log(en/k)}$, we obtain that

$$\mathbf{P}\left(M_k \geq \sigma\sqrt{4 \log n + 2k \log(en/k)}\right) \leq \frac{1}{n^2}. \tag{24}$$

Combining (21), (22), (23), we note that the desired condition (20) is implied by the condition

$$\forall k = 2, \ldots, n: \quad \lambda\sqrt{k} \min_{i<j} \|X_i - X_j\| \geq M_k.$$

Using (24) along with the observation that the function $k \mapsto \sigma\sqrt{4 \log(n)/k + 2 \log(en/k)}$ is decreasing in k, a union bound over $k = 2, \ldots, n$, yields that if

$$\min_{i<j} \|X_i - X_j\|_2 \geq \frac{\sigma\sqrt{6 \log n}}{\lambda},$$

the required inequality (20) for cyclic monotonicity holds with probability at least $1 - 1/n$.

Appendix B. Proof of Theorems 6, 7, 8 and 11 and Proposition 9

The proofs of these two theorems involve a few other results, which we first state in the following subsections. These results may be of independent interest.

Let $\nu_n^* := \frac{1}{n}\sum_{i=1}^n \delta_{\theta_i^*} = \frac{1}{n}\sum_{i=1}^n \delta_{f^*(X_i)}$, and let $\widehat{\nu}$ denote the mixing measure associated with the NPMLE (9). Theorem 12 below provides an *upper bound* on the empirical L_2-loss of the barycentric projection estimator $\{\widehat{f}(X_i)\}_{i=1}^n$, obtained from an optimal coupling between ν_n^* and $\widehat{\nu}$ (see (8)), in terms of the 2-Wasserstein distance between ν_n^* and $\widehat{\nu}$.

B.1 Analysis of the Kantorovich problem (12) for general d

The result below is the central technical component in proving Theorem 6; see Appendix C.1 for its proof, cf. also Deb et al. (2021); Manole et al. (2024).

Theorem 12 *Consider the atomic measure $\nu_n^* := \frac{1}{n}\sum_{i=1}^n \delta_{\theta_i^*}$, $\theta_i^* = f^*(X_i)$, $1 \le i \le n$, with $f^* = \nabla\psi_{f^*}$ such that assumptions* **(A1)**-**(A3)** *in §3.2 are satisfied, and let $\widehat{\nu} := \sum_{j=1}^p \widehat{\alpha}_j \delta_{\widehat{\theta}_j}$ be another atomic measure on $\mathbb{R}^d$. Let further $\mu_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, and consider the barycentric projection (8) based on an optimal coupling (7) between μ_n and $\widehat{\nu}$. We then have*

$$\frac{1}{n}\sum_{i=1}^n \|\widehat{f}(X_i) - f^*(X_i)\|_2^2 \le \frac{L}{\lambda}\mathsf{W}_2^2(\nu_n^*, \widehat{\nu}). \tag{25}$$

We next upper bound $\mathsf{W}_2^2(\nu_n^*, \widehat{\nu})$.

B.2 Wasserstein deconvolution rates

The following result provides an upper bound on the 2-Wasserstein distance between an underlying atomic mixing measure $\nu_n^* = \frac{1}{n}\sum_{i=1}^n \delta_{\theta_i^*}$ with uniformly bounded support and a deconvolution estimator $\widehat{\nu}$ in terms of the Hellinger distance between the convolved measures $\nu_n^* \star \varphi_\sigma$ and $\widehat{\nu} \star \varphi_\sigma$, along the route of the proof of Theorem 2 in Nguyen (2013); see Appendix C.3 for a proof.

Theorem 13 *Let $\widehat{\mathsf{f}}_n$ denote the NPMLE (9) given $\{Y_i\}_{i=1}^n$ such that $\{(Y_i, \theta_i^*)\}_{i=1}^n$ are independent random vectors and $Y_i|\theta_i^* \sim \varphi_\sigma(\cdot - \theta_i^*)$ with $\varphi(z) = (2\pi)^{-d/2}\exp(-\|z\|_2^2)$ and θ_i^* contained in the Euclidean ball of radius B centered at the origin almost surely, $1 \le i \le n$. Let $\nu_n^* = \frac{1}{n}\sum_{i=1}^n \delta_{\theta_i^*}$ and let further $\widehat{\nu}$ be the mixing measure associated with the NPMLE, i.e., $\widehat{\nu} \star \varphi_\sigma = \widehat{\mathsf{f}}_n$. Choose $s > k \ge 1$, and suppose that $n \ge C_0(d, B)$. It then holds with probability at least $1 - 5/n$,*

$$\mathsf{W}_k^k(\nu_n^*, \widehat{\nu}) \le C(s, k, d, \sigma, B)\left(\frac{1}{\log n}\right)^{k/2},$$

where C_0 and C are positive constants depending only on the quantities in the parentheses.

Theorem 14 *Let $\widehat{\mathsf{f}}_n$ denote the NPMLE (9) given $\{Y_i\}_{i=1}^n$ such that $\{(Y_i, \theta_i^*)\}_{i=1}^n$ are independent random vectors and $Y_i|\theta_i^* \sim \varphi_\sigma(\cdot - \theta_i^*)$ with φ satisfying conditions* (**D1**)–(**D3**) *in §3 and θ_i^* contained in the Euclidean ball of radius B centered at the origin almost surely,*

$1 \le i \le n$. Let $\nu_n^* = \frac{1}{n}\sum_{i=1}^n \delta_{\theta_i^*}$ and let further $\widehat{\nu}$ be the mixing measure associated with the NPMLE, i.e., $\widehat{\nu} \star \varphi_\sigma = \widehat{\mathsf{f}}_n$. Suppose that $n \ge C_0(d,\alpha,\beta,B)$. It then holds with probability at least $1 - 7/n$

$$\mathsf{W}_k^k(\nu_n^*, \widehat{\nu}) \le C(k,\alpha,\beta,d,\sigma,B)(\log n)^{\frac{d}{\beta(d+6)(2\alpha+d)}} n^{-\frac{1}{3(d+6)(2\alpha+d)}\frac{\alpha-d}{\alpha-1}},$$

where C_0 and C are positive constants depending only on the quantities in the parentheses.

B.3 Analysis of the Kantorovich problem (12) for general d when $m \neq n$

The next result (proved in Appendix C.2) extends Theorem 12 to the unlinked setting based on samples $\mathcal{X}_n = \{X_1, \dots, X_n\}$ and $\mathcal{Y}_m = \{Y_1, \dots, Y_m\}$. The proof requires only one additional ingredient (Lemma 22) to the preceding proof.

Theorem 15 *Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mu$ and $\theta_1^*, \dots, \theta_m^* \overset{\text{i.i.d.}}{\sim} \nu$, where the support of ν is contained in the Euclidean ball of radius B centered at the origin, and let $f^* = \nabla\psi_{f^*}$ be the Brenier map (cf. Theorem 28) transporting μ to ν with f^* satisfying* **(A1)** *and* **(A2)**. *Consider the atomic measures $\nu_n^* = \frac{1}{n}\sum_{i=1}^n \delta_{f^*(X_i)}$, $\nu_m^* = \frac{1}{m}\sum_{i=1}^m \delta_{\theta_i^*}$, and $\widehat{\nu} = \sum_{j=1}^p \widehat{\alpha}_j \delta_{\widehat{\theta}_j}$. Let further $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, and consider the barycentric projection (8) based on an optimal coupling (7) between μ_n and $\widehat{\nu}$. For positive constants $C, c > 0$, we then have, with probability at least $1 - C(n^{-c} + m^{-c})$,*

$$\frac{1}{n}\sum_{i=1}^n \|\widehat{f}(X_i) - f^*(X_i)\|_2^2 \le \frac{2L}{\lambda}\left(\mathsf{W}_2^2(\widehat{\nu}, \nu_m^*) + 2\sqrt{\frac{\log n}{n}} \vee \Big(\frac{\log n}{n}\Big)^{2/d} + 2\sqrt{\frac{\log m}{m}} \vee \Big(\frac{\log m}{m}\Big)^{2/d}\right).$$

B.4 Proof of Theorem 6

Recall that $\nu_n^* = \frac{1}{n}\sum_{i=1}^n \delta_{\theta_i^*} = \frac{1}{n}\sum_{i=1}^n \delta_{f^*(X_i)}$, and note that $\widehat{\nu}$ denotes the mixing measure associated with the NPMLE (9). Theorem 12 above then yields that the barycentric projection estimator $\{\widehat{f}(X_i)\}_{i=1}^n$, obtained from an optimal coupling between ν_n^* and $\widehat{\nu}$ (see (8)), obeys the bound (25).

Next, note that under the stated condition on n, the squared 2-Wasserstein distance between ν_n^* and $\widehat{\nu}$ can be bounded as $\mathsf{W}_2^2(\nu_n^*, \widehat{\nu}) \lesssim_{d,\sigma,B} \frac{1}{\log n}$ according to Theorem 13 with the stated probability. This completes the proof of the theorem. ∎

B.5 Proof of Theorem 7

The proof is very similar to that of Theorem 6 in Appendix C.1 which proceeds via Theorem 12. We only point out the main differences. We argue as in the proof of Theorem 12 to obtain (29). Now observe that

$$\begin{aligned}
\int x^\top \widehat{f}(x)\, d\mu_n(x) &= \int x^\top \Big\{\sum_{j=1}^p \pi_j(x)\widehat{\theta}_j\Big\}\, d\mu_n(x) = \int x^\top \theta\, d\widehat{\gamma}(x,\theta) \\
&= \sup_{\gamma \in \Pi(\mu_n, \widehat{\nu})} \int x^\top \theta\, d\gamma(x,\theta) = \inf_{\psi}\left\{\int \psi\, d\mu_n + \int \psi^\star\, d\widehat{\nu}\right\} \\
&= \int \widehat{\psi}\, d\mu_n + \int \widehat{\psi^\star}\, d\widehat{\nu} \qquad (26)
\end{aligned}$$

where we have used Kantorovich duality (see e.g., Villani (2003, Chapter 2.1)). Similarly, as f^* pushes forward μ_n to ν_n^*, we have

$$\begin{aligned}\int x^\top f^*(x)\, d\mu_n(x) &= \inf_{\psi}\left\{\int \psi\, d\mu_n + \int \psi^\star\, d\nu_n^*\right\} \\ &= \int \psi_{f^*}\, d\mu_n + \int \psi^\star_{f^*}\, d\nu_n^* \\ &\leq \int \widehat{\psi}\, d\mu_n + \int \widehat{\psi}^\star\, d\nu_n^*. \end{aligned} \tag{27}$$

Combining (26) and (27), we obtain that

$$-\int x^\top(\widehat{f}(x) - f^*(x))\, d\mu_n(x) \leq \int \widehat{\psi}^\star\, d(\nu_n^* - \widehat{\nu}).$$

Now, using (29) in Appendix C.1 below, we obtain

$$\frac{1}{2L}\int \|\widehat{f}(x) - f^*(x)\|_2^2\, d\mu_n(x) \leq \int (\widehat{\psi}^\star - \psi^\star_{f^*})\, d(\nu_n^* - \widehat{\nu}). \tag{28}$$

Note that as $\nabla\psi^\star_{f^*}\#\nu_n^* = \mu_n$ and μ_n has bounded support, we see that $\psi^\star_{f^*}$ is Lipschitz.

Next we show that $\widehat{\psi}^\star$ can also be taken as a Lipschitz function. From Kantorovich duality (see e.g., Villani (2009, Theorem 5.10)) and (26) we know that

$$\widehat{\psi}(x) + \widehat{\psi}^\star(\theta) \geq x^\top\theta \qquad \text{for all } (x,\theta) \text{ with equality } \widehat{\gamma}\text{-a.s.}$$

Thus, for $\theta_0 \in \mathrm{supp}(\widehat{\nu}) \subset \mathbb{R}^d$, there exists $x_0 \in \mathrm{supp}(\mu_n) \subset \mathbb{R}^d$ such that

$$\widehat{\psi}(x_0) + \widehat{\psi}^\star(\theta_0) = x_0^\top\theta_0.$$

Then, for any $\theta \in \mathrm{supp}(\widehat{\nu}) \subset \mathbb{R}^d$, as $\widehat{\psi}(x_0) + \widehat{\psi}^\star(\theta) \geq x_0^\top\theta$, we have:

$$\widehat{\psi}^\star(\theta_0) - \widehat{\psi}^\star(\theta) \leq x_0^\top(\theta_0 - \theta) \leq \left(\max_{i=1,\ldots,n}\|X_i\|_2\right)\|\theta_0 - \theta\|_2 \leq B_{\mathcal{X}}\|\theta_0 - \theta\|_2.$$

Reversing the roles of θ_0 and θ, we can similarly show that

$$\widehat{\psi}^\star(\theta) - \widehat{\psi}^\star(\theta_0) \leq \left(\max_{i=1,\ldots,n}\|X_i\|_2\right)\|\theta_0 - \theta\|_2 \leq B_{\mathcal{X}}\|\theta_0 - \theta\|_2.$$

Combining, we get

$$|\widehat{\psi}^\star(\theta_0) - \widehat{\psi}^\star(\theta)| \leq \left(\max_{i=1,\ldots,n}\|X_i\|_2\right)\|\theta_0 - \theta\|_2 \leq B_{\mathcal{X}}\|\theta_0 - \theta\|_2.$$

As θ_0, θ are arbitrary points in $\mathrm{supp}(\widehat{\nu})$, we see that $\widehat{\psi}^\star$ can be taken as a Lipschitz function.

Thus, using Kantorovich-Rubinstein duality for the Wasserstein-1 distance (e.g., Villani (2009, Remark 6.5)) and (28), we have

$$\int \|\widehat{f}(x) - f^*(x)\|_2^2\, d\mu_n(x) \leq C(B_{\mathcal{X}})\, \mathsf{W}_1(\nu_n^*, \widehat{\nu}).$$

B.6 Proof of Theorem 8

The main modification relative to the proof of Theorem 6 is to consider both $\nu_n^* := \frac{1}{n}\sum_{i=1}^n \delta_{f^*(X_i)}$ and $\nu_m^* := \frac{1}{m}\sum_{i=1}^m \delta_{\theta_i^*}$, where $\{\theta_i^*\}_{i=1}^m \overset{\text{i.i.d.}}{\sim} \nu$. Theorem 15 bounds the mean squared denoising error in terms of the Wasserstein distance $\mathsf{W}_2^2(\nu_m^*, \widehat{\nu})$ and additional lower-order terms, where $\widehat{\nu}$ denotes the mixing measure associated with the NPMLE (9) based on the sample $\{Y_i\}_{i=1}^m$. We finally invoke Theorem 13 to bound $\mathsf{W}_2^2(\nu_m^*, \widehat{\nu})$, with ν_n^* and $\{\theta_i^*\}_{i=1}^n$ replaced by ν_m^* and $\{\theta_i^*\}_{i=1}^m$, respectively. ■

B.7 Proof of Proposition 9

Let us consider the permuted regression setup (1), and consider the two Kantorovich problems

$$\text{(i)} \min_{\gamma \in \Pi(\mu_n, \widehat{\nu})} \int (x-\theta)^2 \, d\gamma(x,\theta), \qquad \text{(ii)} \min_{\gamma \in \Pi(\nu_n^*, \widehat{\nu})} \int (\zeta-\theta)^2 \, d\gamma(\zeta,\theta).$$

Let $\widehat{\gamma}^1$ denote the so-called *Northwest-corner* solution of (i), cf. (Peyré and Cuturi, 2019, §3.4.2), and let $\widetilde{\gamma}^1 = (f^*, \mathsf{id})\#\widehat{\gamma}^1$ the push-forward (cf. Definition 1 in Appendix H) of $\widehat{\gamma}^1$ under the transformation that pushes forward its two marginals to $f^*\#\mu_n = \nu_n^*$ and and $\mathsf{id}\#\widehat{\nu} = \widehat{\nu}$, where id denotes the identity map. Since the $\{X_i\}_{i=1}^n$ and $\{\theta_i^*\}_{i=1}^n$ associated with μ_n and ν_n^* are related by the non-decreasing transformation f^*, $\widetilde{\gamma}^1$ is a minimizer of (ii) as follows, e.g., from Proposition 1 in Cuturi et al. (2019). Consequently, letting $\widetilde{\theta}_i = \mathbf{E}_{(\theta,\zeta)\sim\widetilde{\gamma}^1}[\theta|\zeta = \theta_i^*]$, $1 \le i \le n$, denote the barycentric projections, we have

$$\widetilde{\theta}_i = \frac{\int_\theta \theta \; d\widetilde{\gamma}^1(\theta_i^*, \theta)}{\int_\theta \; d\widetilde{\gamma}^1(\theta_i^*, \theta)} = \frac{\int_\theta \theta \; d\widehat{\gamma}^1(X_i, \theta)}{\int_\theta \; d\widehat{\gamma}^1(X_i, \theta)} = \widehat{f}(X_i), \quad 1 \le i \le n,$$

where the last equality is simply the definition of the $\{\widehat{f}(X_i)\}_{i=1}^n$ (cf. (8)). On the other hand, by Lemma 23, $\frac{1}{n}\sum_{i=1}^n (\widetilde{\theta}_i - \theta_i^*)^2 \le \mathsf{W}_2^2(\nu_n^*, \widehat{\nu})$, which concludes the proof.
The proof for the uncoupled regression setup is analogous to the proof of Theorem 8 (cf. Theorem 15 and its proof) and is hence omitted. ■

B.8 Proof of Theorem 11

The proof parallels the proof of Theorem 6, the only difference being a change in the bound on $\mathsf{W}_2^2(\nu_n^*, \widehat{\nu})$ according to Theorem 14 ■

Appendix C. Proof of Theorems 12, 13, 14 and 15

C.1 Proof of Theorem 12

Proof Consider an optimal coupling $\widehat{\gamma}$ between $\widehat{\nu}$ and μ_n minimizing (7), and let $\widehat{\gamma}_{ij}$ denote the resulting probability mass that is assigned to X_i and $\widehat{\theta}_j$, $1 \le i \le n$, $1 \le j \le p$. Define further $\pi_j(X_i) = \widehat{\Gamma}_{ij} n$, $1 \le i \le n$, $1 \le j \le p$. Accordingly, we have $\widehat{\alpha}_j = \frac{1}{n}\sum_{i=1}^n \pi_j(X_i) =$

$\int \pi_j(x)\, d\mu_n(x)$, $1 \le j \le p$. Recall that $\psi^\star_{f^*}$ denotes the Legendre-Fenchel conjugate of ψ_{f^*}. We first bound $\int \psi^\star_{f^*}(\theta)\, d\widehat{\nu}(\theta) - \int \psi^\star_{f^*}(\theta)\, d\nu^*_n(\theta)$ as

$$\begin{aligned}
&\sum_{j=1}^p \psi^\star_{f^*}(\widehat{\theta}_j)\, \widehat{\alpha}_j - \int \psi^\star_{f^*}(\theta)\, d\nu^*_n(\theta) \\
&= \int \sum_{j=1}^p \pi_j(x) \psi^\star_{f^*}(\widehat{\theta}_j)\, d\mu_n(x) - \int \psi^\star_{f^*}(f^*(x))\, d\mu_n(x) \\
&\ge \int \psi^\star_{f^*}\left(\sum_{j=1}^p \pi_j(x)\widehat{\theta}_j\right)\, d\mu_n(x) - \int \psi^\star_{f^*}(f^*(x))\, d\mu_n(x) \\
&= \int \psi^\star_{f^*}(\widehat{f}(x))\, d\mu_n(x) - \int \psi^\star_{f^*}(f^*(x))\, d\mu_n(x) \\
&\ge \int \nabla\psi^\star_{f^*}(f^*(x))^\top (\widehat{f}(x) - f^*(x))\, d\mu_n(x) \; + \frac{1}{2L}\int \|\widehat{f}(x) - f^*(x)\|_2^2\, d\mu_n(x), \\
&= \int x^\top (\widehat{f}(x) - f^*(x))\, d\mu_n(x) + \frac{1}{2L}\int \|\widehat{f}(x) - f^*(x)\|_2^2\, d\mu_n(x) \qquad (29)
\end{aligned}$$

where the two inequalities follow from convexity and L-smoothness of ψ_{f^*} in virtue of **(A2)**, which implies $\frac{1}{L}$-strong convexity of its conjugate $\psi^\star_{f^*}$ (Kakade et al., 2009); in the same vein, the last equality uses that $\nabla\psi^\star_{f^*}$ is the inverse map of $f^* = \nabla\psi_{f^*}$.

Moreover, the squared 2-Wasserstein distance between $\widehat{\nu}$ and μ_n, i.e., $\mathsf{W}_2^2(\widehat{\nu}, \mu_n)$, can be expressed as

$$\begin{aligned}
\sum_{i=1}^n \sum_{j=1}^p \|\widehat{\theta}_j - X_i\|_2^2 \widehat{\Gamma}_{ij} &= \sum_{j=1}^p \widehat{\alpha}_j \|\widehat{\theta}_j\|_2^2 + \frac{1}{n}\sum_{i=1}^n \|X_i\|_2^2 - 2\sum_{i=1}^n\sum_{j=1}^p \langle \widehat{\theta}_j, X_i\rangle \widehat{\Gamma}_{ij} \\
&= \int \|\theta\|_2^2\, d\widehat{\nu}(\theta) + \int \|x\|_2^2\, d\mu_n(x) - \frac{2}{n}\sum_{i=1}^n \left\langle X_i, \sum_{j=1}^p n\widehat{\Gamma}_{ij}\widehat{\theta}_j \right\rangle \\
&= \int \|\theta\|_2^2\, d\widehat{\nu}(\theta) + \int \|x\|_2^2\, d\mu_n(x) - 2\int x^\top \widehat{f}(x)\, d\mu_n(x). \qquad (30)
\end{aligned}$$

Similarly,

$$\mathsf{W}_2^2(\nu^*_n, \mu_n) = \int \|\theta\|_2^2\, d\nu^*_n(\theta) + \int \|x\|_2^2\, d\mu_n(x) - 2\int x^\top f^*(x)\, d\mu_n(x) \qquad (31)$$

where we note that f^* is the optimal transport map from ν^*_n to μ_n (as $f^*\#\mu_n = \nu^*_n$ and f^* is the gradient of a convex function). Combining (29), (30), (31), we obtain that

$$\begin{aligned}
\int \|\widehat{f}(x) - f^*(x)\|_2^2\, d\mu_n(x) \le L\Big[&\mathsf{W}_2^2(\widehat{\nu}, \mu_n) - \mathsf{W}_2^2(\nu^*_n, \mu_n) + 2\int \psi^\star_{f^*}(\theta)\, d(\widehat{\nu} - \nu^*_n)(\theta) \\
&+ \int \|\theta\|_2^2\, d(\nu^*_n - \widehat{\nu})(\theta)\Big]. \qquad (32)
\end{aligned}$$

Let $\widehat{\eta}$ be an optimal coupling between ν^*_n and $\widehat{\nu}$, and let further $\eta = (\nabla\psi^\star_{f^*}, \mathsf{id})\#\widehat{\eta}$ be the push-forward (cf. Definition 25) of the coupling $\widehat{\eta}$ under the transformation that pushes

forward its two marginals to $\nabla\psi^\star_{f^*}\#\nu^*_n = \mu_n$ and $\mathsf{id}\#\widehat{\nu} = \widehat{\nu}$, where we have used that $\nabla\psi^\star_{f^*}(\theta^*_i) = X_i$, $1 \le i \le n$, by Brenier's theorem, with id denoting the identity map. Accordingly, by the definition of the 2-Wasserstein distance in terms of optimal couplings (cf. Appendix H), we obtain that

$$\mathsf{W}_2^2(\mu_n, \widehat{\nu}) \le \int \|x - \theta\|_2^2 \, d\eta(x, \theta) = \int \|\nabla\psi^\star_{f^*}(\zeta) - \theta\|_2^2 \, d\widehat{\eta}(\zeta, \theta).$$

Adding and subtracting ζ inside the norm on the right hand side and expanding the square, it follows that

$$\begin{aligned}\mathsf{W}_2^2(\mu_n, \widehat{\nu}) &\le \int \|\nabla\psi^\star_{f^*}(\zeta) - \zeta\|_2^2 \, d\nu^*_n(\zeta) + \int \|\theta - \zeta\|_2^2 \, d\widehat{\eta}(\zeta, \theta) + 2\int \langle \nabla\psi^\star_{f^*}(\zeta) - \zeta, \zeta - \theta\rangle \, d\widehat{\eta}(\zeta, \theta) \\ &= \mathsf{W}_2^2(\nu^*_n, \mu_n) + \mathsf{W}_2^2(\nu^*_n, \widehat{\nu}) + 2\int \langle \nabla\psi^\star_{f^*}(\zeta) - \zeta, \zeta - \theta\rangle \, d\widehat{\eta}(\zeta, \theta), \end{aligned} \tag{33}$$

where we have used that $\psi^\star_{f^*}$ is the optimal transport map pushing forward ν^*_n to μ_n, the definition of the 2-Wasserstein distance in terms of optimal transport and optimal couplings, and the definition of $\widehat{\eta}$ as optimal coupling between ν^*_n and $\widehat{\nu}$.

In order to bound the rightmost term in the preceding display, we invoke **(A1)** which implies (Kakade et al., 2009) that the function $\psi^\star_{f^*}$ is $(1/\lambda)$-smooth in the sense of (18). This yields

$$\begin{aligned} 2\int \langle \nabla\psi^\star_{f^*}(\zeta), \zeta - \theta\rangle \, d\widehat{\eta}(\zeta, \theta) &\le 2\int \left\{ \psi^\star_{f^*}(\zeta) - \psi^\star_{f^*}(\theta) + \frac{1}{2\lambda}\|\zeta - \theta\|_2^2 \right\} d\widehat{\eta}(\zeta, \theta) \\ &= 2\int \psi^\star_{f^*}(\zeta) \, d\nu^*_n(\zeta) - 2\int \psi^\star_{f^*}(\theta) \, d\widehat{\nu}(\theta) + \frac{1}{\lambda}\mathsf{W}_2^2(\nu^*_n, \widehat{\nu}), \end{aligned} \tag{34}$$

using the same argument as for the preceding display to obtain the rightmost term.

Finally, we note that

$$\begin{aligned} 2\int \langle -\zeta, \zeta - \theta\rangle \, d\widehat{\eta}(\zeta, \theta) &= \int \left\{ \|\theta\|_2^2 - \|\theta - \zeta\|_2^2 - \|\zeta\|_2^2 \right\} d\widehat{\eta}(\zeta, \theta) \\ &= \int \|\theta\|_2^2 \, d\widehat{\nu}(\theta) - \int \|\zeta\|_2^2 \, d\nu^*_n(\zeta) - \mathsf{W}_2^2(\nu^*_n, \widehat{\nu}). \end{aligned} \tag{35}$$

Combining (33), (34), and (35), we obtain that

$$\begin{aligned}\mathsf{W}_2^2(\mu_n, \widehat{\nu}) \le \mathsf{W}_2^2(\nu^*_n, \mu_n) + \frac{1}{\lambda}\mathsf{W}_2^2(\nu^*_n, \widehat{\nu}) + 2\int \psi^\star_{f^*}(\zeta) \, d\nu^*_n(\zeta) - 2\int \psi^\star_{f^*}(\theta) \, d\widehat{\nu}(\theta) \\ + \int \|\theta\|_2^2 \, d\widehat{\nu}(\theta) - \int \|\zeta\|_2^2 \, d\nu^*_n(\zeta).\end{aligned}$$

Substituting this bound back into (32), we observe that all but the term $\frac{L}{\lambda}\mathsf{W}_2^2(\nu^*_n, \widehat{\nu})$ cancel, yielding the assertion of the theorem. ∎

C.2 Proof of Theorem 15

Proof We first note that the argument in the previous proof continues to apply with $\nu_n^* = \frac{1}{n}\sum_{i=1}^n \delta_{f^*(X_i)}$, which yields

$$\frac{1}{n}\sum_{i=1}^n \|\widehat{f}(X_i) - f^*(X_i)\|_2^2 \leq \frac{L}{\lambda}\mathsf{W}_2^2(\widehat{\nu}, \nu_n^*)$$

We then use the triangle inequality

$$\mathsf{W}_2(\widehat{\nu}, \nu_n^*) \leq \mathsf{W}_2(\widehat{\nu}, \nu_m^*) + \mathsf{W}_2(\nu_m^*, \nu_n^*) \leq \mathsf{W}_2(\widehat{\nu}, \nu_m^*) + \mathsf{W}_2(\nu_n^*, \nu) + \mathsf{W}_2(\nu_m^*, \nu),$$

and accordingly

$$\mathsf{W}_2^2(\widehat{\nu}, \nu_n^*) \leq 2\mathsf{W}_2^2(\widehat{\nu}, \nu_m^*) + 4(\mathsf{W}_2^2(\nu_n^*, \nu) + \mathsf{W}_2^2(\nu_m^*, \nu)).$$

The proof of the result now follows by invoking Lemma 22 with the choices $t = \sqrt{\log n/n} \vee n^{-2/d}(\log n)^{2/d}$ and $t = \sqrt{\log m/m} \vee m^{-2/d}(\log m)^{2/d}$ to control the second and the third term of the above display, respectively, with the stated probability; for the second term, we use that $\{f^*(X_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \nu$ since f^* pushes forward μ to ν (cf. Definition 25 and Theorem 28). ■

C.3 Proof of Theorem 13

Proof The proof is along the lines of the proof of Theorem 2 in Nguyen (2013) combined with an additional truncation argument to address that the support $\widehat{\nu}$ is not assumed to be uniformly bounded.

For a Lebesgue density h on $\mathbb{R}^d$ and $q > 0$, let $\mathsf{M}_h^q := \int \|x\|_2^q \, h(x)\, dx$ denote the q-th moment associated with h.

Let $s > k$ be arbitrary and let $K : \mathbb{R}^d \to (0, \infty)$ be a symmetric PDF such that $\mathsf{M}_K^s := \int_{\mathbb{R}^d} \|x\|_2^s \, K(x)\, dx < \infty$ and such that its Fourier transform $\mathbf{F}[K]$ is continuous with support contained in $[-1, 1]^d$, and for $\delta > 0$, let $K_\delta(\cdot) := \frac{1}{\delta^d} K(\cdot/\delta)$. By the triangle inequality, we have

$$\mathsf{W}_k^k(\nu_n^*, \widehat{\nu}) \leq 2^{2(k-1)} \left\{ \mathsf{W}_k^k(\nu_n^*, \nu_n^* \star K_\delta) + \mathsf{W}_k^k(\widehat{\nu}, \widehat{\nu} \star K_\delta) + \mathsf{W}_k^k(\nu_n^* \star K_\delta, \widehat{\nu} \star K_\delta) \right\}. \tag{36}$$

The first two terms inside the curly brackets are of order $O(\delta^k)$. To see this, consider couplings defined by the pairs of random variables $(X, X + \epsilon)$ and $(\widehat{X}, \widehat{X} + \epsilon)$ with $X \sim \nu_n^*$, $\widehat{X} \sim \widehat{\nu}$, and ϵ (independent of X and $\widehat{X}$) distributed according to the PDF K_δ, and note that $\mathbf{E}[\|X - (X + \epsilon)\|_2^k] = \mathbf{E}[\|\widehat{X} - (\widehat{X} + \epsilon)\|_2^k] = O(\delta^k)$.

In the sequel, the third term $\mathsf{W}_k^k(\nu_n^* \star K_\delta, \widehat{\nu} \star K_\delta)$ will be controlled. By Lemma 20 in Appendix G, we have

$$\mathsf{W}_k^k(\nu_n^* \star K_\delta, \widehat{\nu} \star K_\delta) \leq 2^{k-1} \int_{\mathbb{R}^d} \|x\|_2^k \; d|\nu_n^* \star K_\delta - \widehat{\nu} \star K_\delta|(x). \tag{37}$$

Next, we aim to bound the right hand side of (37) by invoking Lemma 21 in Appendix G. For this purpose, we need to establish first that the s-th moment of $\nu_n^* \star K_\delta$ and $\widehat{\nu} \star K_\delta$ are finite. For $\nu_n^* \star K_\delta$, this follows from

$$\begin{aligned}\int_{\mathbb{R}^d} \|x\|_2^s \, d(\nu_n^* \star K_\delta)(x) &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \|x\|_2^s \, K_\delta(x-\theta)\, dx \, d\nu_n^*(\theta) \\ &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \|x+\theta\|_2^s \, K_\delta(x)\, dx \, d\nu_n^*(\theta) \\ &\leq 2^{s-1} \left(\delta^s \int_{\mathbb{R}^d} \|x\|_2^s \, K(x) \, dx + \int_{\mathbb{R}^d} \|\theta\|_2^s \, d\nu_n^*(\theta) \right) < \infty.\end{aligned}$$

Above, we have used that the s-th moment of K is finite by construction and that the support of ν_n^* is uniformly bounded.

Showing that the s-th moment of $\widehat{\nu} \star K_\delta$ is finite is more intricate since the support of $\widehat{\nu}$ cannot be assumed to be uniformly bounded a priori. A careful truncation argument that relies on tail bounds and the Hellinger rates of the NPMLE is presented in Appendix E. Specifically, consider the two events $\mathcal{H}$ and $\mathcal{M}$ given by

$$\mathcal{H} := \left\{ \mathsf{H}(\mathsf{f}_n, \widehat{\mathsf{f}}_n) \leq C(d,B) \, \frac{(\log n)^{(d+1)/2}}{\sqrt{n}} \right\}, \tag{38}$$

$$\mathcal{M} := \left\{ \int_{\mathbb{R}^d} \|x\|_2^s \, d\widehat{\nu}(x) \leq C'(d,s,\sigma,B) + C''(d,\sigma,s,B) \frac{(\log n)^{(s+d+1)/2}}{\sqrt{n}} \leq C'''(d,\sigma,s,B) \right\},$$

where the constants $C(\ldots)$, $C'(\ldots)$ etc. are specified in the lemmas referenced below. We bound the probability of the complementary event of $\mathcal{H} \cap \mathcal{M}$ as follows:

$$\mathbf{P}(\mathcal{H}^{\mathsf{c}} \cup \mathcal{M}^{\mathsf{c}}) \leq \mathbf{P}(\mathcal{H}^{\mathsf{c}}) + \mathbf{P}(\mathcal{M}^{\mathsf{c}}) \leq 2\,\mathbf{P}(\mathcal{H}^{\mathsf{c}}) + \mathbf{P}(\mathcal{M}^{\mathsf{c}}|\mathcal{H})\,\mathbf{P}(\mathcal{H}).$$

By Lemma 17, we have

$$\mathbf{P}(\mathcal{M}^{\mathsf{c}}|\mathcal{H}) \leq (1/n) \big/ \mathbf{P}(\mathcal{H}).$$

Substituting this into the previous display yields that

$$\mathbf{P}(\mathcal{M}) \geq \mathbf{P}(\mathcal{M} \cap \mathcal{H}) \geq 1 - 2\,\mathbf{P}(\mathcal{H}^{\mathsf{c}}) - 1/n \geq 1 - 5/n,$$

where the last inequality is obtained by using the definition of the event $\mathcal{H}$ and Lemma 16 with $P_n = \nu_n^*$ and the choice $t = 1$.

With these arguments in place, we apply Lemma 21 to the right hand side of (37), which yields

$$\begin{aligned}\mathsf{W}_k^k(\nu_n^* \star K_\delta, \widehat{\nu} \star K_\delta) &\leq C(s,k,d) \, (\mathsf{M}^s_{\nu_n^* \star K_\delta} + \mathsf{M}^s_{\widehat{\nu} \star K_\delta})^{\frac{(s-t)d+k}{s(d+2s)}} \, \|\nu_n^* \star K_\delta - \widehat{\nu} \star K_\delta\|_{L_2}^{\frac{2(s-k)}{d+2s}} \\ &\leq C'(s,k,d,\sigma,B,K) \, \|\nu_n^* \star K_\delta - \widehat{\nu} \star K_\delta\|_{L_2}^{\frac{2(s-k)}{d+2s}} \end{aligned} \tag{39}$$

where C and C' are positive quantities depending only on the quantities in parentheses, assuming for now that δ is uniformly bounded from above.

Consider the Fourier transforms $\mathbf{F}[K_\delta]$ and $\mathbf{F}[\varphi_\sigma]$ of K_δ and φ_σ, respectively, and let $g_\delta := \mathbf{F}^{-1}\big[\mathbf{F}[K_\delta]/\mathbf{F}[\varphi_\sigma]\big]$ be the inverse Fourier transform of $\widetilde{g}_\delta := \mathbf{F}[K_\delta]/\mathbf{F}[\varphi_\sigma]$, which is well-defined since $\mathbf{F}[K_\delta]$ has bounded support by construction and thus so has $\widetilde{g}_\delta$. Furthermore, by the convolution theorem we have $\mathbf{F}[K_\delta] = \widetilde{g}_\delta \cdot \mathbf{F}[\varphi_\sigma] = \mathbf{F}[g_\delta \star \varphi_\sigma]$ and in turn $K_\delta = g_\delta \star \varphi_\sigma$ (cf. Appendix I). It follows that $K_\delta \star \nu_n^* = g_\delta \star \mathsf{f}_n$ and $K_\delta \star \widehat{\nu} = g_\delta \star \widehat{\mathsf{f}}_n$. This yields the following with regard to the term in (39):

$$\begin{aligned}\|\nu_n^* \star K_\delta - \widehat{\nu} \star K_\delta\|_{L_2} &= \|g_\delta \star (\widehat{\mathsf{f}}_n - \mathsf{f}_n)\|_{L_2} \\ &\leq \|\widehat{\mathsf{f}}_n - \mathsf{f}_n\|_{L_1} \|g_\delta\|_{L_2} \\ &\leq 2\mathsf{H}(\widehat{\mathsf{f}}_n, \mathsf{f}_n)\|g_\delta\|_{L_2} \end{aligned} \tag{40}$$

by the distributivity of convolution, Young's inequality, and the fact that $\|\widehat{\mathsf{f}}_n - \mathsf{f}_n\|_{L_1} = 2\mathsf{TV}(\widehat{\mathsf{f}}_n, \mathsf{f}_n) \leq 2\mathsf{H}(\widehat{\mathsf{f}}_n, \mathsf{f}_n)$. It remains to upper bound $\|g_\delta\|_{L_2}$. The Plancherel theorem (cf. Appendix I) yields

$$\begin{aligned}\|g_\delta\|_{L_2}^2 = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \frac{\{\mathbf{F}[K_\delta](\omega)\}^2}{\{\mathbf{F}[\varphi_\sigma](\omega)\}^2}\, d\omega &= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \frac{\{\mathbf{F}[K](\omega\delta)\}^2}{\{\mathbf{F}[\varphi_\sigma](\omega)\}^2}\, d\omega \\ &\leq C(K) \frac{1}{(2\pi)^d} \int_{[-\delta^{-1},\, \delta^{-1}]^d} \frac{1}{\{\mathbf{F}[\varphi_\sigma](\omega)\}^2}\, d\omega.\end{aligned}$$

For the second equality, we have used the definition of the Fourier transformation as integral transform and have a made a change of variables. For the inequality, we have used that $\mathbf{F}[K]$ is supported on $[-1,1]^d$ and bounded by a positive constant $C(K)$. It is well known that

$$\mathbf{F}[\varphi_\sigma](\omega) = \exp\left(-\frac{\sigma^2}{2}\|\omega\|_2^2\right).$$

Combining this with the previous display yields

$$\begin{aligned}\|g_\delta\|_{L_2}^2 &\leq C(K,d) \int_{[-\delta^{-1},\delta^{-1}]^d} \exp\left(\sigma^2\|\omega\|_2^2\right)\, d\omega \\ &\leq C(K,d)\, (2/\delta)^d\, \exp\left(\sigma^2 d\delta^{-2}\right) \\ &\leq C(K,d)\, \exp\left(2\sigma^2 d\delta^{-2}\right). \end{aligned} \tag{41}$$

Combining (36), (39), (40) and (41) then yields

$$\begin{aligned}\mathsf{W}_k^k(\nu_n^*, \widehat{\nu}) &\leq C(s,k,d,\sigma,K,B) \left\{\delta^k + \mathsf{H}(\widehat{\mathsf{f}}_n, \mathsf{f}_n)^{\frac{2(s-k)}{d+2s}}\, \exp\left(\frac{2(s-k)}{d+2s}\sigma^2 d\delta^{-2}\right)\right\} \\ &\leq C'(s,k,d,\sigma,K,B) \left\{\left(\frac{d\sigma^2}{\log\left(\frac{1}{\mathsf{H}(\widehat{\mathsf{f}}_n,\mathsf{f}_n)}\right)}\right)^{k/2} + \mathsf{H}(\widehat{\mathsf{f}}_n, \mathsf{f}_n)^{\frac{(s-k)}{d+2s}}\right\} \end{aligned} \tag{42}$$

by choosing $\delta^{-2} = -\frac{1}{2d\sigma^2}\log \mathsf{H}(\widehat{\mathsf{f}}_n, \mathsf{f}_n)$. Conditional on the event event $\mathcal{H}$ in (38) and the stated condition on the sample size $n \geq C_0(d,B)$, it holds that $\mathsf{H}(\widehat{\mathsf{f}}_n, \mathsf{f}_n) < 1$, and thus the above choice of δ is valid in the sense that $\delta > 0$. Substituting the bound on $\mathsf{H}(\widehat{\mathsf{f}}_n, \mathsf{f}_n)$

under event $\mathcal{H}$ in (38) into (42), absorbing terms depending only on s, k, d, σ and B into a constant, and absorbing the second summand inside the curly brackets in (42) into the first summand at the expense of modified constants yields the assertion (the dependence on the function K can be absorbed into the dependence on d). ■

C.4 Proof of Theorem 14

The proof of Theorem 14 proceeds as the proof of Theorem 13. We here only work out the differences.

The events in display (38) are modified as follows.

$$\mathcal{H} := \left\{ \mathsf{H}(\mathsf{f}_n, \widehat{\mathsf{f}}_n) \leq C(d,\alpha,\beta,B)\,(\log n)^{(d/\beta+1)/2}\, n^{-\frac{1}{6}\frac{\alpha-d}{\alpha-1}} \right\}, \tag{43}$$

$$\mathcal{M} := \left\{ \int_{\mathbb{R}^d} \|x\|_2^s \, d\widehat{\nu}(x) \leq C'(d,\alpha,\beta,\sigma,s,B) + C''(d,\alpha,\beta,\sigma,s,B) \frac{(\log n)^{\frac{d+2s}{2\beta}+\frac{1}{2}}}{n^{\frac{1}{6}\frac{\alpha-d}{\alpha-1}}} \leq C'''(d,\alpha,\beta,\sigma,s,B) \right\}.$$

Note that according to Theorem 10 with $P_n = \nu_n^*$, for $n \geq C_0(d,\alpha,\beta)$, it holds that $\mathbf{P}(\mathcal{H}) \geq 1 - 3/n$. Accordingly, we have that $\mathbf{P}(\mathcal{M}) \geq 1 - 7/n$ by using Lemma 18 and following the reasoning given below display (38) in the proof of Theorem 13.

We then proceed as in the proof of Theorem 13 until (40) and subsequently modify the bound on $\|g_\delta\|_{L_2}^2$. Specifically, in place of (41), we obtain according to condition **(D2)** in §3 that

$$\|g_\delta\|_{L_2}^2 \leq C(K,d,\sigma)\delta^{-(2\alpha+d)},$$

and thus

$$\begin{aligned}
\mathsf{W}_k^k(\nu_n^*, \widehat{\nu}) &\leq C(s,k,\alpha,\beta,d,\sigma,K,B) \left\{ \delta^k + \mathsf{H}(\widehat{\mathsf{f}}_n, \mathsf{f}_n)^{\frac{2(s-k)}{d+2s}}\, \delta^{-(2\alpha+d)(s-k)/(d+2s)} \right\} \\
&\leq C(s,k,\alpha,\beta,d,\sigma,K,B)\, \mathsf{H}(\widehat{\mathsf{f}}_n, \mathsf{f}_n)^{\frac{2k(s-k)}{k(d+2s)+(2\alpha+d)(s-k)}} \\
&= C(s,k,\alpha,\beta,d,\sigma,K,B)\, \mathsf{H}(\widehat{\mathsf{f}}_n, \mathsf{f}_n)^{c(k,s,d,\alpha)},
\end{aligned}$$

where $0 < c(k,s,d,\alpha) := \frac{2k(s-k)}{k(d+2s)+(2\alpha+d)(s-k)} < 1$. Now choosing $s = k+1$ and bounding $\mathsf{H}(\widehat{\mathsf{f}}_n, \mathsf{f}_n)$ according to Theorem 10, we obtain $c(k,s,d,\alpha) = \frac{2k}{k(d+2k+2)+(2\alpha+d)}$ and thus the assertion of the theorem. ■

Appendix D. Rates of convergence of the NPMLE for Gaussian location mixtures

Rates of convergence of the NPMLE for Gaussian location mixtures in Hellinger distance for general dimension $d \geq 1$ were established in Saha and Guntuboyina (2020), generalizing earlier results in Zhang (2009) concerning the case $d = 1$.

Lemma 16 (Theorem 2.1 and Corollary 2.2 in Saha and Guntuboyina (2020)) *Let $\{(Y_i, \zeta_i)\}_{i=1}^n$ be independent pairs of random vectors such that the $\{\zeta_i\}_{i=1}^n$ are distributed according to a probability measure P_n supported in the Euclidean ball of radius R around the*

origin, and $Y_i|\zeta_i \sim \varphi_\sigma(\cdot - \zeta_i)$, $1 \le i \le n$, *with* $\varphi(z) := (2\pi)^{-d/2}\exp(-\|z\|_2^2)$. *Let* $\widehat{\mathsf{f}}_n$ *denote the Kiefer-Wolfowitz NPMLE* (9) *of* $\mathsf{f}_n = \varphi_\sigma \star P_n$ *given data* $\{Y_i\}_{i=1}^n$. *Then for all* $t \ge 1$

$$\mathbf{P}(\mathsf{H}(\mathsf{f}_n, \widehat{\mathsf{f}}_n) > r_n t) \le 2n^{-t^2}, \qquad r_n \asymp \frac{(\log n)^{(d+1)/2}}{\sqrt{n}},$$

where $\asymp$ *involves hidden constants depending (only) on* d *and* R.

Appendix E. Truncation argument

Lemma 17 *Consider the setup of Lemma 16 with* $P_n = \nu_n^*$, *and denote by* $\widehat{\nu}$ *the mixing measure associated with the NPMLE* $\widehat{\mathsf{f}}_n$. *Consider the event* $\mathcal{E} = \{\mathsf{H}(\mathsf{f}_n, \widehat{\mathsf{f}}_n) \le \overline{h}\}$ *for some* $\overline{h} > 0$. *Conditional on* $\mathcal{E}$, *for any* $s \ge 1$, *we have* $\int_{\mathbb{R}^d}\|x\|_2^s \, d\widehat{\nu}(x) \le C_1(d, s, \sigma, B) + C_2(d, s, \sigma, B)(\log n)^{s/2} \cdot \overline{h}$ *with probability at least* $1 - 1/\{n \cdot \mathbf{P}(\mathcal{E})\}$, *where* C_1 *and* C_2 *are positive constants depending only on the quantities in the parentheses.*

Proof We first note that in order to show that the s-th moment of $\widehat{\nu}$ is finite it suffices to show that the s-th moment of $\widehat{\nu} \star \varphi_\sigma$ is finite. In fact, consider random variables $\widehat{X}$, ϵ such that $\widehat{X} \sim \widehat{\nu}$ and $\epsilon \sim \varphi_\sigma$ where $\widehat{X}$ and ϵ are independent. We then have

$$\mathbf{E}[\|\widehat{X}\|_2^s] = \mathbf{E}[\|\widehat{X} - \epsilon + \epsilon\|_2^s] \le 2^{s-1}(\mathbf{E}[\|\widehat{X} + \epsilon\|_2^s] + \mathbf{E}[\|\epsilon\|_2^s]).$$

In order to show that the s-th moment of $\widehat{\nu} \star \varphi_\sigma$ is finite, we will use Lemma 16 regarding the Hellinger rates of convergence between $\nu_n^* \star \varphi_\sigma$ and $\widehat{\nu} \star \varphi_\sigma$ and the fact that the support of ν_n^* is contained in an Euclidean ball of radius B by assumption.

First note that according to established properties of the NPMLE (e.g., Lindsay (1983); Koenker and Mizera (2014)), $\widehat{\nu}$ is an atomic measure, i.e., it can be written as $\widehat{\nu} = \sum_{j=1}^p \widehat{\alpha}_j \delta_{\widehat{\theta}_j}$ for non-negative coefficients $\{\widehat{\alpha}_j\}_{j=1}^p \subset \mathbb{R}_+$ summing to one and atoms $\{\widehat{\theta}_j\}_{j=1}^p \subset \mathbb{R}^d$. Let

$$\widehat{B} = \max_{1 \le j \le p}\|\widehat{\theta}_j\|_2, \quad \rho = \widehat{B} + \sigma R, \quad \mathcal{R} = \mathbb{B}_2^d(\rho), \quad \mathcal{R}_0 = \bigcup_{j=1}^p (\mathbb{B}_2^d(\sigma R) + \widehat{\theta}_j) \tag{44}$$

for $R > 0$ to be chosen later. Observe that $\mathcal{R}_0 \subset \mathcal{R}$ and hence $\mathcal{R}^{\mathsf{c}} \subset \mathcal{R}_0^{\mathsf{c}}$. We have

$$\begin{aligned}
\int_{\mathbb{R}^d}\|x\|_2^s \, d(\varphi_\sigma \star \widehat{\nu})(x) &= \int_{\mathcal{R}}\|x\|_2^s \, d(\varphi_\sigma \star \widehat{\nu})(x) + \int_{\mathcal{R}^{\mathsf{c}}}\|x\|_2^s \, d(\varphi_\sigma \star \widehat{\nu})(x) \\
&\le \int_{\mathcal{R}}\|x\|_2^s \, d(\varphi_\sigma \star \nu_n^*)(x) + \int_{\mathcal{R}}\|x\|_2^s \, d|\varphi_\sigma \star \widehat{\nu} - \varphi_\sigma \star \nu_n^*|(x) \\
&\quad + \int_{\mathcal{R}^{\mathsf{c}}}\|x\|_2^s \, d(\varphi_\sigma \star \widehat{\nu})(x) \\
&\le C_1(d, s, B, \sigma) + 2\rho^s \underbrace{\mathsf{H}(\varphi_\sigma \star \widehat{\nu}, \varphi_\sigma \star \nu_n^*)}_{= \mathsf{H}(\widehat{\mathsf{f}}_n, \mathsf{f}_n) \le \overline{h} \text{ on } \mathcal{E}} + \int_{\mathcal{R}^{\mathsf{c}}}\|x\|_2^s \, d(\varphi_\sigma \star \widehat{\nu})(x)
\end{aligned} \tag{45}$$

for some constant $C_1 > 0$ depending only on the quantities given in parentheses. In order to obtain the bound on the middle term, we use that the integral is over $\mathbb{B}_2^d(\rho)$ and that

the total variation distance can be bounded by twice the Hellinger distance. We now turn our attention to the third term in (45). We have

$$\begin{aligned}
\int_{\mathcal{R}^{\mathsf{c}}} \|x\|_2^s (\varphi_\sigma \star \widehat{\nu})(x)\, dx &\leq \int_{\mathcal{R}_0^{\mathsf{c}}} \|x\|_2^s (\varphi_\sigma \star \widehat{\nu})(x)\, dx \\
&= \sum_{j=1}^p \widehat{\alpha}_j \int_{\sigma^{-1}(\mathcal{R}_0^{\mathsf{c}} - \widehat{\theta}_j)} \|\sigma z + \widehat{\theta}_j\|_2^s\, \varphi(z)\, dz \\
&\leq \sum_{j=1}^p \widehat{\alpha}_j 2^{s-1} \left\{ \sigma^s \int_{\mathbb{R}^d} \|z\|_2^s\, \varphi(z)\, dz + \|\widehat{\theta}_j\|_2^s \int_{\sigma^{-1}(\mathcal{R}_0^{\mathsf{c}} - \widehat{\theta}_j)} \varphi(z)\, dz \right\} \\
&\leq 2^{s-1} \left\{ \sigma^s \int_{\mathbb{R}^d} \|z\|_2^s\, \varphi(z)\, dz + \max_{1\leq j\leq p} \|\widehat{\theta}_j\|_2^s \int_{\mathbb{R}^d \setminus \mathbb{B}_2^d(R)} \varphi(z)\, dz \right\} \\
&\leq C_2(d,s,\sigma) + \widehat{B}^s\, \mathbf{P}(\|Z\|_2 \geq R), \quad Z \sim N(0, I_d) \\
&\leq C_2(d,s,\sigma) + (\widehat{B}/n)^s
\end{aligned} \tag{46}$$

by choosing $R = \sqrt{2s \log n}$, as follows from standard concentration of measure results. In the third inequality from the bottom, we have used that for any j

$$\sigma^{-1}(\mathcal{R}_0^{\mathsf{c}} - \widehat{\theta}_j) = \sigma^{-1} \left(\bigcap_{j=1}^p \{\mathbb{B}_2^d(\sigma R) + \widehat{\theta}_k\}^{\mathsf{c}} - \widehat{\theta}_j \right) \subseteq \sigma^{-1} \left[\{\mathbb{B}_2^d(\sigma R) + \widehat{\theta}_j\}^{\mathsf{c}} - \widehat{\theta}_j \right] = \mathbb{R}^d \setminus \mathbb{B}_2^d(R).$$

In order to wrap up this proof, it remains to control $\widehat{B}$ (with high probability). With the same concentration result as used before in combination with the union bound, one shows that

$$\begin{aligned}
\mathbf{P}(\widehat{B} \geq B + \sigma(\sqrt{d} + 2\sqrt{\log n})) &\leq \mathbf{P}\left(\max_{1\leq i\leq n} \|Y_i\|_2 \geq B + \sigma(\sqrt{d} + 2\sqrt{\log n}) \right) \\
&\leq \mathbf{P}\left(\max_{1\leq i\leq n} \|\theta_i^*\|_2 + \max_{1\leq i\leq n} \|\epsilon_i\|_2 \geq B + \sigma(\sqrt{d} + 2\sqrt{\log n} \right) \\
&= \mathbf{P}\left(\max_{1\leq i\leq n} \|\epsilon_i\|_2 \geq \sigma(\sqrt{d} + 2\sqrt{\log n} \right) \leq 1/n.
\end{aligned} \tag{47}$$

Let $\mathcal{A}$ denote the event inside $\mathbf{P}(\ldots)$ in the last line, and observe that $\mathbf{P}(\mathcal{A}|\mathcal{E}) \leq \mathbf{P}(\mathcal{A})/\mathbf{P}(\mathcal{E})$. Combining this with (45), (46), and the above choice of R then yields the assertion.

Note that in the first inequality (47), we have used that $\widehat{B} = \max_{1\leq j\leq p} \|\widehat{\theta}_j\|_2 \leq \max_{1\leq i\leq n} \|Y_i\|_2 =: Q$ since $\varphi(z)$ is radial and decreasing in $\|z\|_2$. Accordingly, we have

$$\sum_{i=1}^n -\log \left(\sum_{j=1}^p \widehat{\alpha}_j \varphi_\sigma \left(Y_i - P_{\mathbb{B}_2^d(Q)}(\widehat{\theta}_j) \right) \right) \leq \sum_{i=1}^n -\log \left(\sum_{j=1}^p \widehat{\alpha}_j \varphi_\sigma \left(Y_i - \widehat{\theta}_j \right) \right),$$

where P denotes the Euclidean projection, which is a non-expansive operator for convex sets. The latter property implies that for $1 \leq i \leq n$ and $1 \leq j \leq p$, it holds that

$$\|Y_i - P_{\mathbb{B}_2^d(Q)}(\widehat{\theta}_j)\|_2 = \|P_{\mathbb{B}_2^d(Q)}(Y_i) - P_{\mathbb{B}_2^d(Q)}(\widehat{\theta}_j)\|_2 \leq \|Y_i - \widehat{\theta}_j\|_2.$$

■

Lemma 18 *Consider the setup of Theorem 10 with $P_n = \nu_n^*$, and denote by $\widehat{\nu}$ the mixing measure associated with the NPMLE $\widehat{\mathsf{f}}_n$. Consider the event $\mathcal{E} = \{\mathsf{H}(\mathsf{f}_n, \widehat{\mathsf{f}}_n) \leq \overline{h}\}$ for some $\overline{h} > 0$. Conditional on $\mathcal{E}$, for any $s \geq 1$, we have $\int_{\mathbb{R}^d} \|x\|_2^s \, d\widehat{\nu}(x) \leq C_1(d, \alpha, \beta, \sigma, s, B) + +C_2(d, \alpha, \beta, \sigma, s, B)(\log n)^{s/\beta} \cdot \overline{h}$ with probability at least $1 - 1/\{n \cdot \mathbf{P}(\mathcal{E})\}$, where C_1 and C_2 are positive constants depending only on the quantities in the parentheses.*

Proof The proof proceeds as the proof of the previous lemma. We only sketch the main differences, which result from different tail bounds that are here obtained according to **(D3)** in §3. Specifically, the radius $\widehat{B}$ of the ball containing the support of $\widehat{\nu}$ can be chosen such that $\widehat{B} \lesssim_{d,\alpha,B} (\log n)^{1/\beta}$. Similarly, R as appearing in (46) can be chosen as $R \lesssim_{s,\alpha,\beta} (\log n)^{1/\beta}$ so that failure probabilities for the events of interest remain unchanged. The final result is then obtained by combining the pieces corresponding to those in display (45). ■

Appendix F. Proof of Theorem 10

Overview. The proof is a based on a combination of techniques developed in Zhang (2009), Saha and Guntuboyina (2020), Ignatiadis and Sen (2023), and Gao and van der Vaart (2016). The original analysis of the NPMLE for Gaussian location mixtures for $d = 1$ was developed in Zhang (2009) and adapted/simplified in Ignatiadis and Sen (2023) in their analysis of the NPMLE for scale mixtures of χ^2-distributions. Saha and Guntuboyina (2020) extend the approach in Zhang (2009) to general dimension $d \geq 1$. As an important ingredient, we adopt methods from Gao and van der Vaart (2016) to obtain bounds on the $\|\cdot\|_\infty$-covering numbers for the class of location mixtures $\Phi_R := \{\varphi \star P : \ P \in \mathscr{P}_R\}$, where $\mathscr{P}_R$ denotes the class of probability measures supported on a Euclidean ball $\mathbb{B}_2^d(R)$ with radius $R > 0$ around the origin.

Proof

Preliminaries.
1) Without loss of generality, we may assume that $\sigma = 1$. As noted in Zhang (2009), with σ assumed known, it suffices to operate in terms of rescaled data $\{Y_i/\sigma\}_{i=1}^n$ and then invoke the invariance of the Hellinger distance with respect to scale transformations.
2) Observe that $\widehat{\mathsf{f}}_n \in \Phi_{R_n}$ with probability $1 - 1/n$ provided R_n is chosen as $R_n \gtrsim_{d,\alpha,R} \log^{1/\beta}(n)$ so that the event $\{\max_{1 \leq i \leq n} \|Y_i\|_2 \leq R_n\}$ occurs with the stated probability (as a consequence of **(D3)**). To avoid complications that would arise from conditioning on this event, we instead analyze the *constrained* NPMLE $\widetilde{f}_n$ whose mixing measure is required to be contained in $\mathbb{B}_2^d(R_n)$; note that $\{\widehat{\mathsf{f}}_n = \widetilde{\mathsf{f}}_n\}$ with high probability, therefore the triangle inequality yields that

$$\mathbf{P}(\mathsf{H}(\widehat{\mathsf{f}}_n, \mathsf{f}_n) > \delta) \leq \mathbf{P}(\mathsf{H}(\widetilde{\mathsf{f}}_n, \mathsf{f}_n) > \delta/2) + \mathbf{P}(\mathsf{H}(\widehat{\mathsf{f}}_n, \widetilde{\mathsf{f}}_n) > \delta/2), \quad \delta > 0.$$

The main effort will go into bounding the first term on the right hand side. Since $\{\widehat{\mathsf{f}}_n = \widetilde{\mathsf{f}}_n\} \subset \{\mathsf{H}(\widehat{\mathsf{f}}_n, \widetilde{\mathsf{f}}_n) < \delta/2\}$ and $\{\widehat{\mathsf{f}}_n = \widetilde{\mathsf{f}}_n\}$ is a high probability event, the second term is controlled.

Main steps of the proof.
(I) For $g, h \in \Phi_{R_n}$, consider the likelihood ratio

$$L_n(g,h) = \prod_{i=1}^n \frac{g(Y_i)}{h(Y_i)}.$$

Let $\overline{\Phi}_{R_n} := \{\mathsf{f} \in \Phi_{R_n} : \mathsf{H}(\mathsf{f}, \mathsf{f}_n) > t \cdot r_n\}$. We will bound the probability

$$\mathbf{P}\left(\sup_{\mathsf{f} \in \overline{\Phi}_{R_n}} L_n(\mathsf{f}, \mathsf{f}_n) > \exp\left(-2t^2 n r_n^2/15\right) \right). \tag{48}$$

By showing that this probability is small, it is implied that any approximate MLE (as defined in terms of the lower bound on the likelihood ratio in the display) and thus $\widetilde{\mathsf{f}}_n$ in particular must be contained in $\{\mathsf{f} \in \Phi_{R_n} : \mathsf{H}(\mathsf{f}, \mathsf{f}_n) \le t \cdot r_n\}$ with the stated probability.

(II) In order to establish (48), a crucial ingredient is a covering of Φ_{R_n} with respect to $\|\cdot\|_\infty$, which is developed separately in Lemma 19. Note that an η-covering of Φ_{R_n} can be transformed into a 2η-covering (of at most the same cardinality) of the subset $\overline{\Phi}_{R_n}$ using standard arguments (e.g. Vershynin, 2018, §4.2). Denote the resulting covering of $\overline{\Phi}_{R_n}$ by $\{\mathsf{f}_0^j\}_{j=1}^N$, and for $M > R_n$ to be chosen later, define the function $\eta : \mathbb{R}^d \to \mathbb{R}_+$ by

$$\eta(y) = \eta \mathbb{I}(\|y\|_2 \le M) + \eta \left(\frac{M}{\|y\|_2}\right)^{d+1} \mathbb{I}(\|y\|_2 > M), \tag{49}$$

where the η's on the right hand side of (49) represent the number associated with the covering. Observe that for any $\mathsf{f} \in \overline{\Phi}_{R_n}$ there exists $j \in \{1, \dots, N\}$ such that the following holds:

$$\mathsf{f}(y) \le \begin{cases} \mathsf{f}_0^j(y) + 2\eta = \mathsf{f}_0^j(y) + 2\eta(y), & \|y\|_2 \le M, \\ \varphi(0), & \|y\|_2 > M. \end{cases}$$

Consequently, for any $\mathsf{f} \in \overline{\Phi}_{R_n}$ we can upper bound $L_n(\mathsf{f}, \mathsf{f}_n)$ as follows:

$$\begin{aligned}
L_n(\mathsf{f}, \mathsf{f}_n) &= \prod_{i=1}^n \left\{ \frac{\mathsf{f}(Y_i)}{\mathsf{f}_n(Y_i)} \right\} \\
&= \prod_{i:\, \|Y_i\|_2 \le M} \left\{ \frac{\mathsf{f}(Y_i)}{\mathsf{f}_n(Y_i)} \right\} \times \prod_{i:\, \|Y_i\|_2 > M} \left\{ \frac{\mathsf{f}(Y_i)}{\mathsf{f}_n(Y_i)} \right\} \\
&\le \prod_{i=1}^n \left\{ \frac{\mathsf{f}_0^j(Y_i) + 2\eta(Y_i)}{\mathsf{f}_n(Y_i)} \right\} \times \prod_{i:\, \|Y_i\|_2 > M} \left\{ \frac{\mathsf{f}(Y_i)}{\mathsf{f}_0^j(Y_i) + 2\eta(Y_i)} \right\} \\
&\le \prod_{i=1}^n \left\{ \frac{\mathsf{f}_0^j(Y_i) + 2\eta(Y_i)}{\mathsf{f}_n(Y_i)} \right\} \times \prod_{i:\, \|Y_i\|_2 > M} \frac{\varphi(0)}{2\eta(Y_i)} \\
&\le \sup_{1 \le j \le N} L_n(\mathsf{f}_0^j + 2\eta, \mathsf{f}_n) \times \prod_{i:\, \|Y_i\|_2 > M} \frac{\varphi(0)}{2\eta(Y_i)} = T_1 \times T_2.
\end{aligned} \tag{50}$$

In the sequel, the terms T_1 and T_2 are controlled separately according to the argument

$$\mathbf{P}(T_1 \cdot T_2 > \delta) \leq \mathbf{P}(T_1 > \delta_1) + \mathbf{P}(T_2 > \delta_2)$$

for any choice of $\delta_1 > 0, \delta_2 > 0$ such that $\delta_1 \cdot \delta_2 = \delta$. Specifically, we will choose

$$\delta = \exp\left(-2t^2 n r_n^2/15\right), \quad \delta_1 = \exp(-4t^2 n r_n^2/5), \quad \delta_2 = \exp(2t^2 n r_n^2/3).$$

Regarding the first term T_1, denote $L_{n,j,i} := (\mathsf{f}_0^j(Y_i) + 2\eta(Y_i))/\mathsf{f}_n(Y_i)$. We obtain that

$$\begin{aligned}
\mathbf{P}\left(\prod_i L_{n,j,i} \geq \delta_1\right) &= \mathbf{P}\left(\sqrt{\prod_i L_{n,j,i}} \geq \delta_1^{1/2}\right) \\
&= \mathbf{P}\left(\prod_i \sqrt{L_{n,j,i}} \geq \delta_1^{1/2}\right) \\
&\leq \delta_1^{-1/2}\, \mathbf{E}\left[\prod_i \sqrt{L_{n,j,i}}\right] \\
&= \delta_1^{-1/2} \prod_i \mathbf{E}[\sqrt{L_{n,j,i}}] \\
&\leq \exp\left(\frac{2t^2 n r_n^2}{5} + \sum_{i=1}^n \mathbf{E}[\sqrt{L_{n,j,i}} - 1]\right), \qquad (51)
\end{aligned}$$

where we have used Markov's inequality and the elementary inequality $z \leq \exp(z-1)$. Using that the terms inside the summation are i.i.d., we obtain

$$\sum_{i=1}^n \mathbf{E}[\sqrt{L_{n,j,i}} - 1] = n\, \mathbf{E}[\sqrt{L_{n,j,1}} - 1] = n\, \mathbf{E}[\sqrt{L_{n,j}} - 1],$$

say, after dropping the third subscript in $L_{\ldots}$. We now have

$$\begin{aligned}
\mathbf{E}[\sqrt{L_{n,j}} - 1] &= \int \sqrt{\frac{\mathsf{f}_0^j + 2\eta}{\mathsf{f}_n}} \mathsf{f}_n - 1 \\
&= \int \sqrt{(\mathsf{f}_0^j + 2\eta) \cdot \mathsf{f}_n} - 1 \\
&\leq \left(\int \sqrt{\mathsf{f}_0^j \mathsf{f}_n} - 1\right) + \int \sqrt{2\eta \mathsf{f}_n}.
\end{aligned}$$

By the definition of the Hellinger distance, we have for any pair of two densities g and h:

$$\mathsf{H}^2(g,h) = \int (\sqrt{g} - \sqrt{h})^2 = 2\left(1 - \int \sqrt{g \cdot h}\right).$$

Inserting this relationship into the preceding display, we thus obtain the bound

$$\begin{aligned}\mathbf{E}[\sqrt{L_{n,j}} - 1] &\leq -\frac{1}{2}\mathsf{H}^2(\mathsf{f}_n, \mathsf{f}_0^j) + \int \sqrt{2\eta \mathsf{f}_n} \\ &\leq -\frac{1}{2}\mathsf{H}^2(\mathsf{f}_n, \mathsf{f}_0^j) + \left(2\int \eta\right)^{1/2} \cdot \underbrace{\left(\int \mathsf{f}_n\right)^{1/2}}_{=1} \\ &= -\frac{1}{2}\mathsf{H}^2(\mathsf{f}_n, \mathsf{f}_0^j) + C_d \eta M^d \\ &\leq -\frac{1}{2}t^2 r_n^2 + \sqrt{C_d \cdot \eta \cdot M^d}, \end{aligned} \tag{52}$$

where the first inequality is Cauchy-Schwarz and the last inequality results from $\{\mathsf{H}(\mathsf{f}_n, \mathsf{f}_0^j) > t \cdot r_n\}$ for all j. The third line from the top is obtained by evaluating the following integral (cf. the definition of the function η in (49)) using polar coordinates:

$$\begin{aligned}\int_{\mathbb{R}^d} \eta(y)\, dy &= \eta M^d \cdot \mathrm{vol}_d(\mathbb{B}_2^d) + \eta M^{d+1} \int_{\mathbb{R}^d \setminus M\mathbb{B}_2^d} \|y\|_2^{-(d+1)}\, dy \\ &= \eta M^d \cdot \mathrm{vol}_d(\mathbb{B}_2^d) + \eta M^d \mathrm{vol}_{d-1}(\mathbb{S}^d) = C_d \eta M^d, \end{aligned}$$

with $\mathbb{S}^d$ denoting the d-dimensional unit sphere. Combining (51), (52) and the union bound, we obtain

$$\mathbf{P}\left(\sup_{1\leq j\leq N} L_n(\mathsf{f}_0^j + 2\eta, \mathsf{f}_n) > \delta_1\right) \leq \exp\left(-\frac{nt^2 r_n^2}{10} + n\sqrt{C_d \eta M^d} + \log(N)\right) \tag{53}$$

Now choose $M \asymp_d R_n \asymp_{B,d} \log^{1/\beta}(n)$, $\eta = n^{-\frac{2}{3}\frac{\alpha-d}{d}}$, and $r_n = \log(n)^{(d/\beta+1)/2}\, n^{-\frac{1}{6}\frac{\alpha-d}{\alpha-1}}$. Invoking Lemma 19 with $R = R_n$ and $\epsilon \asymp \eta$, these choices yield the following for the exponent in (53):

$$\begin{aligned}&\log N + n\sqrt{C_d \eta M^d} - \frac{nt^2 r_n^2}{10} \\ &\lesssim_{B,d} (1/\eta)^{\frac{d}{\alpha-d}} \log(1/\eta) \log^{d/\beta}(n) + n\eta^{1/2} \log^{d/(2\beta)}(n) - \frac{\log^{d/\beta+1}(n)\, n^{1-\frac{1}{3}\frac{\alpha-d}{\alpha-1}} t^2}{10} \\ &\lesssim n^{2/3} \log^{d/\beta+1}(n) + n^{1-\frac{1}{3}\frac{\alpha-d}{d}} \log^{d/(2\beta)}(n) - \frac{\log^{d/\beta+1}(n)\, n^{1-\frac{1}{3}\frac{\alpha-d}{\alpha-1}} t^2}{10} \\ &\leq \exp\left(-\frac{nt^2 r_n^2}{20}\right) \end{aligned}$$

for all $t \geq t_*$ large enough, noting that $1 - \frac{1}{3}\frac{\alpha-d}{\alpha-1} \geq \max\{1 - \frac{1}{3}\frac{\alpha-d}{d}, 2/3\}$ since $\alpha \geq d+1$.

The above concludes the analysis associated with the term T_1 in (50). We now turn to term T_2. We have

$$\begin{aligned}
\mathbf{P}\left(\prod_{i:\,\|Y_i\|_2>M}\frac{\varphi(0)}{2\eta(Y_i)}>\delta_2\right) &\leq \delta_2^{-1}\prod_{i=1}^n \mathbf{E}\left[\left\{\frac{\varphi(0)}{2\eta(y_i)}\right\}^{\mathbb{I}(\|Y_i\|_2>M)}\right] \\
&\leq \delta_2^{-1}\prod_{i=1}^n \mathbf{E}\left[\left(1+\mathbb{I}(\|Y_i\|_2>M)\frac{\varphi(0)}{2\eta(Y_i)}\right)\right] \\
&= \delta_2^{-1}\prod_{i=1}^n \mathbf{E}\left[\left(1+\mathbb{I}(\|Y_i\|_2>M)\frac{\varphi(0)}{2}\frac{\|Y_i\|_2^{d+1}}{M^{d+1}\eta}\right)\right] \\
&\leq \delta_2^{-1}\exp\left[\frac{\varphi(0)}{2M^{d+1}\eta}\sum_{i=1}^n \mathbf{E}[\mathbb{I}(\|Y_i\|_2>M)\|Y_i\|_2^{d+1}]\right], \qquad (54)
\end{aligned}$$

where the last inequality results from the elementary inequality $z+1\leq\exp(z)$ applied to $z=\mathbf{E}\left[\mathbb{I}(\|Y_i\|_2>M)\frac{\varphi(0)}{2}\frac{\|Y_i\|_2^{d+1}}{M^{d+1}\eta}\right]$, $1\leq i\leq n$. In the sequel, we bound the expectation $\mathbf{E}[\mathbb{I}(\|Y\|_2>M)\|Y\|_2^{d+1}]$, where Y has the same distribution as the $\{Y_i\}_{i=1}^n$. We have

$$\begin{aligned}
\mathbf{E}[\mathbb{I}(\|Y\|_2>M)\|Y\|_2^{d+1}] &= \int_0^\infty \mathbf{P}(\mathbb{I}(\|Y\|_2>M)\|Y\|_2^{d+1}>t)\,dt \\
&= (d+1)\int_0^\infty u^d\,\mathbf{P}(\mathbb{I}(\|Y\|_2>M)\|Y\|_2\geq u)\,du
\end{aligned}$$

by a change of variables. We now split the integral as follows:

$$\begin{aligned}
&\int_0^\infty u^d\,\mathbf{P}(\mathbb{I}(\|Y\|_2>M)\|Y\|_2\geq u)\,du \\
&= \int_0^M u^d\,\mathbf{P}(\|Y\|_2>M)\,du + \int_M^\infty u^d\,\mathbf{P}(\|Y\|_2\geq u)\,du \\
&\leq M^{d+1}\,\mathbf{P}(\|Y\|_2>M) + \int_M^\infty u^d\exp(-(u-R_n)^\beta)\,du \\
&\lesssim_{\beta,d} M^{d+1}\exp(-c(M-R_n)^\beta) + M^{d+(1-\beta)}\exp(-c(M-R_n)^\beta),
\end{aligned}$$

where the first term on the right hand side follows from the assumptions in φ given an appropriate choice of $M\asymp_{d,\alpha} R_n \asymp_{B,d,\alpha}(\log n)^{1/\beta}$ (as above), and the second term on the right hand side follows from Lemma 24. Inserting the above into (54), we obtain

$$\begin{aligned}
&\exp\left(-\frac{2t^2nr_n^2}{3}+(1/\eta)C_{\beta,d}\,n\exp(-c(M-R_n)^\beta)\right) \\
&= \exp\left(-\frac{2t^2nr_n^2}{3}+C_{\beta,d}n^{1+\frac{2}{3}\frac{\alpha-d}{d}}\exp(-c(M-R_n)^\beta)\right) \\
&\leq \exp\left(-\frac{t^2nr_n^2}{3}\right)
\end{aligned}$$

for all $t\geq t_*$ large enough by choosing $M\asymp_{B,\alpha,\beta,d}(\log n)^{1/\beta}$ large enough. ■

F.1 Covering Numbers

In this subsection, we obtain bounds on the $\|\cdot\|_\infty$-covering numbers of the class of location mixtures $\Phi_R := \{\varphi \star P : P \in \mathscr{P}_R\}$ with φ satisfying **(D1)**–**(D3)** in §3, where we recall that $\mathscr{P}_R$ denotes the class of probability measures supported on $\mathbb{B}_2^d(R)$ for $R > 0$. Our bounds are obtained by transferring Fourier-based techniques (cf. Appendix I) in Gao and van der Vaart (2016) to general dimension.

Lemma 19 *Fix $\epsilon \in (0,1)$ and suppose that $R \gtrsim 1$. There exists $\{\mathsf{f}_0^1, \ldots, \mathsf{f}_0^N\} \subset \Phi_R$ such that $\min_{1\le j\le N} \|\mathsf{f} - \mathsf{f}_0^j\|_\infty \lesssim_{d,\alpha} \epsilon$ and $\log N \lesssim_d R^d \epsilon^{-d\cdot(\alpha-d)^{-1}} \log(R/\epsilon)$.*

Proof Consider a mixing measure P so that $\mathsf{f} = \varphi \star P \in \Phi_R$. Given P, we first choose another mixing measure P' so that all its moments up to order k (to be chosen below) agree with those of P, i.e.,

$$\int_{\mathbb{R}^d} z^{\mathbf{j}} d(P - P')(z) = 0, \tag{55}$$

where $\mathbf{j} = (j_1, \ldots, j_d)$ is a multi-index such that $\sum_{l=1}^d j_l \le k$. It is easy to check that the cardinality of the set of such multi-indices is bounded by k^d. By Caratheodory's theorem (e.g. Ziegler, 1995, §1.6), P' can be chosen such that it is supported on k^d+1 atoms (indeed, stack all moments under consideration into a long vector and express this vector as a convex combination of points).

Let $\mathsf{f}' = \varphi \star P'$. By the convolution theorem, we have $\mathbf{F}^{-1}[\mathsf{f} - \mathsf{f}'] = (2\pi)^d \mathbf{F}^{-1}[\varphi] \cdot \mathbf{F}^{-1}[P - P']$. Application of the Fourier inversion theorem and the Hausdorff-Young inequality yields

$$\begin{aligned}
\|\mathsf{f} - \mathsf{f}'\|_{L_\infty} &\le (2\pi)^d \|\mathbf{F}^{-1}[\varphi] \cdot \mathbf{F}^{-1}[P - P']\|_{L_1} \\
&= (2\pi)^d \int_{\mathbb{R}^d} |\mathbf{F}^{-1}[\varphi](\omega)| \cdot |\mathbf{F}^{-1}[P - P'](\omega)| \, d\omega \\
&\lesssim (2\pi)^d \left(\int_{\mathbb{R}^d \setminus \mathbb{B}_2^d(L)} |\mathbf{F}^{-1}[\varphi](\omega)| \, d\omega + \int_{\mathbb{B}_2^d(L)} |\mathbf{F}^{-1}[P - P'](\omega)| \, d\omega \right),
\end{aligned} \tag{56}$$

for $L > 0$ to be chosen later, where the last inequality follows from the fact that $\mathbf{F}^{-1}[\varphi]$ and $\mathbf{F}^{-1}[P - P']$ are uniformly bounded (by the Riemann-Lebesgue lemma).

In the sequel, we will bound the two terms on the right hand separately. In order to bound the second term, we invoke (55) and then use the following:

$$\begin{aligned}
|\mathbf{F}^{-1}[P](\omega) - \mathbf{F}^{-1}[P'](\omega)| &= (2\pi)^{-d}\left|\int \exp(i\,\langle \omega, z\rangle) d(P-P')(z)\right| \\
&\leq (2\pi)^{-d}\int \sum_{j=0}^{k} \frac{(i\,\langle \omega, z\rangle)^j}{j!} d(P-P')(z) + \\
&\quad + \int \left|\exp(i\langle \omega, z\rangle) - \sum_{j=0}^{k} \frac{(i\,\langle \omega, z\rangle)^j}{j!}\right| d(P+P')(z) \\
&\leq (2\pi)^{-d}\int \frac{|\langle \omega, z\rangle|^{k+1}}{(k+1)!} d(P+P')(z) \\
&\leq (2\pi)^{-d}\, 2\left(\frac{e\|\omega\|_2 R}{k+1}\right)^{k+1}, \quad \omega \in \mathbb{R}^d.
\end{aligned}$$

The first inequality uses the triangle inequality. The second inequality results from the fact that $|\exp(ix) - \sum_{j=k+1}^{\infty}(ix)^j/j!| \leq |x|^{k+1}/((k+1)!)$ for all $x \in \mathbb{R}$ (Gao and van der Vaart, 2016, p. 614). For the third inequality, we have used that $|\langle \omega, z\rangle| \leq \|\omega\|_2 R$ for all z in the support of P and P' and the basic inequality $(k+1)! \geq ((k+1)/e)^{k+1}$ for all integers k.

Integration then yields

$$\begin{aligned}
\int_{\mathbb{B}_2^d(L)} \left(\frac{e\|\omega\|_2 R}{k+1}\right)^{k+1} d\omega &= \int_{\mathbb{S}^d}\int_0^L \left(\frac{erR}{k+1}\right)^{k+1} r^{d-1}\, dr\, du \\
&= \mathrm{vol}_{d-1}(\mathbb{S}^d)\left(\frac{eRL}{k+1}\right)^{k+1} L^d \frac{1}{k+d+1} \\
&\lesssim_d \left(\frac{eRL}{k+1}\right)^{k+1} L^d. \qquad (57)
\end{aligned}$$

For the first term in (56), we obtain

$$\begin{aligned}
\int_{\mathbb{R}^d\setminus\mathbb{B}_2^d(L)} |\mathbf{F}^{-1}[\varphi](\omega)|\, d\omega &= (2\pi)^{-d}\int_{\mathbb{R}^d\setminus\mathbb{B}_2^d(L)} |\mathbf{F}[\varphi](\omega)|\, d\omega \\
&\leq (2\pi)^{-d}\int_{\mathbb{R}^d\setminus\mathbb{B}_2^d(L)} \frac{1}{\|\omega\|_2^{\alpha}}\, d\omega \\
&\lesssim_d (2\pi)^{-d}\int_L^{\infty} \frac{1}{r^{\alpha}} r^{d-1}\, dr = (2\pi)^{-d}\left(\frac{1}{L}\right)^{\alpha-d},
\end{aligned}$$

since $\alpha \geq d+1$ by assumption. Consider the choice $L = \epsilon^{-(\alpha-d)^{-1}}$ so that the above right hand evaluates to $(2\pi)^{-d}\epsilon$. At the same, choose k such that $k+1 = 2eRL = 2eR\epsilon^{-(\alpha-d)^{-1}}$. Evaluating (57) accordingly, we obtain

$$\begin{aligned}
\left(\frac{1}{2}\right)^{2eR\epsilon^{-(\alpha-d)^{-1}}} \cdot L^d &= \exp\left((1/\epsilon)^{(\alpha-d)^{-1}}(2e\cdot R)\underbrace{\log(1/2)}_{<0} + d(\alpha-d)^{-1}\log(1/\epsilon)\right) \\
&\lesssim_{d,\alpha} \epsilon.
\end{aligned}$$

In summary, we have thus shown that f can be approximated with the desired accuracy by a location mixture whose mixing measure is supported on at most $K := k^d + 1 \lesssim R^d \epsilon^{-d(\alpha-d)^{-1}}$ atoms. It thus suffices to provide a covering of the set $\Phi_R^K = \{\mathsf{f} \star P : P \in \mathscr{P}_R^K\}$, where $\mathscr{P}_R^K$ denote the class of probability measures supported on at most K atoms contained in $\mathbb{B}_2^d(R)$. In order to obtain such covering we modify the proof of Proposition 1 in Gao and van der Vaart (2016) so it becomes applicable to general dimension. Specifically, let now $\mathsf{f}(\cdot) = \sum_{i=1}^K \lambda_i \varphi(\cdot - \mu_i)$ and $\mathsf{f}' = \sum_{i=1}^K \lambda_i' \varphi(\cdot - \mu_i')$ be two elements in Φ_R^K, where $\lambda = (\lambda_i)$ and $\lambda' = (\lambda_i')$ are contained in the probability simplex $\{w \in \mathbb{R}^K : \sum_{i=1}^K w_i = 1,\, w_i \geq 0,\, 1 \leq i \leq K\}$. We then have

$$\begin{aligned}
\|\mathsf{f}(\cdot) - \mathsf{f}'(\cdot)\|_\infty &= \left\| \sum_{i=1}^K \lambda_i \varphi(\cdot - \mu_i) - \sum_{i=1}^K \lambda_i' \varphi(\cdot - \mu_i') \right\|_\infty \\
&\leq \left\| \sum_{i=1}^K \lambda_i \{\varphi(\cdot - \mu_i) - \varphi(\cdot - \mu_i')\} \right\|_\infty + \left\| \sum_{i=1}^K (\lambda_i - \lambda_i') \varphi(\cdot - \mu_i') \right\|_\infty \\
&\leq \max_{1 \leq i \leq K} \|\varphi(\cdot - \mu_i) - \varphi(\cdot - \mu_i')\|_\infty + \|\varphi\|_\infty \|\lambda - \lambda'\|_1 \\
&\lesssim_{d,\alpha,\beta} \max_{1 \leq i \leq K} \|\mu_i - \mu_i'\|_2 + \|\lambda - \lambda'\|_1,
\end{aligned}$$

where in the last line we have used that φ is uniformly bounded and Lipschitz by asssumption. We thus obtain a covering of the desired accuracy via an ϵ-covering with respect to $\|\cdot\|_2$ for each of the K support points in $\mathbb{B}_2^d(R)$ and an ϵ-covering with respect to $\|\cdot\|_1$ of the associated probability simplex. Covering numbers for the latter two are well-studied (Vershynin, 2018, §4); (Gao and van der Vaart, 2016, Proof of Proposition 1). It follows that the covering number N of interest can be bounded as

$$N \leq \left(\frac{2+R}{\epsilon}\right)^{K \cdot d} \cdot \left(\frac{5}{\epsilon}\right)^K.$$

Inserting the expression for K given above, we thus obtain the conclusion of the lemma:

$$\log N \lesssim K(d+1) \log(R/\epsilon) \lesssim_d R^d \epsilon^{-d \cdot (\alpha-d)^{-1}} \log(R/\epsilon).$$

■

Appendix G. Miscellaneous technical lemmas

The following result for controlling the p-Wasserstein distance in terms of the total variation distance can be found in Villani (2009).

Lemma 20 (Theorem 6.15 in Villani (2009)) *Let μ and ν be two probability measures on $\mathbb{R}^d$. Then for any $1 \leq p < \infty$, we have*

$$\mathsf{W}_p^p(\mu, \nu) \leq 2^{p/q} \int_{\mathbb{R}^d} \|x\|_2^p \, d|\mu - \nu|(x), \qquad \frac{1}{p} + \frac{1}{q} = 1.$$

The next result, which is taken from Nguyen (2013), in turn bounds the right hand side of Lemma 20 if μ and ν have densities.

Lemma 21 (Lemma 6 in Nguyen (2013)) *Let f and g be probability density functions on $\mathbb{R}^d$, $s > 0$, and suppose that $\mathsf{M}_f^s := \int \|x\|_2^s \, f(x)\, dx$ and $\mathsf{M}_g^s := \int \|x\|_2^s \, g(x) dx$ are finite. We then have for any $0 < t < s$,*

$$\int_{\mathbb{R}^d} \|x\|_2^s \, |f(x) - g(x)| \, dx \leq 4 V_d^{\frac{s-t}{d+2s}} \left(\mathsf{M}_f^s + \mathsf{M}_g^s\right)^{\frac{(s-t)d+t}{s(d+2s)}} \|f - g\|_{L_2}^{\frac{2(s-t)}{d+2s}},$$

where $V_d := \pi^{d/2}/\Gamma(d/2+1)$ denotes the volume of the unit Euclidean ball in $\mathbb{R}^d$.

The next result, which is a special case of Theorem 2 in Fournier and Guillin (2015), yields a concentration inequality between the squared 2-Wasserstein distance of a measure and its empirical counterpart constructed from n i.i.d. samples.

Lemma 22 (Fournier and Guillin (2015)) *Let $\{X_i\}_{i=1}^n \overset{i.i.d.}{\sim} \nu$, where ν is a measure in $\mathbb{R}^d$ with compact support. Let $\nu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$. We then have, for all $t > 0$,*

$$\mathbf{P}(\mathsf{W}_2^2(\nu_n, \nu) \geq t) \leq C \begin{cases} \exp(-cnt^2) & \text{if } d \leq 3, \\ \exp(-cn(t/\log(2+1/t))^2) & \text{if } d = 4, \\ \exp(-cnt^{d/2}) & \text{if } d > 4. \end{cases}$$

The following result is a key ingredient in the proof of Proposition 9.

Lemma 23 *Let $P = \sum_{i=1}^n \alpha_i \delta_{x_i}$ and $Q = \sum_{j=1}^m \beta_j \delta_{x'_j}$ be two atomic probability measures on $\{x_i\}_{i=1}^n \subset \mathbb{R}^d$ and $\{x'_j\}_{j=1}^m \subset \mathbb{R}^d$, and suppose that $\Gamma = (\Gamma_{ij})_{1 \leq i \leq n,\, 1 \leq j \leq m}$ specifies an optimal coupling between P and Q with respect to any cost function c of the form $c(x, x') = h(\|x - x'\|)$, for some norm $\|\cdot\|$ and $h : \mathbb{R} \to \mathbb{R}$ convex. Let $\widetilde{x}_i := \sum_{j=1}^m \frac{\Gamma_{ij}}{\alpha_i} x'_j$, $1 \leq i \leq n$. It then holds that*

$$\mathsf{W}_c(P, Q) := \sum_{i=1}^n \sum_{j=1}^m \Gamma_{ij} c(x_i, x'_j) \geq \sum_{i=1}^n \alpha_i c(x_i, \widetilde{x}_i).$$

Proof Define $\lambda_{ij} = \frac{\Gamma_{ij}}{\alpha_i}$, and note that by construction $\sum_{j=1}^m \lambda_{ij} = 1$, for each i. Furthermore, observe that c is convex in either of its arguments. We hence have by Jensen's inequality that

$$\sum_{i=1}^n \sum_{j=1}^m \Gamma_{ij} c(x_i, x'_j) = \sum_{i=1}^n \alpha_i \sum_{j=1}^m \lambda_{ij} c(x_i, x'_j) \geq \sum_{i=1}^n \alpha_i c\Big(x_i, \sum_{j=1}^m \lambda_{ij} x'_j\Big) = \sum_{i=1}^n \alpha_i c(x_i, \widetilde{x}_i).$$

■

The result below is an ingredient in the proof of Theorem 10.

Lemma 24 *Let $d \geq 1$ be an integer, $a \geq 1, \beta > 0$. Then:*

$$\int_a^\infty v^d \exp(-v^\beta) \lesssim_{\beta, d} a^{d+(1-\beta)} \exp(-a^\beta).$$

Proof By a change of variables, we have

$$\int_a^\infty v^d \exp(-v^\beta)\, dv = \frac{1}{\beta} \int_{a^\beta}^\infty u^{\frac{d+(1-\beta)}{\beta}} \exp(-u)\, du. \tag{58}$$

For $\gamma \in \mathbb{R}$ and $\delta \geq 1$, consider the function

$$\mathcal{I}_\delta(\gamma) = \int_\delta^\infty u^\gamma \exp(-u)\, du.$$

If $\gamma < 0$, then $\mathcal{I}_\delta(\gamma) \leq \exp(-\delta)$ since $\delta \geq 1$ and the claim of the lemma follows from (58) after setting $\delta = a^\beta$. Conversely, suppose $\gamma > 0$. Using integration by parts, we find the recursion

$$\mathcal{I}_\delta(\gamma) = \delta^\gamma \exp(-\delta) + \gamma \cdot \mathcal{I}_\delta(\gamma - 1). \tag{59}$$

Let $\underline{\gamma} := \lfloor \gamma \rfloor$. The recursion (59) yields

$$\begin{aligned}
\mathcal{I}_\delta(\gamma) &= \exp(-\delta)\{\delta^\gamma + \gamma\delta^{\gamma-1} + \ldots + \gamma \cdot \ldots \cdot (\gamma - \underline{\gamma} + 1)\delta^{\gamma-\underline{\gamma}} \mathcal{I}_\delta(\gamma - \underline{\gamma})\} + \gamma \cdot \ldots \cdot (\gamma - \underline{\gamma}) \mathcal{I}_\delta(\gamma - \underline{\gamma} - 1) \\
&\leq \exp(-\delta)\{\delta^\gamma + \gamma\delta^{\gamma-1} + \ldots + \gamma \cdot \ldots \cdot (\gamma - \underline{\gamma} + 1)\delta^{\gamma-\underline{\gamma}} \mathcal{I}_\delta(\gamma - \underline{\gamma})\} + \gamma \cdot \ldots \cdot (\gamma - \underline{\gamma}) \exp(-\delta) \\
&\leq \exp(-\delta)\delta^\gamma(\lceil \gamma \rceil + 1)!
\end{aligned}$$

since there are most $\lceil \gamma \rceil + 1$ terms whose to be multiplied by $\exp(-\delta)$ and varying powers of δ, with each of these terms being no larger than $\lceil \gamma \rceil$. Setting $\delta = a^\beta$ and $\gamma = \frac{d+(1-\beta)}{\beta}$ yields the assertion. ■

Appendix H. Notions and Results from Optimal Transport

To make this paper self-contained, we here present notions and results from the theory of optimal transport as far as needed for the purpose of the paper. This material or slight modifications thereof are accessible from popular monographs and lecture notes on the subject, e.g., Peyré and Cuturi (2019); Villani (2009, 2003); Santambrogio (2015); McCann and Guillen (2011).

Definition 25 (Push-forward) *Let μ and ν be two Borel probability measures on measurable spaces $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ and $(\mathcal{Y}, \mathcal{B}_\mathcal{Y})$ respectively, and let T be a measurable map from $\mathcal{X}$ to $\mathcal{Y}$. The map T is said to **push forward** μ to ν, in symbols $T\#\mu = \nu$ if $T\#\mu(B) \equiv \mu(T^{-1}(B)) = \nu(B)$ for all $B \in \mathcal{B}_\mathcal{Y}$.*

Definition 26 (Optimal transport problem; Monge's problem) *Let μ and ν be as in the previous definition, and let $c : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$ be a measurable function ("cost function"). The optimal transport problem (Monge's problem) with μ, ν, and c is given by*

$$\inf_T \int_\mathcal{X} c(x, T(x))\, d\mu(x) \qquad \text{subject to} \quad T\#\mu = \nu.$$

Any minimizer of the above problem is called an optimal transport map.

The following optimization problem is in general a relaxation of the above problem; under certain conditions, both problems are equivalent.

Definition 27 (Kantorovich problem) *Let μ and ν be as in Definition 25, and let c be a cost function as in Definition 26. Let further $\Pi(\mu,\nu)$ denote the set of all couplings between μ and ν, i.e., probability measures on $\mathcal{X}\times\mathcal{Y}$ whose marginals equal to μ and ν. The Kantorovich problem is given by the optimization problem*

$$\inf_{\gamma\in\Pi(\mu,\nu)}\int_{\mathcal{X}}\int_{\mathcal{Y}} c(x,y)\,d\gamma(x,y).$$

Any minimizer of the above problem is called an optimal transport plan.

For measures μ and ν on $\mathbb{R}^d$ with finite k-th moments ($k\geq 1$), i.e., $\int\|x\|_2^k\,d\mu(x)<\infty$ and $\int\|x\|_2^k\,d\nu(x)<\infty$, the k-Wasserstein distance between μ and ν is defined via the above Kantorovich problem with cost function $c(x,y)=\|x-y\|_2^k$, i.e.,

$$\mathsf{W}_k(\mu,\nu):=\left(\inf_{\gamma\in\Pi(\mu,\nu)}\int\int\|x-y\|_2^k\,d\gamma(x,y)\right)^{1/k}. \tag{60}$$

A celebrated result due to Brenier characterizes optimal transport maps in the sense of Definition 26 for $\mathcal{X}=\mathcal{Y}=\mathbb{R}^d$ and quadratic cost, i.e., $c(x,y)=\|x-y\|_2^2$ and μ absolutely continuous with respect to the Lebesgue measure. In the sequel, we let $g^\star(x):=\sup_{y\in\mathbb{R}^d}\{\langle y,x\rangle-g(y)\}$ denote the Legendre-Fenchel conjugate of a convex function $g:\mathbb{R}^d\to\mathbb{R}\cup\{+\infty\}$.

Theorem 28 (Brenier) *Suppose that μ and ν are Borel probability measures on $\mathbb{R}^d$ with finite second moments, and suppose further that μ is absolutely continuous with respect to the Lebesgue measure. Then the optimal transport problem has a (μ-a.e.) unique minimizer $T=\nabla\psi$ for a convex function $\psi:\mathbb{R}^d\to\mathbb{R}\cup\{+\infty\}$. Furthermore, the optimal transport problem and its Kantorovich relaxation are equivalent in the sense that the optimal coupling in Definition 27 is of the form $(\mathrm{id}\times T)\#\mu$. Moreover, if in addition ν is absolutely continuous, then $\nabla\psi^\star$ is the (ν-a.e.) minimizer of the Monge problem transporting ν to μ, and it holds that $\nabla\psi^\star\circ\nabla\psi(x)=x$ (μ-a.e.), and $\nabla\psi\circ\nabla\psi^\star(y)=y$ (ν-a.e.).*

Appendix I. Fourier transform on $\mathbb{R}^d$

For a function $g\in L^1(\mathbb{R}^d)$, we define its Fourier and inverse Fourier transform by

$$\mathbf{F}[g](\omega):=\int_{\mathbb{R}^d}\exp(-i\langle\omega,x\rangle)\,g(x)\,dx,\qquad \mathbf{F}^{-1}[g](x):=\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\exp(i\langle\omega,x\rangle)\,g(\omega)\,d\omega,$$

for $\omega,x\in\mathbb{R}^d$. According to the Fourier inversion theorem, we have $\mathbf{F}^{-1}[\mathbf{F}[g]]=g=\mathbf{F}[\mathbf{F}^{-1}[g]]$ if $g\in L_1(\mathbb{R}^d)$ and $\mathbf{F}[g]\in L_1(\mathbb{R}^d)$. Other important properties that are used herein are as follows:

Plancherel theorem: $\langle\mathbf{F}[g],\mathbf{F}[h]\rangle=\frac{1}{(2\pi)^d}\langle g,h\rangle$,

Convolution theorem: $\mathbf{F}[(f\star g)]=\mathbf{F}[f]\cdot\mathbf{F}[g]$, $\mathbf{F}^{-1}[(f\star g)]=(2\pi)^d\mathbf{F}^{-1}[f]\cdot\mathbf{F}^{-1}[g]$

Haussdorf-Young inequality: $\|\mathbf{F}[g]\|_{L_q}\leq\|f\|_{L_p}(2\pi)^{d(1-1/p)}$, $1\leq p\leq 2$, $\frac{1}{p}+\frac{1}{q}=1$,

where the symbol $\langle \cdot, \cdot \rangle$ here refers to the inner product on $L^2(\mathbb{R}^d)$ and $\star$ denotes convolution.

Appendix J. Properties of the Generalized Multivariate Laplace distribution

The PDF of the generalized multivariate Laplace distribution is given by (Kozubowski et al., 2013)

$$\varphi(z) = \frac{2}{(2\pi)^{d/2}\Gamma(\kappa)} \|z\|_2^{\kappa - d/2} K_{\kappa - d/2}(\|z\|_2 \cdot C'),$$

for certain constants $C, C' > 0$. Here, for $\eta > 0$, K_η denotes the modified Bessel function of the third kind with index η:

$$K_\eta(u) = \frac{1}{2}(u/2)^\eta \int_0^\infty \frac{1}{t^{\eta+1}} \exp\left(-t - \frac{u^2}{4t}\right)\, dt, \quad u > 0.$$

In light of the above representation, we have

$$\varphi(z) \propto \psi(\|z\|_2), \quad \psi(r) = r^\lambda K_\lambda(r \cdot C'),\ r \geq 0,$$

where $\lambda := \kappa - d/2 \geq 1/2$ given the condition on κ preceding (19).

Verification of Property **(D1)**. Note that in order to show that φ is bounded at the origin, it suffices to show that the following function $\widetilde{\psi}$ is bounded:

$$\widetilde{\psi}(u) = u^{2\lambda} \int_0^1 \frac{1}{t^{\lambda+1}} \exp\left(-t - \frac{u^2}{4t}\right)\, dt,$$

Using the substitution $z = 1/t$, we obtain that

$$\widetilde{\psi}(u) = u^{2\lambda} \int_1^\infty z^{\lambda-1} \exp\left(-\frac{u^2}{4} \cdot z - 1/z\right)\, dz.$$

Using a second substitution $y = (u^2/4) \cdot z$, we find that

$$\widetilde{\psi}(u) = \int_{u^2/4}^\infty 4^\lambda y^{\lambda-1} \exp\left(-y - (u^2/4)(1/y)\right)\, dy,$$

which yields that $\widetilde{\psi}$ is upper bounded by $4^\lambda \Gamma(\lambda)$, which is a finite constant since $\lambda \geq 1/2$.

In order to show that φ is Lipschitz, it suffices to show that $\widetilde{\psi}'$ is bounded. The Leibniz integral rule yields

$$\begin{aligned}
\widetilde{\psi}'(u) &= -2u^{2\lambda-1} \exp(-u^2/4 - 1) + 4^\lambda \int_{u^2/4}^\infty \frac{d}{du} y^{\lambda-1} \exp\left(-y - \frac{u^2}{4}(1/y)\right)\, dy \\
&= -2u^{2\lambda-1} \exp(-u^2/4 - 1) + 4^\lambda \int_{u^2/4}^\infty y^{\lambda-2} \left(-\frac{1}{2}u\right) \exp\left(-y - \frac{u^2}{4}(1/y)\right)\, dy.
\end{aligned}$$

Note that since $\lambda \geq 1/2$, it remains to be shown that the integral on the right hand side is finite in the limit $u \to 0$. This is obvious whenever $\lambda > 1$, hence it suffices to consider

$\frac{1}{2} \leq \lambda \leq 1$. In that range, we have for any $u > 0$

$$\begin{aligned}\int_{u^2/4}^{\infty} y^{\lambda-2} u \exp\left(-y - \frac{u^2}{4}(1/y)\right)\, dy &\leq \int_{u^2/4}^{1} y^{\lambda-2} u \exp\left(-y - \frac{u^2}{4}(1/y)\right)\, dy + \\ &\quad + \int_{1}^{\infty} y^{\lambda-2} u \exp\left(-y - \frac{u^2}{4}(1/y)\right)\, dy \\ &\leq \int_{u^2/4}^{1} y^{-\frac{3}{2}} u \exp\left(-y - \frac{u^2}{4}(1/y)\right)\, dy + \\ &\quad + u \int_{1}^{\infty} \exp\left(-y - \frac{u^2}{4}(1/y)\right)\, dy,\end{aligned}$$

where we have used that $1/2 \leq \lambda \leq 1$. The second integral is clearly finite. Regarding the first integral, we have

$$\begin{aligned}\int_{u^2/4}^{1} y^{-\frac{3}{2}} u \exp\left(-y - \frac{u^2}{4}(1/y)\right)\, dy &\leq u \int_{u^2/4}^{1} y^{-3/2}\, dy \\ &= u \cdot (-2)\left(\frac{1}{y^{1/2}}\Big|_{u^2/4}^{1}\right) = 2u(2/u - 1) = (4 - 2u),\end{aligned}$$

which shows that the second integral is finite as well, and as a result yields the assertion that $\widetilde{\psi}'$ is bounded.

Verification of Property (**D3**). Note that a random variable Z with PDF φ has the following characterization (Kozubowski et al., 2013)

$$Z \stackrel{\mathrm{D}}{=} G\sqrt{\Gamma_\kappa},$$

where $G \sim N(0, I_d)$ and Γ_κ is a random variable having a Gamma distribution with shape parameter κ and scale parameter one. The moment-generating function of Γ_κ is given by

$$M_{\Gamma_\kappa}(t) = \left(\frac{1}{1-t}\right)^{\kappa}, \quad t \in (0, 1).$$

Therefore, by Markov's inequality, for any $z \geq 0$,

$$\mathbf{P}(\Gamma_\kappa > z) \leq \exp(-z/2)\, \mathbf{E}[\exp(\Gamma_\kappa/2)] = 2^\kappa \exp\left(-\frac{1}{2}z\right) = \exp\left(\kappa \log(2) - \frac{1}{2}z\right).$$

In particular,

$$\mathbf{P}(\Gamma_\kappa > 2\log(2)\kappa + t) \leq \exp(-t).$$

Moreover, by Lipschitz concentration of Gaussian random vectors, we have for any $\delta \geq 0$

$$\mathbf{P}(\|G\|_2 \geq \sqrt{d} + \delta) \leq \exp(-\delta^2/2).$$

Combining the above two concentration inequalities with the elementary inequality $\mathbf{P}(A \cdot B > a \cdot b) \leq \mathbf{P}(A > a) + \mathbf{P}(B > b)$ for non-negative random variables A and B and $a, b > 0$, setting $\delta = \sqrt{2t}$ yields

$$\mathbf{P}\left(\|Z\|_2 \geq \sqrt{d}\sqrt{2\log 2\kappa} + \sqrt{d}\sqrt{t} + 2\sqrt{\log(2)\kappa t} + \sqrt{2}t\right) \leq 2\exp(-t).$$

In particular, for any $u \geq 1$,

$$\mathbf{P}\left(\|Z\|_2 \geq C_{d,\kappa} u\right) \leq 2\exp(-u/\sqrt{2}),$$

where $C_{d,\kappa} = \sqrt{d}\sqrt{2\log 2\kappa} + \sqrt{d/2} + 2\sqrt{\log(2)\kappa}$.