MIT Open Access Articles

Uncoupled isotonic regression via minimum Wasserstein deconvolution

Uncoupled isotonic regression via minimum Wasserstein deconvolution

Philippe Rigollet[*] and Jonathan Weed[†]

Massachusetts Institute of Technology

Abstract. Isotonic regression is a standard problem in shape-constrained estimation where the goal is to estimate an unknown nondecreasing regression function f from independent pairs (x_i, y_i) where $\mathbb{E}[y_i] = f(x_i), i = 1, \ldots n$. While this problem is well understood both statistically and computationally, much less is known about its uncoupled counterpart where one is given only the unordered sets $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_n\}$. In this work, we leverage tools from optimal transport theory to derive minimax rates under weak moments conditions on y_i and to give an efficient algorithm achieving optimal rates. Both upper and lower bounds employ moment-matching arguments that are also pertinent to learning mixtures of distributions and deconvolution.

AMS 2000 subject classifications: 62G08.
Key words and phrases: Isotonic regression, Coupling, Moment matching, Deconvolution, Minimum Kantorovich distance estimation.

1. INTRODUCTION

Optimal transport distances have proven valuable for varied tasks in machine learning, computer vision, computer graphics, computational biology, and other disciplines; these recent developments have been supported by breakneck advances in computational optimal transport in the last few years [Cut13, AWR17, PC18, ABRW18]. This increasing popularity in applied fields has led to a corresponding increase in attention to optimal transport as a tool for theoretical statistics [FHN+19, RW18, ZP18]. In this paper, we show how to leverage techniques from optimal transport to solve the problem of *uncoupled isotonic regression*, defined as follows.

Let f be an unknown nondecreasing regression function from $[0, 1]$ to $\mathbb{R}$, and for $i = 1, \ldots, n$, let

$$y_i = f(x_i) + \xi_i \,,$$

where $\xi_i \sim \mathcal{D}$ are i.i.d. from some known distribution $\mathcal{D}$ and x_i are fixed (deterministic) design points. We note that the location of the design points is immaterial as long as $x_1 < \cdots < x_n$. Given $p \geq 1$, the goal of isotonic regression is to produce an estimator $\hat{f}_n$ that is close to f in the sense

that $\mathbb{E}\|\hat{f}_n - f\|_p^p$ is small, where for any g from $[0,1]$ to $\mathbb{R}$ we define

$$\|g\|_p^p := \frac{1}{n} \sum_{i=1}^n |g(x_i)|^p. \tag{1}$$

The key novelty in *uncoupled* isotonic regression is that the data at hand to construct $\hat{f}_n$ is given by the unordered sets $\{y_1, \ldots, y_n\}$ and $\{x_1, \ldots, x_n\}$. Informally, one does not know "which x corresponds to which y." In contrast, for *standard* isotonic regression, estimation is performed on the basis of the *coupled* data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$. To our best knowledge, uncoupled isotonic regression was introduced in [CS16] as a natural model for situations arising in the social sciences where uncoupled data is a common occurrence. For instance, the authors of [CS16] give the example of analyzing data collected by two different organizations, such as wage data collected by a governmental agency and housing price data collected by a bank. The relationship between wages and housing prices can naturally be assumed to be monotonic. Though these data sets involve the same individuals, the data is uncoupled, and no paired information exists. Our results indicate that despite the lack of paired data, a relationship between the data sets *can* be learned. In addition to raising obvious privacy issues, this result also has drastic implications for sample sizes, since it suggests that it is possible to integrate extremely large datasets such a census data or public real estate data even in the absence of coupled data.

While standard isotonic regression is a well understood and classical problem in shape-constrained estimation [vdG90,Mam91,vdG93,RWD88,MW00,Zha02,BBBB72,NPT85,BT15,FMR16,Bel18], it is not even clear *a priori* that consistent estimators for its uncoupled version exist. In absence of the noise random variables, $\xi_i, i = 1, \ldots, n$, the regression function is easy to estimate using monotonicity: After ordering the sets $\{y_1, \ldots, y_n\}$ and $\{x_1, \ldots, x_n\}$ as $y_{(1)} \leq \ldots \leq y_{(n)}$ and $x_{(1)} \leq \ldots \leq x_{(n)}$), it is clear that $y_{(i)} = f(x_{(i)})$, $i = 1, \ldots, n$. In the presence of noise, however, this naïve scheme fails, and the problem appears to be much more difficult—see Figure 1.

In this paper, we show that, quite surprisingly, a consistent estimator for f exists under general moment conditions on the noise distribution $\mathcal{D}$. We define an estimator by leveraging connections with optimal transport and show that it is minimax optimal simultaneously for all choices of p in the performance measure (1). As noted in [CS16], uncoupled isotonic regression is closely connected to deconvolution, which is a much harder problem than regression from a statistical perspective. Consequently, as our results show, minimax rates for this problem are *exponentially* worse than for the standard isotonic regression problem. A practical implication is that while uncoupled datasets may be integrated, their size should be exponentially larger in order to lead at least as good statistical accuracy.

Notation. Given quantities a and b, we write $a \lesssim b$ to indicate that $a \leq Cb$ for some universal constant $C > 0$ and define $a \vee b := \max(a, b)$. The notation $a \asymp b$ is used to indicate that $a \lesssim b$ and $b \lesssim a$. Throughout, log refers to the natural logarithm, and $\log_+ x := (\log x) \vee 0$. The terminology "$\ell_p$ norm" refers always to the empirical ℓ_p norm defined in (1). $\mathcal{F}$ denotes the class of nondecreasing functions from $[0,1]$ to $\mathbb{R}$ and for any $V > 0$, $\mathcal{F}_V \subset \mathcal{F}$ denotes the subset of functions $f \in \mathcal{F}$ such that $|f(x)| \leq V$ for $x \in [0,1]$.

1.1 Prior work

Isotonic regression is a fundamental problem in nonparametric statistics. As such, the literature on this topic is vast and very well established. A representative result is the following.
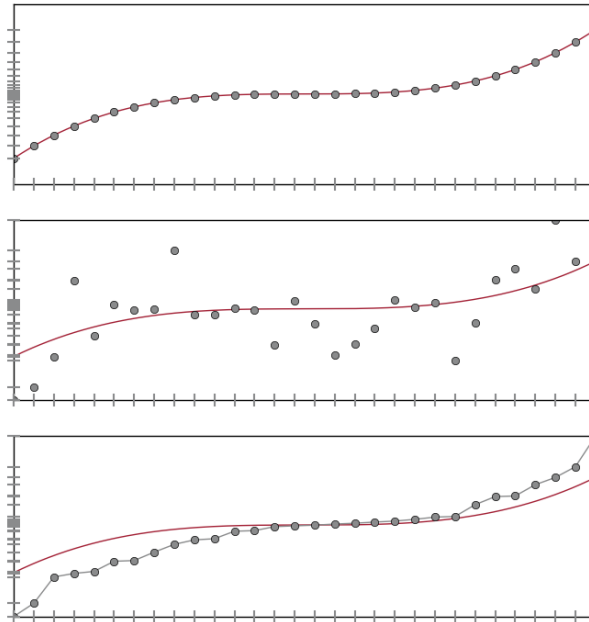
Figure 1: In the noiseless case (top figure), either coupled data (gray dots) or uncoupled data (gray tick marks on axes) suffice to recover the regression function (magenta curve). When noise is added (middle figure), uncoupling changes the problem considerably. Estimating the regression function by ordering the sets $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_n\}$ does not yield a consistent estimator (bottom figure).

THEOREM 1.   *[NPT85] If $\mathcal{F}_V$ is the class of nondecreasing functions from $[0,1]$ to $\mathbb{R}$ satisfying $|f(x)| \leq V$ for $x \in [0,1]$, then*

$$\inf_{g_n} \sup_{f \in \mathcal{F}_V} (\mathbb{E}\|f - g_n\|_2^2)^{1/2} \asymp \frac{\sigma^{2/3}V^{1/3}}{n^{1/3}} \, ,$$

where the infimum is taken over all measurable functions of the data. Moreover the minimax rate is achieved by the least squares estimator over $\mathcal{F}$ for which efficient algorithms such that the pool-adjacent-violators algorithm are well developed [RWD88].

While there are many refinements of this result, the $n^{-1/3}$ rate is a common feature of isotonic regression problems in a variety of contexts. By contrast, our results indicate that the minimax rate for the uncoupled problem is of order $\frac{\log \log n}{\log n}$. In other words, the number of samples required to obtain a certain level of accuracy in the uncoupled setting is exponentially larger than the number required for isotonic regression. This gap illustrates the profound difference between the coupled and uncoupled models.

In [CS16], the authors propose an estimator of f for uncoupled regression under smoothness assumptions. Crucially, this work draws an important connection between uncoupled isotonic re-

gression and deconvolution and their estimator actually uses deconvolution as a black box. Under smoothness assumptions, rates of convergence may be obtained by combining the results of [CS16] and rates of convergence for the cumulative distribution function (CDF) in deconvolution as in [DGJ11]. Whether the rates obtained in this way would be optimal over smooth classes of functions is unknown, but this question falls beyond the scope of standard shape-constrained estimation. Instead our results show that, as in standard isotonic regression, the function f can be consistently estimated in the uncoupled isotonic regression model, *without smoothness assumptions*. Furthermore, we prove matching upper and lower bounds on the optimal rates of estimation with respect to the empirical ℓ_p distance, for $1 \le p < \infty$.

The connection with deconvolution is not hard to see in hindsight: obtaining the function f from the data $\{y_1, \ldots, y_n\}$ resembles the problem of obtaining an estimate of a measure μ on the basis of samples from the convolution $\mu * \mathcal{D}$, where $\mathcal{D}$ is a known noise distribution. As we note below, the metric of interest in our case is the Wasserstein distance between univariate distributions. While question has been recently considered in the deconvolution literature [CCDM11, DFM15, DM13], our work present the following specificities. First, we make no smoothness assumptions on the noise distribution, and we assume, as is common in the isotonic regression literature, that the regression function f has bounded variation. This leads to different rates of estimation than those appearing in the deconvolution context. Moreover, we employ a simple minimum distance estimator (see Section 3) as opposed to the kernel estimators common in deconvolution. In short, our assumptions, estimator, and results are quite different from those appearing in the deconvolution literature, despite the similarities in the problem setting.

Our techniques leverage moment matching arguments, which have proven powerful in mixture estimation [MV10, BRW17, WY18] and nonparametric statistics [LNS99, JN02, CL11, CV17]. As in those works, our lower bounds are constructed by leveraging moment comparison theorems, which connect the moments of two distributions to their total variation distance. Our upper bounds are based on a novel result showing that the Wasserstein distance of any order between univariate measures can be controlled by moment matching. This result significantly extends and generalizes several similar results in the literature [KV17, WY18].

Finally, it is worth noting that uncoupled isotonic regression bears comparison to a similar problem in which the regression function is assumed to be *linear* instead of isotonic. This model, which goes under the names "estimation from a broken sample" [DG80], "shuffled linear regression" [APZ17], "linear regression without correspondence" [HSS17], "regression with permuted data" [PWC17, SBD17], and "unlabeled sensing" [UHV18], has been explored from both algorithmic and statistical perspectives. On the algorithmic side, the core question in these works is how to design efficient estimators for multivariate regression problems, which is nontrivial even in the noiseless setting (i.e., when $\xi_i \equiv 0$). On the statistical side, several computationally efficient estimators have been proposed [PWC17, APZ17] with provable guarantees. However, these estimators rely heavily on the linear model and do not extend to the isotonic case.

1.2 Model and assumptions

We focus on the fixed design case, as is common in the literature on isotonic regression. We assume the existence of a nondecreasing function $f \in \mathcal{F}_V$ such that

$$y_i = f(x_i) + \xi_i \qquad 1 \le i \le n, \tag{2}$$

where $\xi_i \sim \mathcal{D}$ i.i.d. We observe the design points $\{x_1, \ldots, x_n\}$, which we assume to be distinct, and the (unordered) set of points $\{y_1, \ldots, y_n\}$.

We make the following assumptions.

ASSUMPTION 1. *Both $\mathcal{D}$ and V are known.*

The assumption that $\mathcal{D}$ is known is essential and is ubiquitous in the deconvolution literature: if $\mathcal{D}$ is unknown, then no consistent estimator of f exists. For instance, if $\mathcal{D}$ is unknown, then it is impossible to reject the hypothesis that f is identically 0, and that all the variation in the set $\{y_1, \ldots, y_n\}$ is due to noise. By contrast, the assumption that V is known is for convenience only, since an upper bound on V can be estimated from the data.

We also require that $\mathcal{D}$ is sub-exponential [Ver18], a concept that we define rigorously via Orlicz norms.

DEFINITION 1. *Let $\psi_1(x) := e^x - 1$. We define an Orlicz norm $\|X\|_{\psi_1}$ of a random variable X by*

$$\|X\|_{\psi_1} := \inf\{t > 0 : \mathbb{E}\psi_1(|X|/t) \leq 1\}.$$

We say that a distribution $\mathcal{D}$ is sub-exponential *if $\xi \sim \mathcal{D}$ satisfies $\|\xi\|_{\psi_1} < \infty$ and we write by extension $\|\mathcal{D}\|_{\psi_1} := \|\xi\|_{\psi_1}$.*

It can be shown that $\|\cdot\|_{\psi_1}$ defines a norm on the space of random variables satisfying $\|X\|_{\psi_1} < \infty$, and that a random variable has a finite moment generating function in a neighborhood of the origin if and only if $\|X\|_{\psi_1} < \infty$. We note also that if $X \in [-V, V]$ almost surely, then $\|X\|_{\psi_1} \leq 2V$.

ASSUMPTION 2. *The noise distribution $\mathcal{D}$ is centered sub-exponential.*

We note that, in particular, Assumption 2 implies that $\mathcal{D}$ has finite moments of all orders. Nevertheless, this restriction is quite mild, as this encompasses most distributions which arise in practice and in theory.

Our only use of Assumption 2 will be to provide a bound on the moments of $\mathcal{D}$, which we obtain via the following well known lemma (see, e.g., [Ver18]). We reproduce a proof in Section A that exhibits an explicit constant.

LEMMA 1. *For all $p \geq 1$,*

$$(\mathbb{E}|X|^p)^{1/p} \leq p\|X\|_{\psi_1}.$$

1.3 Main results

Our main results are matching upper and lower minimax bounds for the problem of estimating the regression function in the ℓ_p distance, for any $1 \leq p < \infty$.

THEOREM 2 (Upper bound). *Assume that $\mathcal{D}$ is sub-exponential. There exists an estimator $\hat{f}_n$ and a universal constant C, such that, for all $1 \leq p < \infty$, the risk of $\hat{f}_n$ over $\mathcal{F}_V$ satisfies*

$$\sup_{f \in \mathcal{F}_V} (\mathbb{E}\|f - \hat{f}_n\|_p^p)^{1/p} \leq CpV\frac{\log\log n}{\log n}(1 + o_{V,\mathcal{D},p}(1)),$$

Where $o_{V,\mathcal{D},p}(1)$ indicates a quantity depending on V, $\|\mathcal{D}\|_{\psi_1}$, and p that goes to 0 as $n \to \infty$.

The estimator $\hat{f}_n$ appearing in Theorem 2 is a minimum distance estimator with respect to the Wasserstein distance, which we call a *minimum Wasserstein deconvolution estimator* (see Section 2). Surprisingly, the same estimator achieves the above bound for all $1 \leq p < \infty$. An analysis of this estimator appears in Section 3.

We complement this result with the following lower bound, which holds already in the case when $\mathcal{D}$ is the standard Gaussian distribution.

THEOREM 3 (Lower bound). *Let* $\mathcal{D} = N(0,1)$. *Under the same conditions as Theorem 2, there exists a universal constant* C' *such that the estimation risk over the class* $\mathcal{F}_V$ *satisfies*

$$\inf_{g_n} \sup_{f \in \mathcal{F}_V} \mathbb{E}\|f - g_n\|_p \geq C'V \frac{\log\log n}{\log n}(1 + o_V(1)),$$

where the infimum is taken over all measurable functions of the data.

The proofs of Theorems 2 and 3 rely on hitherto unexplored connections between isotonic regression and optimal transport between probability measures [Vil08]. To exploit this connection, we establish a novel result connecting the Wasserstein p-distance between two univariate distributions with the differences in the moments of the two distributions (Theorem 4). Since we believe this connection will prove useful for other works, we prove a more general version than is needed to obtain Theorems 2 and 3. While similar results have appeared elsewhere in the literature for the W_1 distance [KV17, WY18], our general version is the first to our knowledge to apply to W_p for $p > 1$ and to unbounded measures.

2. UNCOUPLED REGRESSION VIA OPTIMAL TRANSPORT

The observation that forms the core of our work is that the uncoupled regression model naturally relates to the Wasserstein distance between univariate measures.

2.1 Minimum Wasserstein deconvolution

We first recall the following definition.

DEFINITION 2. *For* $1 \leq p < \infty$, *the* Wasserstein-p distance *between two probability distributions* μ *and* ν *is defined by*

$$W_p(\mu, \nu) := \inf_{\gamma \in \mathcal{C}(\mu,\nu)} \left(\int_\infty^\infty |x - y|^p \, \mathrm{d}\gamma(x,y) \right)^{1/p}, \tag{3}$$

where the infimum is taken over the set $\mathcal{C}(\mu, \nu)$ *of all joint distributions on* $\mathbb{R} \times \mathbb{R}$ *with first marginal* μ *and second marginal* ν.

For all $p \geq 1$, the space $\mathcal{M}$ of probability measures having finite moments of all orders equipped with the distance W_p defines a metric space denoted by $(\mathcal{M}, W_p)$. The key observation is that the risk in isotonic regression can be controlled via the Wasserstein distance. To see this, we need the following definition.

DEFINITION 3. *Let* $x_1, \ldots, x_n$ *be fixed. For any nondecreasing function* $g : [0,1] \to \mathbb{R}$, *denote by* π_g *the measure*

$$\frac{1}{n} \sum_{i=1}^n \delta_{g(x_i)}.$$

We call π_g *a* pushforward measure *(of the uniform measure on* $\{x_1, \ldots, x_n\}$ *through* g*).*

The following proposition establishes the central connection between isotonic regression functions and the Wasserstein distance.

PROPOSITION 1. *Let $\mathcal{F}$ be the class of nondecreasing functions from $[0,1]$ to $\mathbb{R}$. For all $1 \le p < \infty$, the map $f \mapsto \pi_f$ is an isometry between $(\mathcal{F}, \ell_p)$ and $(\mathcal{M}, W_p)$. In other words, the empirical ℓ_p distance corresponds to the Wasserstein distance between the pushforward measures:*

$$\|f - g\|_p = W_p(\pi_f, \pi_g).$$

PROOF. Let $\gamma = \frac{1}{n} \sum_{i=1}^{n} \delta_{(f(x_i), g(x_i))}$. Clearly $\|f - g\|_p^p = \int_{-\infty}^{\infty} |x - y|^p \, \mathrm{d}\gamma(x, y)$, and γ is a coupling between π_f and π_g. It suffices to show that this coupling is optimal in the sense that it realizes the minimum definition (3). For $i, j \in [n]$, the monotonicity of f and g implies

$$(f(x_i) - f(x_j))(g(x_i) - g(x_j)) \ge 0.$$

Therefore, the support $\{(f(x_i), g(x_i)), i = 1, \ldots, n\}$ of γ is *monotone*, meaning that for any $(a, b), (c, d) \in \operatorname{supp}(\gamma)$, the implication $a < c \implies b \le d$ holds. Standard facts [San15, Theorem 2.9] then imply that it is optimal. □

Denote by $\hat{\pi}$ the empirical distribution of the observation $\{y_1, \ldots, y_n\}$. A sample from $\hat{\pi}$ is marginally distributed as $\pi_f * \mathcal{D}$, the convolution of the pushforward measure π_f with the noise distribution $\mathcal{D}$. Thus, finding π_f can be viewed as a deconvolution problem, or equivalently as a mixture learning problem whose centers are given by the distribution π_f. Consequently, our estimator is similar to estimators proposed in the mixture learning literature. One common choice is to choose the parameter that minimizes the distance to the empirical distribution in the Kolomogorov-Smirnov distance [DK68, Che95, HK15]; however, Proposition 1 suggests as an estimator a minimizer of $g \mapsto W_p(\pi_g * \mathcal{D}, \hat{\pi})$ over a suitable function class. Such estimators were introduced in [BBR06] under the name *minimum Kantorovich distance estimators* and shown to be consistent under regularity assumptions. By analogy, we call our technique *minimum Wasserstein deconvolution*.[1]

We focus on the following estimator:

$$\hat{f} \in \underset{g \in \mathcal{F}_V}{\operatorname{argmin}} \, W_2^2(\pi_g * \mathcal{D}, \hat{\pi}). \tag{4}$$

As Theorem 2 shows, the estimator $\hat{f}$ is *adaptive* to p, in the sense that it converges to f at the same rate in all ℓ_p metrics. Furthermore, by Theorem 3, this rate is minimax optimal.

The definition of our estimator involves the distance W_2. However, our analysis reveals that W_2 can be replaced by W_r for any $r \in (1, \infty)$ to obtain an estimator with the same performance. Indeed, the interested reader may check that the only results which need to be updated are Theorem 5 and Proposition 4. Theorem 5 can be replaced by a similar argument following [PP14, Proposition 2.21, (ii)]. Likewise, Proposition 4 holds with exactly the same proof, since it relies only on the triangle inequality (which holds for all W_r) and Lemma 10, a more general version of which can be found in [BL16].

---

[1] Note that the shorthand *Wassertein deconvolution* has appeared in the deconvolution literature [DM13, DFM15] to refer to deconvolution problems in which the Wasserstein distance is used as a measure of success. We emphasize that *minimum Wasserstein deconvolution* refers here to a novel method to perform deconvolution based on Wasserstein distances. Moreover, in light of Proposition 1, this method also achieves good performance in the Wasserstein metric.

2.2 A computationally efficient estimator

A priori, it is unclear how to optimize the function $f \mapsto W_2^2(\pi_f * \mathcal{D}, \hat{\pi})$ explicitly. In order to obtain an estimator which can be computed in polynomial time, we propose in this section a computationally efficient version of (4), which enjoys the same theoretical guarantees. We first relax (4) and consider instead the program

$$\operatorname*{argmin}_{\mu \in \mathcal{M}_V} W_2^2(\mu * \mathcal{D}, \hat{\pi}),$$

where the minimization is taken over *all* measures with support in $[-V, V]$. This is now a convex program, albeit an infinite dimensional one. However, we show below that it suffices to optimize over a finite-dimensional subset of $\mathcal{M}_V$, which yields a tractable convex program. Finally, we show how to round the resulting solution $\hat{\mu}$ to a pushforward measure in the sense of Definition 3.

We first consider the following quantization of the real line. Assume $n \geq 3$. Let $\alpha_0 := -(V + \sigma) \log n$ and set

$$\alpha_i := \alpha_0 + i \cdot \frac{V + \sigma}{n^{1/4}} \qquad \text{for } 1 \leq i \leq N := \lceil 2n^{1/4} \log n \rceil.$$

Let $\mathcal{A} := \{\alpha_i\}_{i=0}^N$, and denote by $\mathcal{M}_{\mathcal{A},V}$ the set of measures supported on $\mathcal{A} \cap [-V, V]$, which is a discrete set of cardinality $O(n^{1/4})$. Finally, define the projection operator $\Pi_{\mathcal{A}} : \mathbb{R} \to \mathcal{A}$ by

$$\Pi_{\mathcal{A}}(x) := \begin{cases} \alpha_0 & \text{if } x < \alpha_0 \\ \alpha_i & \text{if } \alpha_i \leq x < \alpha_{i+1} \text{ for } 0 \leq i \leq N-1 \\ \alpha_N & \text{if } x \geq \alpha_N. \end{cases}$$

We propose the following computationally efficient estimator:

$$\hat{\mu} \in \operatorname*{argmin}_{\mu \in \mathcal{M}_{\mathcal{A},V}} W_2^2(\Pi_{\mathcal{A}\sharp}(\mu * \mathcal{D}), \hat{\pi}), \tag{5}$$

where $\Pi_{\mathcal{A}\sharp}(\mu * \mathcal{D})$ is the pushforward of the measure $\mu * \mathcal{D}$ by the projection operator. The map $\mu \mapsto W_2^2(\Pi_{\mathcal{A}\sharp}(\mu * \mathcal{D}), \hat{\pi})$ is convex, and subgradients can be obtained by standard methods in computational optimal transport [PC18]. The measure $\hat{\mu}$ can therefore be obtained efficiently.

In general, the solution $\hat{\mu}$ to (5) will not be of the form π_g for some isotonic function g. However, a sufficiently close function can easily be obtained. Given a measure μ, denote by $\mathcal{Q}_\mu$ the quantile function of μ; we then define $\hat{g}$ by

$$\hat{g}(x_i) := \mathcal{Q}_{\hat{\mu}}(i/n) \qquad \text{for } 1 \leq i \leq n,$$

and extend $\hat{g}$ to other values in $[0, 1]$ arbitrarily so that the resulting function lies in $\mathcal{F}_V$.

PROPOSITION 2.   *The estimator $\hat{g}$ achieves the same rate as the estimator $\hat{f}$ defined in (4).*

The proof is deferred to Appendix A.1.

2.3 From Wasserstein distances to moment-matching, and back

Both the upper and lower bounds for the uncoupled regression problem (Theorems 2 and 3) depend on moment-matching arguments that we gather here. The core of our approach is Theorem 4, which establishes that the Wasserstein distance between univariate measures can be controlled by

comparing the moments of the two measures. In Proposition 3, we give examples establishing that Theorem 4 cannot be improved in general.

Similar moment-matching results for the Wasserstein-1 distance W_1 have appeared in other works [KV17, WY18], but in general these results rely on arguments via polynomial approximation of Lipschitz functions combined with the dual representation of W_1 [Vil03]. This approach breaks down for measures with unbounded support, and cannot establish tight bounds W_p for $p > 1$. By contrast, Theorem 4 applies to all measures with convergent moment generating functions, and yields bounds for W_p for all $1 \leq p < \infty$.

DEFINITION 4. *For any distributions μ and ν on $\mathbb{R}$ and $\ell \geq 1$, define*

$$\Delta_\ell(\mu, \nu) := \left| \mathbb{E}[X^\ell] - \mathbb{E}[Y^\ell] \right|^{1/\ell} \qquad X \sim \mu, Y \sim \nu \,.$$

When μ and ν are clear from context, we abbreviate $\Delta_\ell(\mu, \nu)$ by Δ_ℓ.

We are now in a position to state the main result of this section: it shows that two distributions with similar moments are close in Wasserstein distance. Its proof is postponed to Appendix A.2.

THEOREM 4. *Let μ and ν be two distributions on $\mathbb{R}$ whose moment generating functions are finite everywhere. There exists a universal constant $C > 0$ such that, for $1 \leq p < \infty$,*

$$W_p(\mu, \nu) \leq C p \sup_{\ell \geq 1} \frac{\Delta_\ell(\mu, \nu)}{\ell} \,.$$

Theorem 4 includes as a corollary the following result for bounded measures, a version of which appeared in [KV17, Proposition 1] for the $p = 1$ case.

COROLLARY 1. *Let μ and ν be two measures supported on $[-1, 1]$. For any $k \geq 1$, if $\max_{\ell \leq k} \Delta_\ell^\ell(\mu, \nu) \leq \varepsilon < 1$, then*

$$W_p(\mu, \nu) \lesssim p \left(\frac{1}{\log(1/\varepsilon)} \vee \frac{1}{k} \right) \,.$$

PROOF. For $\ell \leq k$, we have by assumption the bound $\Delta_\ell \leq \varepsilon^{1/\ell}$, whereas for $\ell > k$, we have the bound $\Delta_\ell \lesssim 1$ because μ and ν are supported on $[-1, 1]$. Applying Theorem 4 and noting that $\ell \mapsto \varepsilon^{1/\ell}/\ell$ is maximized at $\ell = \log(1/\varepsilon)$ yields the claim. $\square$

Our results imply a similar simple result for sub-Gaussian measures. We state it as a result of independent interest but will not need it to analyze uncoupled isotonic regression.

COROLLARY 2. *Let μ and ν be two sub-Gaussian measures. For any $k \geq 1$, if $\max_{\ell \leq k} \Delta_\ell^\ell(\mu, \nu) \leq \varepsilon < 1$, then*

$$W_p(\mu, \nu) \lesssim p \left(\frac{1}{\log(1/\varepsilon)} \vee \frac{1}{\sqrt{k}} \right) \,.$$

PROOF. The proof is the same as the proof of Corollary 1, except that we replace the estimate $\Delta_\ell \lesssim 1$ for $\ell > k$ by the estimate $\Delta_\ell \lesssim \sqrt{\ell}$. $\square$

As the following proposition makes clear, Theorem 4 is essentially tight.

PROPOSITION 3. *There exists a universal constant $c > 0$ such that, for any $k \geq 1$, there exist two measures μ and ν on $[-1, 1]$ such that $\Delta_\ell = 0$ for $1 \leq \ell < k$ but*

$$W_1(\mu, \nu) \geq \frac{c}{k}.$$

In other words, the dependence on $\sup_{\ell \geq 1} \frac{\Delta_\ell}{\ell}$ cannot be improved.

Moreover, there exists a universal constant $c > 0$ such that, for all $\varepsilon > 0$ sufficiently small, there exist two measures μ and ν whose moment generating functions are finite everywhere and

$$W_p(\mu, \nu) \geq cp^{1-\varepsilon} \qquad \forall p \geq 1.$$

In other words, the dependence on p cannot be improved.

A proof of Proposition 3 appears in Appendix A.5.

The following result complements Theorem 4 by showing that if two probability measures μ and ν are close in Wasserstein-2 distance, then their moments are close. This direction is much easier than that of Theorem 4 and illustrates that Wasserstein distances are strong distances.

THEOREM 5. *For any two subexponential probability measures μ and ν on $\mathbb{R}$ and any integer $\ell \geq 1$, it holds*

$$\Delta_\ell^\ell(\mu, \nu) \leq (2\ell)^\ell (\|\mu\|_{\psi_1} \vee \|\nu\|_{\psi_1})^{\ell-1} W_2(\mu, \nu).$$

PROOF. We employ the following bound [PP14], Proposition 2.21, (ii), valid for any random variables $X \sim \mu$ and $Y \sim \nu$ and positive integer ℓ:

$$\mathbb{E}[X^\ell - Y^\ell] \leq \ell \cdot W_2(\mu, \nu) \left(\left(\mathbb{E}|X|^{2(\ell-1)} \right)^{1/2} + \left(\mathbb{E}|Y|^{2(\ell-1)} \right)^{1/2} \right)$$

Lemma 1 implies

$$\ell \left(\left(\mathbb{E}|X|^{2(\ell-1)} \right)^{1/2} + \left(\mathbb{E}|Y|^{2(\ell-1)} \right)^{1/2} \right) \leq 2\ell \left(2(\ell-1)(\|X\|_{\psi_1} \vee \|Y\|_{\psi_1}) \right)^{\ell-1}$$
$$\leq (2\ell)^\ell (\|X\|_{\psi_1} \vee \|Y\|_{\psi_1})^{\ell-1}.$$

Combining these bounds yields the claim. $\square$

A similar result showing that Δ_ℓ may be controlled by the Wasserstein-1 distance follows directly from the dual representation of W_1 as a supremum over Lipschitz functions (see, e.g., [Vil03]) when the measures μ and ν have *bounded support*. As we will see in the proof of Theorem 2, we apply Theorem 5 to convolved distributions of the form $\mu = \pi_g * \mathcal{D}$, which have unbounded support whenever $\mathcal{D}$ does. Our proof techniques therefore require the use of a stronger metric than W_1 whenever the noise distribution $\mathcal{D}$ has unbounded support.

3. PROOF OF THE UPPER BOUND

In this section, we show that the minimum Wasserstein deconvolution estimation (4) achieves the upper bound of Theorem 2. The proof employs the following steps.

1. We show that it follows from the fact that $\hat{f}$ is a minimum Wasserstein distance estimator that $W_2(\pi_{\hat{f}} * \mathcal{D}, \pi_f * \mathcal{D})$ is small (Proposition 4).
2. In light of Theorem 5, this implies that the sequence $\{\Delta_\ell(\pi_{\hat{f}} * \mathcal{D}, \pi_f * \mathcal{D})\}_{\ell \geq 1}$ is uniformly controlled.
3. A simple lemma (Lemma 2) induces a weaker control for the deconvolved measures so that $\{\Delta_\ell(\pi_{\hat{f}}, \pi_f)\}_{\ell \geq 1}$ is also controlled.
4. Finally, we use Theorem 4 to control $W_p(\pi_{\hat{f}}, \pi_f)$ for all $p \geq 1$.

We collect steps 2–4 into Proposition 5, a deconvolution result which may be of independent interest.

Throughout this section, we assume $\|\mathcal{D}\|_{\psi_1} \leq \sigma$. We first carry out step 1, and show that $\hat{f}$ satisfies the following "convolved" guarantee as a simple consequence of its definition.

PROPOSITION 4. *The estimator $\hat{f}$ defined in (4) satisfies*

$$\mathbb{E}W_2(\pi_{\hat{f}} * \mathcal{D}, \pi_f * \mathcal{D}) \lesssim (\sigma + V)n^{-1/4}.$$

PROOF. The triangle inequality and the definition of $\hat{f}$ imply

$$W_2(\pi_{\hat{f}} * \mathcal{D}, \pi_f * \mathcal{D}) \leq W_2(\pi_{\hat{f}} * \mathcal{D}, \hat{\pi}) + W_2(\pi_f * \mathcal{D}, \hat{\pi}) \leq 2W_2(\pi_f * \mathcal{D}, \hat{\pi}). \tag{6}$$

By definition, the support of $\hat{\pi}$ is $\{f(x_1) + \xi_1, \ldots, f(x_n) + \xi_n\}$. Let $w_1, \ldots, w_n$ be i.i.d. samples from π_f, independent of all other randomness, and denote by $w_{(1)}, \ldots, w_{(n)}$ their increasing rearrangement. Since $\{w_i\}$ and $\{\xi_i\}$ are independent, the set $\{w_{(i)} + \xi_i, i = 1, \ldots, n\}$ comprises i.i.d. samples from $\pi_f * \mathcal{D}$. Applying the triangle inequality, we get that

$$\mathbb{E}[W_2(\pi_f * \mathcal{D}, \hat{\pi})] \leq \mathbb{E}[W_2(\pi_f * \mathcal{D}, \frac{1}{n} \sum_{i=1}^n \delta_{w_{(i)} + \xi_i})] + \mathbb{E}[W_2(\hat{\pi}, \frac{1}{n} \sum_{i=1}^n \delta_{w_{(i)} + \xi_i})]. \tag{7}$$

It follows from Lemma 10 that

$$\mathbb{E}[W_2(\pi_f * \mathcal{D}, \frac{1}{n} \sum_{i=1}^n \delta_{w_{(i)} + \xi_i})] \lesssim \frac{\sigma + V}{n^{1/4}}. \tag{8}$$

We now control the second term in the right-hand side of (7). A simple coupling between the two measures $\hat{\pi}$ and $\frac{1}{n} \sum_{i=1}^n \delta_{w_{(i)} + \xi_i}$ yields

$$W_2^2(\hat{\pi}, \frac{1}{n} \sum_{i=1}^n \delta_{w_{(i)} + \xi_i}) \leq \frac{1}{n} \sum_{i=1}^n |f(x_i) + \xi_i - w_{(i)} - \xi_i|^2 = \frac{1}{n} \sum_{i=1}^n |f(x_i) - w_{(i)}|^2 = W_2^2(\pi_f, \frac{1}{n} \sum_{i=1}^n \delta_{w_i}).$$

Thus, applying again, Lemma 10, we get

$$\mathbb{E}[W_2(\hat{\pi}, \frac{1}{n} \sum_{i=1}^n \delta_{w_{(i)} + \xi_i})] \lesssim \frac{V}{n^{1/4}} \tag{9}$$

Combining (6)–(9) completes the proof. □

The following uses steps 2–4 to obtain a deconvolution result. It implies that a bound on $W_2(\mu * \mathcal{D}, \nu * \mathcal{D})$ can yield a bound on $W_p(\mu, \nu)$ for all $p \in [1, \infty)$, as long as μ and ν have bounded support.

PROPOSITION 5. *If μ and ν have support lying in $[-V, V]$ and $W_2(\mu * \mathcal{D}, \nu * \mathcal{D}) \leq (\sigma + V)e^{-e^{1/2}}$,*
then

$$W_p(\mu, \nu) \lesssim pV \frac{\log\log \frac{\sigma + V}{W_2(\mu * \mathcal{D}, \nu * \mathcal{D})} + \log_+ \frac{\sigma(\sigma + 2V)}{V}}{\log \frac{\sigma + V}{W_2(\mu * \mathcal{D}, \nu * \mathcal{D})}}.$$

PROOF. As mentioned above, this proofs goes via a moment-matching argument. Since μ and ν
have bounded support, their moment generating functions converge everywhere; hence, Theorem 4
implies that it suffices to control $\sup_{\ell \geq 1} \frac{\Delta_\ell(\mu, \nu)}{\ell}$ to obtain a bound on $W_p(\mu, \nu)$.

Define

$$\varepsilon := \frac{W_2(\mu * \mathcal{D}, \nu * \mathcal{D})}{\sigma + V}.$$

Note that $\|\mu * \mathcal{D}\|_{\psi_1} + \|\nu * \mathcal{D}\|_{\psi_1} \leq 2(\sigma + 2V)$, so that Theorem 5 yields

$$\Delta_\ell^\ell(\mu * \mathcal{D}, \nu * \mathcal{D}) \leq \left(4\ell(\sigma + 2V)\right)^\ell \varepsilon \tag{10}$$

We now use the following deconvolution Lemma. Its proof is postponed to Appendix A.6.

LEMMA 2. *For any two subexponential probability measures μ and ν on $\mathbb{R}$ and any integer $\ell \geq 1$,*
it holds

$$\Delta_\ell^\ell(\mu, \nu) \leq (4\ell\|\mathcal{D}\|_{\psi_1})^\ell \cdot \sup_{m \leq \ell} \Delta_m^m(\mu * \mathcal{D}, \nu * \mathcal{D}).$$

Together with (10), it yields

$$\Delta_\ell(\mu, \nu) \leq 16\ell^2 \sigma(\sigma + 2V)\varepsilon^{1/\ell}, \qquad \ell \geq 1 \tag{11}$$

We now split the analysis into small and large ℓ. Assume first that

$$\ell < \frac{\log(1/\varepsilon)}{2\log\log(1/\varepsilon) + \log_+ \frac{\sigma(\sigma + 2V)}{V}},$$

Then, (11) yields

$$\frac{\Delta_\ell}{\ell} \leq 16\ell\sigma(\sigma + 2V)\varepsilon^{1/\ell} \leq 16\ell \frac{V}{(\log(1/\varepsilon))^2} \lesssim V\frac{\log\log(1/\varepsilon) + \log_+ \frac{\sigma(\sigma + V)}{V}}{\log(1/\varepsilon)},$$

where we have used the fact that $2\log\log(1/\varepsilon) + \log_+ \frac{\sigma(\sigma + 2V)}{V} \geq 2\log\log(1/\varepsilon) \geq 1$, by assumption.
Next assume that

$$\ell \geq \frac{\log(1/\varepsilon)}{2\log\log(1/\varepsilon) + \log_+ \frac{\sigma(\sigma + 2V)}{V}}.$$

Since μ and ν have bounded support, clearly $\Delta_\ell(\mu, \nu) \lesssim V$ for all $\ell \geq 1$. Therefore,

$$\frac{\Delta_\ell(\mu, \nu)}{\ell} \lesssim V\frac{\log\log(1/\varepsilon) + \log_+ \frac{\sigma(\sigma + 2V)}{V}}{\log(1/\varepsilon)}.$$

Combining small and large ℓ, we obtain

$$\sup_{\ell \geq 1} \frac{\Delta_\ell(\mu, \nu)}{\ell} \lesssim V\frac{\log\log(1/\varepsilon) + \log_+ \frac{\sigma(\sigma + V)}{V}}{\log(1/\varepsilon)}.$$

The proof of Proposition 5 then follows by applying Theorem 4. □

We are now in a position to conclude the proof of the upper bound in Theorem 2. Let $W :=$ $W_2(\pi_{\hat{f}} * \mathcal{D}, \pi_f * \mathcal{D})$. Assume that n is large enough that $n^{1/8} \geq e^{e^{1/2}}$. Denote by $\mathcal{E}$ the event on which the inequality $W \leq (\sigma + V)n^{-1/8}$ holds.

$$
\begin{aligned}
(\mathbb{E}\|f - \hat{f}\|_p^p)^{1/p} &= (\mathbb{E}W_p^p(\pi_f, \pi_{\hat{f}}))^{1/p} \\
&\leq (\mathbb{E}[W_p^p(\pi_f, \pi_{\hat{f}})\mathbb{1}_{\mathcal{E}}])^{1/p} + (\mathbb{E}[W_p^p(\pi_f, \pi_{\hat{f}})\mathbb{1}_{\mathcal{E}^C}])^{1/p} \, .
\end{aligned}
$$

On $\mathcal{E}$, Proposition 5 yields

$$
W_p(\pi_f, \pi_{\hat{f}}) \lesssim pV \frac{\log \log n^{1/8} + \log_+ \frac{\sigma(\sigma + 2V)}{V}}{\log n^{1/8}} \, .
$$

On the other hand, since $f, \hat{f} \in \mathcal{F}_V$, we have the trivial bound $W_p^p(\pi_f, \pi_{\hat{f}}) \leq (2V)^p$, so Markov's inequality combined with Proposition 4 yields

$$
(\mathbb{E}[W_p^p(\pi_f, \pi_{\hat{f}})\mathbb{1}_{\mathcal{E}^C}])^{1/p} \leq 2V\mathbb{P}[W > (\sigma + V)n^{-1/8}]^{1/p} \lesssim Vn^{-1/8p} \, .
$$

We obtain

$$
\begin{aligned}
(\mathbb{E}\|f - \hat{f}\|_p^p)^{1/p} &\lesssim pV \frac{\log \log n^{1/8} + \log_+ \frac{\sigma(\sigma + 2V)}{V}}{\log n^{1/8}} + Vn^{-1/8p} \\
&= pV \frac{\log \log n}{\log n}(1 + o_{V,\sigma,p}(1)) \, .
\end{aligned}
$$

4. PROOF OF THE LOWER BOUND

In this section, we prove Theorem 3. To that end, we employ the "method of fuzzy hypotheses" [Tsy09] and define two prior probability distributions on the space of nondecreasing functions.

Our construction is based on the following lemma which has appeared before in the moment-matching literature.

LEMMA 3.   *[CL11, WY18] There exists a universal constant c such that, for any $k \geq 1$, there exist two centered probability distributions P and Q on $[-V, V]$ such that*

$$
\Delta_\ell(P, Q) = 0 \quad \text{for } \ell = 1, \ldots, k - 1,
$$

and such that $W_1(P, Q) \geq \frac{cV}{k}$.

PROOF OF THEOREM 3. By the monotonicity of ℓ_p norms, it suffices to prove the claim for $p = 1$. First, we show how a measure on $[-V, V]$ can be reduced to a sample $\{y_1, \ldots y_n\}$ from an uncoupled regression model, with a possibly random regression function.

Let μ be any measure on $[-V, V]$. Let $Z_1, \ldots, Z_n$ be i.i.d from μ, and denote by $\{Z_{(i)}\}$ the sorted version of $\{Z_i\}$ such that $Z_{(1)} \leq \cdots \leq Z_{(n)}$. Let $F = F_{Z_1,\ldots,Z_n}$ from $[0, 1]$ to $[-V, V]$ be a *random* monotonically non decreasing function such that

$$
F(x_{(i)}) = Z_{(i)} \, .
$$

Finally, let $y_i = Z_i + \xi_i$ where ξ_i are i.i.d $N(0, 1)$, and let the pair of unordered sets $\mathcal{X} = \{x_1, \ldots, x_n\}$, $\mathcal{Y} = \{y_1, \ldots, y_n\}$ be the uncoupled observations: The y_i's are i.i.d. from $\mu * N(0, 1)$ and we denote

by $\mathbb{P}_F$ their joint distribution. Similarly, denote by $\mathbb{P}_f$ the joint distribution of $y_1, \ldots, y_n$ when $y_i = f(x_i) + \xi_i$ and note that $\mathbb{P}_f$ need not be a *product* distribution: it is, in general, different from $(\pi_f * N(0,1))^{\otimes n}$. This is because the sampling mechanism of uncoupled isotonic regression that does not allow for replacement when sampling from the x_i's.

Let $\tilde{f}$ be any measurable function of $y_1, \ldots, y_n$. Fix a k to be chosen later, and let P and Q be the two distributions from Lemma 3. Then for any $r_n > 0$, recalling that F is a random function since it depends on $Z_1, \ldots, Z_n$, it holds

$$\sup_g \mathbb{P}_g\left(W_1(\pi_g, \pi_{\tilde{f}}) > r_n\right) \geq \max\left\{\int \mathbb{P}_F\left(W_1(\pi_F, \pi_{\tilde{f}}) > r_n\right) dP^{\otimes n}(Z_1, \ldots, Z_n),\right.$$
$$\left.\int \mathbb{P}_F\left(W_1(\pi_F, \pi_{\tilde{f}}) > r_n\right) dQ^{\otimes n}(Z_1, \ldots, Z_n)\right\}, \qquad (12)$$

where the supremum is taken over all non-decreasing functions g from $[0,1]$ to $[-V, V]$.

Observe first that the two mixture distributions that appear above are, in fact, product distributions: for any event $\mathcal{A}$ in the sigma-algebra generated by $y_1, \ldots, y_n$,

$$\int \mathbb{P}_F(\mathcal{A}) dP^{\otimes n}(Z_1, \ldots, Z_n) = P_*^{\otimes n}(\mathcal{A}) \qquad \text{and} \qquad \int \mathbb{P}_F(\mathcal{A}) dQ^{\otimes n}(Z_1, \ldots, Z_n) = Q_*^{\otimes n}(\mathcal{A}),$$

where $P_* = P * N(0,1)$ and $Q_* = Q * N(0,1)$.

For any measure μ on $[-V, V]$, note that

$$\pi_F = \frac{1}{n} \sum_{i=1}^n \delta_{Z_i},$$

where the Z_is are i.i.d from μ. Thus by [BL16, Theorem 3.2] we obtain

$$\int W_1(\pi_F, \mu) d\mu^{\otimes n}(Z_1, \ldots, Z_n) \leq \frac{V}{\sqrt{n}},$$

which yields via Markov's inequality and the triangle inequality that

$$\int \mathbb{P}_F(W_1(\pi_F, \pi_{\hat{f}}) > r_n) d\mu^{\otimes n} \geq \int \mathbb{P}_F(W_1(\mu, \pi_{\hat{f}}) > 2r_n) d\mu^{\otimes n} - \int \mathbb{1}\{(W_1(\pi_F, \mu) > r_n)\} d\mu^{\otimes n}$$
$$\geq \int \mathbb{P}_F(W_1(\mu, \pi_{\hat{f}}) > 2r_n) d\mu^{\otimes n} - \frac{V}{r_n\sqrt{n}}.$$

Next, if $W_1(P, Q) \geq 4r_n$, we get from the triangle inequality that

$$\mathbb{P}_F\left(W_1(Q, \pi_{\tilde{f}}) > 2r_n\right) \geq \mathbb{P}_F\left(W_1(P, \pi_{\tilde{f}}) \leq 2r_n\right).$$

Combining (12) with the above two displays yields

$$\sup_g \mathbb{P}_g\left(W_1(\pi_g, \pi_{\tilde{f}}) > r_n\right) \geq \max\left\{P_*^{\otimes n}\left(W_1(P, \pi_{\tilde{f}}) > 2r_n\right), Q_*^{\otimes n}\left(W_1(P, \pi_{\tilde{f}}) \leq 2r_n\right)\right\} - \frac{2V}{r_n\sqrt{n}}$$
$$= \frac{1}{2}\left(1 - \text{TV}\left(P_*^{\otimes n}, Q_*^{\otimes n}\right)\right) - \frac{2V}{r_n\sqrt{n}}.$$

By Lemma 11, we have that

$$\text{TV}(P_*^{\otimes n}, Q_*^{\otimes n})^2 \leq \left(1 + e^{5V^2/2}(eV^2/k)^k\right)^n - 1.$$

Choosing $k = c_1 \frac{\log n}{\log \log n}$ and $r_n = c_2 V/k$ for suitable constants c_1 and c_2, we obtain

$$\sup_g \mathbb{P}_g \left(W_1(\pi_g, \pi_{\hat{f}}) > c_1^{-1} c_2 V \frac{\log \log n}{\log n} \right) \geq \frac{1}{2} - o_V(1),$$

and the claim follows. □

5. CONCLUSION

Our results establish that uncoupled isotonic regression can surprisingly be solved without further assumptions on the regression function f. However, as in nonparametric deconvolution, minimax rates are much slower than standard isotonic regression. One conclusion of the mixture learning literature is that significantly better results are possible when the original measure has small support [HK15, WY18]. In the context of uncoupled regression, this suggests that better rates may be available when the regression function f is piecewise constant with a small number of pieces, an assumption which also improves rates of estimation under the standard isotonic regression model [BT15]. Additional smoothness assumptions or more restrictive shape constraints may also lead to better rates. We leave this question to future work.

In this work, we have restricted ourselves to the univariate problem. Recent work [HWCS17] has considered the generalization of isotonic regression in which the regression function is a coordinate-wise nondecreasing function on $[0, 1]^d$. Extending our results to the multidimensional setting is another interesting future research direction.

FUNDING

ACKNOWLEDGEMENTS

APPENDIX A: OMITTED PROOFS

A.1 Proof of Proposition 2

In the following proof, the symbol C will represent a universal constant whose value may change from line to line. We will show that the estimator $\hat{g}$ satisfies

$$W_2(\pi_{\hat{g}} * \mathcal{D}, \hat{\pi}) \leq W_2(\pi_{\hat{f}} * \mathcal{D}, \hat{\pi}) + C(\sigma + V)n^{-1/4}.$$

Following the proof of Proposition 4, this implies

$$\mathbb{E}W_2(\pi_{\hat{g}} * \mathcal{D}, \pi_f * \mathcal{D}) \leq 2\mathbb{E}W_2(\pi_f * \mathcal{D}, \hat{\pi}) + C(\sigma + V)n^{-1/4} \lesssim (\sigma + V)n^{-1/4},$$

or, in other words, that $\hat{g}$ satisfies the same inequality as $\hat{f}$ does (Proposition 4), up to constants. Since the inequality in Proposition 4 is the only fact about $\hat{f}$ used in the proof of the upper bound, this will serve to establish the claim.

We first analyze the solution $\hat{\mu}$ to (5). Let $\hat{\nu} \in \text{argmin}_{\nu \in \mathcal{M}_V} W_2^2(\nu * \mathcal{D}, \hat{\pi})$, where the minimization is taken over the set $\mathcal{M}_V$ of all measures on $[-V, V]$ rather than over the set $\mathcal{M}_{\mathcal{A},V}$. By Lemma 5, there exists a $\hat{\nu}' \in \mathcal{M}_{\mathcal{A},V}$ such that

$$W_2(\hat{\nu}', \hat{\nu}) \le (V + \sigma) n^{-1/4}. \tag{13}$$

Moreover, by Lemma 6, we have that for all $\mu \in \mathcal{M}_{\mathcal{A},V}$,

$$W_2(\Pi_{\mathcal{A}\sharp}(\mu * \mathcal{D}), \mu * \mathcal{D}) \le C(V + \sigma) n^{-1/4}. \tag{14}$$

Combining these inequalities yields

$$
\begin{aligned}
W_2(\hat{\mu} * \mathcal{D}, \hat{\pi}) &\le W_2(\Pi_{\mathcal{A}\sharp}(\hat{\mu} * \mathcal{D}), \hat{\pi}) + C(V + \sigma) n^{-1/4} && \text{(triangle inequality and (14))} \\
&\le W_2(\Pi_{\mathcal{A}\sharp}(\hat{\nu}' * \mathcal{D}), \hat{\pi}) + C(V + \sigma) n^{-1/4} && \text{(optimality of } \hat{\mu}) \\
&\le W_2(\hat{\nu}' * \mathcal{D}, \hat{\pi}) + C(V + \sigma) n^{-1/4} && \text{(triangle inequality and (14))} \\
&\le W_2(\hat{\nu} * \mathcal{D}, \hat{\pi}) + C(V + \sigma) n^{-1/4} && \text{(triangle inequality and (13))} \\
&\le W_2(\pi_{\hat{f}} * \mathcal{D}, \hat{\pi}) + C(V + \sigma) n^{-1/4}, && \text{(optimality of } \hat{\nu})
\end{aligned}
$$

where we have used in the fourth step the fact that, for *any* two measures α and β,

$$W_2(\alpha * \mathcal{D}, \beta * \mathcal{D}) \le W_2(\alpha, \beta).$$

(See, e.g., [San15], Lemma 5.2.)

Finally, by Lemma 7, we have

$$W_2(\hat{\mu}, \pi_{\hat{g}}) \lesssim 2V n^{-1/2}.$$

Therefore, by another application of the triangle inequality, we obtain

$$W_2(\pi_{\hat{g}} * \mathcal{D}, \hat{\pi}) \le W_2(\hat{\mu} * \mathcal{D}, \hat{\pi}) + 2V n^{-1/2} \le W_2(\pi_{\hat{f}} * \mathcal{D}, \hat{\pi}) + C(V + \sigma) n^{-1/4},$$

as claimed.

A.2 Proof of Theorem 4

The proof of Theorem 4 depends on convolving the measures μ and ν with a kernel with specific smoothness and decay properties. This kernel is related to the well-known sinc kernel [Tsy09], and coincides with a kernel proposed for deconvolution with respect to Wasserstein distance [DM13].

We define the kernel as follows. For any positive integer m, denote by $\mathcal{S}_m$ the distribution on $\mathbb{R}$ with density function

$$f_m(t) := \begin{cases} C_m \left(\frac{\sin(t/4em)}{t/4em} \right)^{2m} & \text{if } t \ne 0, \\ C_m & \text{if } t = 0, \end{cases} \tag{15}$$

where C_m is positive a constant chosen so that $\int_{-\infty}^{\infty} f_m(t) \, dt = 1$. Lemma 8 establishes $C_m \le 1$.

We require two properties of the distribution $\mathcal{S}_m$:

(i) that it possesses sufficiently many moments, and
(ii) that the successive derivatives of the density f_m decay sufficiently quickly.

To see that (i) holds, note that for any $p \leq 2m - 2$, since $f_m(t) \leq 1 \wedge (t/4em)^{-2m}$, it holds

$$\mathbb{E}[|S|^p] = 2 \int_0^\infty t^p f_m(t)\, \mathrm{d}t \leq 2 \int_0^{4em} t^p\, \mathrm{d}t + 2(4em)^{2m} \int_{4em}^\infty t^{p-2m}\, \mathrm{d}t \leq 4(4em)^{p+1} \leq 4(4em)^{p+1},$$
(16)

The requirement (ii) on the successive derivatives of f_m is ensured by the following lemma. Its proof is presented in Appendix A.4.

LEMMA 4. *The function f_m is analytic on $\mathbb{R}$ and satisfies*

$$|f_m^{(n)}(t)| \leq \frac{(8em)^{2m}}{(2e)^n (4em + |t|)^{2m}} \qquad \forall n \geq 0\,.$$

With these two lemmas, we can establish the claimed result.

PROOF OF THEOREM 4. The assumption that μ and ν have finite moment generating functions implies that $\sup_{\ell \geq 1} \frac{\Delta_\ell}{\ell} < \infty$. Since the statement of the theorem is scale-invariant, it suffices to prove the claim in the case that $\sup_{\ell \geq 1} \frac{\Delta_\ell}{\ell} = 1$, where the claimed bound simplifies to $W_p(\mu, \nu) \lesssim p$. Also, because $W_q \leq W_p$ for $q \leq p$, we can assume without loss of generality that p is a positive even integer.

Set $m := p/2 + 1$. Let $X \sim \mu$ and $Y \sim \nu$, and let $S \sim \mathcal{S}_m$ be independent of X and Y. We denote by $\tilde{\mu}$ the distribution of $X + S$ and by $\tilde{\nu}$ the distribution of $Y + S$.

By the triangle inequality applied to W_p,

$$W_p(\mu, \nu) \leq W_p(\mu, \tilde{\mu}) + W_p(\nu, \tilde{\nu}) + W_p(\tilde{\mu}, \tilde{\nu})\,.$$
(17)

Since $(X, X + S)$ is a valid coupling between μ and $\tilde{\mu}$, by (16), it holds

$$W_p(\mu, \tilde{\mu}) \leq (\mathbb{E}|X - X - S|^p)^{1/p} \leq (4(4em)^{p+1})^{1/p} \lesssim p\,.$$
(18)

It remains to bound the final term. Denote by $f_{\tilde{\mu}}$ and $f_{\tilde{\nu}}$ the densities of $\tilde{\mu}$ and $\tilde{\nu}$, respectively. By [Vil08], Theorem 6.15,

$$W_p^p(\tilde{\mu}, \tilde{\nu}) \leq 2^{p-1} \int_{-\infty}^\infty |t|^p |f_{\tilde{\mu}}(t) - f_{\tilde{\nu}}(t)|\, \mathrm{d}t\,.$$

The definitions of f_μ and f_ν imply

$$f_{\tilde{\mu}}(t) - f_{\tilde{\nu}}(t) = \mathbb{E}\left[f_m(t - X) - f_m(t - Y) \right] = \mathbb{E} \sum_{\ell=1}^\infty \frac{f_m^{(\ell)}(t)(X^\ell - Y^\ell)}{\ell!} = \sum_{\ell=1}^\infty \frac{f_m^{(\ell)}(t)\mathbb{E}(X^\ell - Y^\ell)}{\ell!}\,,$$

where in the last step we used Fubini's theorem since μ and ν have moment generating functions that are finite everywhere. By applying successively the assumption that $\sup_{\ell \geq 1} \Delta_\ell/\ell \leq 1$, Lemma 4, and Stirling's approximation, we obtain

$$|f_{\tilde{\mu}}(t) - f_{\tilde{\nu}}(t)| \leq \sum_{\ell=1}^\infty \frac{|f_m^{(\ell)}(t)|\ell^\ell}{\ell!} \leq \frac{(8em)^{2m}}{(4em + |t|)^{2m}} \sum_{\ell=1}^\infty \frac{\ell^\ell}{(2e)^\ell \ell!} \leq \frac{(8em)^{2m}}{(4em + |t|)^{2m}}\,.$$

Therefore, recalling that $2m = p + 2$, we obtain

$$W_p^p(\tilde{\mu}, \tilde{\nu}) \leq 2^{p-1}(4\mathrm{e}(p+2))^{p+2} \int_{-\infty}^{\infty} \frac{|t|^p}{(2\mathrm{e}(p+2) + |t|)^{p+2}} \, \mathrm{d}t$$

$$= \frac{2^p(4\mathrm{e}(p+2))^{p+2}}{2\mathrm{e}(p+1)(p+2)} \leq (cp)^p \,, \tag{19}$$

where c is a universal constant.

Combining (18) and (19) with (17) yields

$$W_p(\mu, \nu) \lesssim p = p\left(\sup_{\ell \geq 1} \frac{\Delta_\ell}{\ell}\right),$$

as claimed. □

A.3 Proof of Lemma 1

The claim is trivial if $\|X\|_{\psi_1} = \infty$, so we assume $\|X\|_{\psi_1} < \infty$, and indeed, by homogeneity, we may assume $\|X\|_{\psi_1} = 1$. We have

$$\left(\frac{|X|}{p}\right)^p \leq (e^{|X|/p} - 1)^p \leq e^{|X|} - 1 = \psi_1(X),$$

so

$$\mathbb{E}|X|^p \leq p^p \mathbb{E}\psi_1(X) \leq p^p.$$

□

A.4 Proof of Lemma 4

The analyticity of f_m follows immediately from the well known fact that $\frac{\sin t}{t}$ is analytic, so it suffices to prove the derivative bound. The claim will follow from the fact that

$$\left|\frac{d^n}{dt^n}\left(\frac{\sin t}{t}\right)^{2m}\right| \leq \frac{2^{2m}(2m)^n}{(1 + |t|)^{2m}},$$

which we prove by induction on m. Recall that, for function f and g, the general Leibniz rule states

$$\frac{d^n}{dt^n} f(t)g(t) = \sum_{k=0}^n \binom{n}{k} f^{(n-k)}(t)g^{(k)}(t).$$

We therefore have

$$\left|\frac{d^n}{dt^n}\left(\frac{\sin t}{t}\right)^2\right| \leq \sum_{k=0}^n \binom{n}{k}\left|\frac{d^{n-k}}{dt^{n-k}}\frac{\sin t}{t}\right|\left|\frac{d^k}{dt^k}\frac{\sin t}{t}\right| = \frac{4}{(1+|t|)^2}\sum_{k=0}^n \binom{n}{k} = \frac{4 \cdot 2^n}{(1 + |t|)^2} \,,$$

where we have used Lemma 9 to bound the derivatives of $\frac{\sin t}{t}$. This proves the base case $m = 1$. By induction, for $m > 1$, we have

$$\left| \frac{d^n}{dt^n} \left(\frac{\sin t}{t} \right)^{2m} \right| \leq \sum_{k=0}^{n} \binom{n}{k} \left| \frac{d^{n-k}}{dt^{n-k}} \left(\frac{\sin t}{t} \right)^2 \right| \left| \frac{d^k}{dt^k} \left(\frac{\sin t}{t} \right)^{2m-2} \right|$$

$$\leq \sum_{k=0}^{n} \binom{n}{k} \frac{4 \cdot 2^{n-k}}{(1 + |t|)^2} \frac{2^{2m-2}(2m-2)^k}{(1 + |t|)^{2m-2}}$$

$$= \frac{2^{2m}}{(1 + |t|)^{2m}} \sum_{k=0}^{n} \binom{n}{k} 2^{n-k}(2m-2)^k$$

$$= \frac{2^{2m}(2m)^n}{(1 + |t|)^{2m}} .$$

The function $f_m(t)$ therefore satisfies

$$\left| \frac{d^n}{dx^n} f_m(t) \right| = C_m \left| \frac{d^n}{dx^n} \left(\frac{\sin(t/4em)}{t/4em} \right)^{2m} \right| \leq C_m(4em)^{-n} \frac{2^{2m}(2m)^n}{(1 + |t/4em|)^{2m}} \leq \frac{(8em)^{2m}}{(2e)^n(4em + |t|)^{2m}} ,$$

which concludes the proof. $\qquad \square$

A.5 Proof of Proposition 3

The first part is the content of Lemma 3. For the second part, for a given $\varepsilon \in (0, 1/2]$, denote by $\mathcal{P}_\varepsilon$ the distribution on $\mathbb{R}$ with density

$$f_\varepsilon(x) := c_\varepsilon \mathrm{e}^{-|x|^{\frac{1}{1-\varepsilon}}} ,$$

where $c_\varepsilon \leq 1$ is a suitable normalizing constant. Note that the moment generating function of $\mathcal{P}_\varepsilon$ is finite everywhere. Integrating f_ε implies that if $X \sim \mathcal{P}_\varepsilon$, then for all positive integers p,

$$(\mathbb{E}|X|^p)^{1/p} \asymp p^{1-\varepsilon} .$$

Denote by $\mathcal{P}'_\varepsilon$ the distribution of $2X$. The coupling $(X, 2X)$ is a monotone coupling between $\mathcal{P}_\varepsilon$ and $\mathcal{P}'_\varepsilon$, so by [San15, Theorem 2.9] we have

$$W_p(\mathcal{P}_\varepsilon, \mathcal{P}'_\varepsilon) = (\mathbb{E}|X - 2X|^p)^{1/p} \asymp p^{1-\varepsilon} ,$$

as claimed. $\qquad \square$

A.6 Proof of Lemma 2

We assume that $\|\mathcal{D}\|_{\psi_1} < \infty$, since otherwise the claim is vacuous. Write M_μ, M_ν, and $M_\mathcal{D}$ for the moment generating functions of μ, ν, and $\mathcal{D}$, respectively. We have

$$\Delta_\ell^\ell(\mu, \nu) = \left| \frac{d^\ell}{dt^\ell} M_\mu(t) - M_\nu(t) \right|_{t=0}$$

$$= \left| \frac{d^\ell}{dt^\ell} \frac{M_\mu(t)M_\mathcal{D}(t) - M_\nu(t)M_\mathcal{D}(t)}{M_\mathcal{D}(t)} \right|_{t=0}$$

$$\leq \sum_{m=0}^{\ell} \binom{\ell}{m} \left| \frac{d^{\ell-m}}{dt^{\ell-m}} M_\mu(t)M_\mathcal{D}(t) - M_\nu(t)M_\mathcal{D}(t) \right|_{t=0} \left| \frac{d^m}{dt^m} \frac{1}{M_\mathcal{D}(t)} \right|_{t=0}$$

$$\leq \left(\sum_{m=0}^{\ell} \binom{\ell}{m} \left| \frac{d^m}{dt^m} \frac{1}{M_\mathcal{D}(t)} \right|_{t=0} \right) \cdot \sup_{m \leq \ell} \Delta_m^m(\mu * \mathcal{D}, \nu * \mathcal{D})$$

If $X \sim \mathcal{D}$, then

$$|M_{\mathcal{D}}(t) - 1| = |\mathbb{E}e^{tX} - 1| \leq \mathbb{E}|e^{tX} - 1| \leq \mathbb{E}e^{t|X|} - 1 = \mathbb{E}\psi_1(t|X|),$$

which implies in particular that

$$|M_{\mathcal{D}}(t) - 1| \leq \frac{1}{2} \qquad \forall t \leq (2\|\mathcal{D}\|_{\psi_1})^{-1}.$$

The function $\frac{1}{M_{\mathcal{D}}(t)}$ is therefore analytic and bounded in norm by 2 on a disk of radius $(2\|\mathcal{D}\|_{\psi_1})^{-1}$ around the origin. Standard results from complex analysis (see, e.g., [FS09], Proposition IV.1) then imply that

$$\left| \frac{d^m}{dt^m} \frac{1}{M_{\mathcal{D}}(t)} \right|_{t=0} \leq m!(2\|\mathcal{D}\|_{\psi_1})^m.$$

Combining this with the above bound yields

$$\Delta_\ell^\ell(\mu, \nu) \leq \ell!(4\|\mathcal{D}\|_{\psi_1})^\ell \cdot \sup_{m \leq \ell} \Delta_m^m(\mu * \mathcal{D}, \nu * \mathcal{D}),$$

and the claim follows. □

APPENDIX B: SUPPLEMENTAL LEMMAS

LEMMA 5. *For all $\nu \in \mathcal{M}_V$, there exists a $\nu' \in \mathcal{M}_{\mathcal{A},V}$ such that*

$$W_2(\nu, \nu') \leq (V + \sigma)n^{-1/4}.$$

PROOF. Let $\Pi_{\mathcal{A},V}$ be the map sending each point in $[-V, V]$ to the nearest point in $\mathcal{A} \cap [-V, V]$, and set $\nu' := \Pi_{\mathcal{A},V\sharp}\nu$. Clearly $\nu' \in \mathcal{M}_{\mathcal{A},V}$, and

$$W_2^2(\nu, \nu') \leq \sup_{x \in [-V,V]} |x - \Pi_{\mathcal{A},V}(x)|^2 \leq \frac{(V + \sigma)^2}{n^{1/2}},$$

which proves the claim. □

LEMMA 6. *For all $\mu \in \mathcal{M}_V$,*

$$W_2(\Pi_{\mathcal{A}\sharp}(\mu * \mathcal{D}), \mu * \mathcal{D}) \lesssim (V + \sigma)n^{-1/4}.$$

PROOF. By the definition of the Wasserstein distance, we have

$$W_2^2(\Pi_{\mathcal{A}\sharp}(\mu * \mathcal{D}), \mu * \mathcal{D}) \leq \mathbb{E}|\Pi_{\mathcal{A}}(X + \xi) - X - \xi|^2 \qquad X \sim \mu, \xi \sim \mathcal{D}.$$

If $|(X + \xi)| \leq (V + \sigma)\log n$, then $|\Pi_{\mathcal{A}}(X + \xi) - X - \xi|^2 \leq (V + \sigma)^2 n^{-1/2}$, which implies

$$\mathbb{E}|\Pi_{\mathcal{A}}(X + \xi) - X - \xi|^2 \leq (V + \sigma)n^{-1/4} + \mathbb{E}[|\alpha_N - X - \xi|^2 \mathbb{1}_{X+\xi > \alpha_N}] + \mathbb{E}[|\alpha_0 - X - \xi|^2 \mathbb{1}_{X+\xi < \alpha_0}]$$

$$\leq (V + \sigma)^2 n^{-1/2} + \mathbb{E}[|X + \xi|^2 \mathbb{1}_{|X+\xi| > (V+\sigma)\log n}]$$

$$\leq (V + \sigma)^2 n^{-1/2} + \mathbb{E}[|X + \xi|^4]^{1/2} \mathbb{P}[|X| > \sigma \log n]^{1/2},$$

where the last step uses the Cauchy-Schwarz inequality.

The assumption that $\|X\|_{\psi_1} \leq \sigma$ implies

$$\mathbb{E}[|X + \xi|^4]^{1/2} \lesssim (V + \sigma)^2$$

and

$$\mathbb{P}[|X| > \sigma \log n] \leq \mathbb{P}[e^{|X|/\sigma} > n] \leq \frac{2}{n}\,.$$

Combining the above three displays yields

$$\mathbb{E}|\Pi_{\mathcal{A}}(X + \xi) - X - \xi|^2 \lesssim (V + \sigma)^2 n^{-1/2}\,,$$

and this implies the stated bound.   □

LEMMA 7.   *Let μ be any measure on $[-V, V]$ with quantile function $\mathcal{Q}_\mu$, and let $g \in \mathcal{F}_V$ satisfy $g(x_i) = \mathcal{Q}_\mu(i/n)$ for $1 \leq i \leq n$. Then*

$$W_2(\mu, \pi_g) \leq 2V n^{-1/2}\,.$$

PROOF.   The definition of π_g implies that the quantile function $\mathcal{Q}_{\pi_g}$ of π_g satisfies

$$\mathcal{Q}_{\pi_g}(x) = \mathcal{Q}_\mu(i/n) \qquad \text{where } (i-1)/n < x \leq i/n\,.$$

Since μ is supported on $[-V, V]$, we set $\mathcal{Q}_\mu(0) := \lim_{p \to 0+} \mathcal{Q}_\mu(p) \geq -V$. By the explicit representation for the Wasserstein distance between one-dimensional measures [BL16, Theorem 2.10], we have

$$
\begin{aligned}
W_2^2(\mu, \pi_g) &= \int_0^1 |\mathcal{Q}_\mu(x) - \mathcal{Q}_{\pi_g}(x)|^2\, \mathrm{d}x \\
&= \sum_{i=1}^n \int_{(i-1)/n}^{i/n} |\mathcal{Q}_\mu(x) - \mathcal{Q}_\mu(i/n)|^2\, \mathrm{d}x \\
&\leq \sum_{i=1}^n \int_{(i-1)/n}^{i/n} |\mathcal{Q}_\mu((i-1)/n) - \mathcal{Q}_\mu(i/n)|^2\, \mathrm{d}x \\
&\leq \frac{1}{n} \sum_{i=1}^n |\mathcal{Q}_\mu((i-1)/n) - \mathcal{Q}_\mu(i/n)|^2 \\
&\leq \frac{2V}{n} \sum_{i=1}^n \mathcal{Q}_\mu(i/n) - \mathcal{Q}_\mu((i-1)/n) \leq \frac{(2V)^2}{n}\,.
\end{aligned}
$$

                                                                        □

LEMMA 8.   *Let f_m be defined as in* (15). *If $\int_{-\infty}^\infty f_m(t)\, \mathrm{d}t = 1$, then*

$$C_m \leq 1\,.$$

PROOF.   It suffices to show that $\int_{-\infty}^\infty \left(\frac{\sin(t/4em)}{t/4em} \right)^{2m} \mathrm{d}t \geq 1$. The inequality $|\sin(t)| \geq |t| - \frac{|t|^3}{6}$ implies

$$\left(\frac{\sin(t/4em)}{t/4em} \right)^{2m} \geq \left(1 - \frac{t^2}{6(4em)^2} \right)^{2m} \geq 1 - \frac{t^2}{48e^2 m} \geq \frac{1}{2} \quad \text{if } t^2 \leq 24e^2 m\,.$$

Therefore

$$\int_{-\infty}^{\infty} \left(\frac{\sin(t/4em)}{t/4em} \right)^{2m} \mathrm{d}t \geq \int_{-1}^{1} \frac{1}{2} \mathrm{d}t \geq 1 \,.$$

$\square$

LEMMA 9.

$$\left| \frac{d^n}{dt^n} \frac{\sin t}{t} \right| \leq \frac{2}{1 + |t|} \qquad \forall n \geq 0 \,.$$

PROOF. Recall that

$$\frac{\sin t}{t} = \int_0^1 \cos(tx) \, \mathrm{d}x \,,$$

which implies after differentiating under the integral that

$$\frac{d^n}{dt^n} \frac{\sin t}{t} = \int_0^1 x^n \cos^{(n)}(tx) \, \mathrm{d}x \,.$$

Since $|\cos^{(n)}(tx)| \leq 1$, we obtain immediately that

$$\left| \frac{d^n}{dt^n} \frac{\sin t}{t} \right| \leq \int_0^1 x^n \, \mathrm{d}x = \frac{1}{n+1} \,,$$

which proves the claim when $|t| \leq 2n + 1$.

To prove the claim when $|t| > 2n + 1$, we proceed by induction. When $n = 0$ and $|t| > 2n + 1 = 1$, the bound $|\sin(t)| \leq 1$ implies

$$\left| \frac{\sin t}{t} \right| \leq \frac{1}{t} \leq \frac{2}{1 + |t|} \,.$$

We now assume that the bound in question holds for $n - 1$ and all t. Integrating by parts and applying the induction hypothesis yields

$$\left| \int_0^1 x^n \cos^{(n)}(tx) \right| = \left| \frac{\cos^{(n-1)} t}{t} - \frac{n}{t} \int_0^1 x^{n-1} \cos^{(n-1)}(tx) \, \mathrm{d}x \right|$$

$$\leq \frac{1}{|t|} + \frac{n}{|t|} \frac{2}{1 + |t|} = \frac{|t| + 2n + 1}{|t|(1 + |t|)} \,.$$

Since $|t| > 2n + 1$, this quantity is smaller than $\frac{2}{1+|t|}$, as claimed. $\square$

LEMMA 10. *Let μ be any distribution satisfying $\|\mu\|_{\psi_1} \leq K$, and let $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$, where $X_i \sim \mu$ are i.i.d. Then*

$$\mathbb{E}[W_2^2(\mu, \hat{\mu})] \leq \frac{16K^2}{\sqrt{n}} \,.$$

PROOF. We assume without loss of generality that μ is centered. By [BL16, Theorem 7.16],

$$\mathbb{E}[W_2^2(\mu, \hat{\mu})] \leq \frac{4}{\sqrt{n}} \int_{-\infty}^{\infty} |x| \sqrt{F(x)(1 - F(x))} \, \mathrm{d}x \,,$$

where F is the CDF of the measure μ. Let $X \sim \mu$. Then

$$F(x)(1 - F(x)) = \mathbb{P}[X \leq x]\mathbb{P}[X > x] \leq \mathbb{P}[|X| \geq |x|] \leq 2e^{-|x|/K} \,,$$

where in the last step we have used the fact that $\|\mu\|_{\psi_1} \le K$. We obtain

$$\mathbb{E}[W_2^2(\mu, \hat{\mu})] \le \frac{16}{\sqrt{n}} \int_0^\infty x e^{-x/K} \, \mathrm{d}x = \frac{16K^2}{\sqrt{n}} \,.$$

$\square$

LEMMA 11.   *[CL11] If P and Q are two centered measures supported on $[-V, V]$ such that $\Delta_\ell(P, Q) = 0$ for $\ell = 1, \dots, k-1$, then*

$$\mathrm{TV}((P * N(0,1))^{\otimes n}, (Q * N(0,1))^{\otimes n})^2 \le \left(1 + e^{5V^2/2} \frac{(2V^2)^k}{k!}\right)^n - 1\,.$$

PROOF.   By [CL11], proof of Theorem 3, (see also [WY18], Lemma 14), if P and Q are supported on $[-V, V]$, then

$$\chi^2(P * N(0,1), Q * N(0,1)) \le e^{V^2/2} \sum_{\ell=1}^\infty \frac{\Delta_\ell^{2\ell}}{\ell!} \,.$$

By assumption, $\Delta_\ell(P, Q) = 0$ for $\ell < k$, and for $\ell \ge k$ the fact that P and Q are supported on $[-V, V]$ implies $\Delta_\ell^\ell \le (2V)^\ell$. Combining these bounds yields

$$\chi^2(P * N(0,1), Q * N(0,1)) \le e^{V^2/2} \sum_{\ell \ge k} \frac{(2V)^{2\ell}}{\ell!} \le e^{5V^2/2} \frac{(2V)^{2k}}{k!} \le e^{5V^2/2} \left(\frac{4eV^2}{k}\right)^k \,,$$

where in the last step we have applied Stirling's approximation.

The claim then follows from standard properties of the χ^2-divergence [Tsy09]. $\square$

REFERENCES

[ABRW18] Jason Altschuler, Francis Bach, Alessandro Rudi, and Jonathan Weed, *Massively scalable sinkhorn distances via the nyström method*, arXiv preprint arXiv:1812.05189 (2018).

[APZ17]   Abubakar Abid, Ada Poon, and James Zou, *Linear regression with shuffled labels*, arXiv preprint arXiv:1705.01342 (2017).

[AWR17]   Jason Altschuler, Jonathan Weed, and Philippe Rigollet, *Near-linear time approximation algorithms for optimal transport via sinkhorn iteration*, Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, 2017, pp. 1961–1971.

[BBBB72]   R. E. Barlow, D. J. Bartholomew, J. M. Bremner, and H. D. Brunk, *Statistical inference under order restrictions. The theory and application of isotonic regression*, John Wiley & Sons, London-New York-Sydney, 1972, Wiley Series in Probability and Mathematical Statistics. MR0326887

[BBR06]   Federico Bassetti, Antonella Bodini, and Eugenio Regazzini, *On minimum kantorovich distance estimators*, Statistics & Probability Letters **76** (2006), no. 12, 1298–1302.

[Bel18]   Pierre C. Bellec, *Sharp oracle inequalities for Least Squares estimators in shape restricted regression*, Ann. Statist. **46** (2018), no. 2, 745–780. MR3782383

[BL16]   Sergey Bobkov and Michel Ledoux, *One-dimensional empirical measures, order statistics and kantorovich transport distances*, Preprint, to appear in the Memoirs of the Amer. Math. Soc. (2016).

[BRW17]   Afonso S. Bandeira, Philippe Rigollet, and Jonathan Weed, *Optimal rates of estimation for multi-reference alignment*, arXiv:1702.08546 (2017).

[BT15]   Pierre C. Bellec and Alexandre B. Tsybakov, *Sharp oracle bounds for monotone and convex regression through aggregation*, J. Mach. Learn. Res. **16** (2015), 1879–1892. MR3417801

[CCDM11] Claire Caillerie, Frédéric Chazal, Jérôme Dedecker, and Bertrand Michel, *Deconvolution for the Wasserstein metric and geometric inference*, Electron. J. Stat. **5** (2011), 1394–1423. MR2851684

[Che95]   Jiahua Chen, *Optimal rate of convergence for finite mixture models*, The Annals of Statistics (1995), 221–233.

[CL11]   T Tony Cai and Mark G. Low, *Testing composite hypotheses, hermite polynomials and optimal estimation of a nonsmooth functional*, The Annals of Statistics **39** (2011), no. 2, 1012–1041.

[CS16]   Alexandra Carpentier and Teresa Schlueter, *Learning relationships between data obtained independently*, Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (Cadiz, Spain) (Arthur Gretton and Christian C. Robert, eds.), Proceedings of Machine Learning Research, vol. 51, PMLR, 09–11 May 2016, pp. 658–666.

[Cut13]   Marco Cuturi, *Sinkhorn distances: Lightspeed computation of optimal transport*, Advances in Neural Information Processing Systems 26 (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, eds.), Curran Associates, Inc., 2013, pp. 2292–2300.

[CV17]   A. Carpentier and N. Verzelen, *Adaptive estimation of the sparsity in the Gaussian vector model*, ArXiv: 1703.00167 (2017).

[DFM15]   Jérôme Dedecker, Aurélie Fischer, and Bertrand Michel, *Improved rates for Wasserstein deconvolution with ordinary smooth error in dimension one*, Electron. J. Stat. **9** (2015), no. 1, 234–265. MR3314482

[DG80]   Morris H. DeGroot and Prem K. Goel, *Estimation of the correlation coefficient from a broken random sample*, Ann. Statist. **8** (1980), no. 2, 264–278. MR560728

[DGJ11]   I. Dattner, A. Goldenshluger, and A. Juditsky, *On deconvolution of distribution functions*, Ann. Statist. **39** (2011), no. 5, 2477–2501. MR2906875

[DK68]   J. J. Deely and R. L. Kruse, *Construction of sequences estimating the mixing distribution*, Ann. Math. Statist. **39** (1968), 286–288. MR0220377

[DM13]   Jérôme Dedecker and Bertrand Michel, *Minimax rates of convergence for Wasserstein deconvolution with supersmooth errors in any dimension*, J. Multivariate Anal. **122** (2013), 278–291. MR3189324

[FHN+19] Aden Forrow, Jan-Christian Hütter, Mor Nitzan, Philippe Rigollet, Geoffrey Schiebinger, and Jonathan Weed, *Statistical Optimal Transport via Factored Couplings*, AISTATS (2019).

[FMR16]   Nicolas Flammarion, Cheng Mao, and Philippe Rigollet, *Optimal rates of statistical seriation*, Bernoulli (to appear) (2016).

[FS09]   Philippe Flajolet and Robert Sedgewick, *Analytic combinatorics*, Cambridge University Press, Cambridge, 2009. MR2483235

[HK15]   Philippe Heinrich and Jonas Kahn, *Optimal rates for finite mixture estimation*, arXiv preprint arXiv:1507.04313 (2015).

[HSS17]   Daniel J. Hsu, Kevin Shi, and Xiaorui Sun, *Linear regression without correspondence*,

Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA (Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, eds.), 2017, pp. 1530–1539.

[HWCS17] Qiyang Han, Tengyao Wang, Sabyasachi Chatterjee, and Richard J Samworth, *Isotonic regression in general dimensions*, arXiv preprint arXiv:1708.09468 (2017).

[JN02]   Anatoli Juditsky and Arkadi Nemirovski, *On nonparametric tests of positivity/monotonicity/convexity*, Ann. Statist. **30** (2002), no. 2, 498–527.

[KV17]   Weihao Kong and Gregory Valiant, *Spectrum estimation from samples*, Ann. Statist. **45** (2017), no. 5, 2218–2247. MR3718167

[LNS99]   O. Lepski, A. Nemirovski, and V. Spokoiny, *On estimation of the L_r norm of a regression function*, Probab. Theory Related Fields **113** (1999), no. 2, 221–253. MR1670867

[Mam91]   Enno Mammen, *Estimating a smooth monotone regression function*, Ann. Statist. **19** (1991), no. 2, 724–740. MR1105841

[MV10]   Ankur Moitra and Gregory Valiant, *Settling the polynomial learnability of mixtures of gaussians*, Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on, IEEE, 2010, pp. 93–102.

[MW00]   Mary Meyer and Michael Woodroofe, *On the degrees of freedom in shape-restricted regression*, Ann. Statist. **28** (2000), no. 4, 1083–1104.

[NPT85]   A. S. Nemirovskiĭ, B. T. Polyak, and A. B. Tsybakov, *The rate of convergence of nonparametric estimates of maximum likelihood type*, Problemy Peredachi Informatsii **21** (1985), no. 4, 17–33. MR820705

[PC18]   G. Peyré and M. Cuturi, *Computational Optimal Transport*, ArXiv:1803.00567 (2018).

[PP14]   Georg Ch. Pflug and Alois Pichler, *Multistage stochastic optimization*, Springer Series in Operations Research and Financial Engineering, Springer, Cham, 2014. MR3288310

[PWC17]   Ashwin Pananjady, Martin J. Wainwright, and Thomas A. Courtade, *Denoising linear models with permuted data*, 2017 IEEE International Symposium on Information Theory, ISIT 2017, Aachen, Germany, June 25-30, 2017, IEEE, 2017, pp. 446–450.

[RW18]   Philippe Rigollet and Jonathan Weed, *Entropic optimal transport is maximum-likelihood deconvolution*, Comptes Rendus Mathematique **356** (2018), no. 11, 1228 – 1235.

[RWD88]   Tim Robertson, F. T. Wright, and R. L. Dykstra, *Order restricted statistical inference*, Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons, Ltd., Chichester, 1988. MR961262

[San15]   Filippo Santambrogio, *Optimal transport for applied mathematicians*, Birkäuser, NY (2015).

[SBD17]   Martin Slawski and Emanuel Ben-David, *Linear regression with sparsely permuted data*, arXiv preprint arXiv:1710.06030 (2017).

[Tsy09]   Alexandre B. Tsybakov, *Introduction to nonparametric estimation*, Springer Series in Statistics, Springer, New York, 2009, Revised and extended from the 2004 French original, Translated by Vladimir Zaiats. MR2724359 (2011g:62006)

[UHV18]   Jayakrishnan Unnikrishnan, Saeid Haghighatshoar, and Martin Vetterli, *Unlabeled sensing with random linear measurements*, IEEE Trans. Information Theory **64** (2018), no. 5, 3237–3253.

[vdG90]   Sara van de Geer, *Estimating a regression function*, Ann. Statist. **18** (1990), no. 2, 907–924. MR1056343

[vdG93] _____ , *Hellinger-consistency of certain nonparametric maximum likelihood estimators*, Ann. Statist. **21** (1993), no. 1, 14–44. MR1212164

[Ver18] Roman Vershynin, *High-dimensional probability*, Cambridge University Press (to appear), 2018.

[Vil03] Cédric Villani, *Topics in optimal transportation*, Graduate Studies in Mathematics, vol. 58, American Mathematical Society, Providence, RI, 2003. MR1964483 (2004e:90003)

[Vil08] Cédric Villani, *Optimal transport: old and new*, vol. 338, Springer Science & Business Media, 2008.

[WY18] Yihong Wu and Pengkun Yang, *Optimal estimation of gaussian mixtures via denoised method of moments*, Tech. report, working paper, 2018.

[Zha02] Cun-Hui Zhang, *Risk bounds in isotonic regression*, Ann. Statist. **30** (2002), no. 2, 528–555. MR1902898

[ZP18] Y. Zemmel and V. Panaretos, *Statistical aspects of wasserstein distances*, Annual Reviews of Statistics and its Applications, to appear. (2018).

Philippe Rigollet
Department of Mathematics
Statistics and Data Science Center
Massachusetts Institute of Technology
77 Massachusetts Avenue,
Cambridge, MA 02139-4307, USA
(rigollet@math.mit.edu)

Jonathan Weed
Department of Mathematics
Statistics and Data Science Center
Massachusetts Institute of Technology
77 Massachusetts Avenue,
Cambridge, MA 02139-4307, USA
(jweed@math.mit.edu)