Minimax estimation of smooth optimal transport maps

As Published: 10.1214/20-aos1997

Publisher: Institute of Mathematical Statistics

Persistent URL: https://hdl.handle.net/1721.1/133394

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing arXiv: 1905.05828

Minimax estimation of smooth optimal transport maps

Jan-Christian Hütter and Philippe Rigollet

Broad Institute and Massachusetts Institute of Technology

Abstract. Brenier's theorem is a cornerstone of optimal transport that guarantees the existence of an optimal transport map T between two probability distributions P and Q over $\mathbb{R}^d$ under certain regularity conditions. The main goal of this work is to establish the minimax estimation rates for such a transport map from data sampled from P and Q under additional smoothness assumptions on T. To achieve this goal, we develop an estimator based on the minimization of an empirical version of the semi-dual optimal transport problem, restricted to truncated wavelet expansions. This estimator is shown to achieve near minimax optimality using new stability arguments for the semi-dual and a complementary minimax lower bound. Furthermore, we provide numerical experiments on synthetic data supporting our theoretical findings and highlighting the practical benefits of smoothness regularization. These are the first minimax estimation rates for transport maps in general dimension.

AMS 2000 subject classifications: 62G05.
Key words and phrases: Optimal transport, nonparametric estimation, minimax rates, wavelet estimator.

1. INTRODUCTION

Wasserstein distances and the associated problem of optimal transport date back to the work of Gaspard Monge [Mon81] and have since then become important tools in pure and applied mathematics [Vil03, Vil09, San15]. Tools from optimal transport have been successfully employed in machine learning [AGCJ18, PC19, SHB+18, FCCR18, ACB17, GPC18, JCG18, MMC16, RCP16, SHB+17, GJB19, SCSJ17, AJJ18, ACB17, DBGGL19, CR12] computer graphics [LCCS18, SdGP+15, SPKS16, FCVP17], statistics [SC15, AGP18, RW18, WB19, ZP19, PZ19, BGK+17, CSB+18, RTC17, CS18, TM18, KTM18, BCP17, dBGLL19, KSS19], and, more recently, computational biology [SST+19, YDV+18].

Monge asked the following question: Given two probability measures P, Q in $\mathbb{R}^d$, how can we transport P to Q while minimizing the total distance traveled by this transport. A classical instantiation of this problem over $\mathbb{R}^d$ is to find a map $T_0 \colon \mathbb{R}^d \to \mathbb{R}^d$ that minimizes the objective

$$\min_T \int_{\mathbb{R}^d} \|T(x) - x\|_2^2 \, \mathrm{d}P(x), \quad \text{s.t. } T_\# P = Q, \tag{1.1}$$

which is known as the *Monge problem*, where $T_{\#}P$ denotes the push-forward of P under T, that is,

$$T_{\#}P(A) = P(T^{-1}(A)), \quad \text{for all Borel sets } A. \tag{1.2}$$

The highly non-linear constraint in (1.2) made the mathematical treatment of the Monge problem seem elusive for a long time, until the seminal work of Kantorovich [Kan42, Kan48], who considered the following relaxation. Instead of looking for a map T_0, look for a transport plan Γ_0 in the set of all possible probablity measures on $\mathbb{R}^d \times \mathbb{R}^d$ whose marginals coincide with P and Q, which we denote by $\Pi(P, Q)$. This leads to the optimization problem

$$\min_{\Gamma} \int \|x-y\|_2^2 \, \mathrm{d}\Gamma(x,y) \quad \text{s.t. } \Gamma \in \Pi(P,Q), \tag{1.3}$$

which is known as the *Kantorovich problem*, and whose value is the square of the 2-Wasserstein distance, $W_2^2(P,Q)$, between the two probability measures P and Q. The two optimization problems are indeed linked: Brenier's Theorem (Theorem 1 below), guarantees that under regularity assumptions on P, a solution Γ_0 to (1.3) is concentrated on the graph of a map T_0. That is, using a suggestive informal notation, $\Gamma_0(x,y) = P(x)\, \delta_{y=T_0(x)}$, where δ denotes a point mass. Moreover, T_0 is the gradient of a convex function f_0. While cost functions other than $\|x-y\|_2^2$ could be of interest, such as $\|x-y\|_2^p$ for $p \geq 1$, this work entirely focuses on the quadratic cost, which allows leveraging the well-established theory of convex functions and formulating key assumptions in terms of strong convexity.

Statistical optimal transport describes a body of questions that arise when the measures P and Q are unknown but samples are available. While the question of estimation of various quantities such as $W_2(P,Q)$, for example, are of central importance, for applications such as domain adaptation and data integration [DKF+18, CFT14, CFHR17, SDF+18, PCFH16, CFTR17, FHN+19, RW19], the main quantity of interest is the transport map T_0 itself since it can be used to push almost every point in the support of P to a point in the support of Q. Moreover, the optimal transport map plays an important role in characterizing the Riemannian geometry that arises from endowing probability measures that have finite second moments with the 2-Wasserstein-distance. In particular, it can be used to define the right-inverse to the exponential map in that space [Gig11], which in turn enables the generalization of PCA (principal component analysis) to spaces of probability measures [BGK+17, MPZ19]. The goal of this paper is to study the rates of estimation of a smooth transport map T_0 from samples.

To fix a concrete setup assume that we have at our disposal $2n$ independent observations $X_1, \dots, X_n$ from P and $Y_1, \dots, Y_n$ from Q, based on which we would like to find an estimator $\hat{T}$ for T_0. This statistical problem poses several challenges:
(i) The most straightforward estimator is obtained by replacing P and Q by their empirical counterparts [MC98]. It leads to a finite-dimensional linear problem that can be approximated very efficiently due to recent algorithmic advances [Cut13, AWR17, PC19, DGK18]. However, even if the resulting optimizer $\hat{\Gamma}$ is actually a map (matching), which it is not in general, it is not defined outside the sample points. In particular, it does not indicate how to transport a point $x \notin \{X_1, \dots, X_n\}$. In contrast, we would like to obtain an estimator $\hat{T}$ with guarantees in $L^2(P)$, that is, with convergence of

$$\|\hat{T} - T_0\|_{L^2(P)}^2 := \int \|\hat{T}(x) - T_0(x)\|_2^2 \, \mathrm{d}P(x)\,. \tag{1.4}$$

Note that such an estimator of T_0 could be obtained by post-processing the above optimizer $\hat{\Gamma}$, for example by interpolation techniques, see [ABM16]. We also employ related techniques in Section

6 to obtain practical estimators for T_0. However, we are not aware of a statistical analysis of these procedures.
(ii) It is known that the estimator $W_2^2(\hat{P}, \hat{Q})$ can be a poor proxy for $W_2^2(P, Q)$ if the underlying dimensionality of the distributions P and Q is large, as it suffers from the so-called *curse of dimensionality*. For example, if both P and Q are absolutely continuous with respect to the Lebesgue measure in $\mathbb{R}^d, d \geq 4$, it is known that up to logarithmic factors, $\mathbb{E}[|W_2(\hat{P}, \hat{Q}) - W_2(P, Q)|] \asymp n^{-1/d}$ (see [NWR19]). In fact, we show in Theorem 6 that without further assumptions on either P, Q, or T_0, no estimator can have an expected squared loss (1.4) uniformly better than $n^{-2/d}$.
(iii) While many heuristic approaches have been brought forward to address the previous point, a thorough statistical analysis of the rate of convergence has so far been lacking. This can be partly attributed to the structure (or lack thereof) of problem (1.3). Being a linear optimization problem, it lacks simple stability estimates that are key to establish statistical guarantees by relating $\|\hat{T} - T_0\|_{L^2(P)}$ to the sub-optimality gap

$$\int_{\mathbb{R}^d} \|\hat{T}(x) - x\|_2^2 \,\mathrm{d}P(x) - \int_{\mathbb{R}^d} \|T_0(x) - x\|_2^2 \,\mathrm{d}P(x).$$

In this paper, we aim to address these problems by imposing additional assumptions on the transport map T_0 that lead to a rate faster than $n^{-2/d}$. One assumption we impose on the transport map T_0 is smoothness, a standard way of alleviating the curse of dimensionality in non-parametric estimation. Another key assumption is based on an observation of Ambrosio published in an article by Gigli [Gig11]. They show that the optimization problem (1.1) has quadratic growth, in the sense of a stability estimate

$$\|T - T_0\|_{L^2(P)}^2 \lesssim \int_{\mathbb{R}^d} \|T(x) - x\|_2^2 \,\mathrm{d}P(x) - \int_{\mathbb{R}^d} \|T_0(x) - x\|_2^2 \,\mathrm{d}P(x), \tag{1.5}$$

provided $T_0 = \nabla f_0$ is Lipschitz continuous on $\mathbb{R}^d$ and $T_\# P = Q$. While this observation does not immediately lend itself to the analysis of an estimator due to the presence of the push-forward constraint, we show in Proposition 10 that under similar assumptions, the so-called semi-dual problem (see (2.5) below) admits a stability estimate.

Due to the rising interest in Optimal Transport as a tool for statistics and machine learning, many empirical regularization techniques have been proposed, ranging from the computationally successful entropic regularization [Cut13, GCPB16, AWR17], ℓ^2-regularization [BSR17], smoothness regularization [PCFH16], to regularization techniques specifically adapted to the application of domain adaptation [CFT14, CFTR17, CFHR17]. Notably, [GCPB16] also consider regularization based on the semi-dual objective and reproducing kernel Hilbert spaces. However, the statistical performance of these regularization techniques to estimate transport maps from sampled data has been largely unanswered, with the following exceptions.

The estimation of transport maps has been studied in the one-dimensional case under the name *uncoupled regression* [RW19] where the sample $Y_1, \ldots, Y_n$ is subject to measurement noise. There, the main statistical difficulty arises from the presence of this additional noise and boils down to obtaining deconvolution guarantees in the Wasserstein distance. Such guarantees were recently obtained under smoothness assumptions on the underlying density [CCDM11, DM13, DFM15] but they do not translate directly into rates of estimation for the optimal transport map beyond the 1D case. Note that in the presence of Gaussian measurement noise, the rates of estimation are likely to become logarithmic rather than polynomial as deconvolution is a statistically difficult task.

Concurrently to this work, [FLF19] study the estimation of linear transport maps and establish a fast rate of convergence in this parametric setup. Moreover, after finishing the first version of this paper, the authors became aware of the parallel work [Gun18]. There, Gunsilius analyzes the asymptotic variance of an estimator for f_0 under smoothness assumptions on the densities of the marginals P and Q and states the problem of obtaining estimation rates for the transport map T_0 explicitly as an open problem, which we address in this paper. A similarity between [Gun18] and this work is that the rates for the variance are obtained by applying empirical process theory to the semi-dual objective function. However, we note the following key differences: First, Gunsilius obtains curvature estimates for the semi-dual objective via variational techniques, while we use strong convexity of the candidate potentials. Note that under slightly stronger regularity assumptions, i.e., uniform convexity of the supports, his assumptions would imply strong convexity for the ground truth potential, as shown in [Gig11]. Second, by assuming smoothness of the transport map instead of the distributions P and Q, our results are more flexible and can be applied in cases where neither distribution possesses Hölder smooth densities. Third, by appealing to Cafarelli's global regularity theory, [Vil09, Theorem 12.50(iii)], [Caf92b, Caf92a, Caf96], we can obtain a variance rate of $n^{-2\alpha/(2\alpha-2+d)}$ (up to log factors) under assumptions quantitatively comparable to Gunsilius's, while his results imply a sub-optimal rate of $n^{-(2\alpha-2)/(2\alpha-2+d)}$, see Section E in the appendix.

We note that both in the application of Caffarelli's theorem in Appendix E and in the statement of our main result, Theorem 2 below, currently available analytical tools limit the extent to which minimax results can be established. In Appendix E, the lack of uniformity in Cafarelli's global regularity theory prevents us from claiming (near) minimax optimality over Hölder smooth densities, see Remark 36. Similarly, in Theorem 2, (near) minimax optimality is attained by considering fixed marginal supports because of the need for uniformly bounded constants in some of the classical inequalities we employ (for example, Poincaré inequality), see Remark 3. In effect, further strengthening these minimax results poses an interesting open problem involving deep analytical questions.

The rest of this paper is organized as follows. In Section 2, we review some important concepts of optimal transport, mainly duality and Brenier's theorem. These are instrumental in the definition of our estimator, which is postponed to Section 5. Indeed, since the main goal of this paper is to establish minimax rates of convergence for smooth transport maps, we present these rates in Section 3 and prove lower bounds in the following Section 4, since this proof illustrates well the source of the nonstandard exponent in the rates. We then proceed to Section 5 where we define a minimax optimal estimator constructed as follows. First, we define an estimator for the optimal Kantorovich potential as the solution to the empirical counterpart of the semi-dual problem restricted to a class of wavelet expansions. Then, our estimator is defined as the gradient of this potential. We prove that it achieves the near-optimal rate in the same section. In Section 6, we present numerical experiments on synthetic data, introducing two estimators that exploit smoothness of the transport map. The first illustrates that a version of the estimator considered in Section 5 can in fact be implemented, at least in low dimensions. The second is heuristically motivated and based on kernel-smoothing the transport plan between empirical distributions, showing that practical gains in higher dimensions can be achieved for smooth transport maps. Finally, some useful facts from convex analysis (Section A), approximation theory for wavelets (Section B), empirical process theory (Section C), and tools for proving lower bounds (Section D), are appendix. Moreover, the appendix also contains a version of our upper bounds based on smoothness assumptions on the densities instead of the transport map (Section E), the deferred proofs (Section F), additional lemmas (Section G), and more details on the numerical experiments (Section H).

NOTATION. For any positive integer m, define $[m] := \{1, \dots, m\}$. We write $|A|$ for the cardinality of a set A. The relation $a \lesssim b$ is used to indicate that two quantities are the same up to a constant

C, $a \le Cb$. The relation $\gtrsim$ is defined analogously, and we write $a \asymp b$ if $a \lesssim b$ and $a \gtrsim b$. We denote by c and C constants that might change from line to line and that may depend on all parameters of the statistical problem except n. We abbreviate with $a \vee b$, $a \wedge b$ the maximum and minimum of $a \in \mathbb{R}$ and $b \in \mathbb{R}$, respectively. For $a \in \mathbb{R}$, the floor and ceiling functions are denoted by $\lfloor a \rfloor$ and $\lceil a \rceil$, indicating rounding a to the next smaller and larger integer, respectively. We use $\operatorname{supp} f$ to denote the support of a function or measure f, and $\operatorname{diam} \Omega$ for the diameter of a set $\Omega \subseteq \mathbb{R}^d$. We denote by B_1 the unit-ball with respect to the Euclidean distance in $\mathbb{R}^d$, where d should be clear from the context. For a real symmetric matrix A and $\lambda \in \mathbb{R}$, we write $\lambda \preceq A$ if all eigenvalues of A are bounded below λ, and similarly for $A \preceq \lambda$. Moreover, we denote the smallest and largest eigenvalues of A by $\lambda_{\min}(A)$, $\lambda_{\max}(A)$, respectively.

For $p \in [1, \infty]$, we denote by ℓ^p either the space $\mathbb{R}^d$ endowed with the usual ℓ^p norms $\| \,.\, \|_p$, or, by abuse of notation, the spaces of multi-dimensional sequences $\gamma : \mathbb{Z}^m \to \mathbb{R}$ with $\|\gamma\|_{\ell^p}^p = \sum_{k \in \mathbb{Z}^m} |\gamma_k|^p$ for $m \in \mathbb{N}$. Further, for $p \in [1, \infty]$, we denote by L^p the Lebesgue spaces of equivalence classes of functions on $\mathbb{R}^d$ or subsets $\Omega \subseteq \mathbb{R}^d$ with respect to the Lebesgue measure λ on $\mathbb{R}^d$, whose norms we denote by $\| \,.\, \|_{L^p(\mathbb{R}^d)}$ and $\| \,.\, \|_{L^p(\Omega)}$, respectively. By abuse of notation, for a different measure P, we denote the associated Lebesgue norms by $\| \,.\, \|_{L^p(P)}$. We abbreviate with "a.e." any statement that holds "almost everywhere" with respect to the Lebesgue measure.

For a differentiable one-dimensional function $f \colon \mathbb{R} \supseteq \Omega \to \mathbb{R}$, we denote its derivative by $\frac{d}{dx} f$. For a function $f \colon \mathbb{R}^d \supseteq \Omega \to \mathbb{R}$, we denote by $\partial_i = \partial / (\partial x_i)$ its weak derivative in the sense of distributions in direction x_i, which coincides with the usual (point-wise) derivative if f is differentiable in Ω. For a multi-index $\flat \in \mathbb{N}^d$, we set

$$\partial^\flat f = \frac{\partial}{\partial_{x_1}^{\flat_1}} \cdots \frac{\partial}{\partial_{x_d}^{\flat_d}} f, \qquad \text{and } |\flat| = \sum_{i=1}^d \flat_i.$$

The symbol ∂f is also used to denote the sub-differential of a convex function f, while we use the symbols ∇f for the gradient of a function f and Dg for the derivative of a vector-valued function $g \colon \mathbb{R}^{d_1} \supseteq \Omega \to \mathbb{R}^{d_2}$, $\nabla f = (\partial_1 f, \ldots, \partial_d f)^\top$ and $Dg = (\nabla g_1, \ldots, \nabla g_{d_2})^\top$ respectively, and $D^2 f = D \nabla f$ denotes the Hessian of f.

If $\Omega \subseteq \mathbb{R}^d$ is a closed set with non-empty interior and $\alpha > 0$, the Hölder spaces on Ω as defined in Appendix B are denoted by $C^\alpha(\Omega)$ and their associated norms by $\| \,.\, \|_{C^\alpha(\Omega)}$. Similarly, the p-Sobolev spaces of order α for $p \in [1, \infty]$ are denoted by $W^{\alpha,p}(\Omega)$ with norms $\| \,.\, \|_{W^{\alpha,p}(\Omega)}$, as defined in Appendix B.

We say that $\Omega \subseteq \mathbb{R}^d$ is a Lipschitz domain if its boundary can be locally expressed as the sublevel set of Lipschitz functions [Tri06, Definition 1.103].

2. BRENIER'S THEOREM AND THE SEMI-DUAL PROBLEM

We begin by recalling the Monge and Kantorovich problems given in Section 1. Let P, Q be two Borel probability measures on $\mathbb{R}^d$ with finite second moments.

The *Monge (primal) problem* is defined as

$$\min_T \mathcal{P}_{\mathrm{M}}(T) \quad \text{s.t. } T_\# P = Q\,,$$
$$\text{where } \mathcal{P}_{\mathrm{M}}(T) := \frac{1}{2} \int_{\mathbb{R}^d} \|T(x) - x\|_2^2 \,\mathrm{d}P(x),$$

and the push-forward $T_\# P$ is defined as $T_\# P(A) = P(T^{-1}(A))$ for all Borel sets A.

Its relaxation, the *Kantorovich (primal) problem*, is given by

$$\min_{\Gamma} \tilde{\mathcal{P}}_{\mathrm{K}}(\Gamma) \quad \text{s.t. } \Gamma \in \Pi(P, Q) \tag{2.1}$$
$$\text{where } \tilde{\mathcal{P}}_{\mathrm{K}}(\Gamma) := \frac{1}{2} \int \|x - y\|_2^2 \, \mathrm{d}\Gamma(x, y),$$

and $\Pi(P, Q)$ denotes the set of couplings between P and Q, that is, the set of probability measures Γ on $\mathbb{R}^d \times \mathbb{R}^d$ such that $\Gamma(A \times \mathbb{R}^d) = P(A)$ and $\Gamma(\mathbb{R}^d \times A) = Q(A)$ for all Borel set $A \subset \mathbb{R}^d$.

The value of problem (2.1) is the square of the 2-Wasserstein distance, denoted by

$$W_2^2(P, Q) := \min_{\Gamma \in \Pi(P,Q)} \tilde{\mathcal{P}}_{\mathrm{K}}(\Gamma).$$

Note that we can expand the objective in (2.1) as

$$\begin{aligned} \tilde{\mathcal{P}}_{\mathrm{K}}(\Gamma) &= \frac{1}{2} \int \|x - y\|_2^2 \, \mathrm{d}\Gamma(x, y) \\ &= \frac{1}{2} \int \|x\|_2^2 \, \mathrm{d}P(x) + \frac{1}{2} \int \|y\|_2^2 \, \mathrm{d}Q(y) - \int \langle x, y \rangle \, \mathrm{d}\Gamma(x, y), \end{aligned}$$

Since the first two terms above do not depend on Γ, we obtain the equivalent optimization problem

$$\max_{\Gamma} \mathcal{P}_{\mathrm{K}}(\Gamma) \quad \text{s.t. } \Gamma \in \Pi(P, Q)\,, \quad \text{where} \quad \mathcal{P}_{\mathrm{K}}(\Gamma) := \int \langle x, y \rangle \, \mathrm{d}\Gamma(x, y)\,. \tag{2.2}$$

We focus on this equivalent formulation for the rest of the paper because it is more convenient to work with.

Problem (2.2) is a linear optimization problem, albeit an infinite-dimensional one. Hence, it is natural to consider its dual problem:

$$\begin{aligned} \min_{f,g} \int f(x) \, \mathrm{d}P(x) + \int g(y) \, \mathrm{d}Q(y) \quad &\text{s.t.} \\ f(x) + g(y) \geq \langle x, y \rangle\,,\ &P \otimes Q\text{-a.e}, \\ f \in L^1(P), g \in L^1(Q).& \end{aligned} \tag{2.3}$$

The dual variables f and g are called *potentials*, and for an optimal pair (f_0, g_0), f_0 is called a *Kantorovich potential*.

The dual problem (2.3) can be further simplified: Assume we are given a candidate function f in (2.3) above. Then, we can formally solve for the corresponding g to get an optimal g given by the Legendre-Fenchel conjugate (see Section A) of f:

$$g_f(y) = \sup_{x \in \mathbb{R}^d} \langle x, y \rangle - f(x) = f^*(y), \tag{2.4}$$

Plugging solution (2.4) back into the optimization problem leads to the so-called *semi-dual problem*,

$$\min_{f} \mathcal{S}(f) = \int f(x) \, \mathrm{d}P(x) + \int f^*(y) \, \mathrm{d}Q(y) \quad \text{s.t. } f \in L^1(P), \tag{2.5}$$

where the supremum in (2.4) is interpreted as an essential supremum with respect to P. By transitioning to the semi-dual, we effectively solved for all constraints in (2.3), leaving us with an unconstrained convex problem that is not linear anymore. Under regularity assumptions, a solution to the semi-dual provides a solution to the Monge problem as indicated by the following theorem, which is a cornerstone of modern optimal transport.

THEOREM 1 (Brenier's theorem, [KS84, Bre91, RR90]). *Assume P is absolutely continuous with respect to the Lebesgue measure and that both P and Q have finite second moments. Then, a unique optimal solution to (2.2) exists and is of the form $\Gamma_0 = (\mathrm{id}, T_0)_\# P$, where $T_0 = \nabla f_0$ is the gradient of a convex function $f_0 \colon \mathbb{R}^d \to \mathbb{R}$. In fact, f_0 can be chosen to be a minimizer of the semi-dual objective in (2.5).*

Brenier's theorem implies that a solution to the semi-dual problem readily gives an optimal transport map. Our strategy is to minimize an approximation of the semi-dual and establish stability results as well as generalization bounds to conclude that the minimizer to the approximation is close to the minimizer of the original problem.

3. MAIN RESULTS

Let $X_{1:n} = (X_1, \ldots, X_n)$ and $Y_{1:n} = (Y_1, \ldots, Y_n)$ be n independent copies of $X \sim P$ and $Y \sim Q = (T_0)_\# P$ respectively. Furthermore, assume that $X_{1:n}$ and $Y_{1:n}$ are mutually independent. Our goal is to estimate T_0. To that end, we consider the following set of assumptions on P, Q and T_0. Throughout, we fix a constant $M \geq 2$.

A1 (Source distribution). *Let $\mathcal{M} = \mathcal{M}(M)$ be the set of all probability measures P with support $\Omega = [0,1]^d$ that admit a density ρ_P with respect to the Lebesgue measure such that $M^{-1} \leq \rho_P(x) \leq M$ for almost all $x \in \Omega$. Assume that the source distribution P is in $\mathcal{M}$.*

A2 (Transport map). *Let $\tilde{\Omega} = [-1,2]^d$ denote the enlargement of Ω by 1 in every direction. Let $\mathcal{T} = \mathcal{T}(M)$ be the set of all differentiable functions $T \colon \tilde{\Omega} \to \mathbb{R}^d$ such that $T = \nabla f$ for some differentiable convex function $f \colon \tilde{\Omega} \to \mathbb{R}^d$ and*

(i) $|T(x)| \leq M$ *for all* $x \in \tilde{\Omega}$,
(ii) $M^{-1} \preceq DT(x) \preceq M$ *for all* $x \in \tilde{\Omega}$,
(iii) $\operatorname{supp} T_\# P = \Omega = [0,1]^d$.

For $R > 1$ and $\alpha > 1$, assume that $T_0 \in \mathcal{T}_\alpha = \mathcal{T}_\alpha(M, R)$, where

$$\mathcal{T}_\alpha(M,R) = \{T \in \mathcal{T}(M) : T \text{ is } \lfloor \alpha \rfloor\text{-times differentiable and } \|T\|_{C^\alpha(\tilde{\Omega})} \leq R\}.$$

Our main result is the following theorem. It characterizes, up to logarithmic factors, the minimax rate of estimation of an α-smooth transport map $T_0 \in \mathcal{T}_\alpha$ in the setup described above.

THEOREM 2. *Fix $\alpha \geq 1$, then*

$$\inf_{\hat{T}} \sup_{P \in \mathcal{M}, T_0 \in \mathcal{T}_\alpha} \mathbb{E}\left[\int \|\hat{T}(x) - T_0(x)\|_2^2 \, \mathrm{d}P(x)\right] \gtrsim n^{-\frac{2\alpha}{2\alpha - 2 + d}} \vee \frac{1}{n}\,. \tag{3.1}$$

where the infimum is taken over all measurable functions $\hat{T}$ of the data $X_{1:n} = (X_1, \ldots, X_n)$, $Y_{1:n} = (Y_1, \ldots, Y_n)$. Moreover, if $P \in \mathcal{M}$ and $T_0 \in \mathcal{T}_\alpha$, there exists an estimator $\hat{T}$, given in Section 5, that is near minimax optimal. More specifically, there exists an integer $n_0 = n_0(d, \alpha, M, R)$ such that for any $n \geq n_0$, it holds,

$$\sup_{P \in \mathcal{M}, T_0 \in \mathcal{T}_\alpha} \mathbb{E}\left[\int \|\hat{T}(x) - T_0(x)\|_2^2 \, \mathrm{d}P(x)\right] \lesssim n^{-\frac{2\alpha}{2\alpha - 2 + d}} (\log(n))^2 \vee \frac{1}{n}\,. \tag{3.2}$$

REMARK 3. *Assumption A1 can be significantly relaxed with respect to the geometry of Ω and the density of P. In fact, the upper bounds are given under more general assumptions in Section 5. Similarly, the assumption A2(iii) that* $\mathrm{supp}(Q) = [0,1]^d$ *can also be relaxed.*

However, the constants in the resulting upper bounds exhibit a dependence on the geometry of the supports of both P and Q as well as on the enclosing set $\tilde{\Omega}$ through functional analytical results used in the proofs. While it may be possible to make this dependence explicit in terms of geometric features of the sets $\mathrm{supp}(P), \mathrm{supp}(Q)$ *and $\tilde{\Omega}$—see for example [Jon81, ENT11, TMS+13] for such estimates under restrictive assumptions—providing a uniform control on these quantities in terms of easily interpretable properties of the sets is beyond the scope of this article. Instead, we chose to present Theorem 2 under these simplified assumptions to make the results more readable.*

To discuss the remaining assumptions, we note that the most essential ones to obtain upper bounds are the following: first, the lower bound in A2 (ii), $M^{-1} \preceq DT(x)$, in particular on the support of P, $x \in \Omega$. This yields convergence estimates for the optimal transport map as shown in [Gig11], see (1.5)*, and might be necessary to obtain fast rates for transport map estimation since it provides curvature estimates commonly needed to prove error bounds for M-estimators [VdV00, Chapter 5]. Second, the Sobolev regularity of T_0 is what governs the approximation rates of T_0 by wavelet expansions (see Section 5.5 below) and thus enables fast rates via a bias-variance trade-off. All remaining assumptions in A1 and A2, including the existence of extensions of T_0 to a superset $\tilde{\Omega}$, are of technical nature and serve to give explicit bounds as needed in the proof of the upper bound. While one might be able to relax these assumptions, especially in specific problem instances, we do not pursue this here beyond the more general versions given in B1 and B2 below.*

We conjecture that the logarithmic terms appearing in the upper bound are superfluous and arise as an artifact of our proof techniques. We briefly make a qualitative comment on the rate $n^{-\frac{2\alpha}{2\alpha-2+d}}$. Note first that it appears from this rate that estimation of transport maps, like the estimation of smooth functions suffers from the curse of dimensionality. However, as $\alpha \to \infty$, this curse of dimensionality may be mitigated by extra smoothness with the parametric rate n^{-1} as a limiting case. Note also that we can formally take the limit $\alpha \to 1$, which corresponds to the case where no additional smoothness condition holds beyond having a strongly convex Kantorovich potential with Lipschitz gradient. This is essentially the minimal structural condition arising from Brenier's theorem with additional bounds on the derivative of T_0. In this case, one formally recovers the rate $n^{-2/d}$ and we conjecture that this is the minimax rate of estimation in the context where T_0 is only assumed to be the gradient of a strongly convex function with Lipschitz gradient. If either of these two additional requirements is not fulfilled, our stability results no longer hold.

REMARK 4. *Since the transport map T_0 is the main focus of our results, our assumptions impose smoothness directly on T_0. In fact, smoothness of T_0 can also be seen as a consequence of smoothness of the source and target distribution using Caffarelli's regularity theory [Caf92b, Caf92a, Caf96]. For completeness, we also give a version of our upper bound results under smoothness assumptions on P and Q in Theorem 33, Section E of the Appendix.*

REMARK 5. *By prescribing a known base measure P, such as the uniform distribution on $[0,1]^d$, and considering $\tilde{Q} = \hat{T}_{\#}P$, estimation rates for $\hat{T}$ immediately translate into rates for estimating Q in the 2-Wasserstein distance [WB17, SP18, WB19]. In fact, $W_2^2(\tilde{Q}, Q)$ can be bounded by* $\mathbb{E}_P[\|\hat{T} - T_0\|_2^2]$, *since $(\hat{T}, T_0)_{\#}P$ is a candidate transport plan between Q and $\tilde{Q}$. Up to log factors, our rates obtained below match those obtained in [WB19] for the estimation of a smooth density on $[0,1]^d$ in the 2-Wasserstein distance.*

4. LOWER BOUND

In this section we begin by proving the lower bound (3.1) as it sheds light on the source of the non-standard exponent $\frac{2\alpha}{2\alpha-2+d}$ in the minimax rate. We prove the following theorem.

THEOREM 6. *Fix $\alpha \geq 1$. It holds that*

$$\inf_{\hat{T}} \sup_{P \in \mathcal{M}, T_0 \in \mathcal{T}_\alpha} \mathbb{E}\left[\int \|\hat{T}(x) - T_0(x)\|_2^2 \, \mathrm{d}P(x)\right] \gtrsim n^{-\frac{2\alpha}{2\alpha-2+d}} \vee \frac{1}{n}\,.$$

where the infimum is taken over all measurable functions of $X_1, \dots, X_n$, $Y_1, \dots, Y_n$.

PROOF. The proof uses standard tools to establish minimax lower bounds, [Tsy09, Theorem 2.5, Lemma 2.9, Theorem 2.2], that we restate in Appendix D for the convenience of the reader as Theorem 30, Lemma 31, and Theorem 32, respectively. It relies on the following construction.

Set $P = \mathsf{Unif}([0,1]^d) \in \mathcal{M}$, the uniform distribution on the hypercube. For $\alpha > 1$, let $\xi \colon \mathbb{R} \to \mathbb{R}$ be a non-zero function in $C^\infty(\mathbb{R})$ with support contained in $[0,1]$ such that there exists $x_0 \in [0,1]$ with $\xi(x_0) \neq 0$, $\frac{d}{dx}\xi(x_0) \neq 0$, for example a bump-function. Define $g \colon \mathbb{R}^d \to \mathbb{R}$ by

$$g(x) = \prod_{i \in [d]} \xi(x_i)\,, \qquad x = (x_1, \dots, x_d)\,,$$

and note that $\nabla g(x_0, \dots, x_0) \neq 0$ and $\mathrm{supp}(g) = [0,1]^d$ by the above assumptions on ξ.

Let $m = \lceil \theta n^{\frac{1}{2\alpha-2+d}} \rceil$ be a positive integer where θ is a universal constant to be chosen later. We form a regular discretization of the space $[0,1]^d$ by defining the collection of vectors $\{x^{(j)} : j \in [m]^d\} \subset [0,1]^d$ to have coordinates $x_i^{(j)} = (j_i - 1)/m$, $i = 1, \dots, d$ and let

$$g_j(x) = \frac{\kappa}{m^{\alpha+1}} g(m(x - x^{(j)})),$$

for a constant $\kappa > 0$ to be chosen later. Note that $\mathrm{supp}(g_j) \subseteq x^{(j)} + [0, 1/m]^d$ and hence that the supports of the functions $\{g_j\}_{j \in [m]^d}$ are pairwise disjoint.

Next, let $\flat \in \mathbb{N}^d$ be a multi-index and observe that the differential operator $\partial^\flat$ applied to g_j yields $\partial^\flat g_j(\cdot) = m^{|\flat|-\alpha-1}\partial^\flat g(m(\cdot - x_j))$. Since $\xi \in C^{\alpha+1}$, if $\alpha > 1$, a second-order Taylor expansion yields that g_j has uniformly vanishing Hessian: $\|D^2 g_j\|_{L^\infty(\mathbb{R}^d)} \to 0$ as $m \to \infty$. In particular, in that case, there exists m_0 such that $\|D^2 g_j(x)\|_{\mathrm{op}} \leq 1/2$ for all $x \in \mathbb{R}^d$, $m \geq m_0$, $j \in [m]^d$. If $\alpha = 1$, the same can be obtained by choosing κ small enough. By the same reasoning, we can also guarantee $\|\nabla g_j\|_{L^\infty(\mathbb{R}^d)} \leq 1$.

For $m^d \geq 8$, the Varshamov-Gilbert lemma, Lemma 31, guarantees the existence of binary vectors $\tau^{(0)}, \tau^{(1)}, \dots, \tau^{(K)} \in \{0,1\}^{[m]^d}$, $\tau^{(0)} = (0, \dots, 0)$, $K \geq 2^{m^d/8}$ such that $\|\tau^{(k)} - \tau^{(k')}\|_2^2 \geq m^d/8$ for $0 \leq k \neq k' \leq K$. With this, we define the following collection of Kantorovich potentials:

$$\phi_k(x) = \frac{1}{2}\|x\|^2 + \sum_{j \in [m]^d} \tau_j^{(k)} g_j(x)\,, \qquad k = 0, \dots, K\,.$$

It is easy to see (Lemma 38) that for any $k = 0, \dots, K$ and $m \geq m_0$, $\nabla\phi_k$ is a bijection from $[0,1]^d$ to $[0,1]^d$. Moreover, by Weyl's inequality and the above bound $\|D^2 g_j(x)\|_{\mathrm{op}} \leq 1/2$, for all k,

$$\lambda_{\min}(D^2\phi_k(x)) \geq 1 - \sum_{j \in [m]^d} \lambda_{\max}(D^2 g_j(x)) \geq \frac{1}{2}.$$

Similarly, we obtain $\|\nabla\phi_k\|_{L^\infty([0,1]^d)} \le 2$ and $\lambda_{\max}(D^2\phi_k(x)) \le 2$ for all $x \in \mathbb{R}^d$. Hence, $T_k := \nabla\phi_k \in \mathcal{T}_\alpha(M,R)$ for $M > 2$ and κ small enough. We now check the conditions of Theorem 30, where we consider the distance measure

$$d(T_k, T_{k'})^2 = \int_{[0,1]^d} \|T_k - T_{k'}\|^2 \,\mathrm{d}x, \quad , 0 \le k, k' \le K.$$

First, observe that for $0 \le k \ne k' \le K$, it holds that

$$\begin{aligned}
&\int_{[0,1]^d} \|\nabla\phi_k(x) - \nabla\phi_{k'}(x)\|^2 \,\mathrm{d}x \\
&\qquad = \frac{\kappa^2}{m^{2\alpha+d}} \sum_{j\in[m]^d} (\tau_j^{(k)} - \tau_j^{(k')})^2 \int_{\mathbb{R}^d} \|\nabla g(x)\|^2 \,\mathrm{d}x \gtrsim \frac{1}{m^{2\alpha}}\,.
\end{aligned}$$

This yields

$$\int_{[0,1]^d} \|\nabla\phi_k(x) - \nabla\phi_{k'}(x)\|^2 \,\mathrm{d}x \gtrsim n^{-\frac{2\alpha}{2\alpha-2+d}},$$

which completes checking the separation condition (i) of Theorem 30.

To check condition (ii) of Theorem 30, recall the Kullback-Leibler (KL) divergence between two measures Q, P such that Q is absolutely continuous with respect to P is defined by

$$D(Q\|P) = \mathbb{E} \log\big(\frac{\mathrm{d}Q}{\mathrm{d}P}(W)\big)\,, \quad W \sim Q.$$

In view of Lemma 38, for any $k = 0,\dots,K$, the measure $Q_k = (\nabla\phi_k)_\# P$ is supported on $[0,1]^d$ and in particular, it is absolutely continuous with respect to P. By the change of variables formula, it admits the density

$$\frac{\mathrm{d}Q_k}{\mathrm{d}P}(y) := \frac{1}{\det D^2\phi_k((\nabla\phi_k)^{-1}(y))} \mathbb{1}((\nabla\phi_k)^{-1}(y) \in [0,1]^d). \tag{4.2}$$

Moreover, let $X \sim P$ and $Y \sim Q_k$ be two random variables. It holds

$$\begin{aligned}
D(Q_k\|P) &= \mathbb{E}\log\left(\frac{\mathrm{d}Q_k}{\mathrm{d}P}(Y)\right) = \mathbb{E}\log\left(\frac{\mathrm{d}Q_k}{\mathrm{d}P}(\nabla\phi_k(X))\right) \\
&= -\int_{[0,1]^d} \log\big(\det D^2\phi_k(x)\big)\,\mathrm{d}x\,.
\end{aligned} \tag{4.3}$$

Recall that $D^2\phi_k = I_d + \sum_{j\in[m]^d} \tau_j^{(k)} D^2 g_j$ where I_d denotes the identity matrix in $\mathbb{R}^d$. Therefore, since the functions g_j have disjoint support, we have for all $x \in [0,1]^d$ that

$$\begin{aligned}
\log\big(\det D^2\phi_k(x)\big) &= \sum_{l=1}^d \log\Big(1 + \lambda_l\big(\sum_{j\in[m]^d} \tau_j^{(k)} D^2 g_j(x)\big)\Big) \\
&= \sum_{l=1}^d \sum_{j\in[m]^d} \log\Big(1 + \tau_j^{(k)} \lambda_l\big(D^2 g_j(x)\big)\Big)\,,
\end{aligned}$$

where $\lambda_l(A)$ denotes the lth eigenvalue of a matrix A. Since $\log(1+z) \geq z - z^2/2$ for all $z > 0$,

$$\log\left(\det D^2\phi_k(x)\right) \geq \sum_{j\in[m]^d} \tau_j^{(k)} \mathsf{Tr}(D^2 g_j(x)) - \frac{1}{2}\sum_{j\in[m]^d} \|D^2 g_j(x)\|_F^2$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Thus,

$$\begin{aligned} D(Q_k\|P) \leq & -\sum_{j\in[m]^d} \tau_j^{(k)} \int_{[0,1]^d} \mathsf{Tr}(D^2 g_j(x))\,\mathrm{d}x \\ & + \frac{1}{2}\sum_{j\in[m]^d} \int_{[0,1]^d} \|D^2 g_j(x)\|_F^2\,\mathrm{d}x. \end{aligned}$$

On the one hand, by the divergence theorem and the fact that g_j has bounded support,

$$\int_{[0,1]^d} \mathsf{Tr}(D^2 g_j(x))\,\mathrm{d}x = \int_{\partial[0,1]^d} \langle v(x), \nabla g_j(x)\rangle\,\mathrm{d}x = 0\,,$$

where $\partial[0,1]^d$ denotes the boundary of the unit hypercube and $v(x)$ its outward-pointing unit normal vector. On the other hand

$$\sum_{j\in[m]^d} \int_{[0,1]^d} \|D^2 g_j(x)\|_F^2\,\mathrm{d}x = \frac{\kappa^2}{m^{2\alpha-2+d}} \sum_{j\in[m]^d} \int_{\mathbb{R}^d} \|D^2 g(x)\|_F^2\,\mathrm{d}x \lesssim \frac{1}{m^{2\alpha-2}}.$$

The above three displays yield

$$D(P^{\otimes n}\otimes Q_k^{\otimes n}\|P^{\otimes n}\otimes P^{\otimes n}) = nD(Q_k\|P) \lesssim \frac{n}{m^{2\alpha-2}} \leq \frac{m^d}{\theta} \leq \frac{\log K}{9}$$

for θ large enough. This completes checking (ii) in Theorem 30 and hence the proof of the first part of the minimax lower bound.

To show the remaining lower bound of $1/n$, repeat the same argument as above with the two potentials $\phi_0(x) = \|x\|_2^2/2$ and $\phi_1(x) = \phi_0(x) + (\tilde\theta/\sqrt{n})g(x)$ for $\tilde\theta$ chosen to ensure $\phi_0, \phi_1 \in \mathcal{T}_\alpha$, applying Theorem 32 in Appendix D. The separation condition is given by

$$\int_{[0,1]^d} \|\nabla\phi_0(x) - \nabla\phi_1(x)\|^2\,\mathrm{d}x = \frac{\tilde\theta^2}{n}\int_{[0,1]^d} \|\nabla g(x)\|_2^2\,\mathrm{d}x \gtrsim \frac{1}{n},$$

and the KL divergence between the associated probability distributions can be estimated by

$$D(P^{\otimes n}\otimes Q_1^{\otimes n}\|P^{\otimes n}\otimes P^{\otimes n}) = nD(Q_1\|P) \lesssim \frac{\tilde\theta^2 n}{n}\int_{[0,1]^d} \|D^2 g(x)\|_F^2\,\mathrm{d}x \lesssim \frac{1}{9},$$

for $\tilde\theta$ large enough. □

Looking back at this proof, we get a better understanding of the exponent in the minimax rate $n^{-\frac{2\alpha}{2\alpha-2+d}}$. Given that $n^{-\frac{2(\beta-k)}{2\beta+d}}$ is the minimax rate of estimation of the kth derivative of a β-smooth density in L^2 [MG79], the rate that we obtain is formally that of an "antiderivative" ($k=-1$) of a $\beta = \alpha - 1$-smooth signal in this model. This is due to the fact that, on the one hand, the information structure, measured in terms of Kullback-Leibler divergence, is governed by the *derivative* of the signal T_0, i.e., the Hessian of the $\alpha+1$-smooth Kantorovich potential, see (4.3), which is $\alpha-1$-smooth. This follows directly from the Monge-Ampère equation (4.2). On the other hand, we measure the performance of the estimator in terms of the $L^2(P)$ distance between $\hat{T}$ and T_0, corresponding to the antiderivative of the Hessian. Of course, in the absence of the classical fundamental theorem of calculus in dimension $d > 1$ for arbitrary maps $T\colon \mathbb{R}^d \to \mathbb{R}^d$, the existence of an antiderivative needs to be guaranteed a priori, as in the case of transport maps by assuming $T = \nabla f$ for $f\colon \mathbb{R}^d \to \mathbb{R}$.

Similar rates arise in the estimation of the invariant measure of a diffusion process when smoothness is imposed on the drift [DR07, Str18]. This is not surprising as the drift is the gradient of the logarithm of the density of the invariant measure in an overdamped Langevin process.

Finally, note that the multivariate case is singularly different from the traditional univariate case where the rate of estimation of linear functionals such as anti-derivatives is known to be parametric regardless of the smoothness of the signal [IH81].

5. UPPER BOUNDS

In this section, we give an estimator $\hat{T}$ that achieves the near-optimal rate (3.2). We present this estimator under the following more general assumptions on the distribution and the geometry of the support of both P and $Q = (T_0)_\# P$. We also need slightly weaker conditions on the regularity of the transport map (Sobolev instead of Hölder regularity). After stating these weaker assumptions, we present our estimator and restate the main upper bound. Its proof relies on a separate control of approximation error and stochastic error, similar to a standard bias-variance tradeoff.

5.1 Assumptions

Throughout, we fix two constants $M \geq 2, \beta > 1$.

B1 (Source distribution). *Let $\mathcal{M} = \mathcal{M}(M)$ be the set of all probability measures P whose support $\Omega_P \subseteq MB_1$ is a bounded and connected Lipschitz domain, and that admit a density ρ_P with respect to the Lebesgue measure such that $M^{-1} \leq \rho_P(x) \leq M$ for almost all $x \in \Omega_P$. Assume that the measure $P \in \mathcal{M}$.*

B2 (Transport map). *For any $P \in \mathcal{M}$ with support Ω_P, let $\tilde{\Omega}_P$ denote a convex set with Lipschitz boundary such that $\tilde{\Omega}_P \subseteq MB_1$, and $\Omega_P + M^{-1}B_1 \subseteq \tilde{\Omega}_P$. Let $\tilde{\mathcal{T}} = \tilde{\mathcal{T}}(M)$ be the set of all differentiable functions $T\colon \tilde{\Omega}_P \to \mathbb{R}^d$ such that $T = \nabla f$ for some differentiable convex function $f\colon \tilde{\Omega}_P \to \mathbb{R}^d$ and*

(i) $|T(x)| \leq M$ for all $x \in \tilde{\Omega}_P$,
(ii) $M^{-1} \preceq DT(x) \preceq M$ for all $x \in \tilde{\Omega}_P$.

For $R > 1$ and $\alpha > 1$, assume that

$$T_0 \in \tilde{\mathcal{T}}_\alpha = \tilde{\mathcal{T}}_\alpha(M, R) = \{T \in \tilde{\mathcal{T}}(M) : \|T\|_{C^\beta(\tilde{\Omega}_P)} \vee \|T\|_{W^{\alpha,2}(\tilde{\Omega}_P)} \leq R\}.$$

These new conditions have two implications. First, they imply regularity of the Kantorovich potential f_0, where $T_0 = \nabla f_0$, and second, they imply some conditions on the pushforward measure $Q = (T_0)_\# P$ that subsume the generalization of Assumption A2(iii). These results are gathered in the following proposition (see Section F.1 for a proof).

PROPOSITION-DEFINITION 7. *Assume that P satisfies B1, T_0 satisfies B2 and let $\mathcal{X} = \mathcal{X}(M)$ be the set of all twice continuously differentiable functions $f : \tilde{\Omega}_P \to \mathbb{R}$ such that*

(i) $|f(x)| \le 2M^2$ *and* $|\nabla f(x)| \le M$ *for all* $x \in \tilde{\Omega}_P$,
(ii) $M^{-1} \preceq D^2 f(x) \preceq M$ *for all* $x \in \tilde{\Omega}_P$.

Then there exists a Kantorovich potential $f_0 \in \mathcal{X}(M)$ such that $T_0 = \nabla f_0$,

$$\|f_0\|_{C^{\beta+1}(\tilde{\Omega}_P)} \vee \|f_0\|_{W^{\alpha+1,2}(\tilde{\Omega}_P)} \le R + 2M^3. \tag{5.1}$$

Further, $Q = \nabla(f_0)_\# P$ has a connected and bounded Lipschitz support $\Omega_Q \subseteq MB_1$ and admits a density ρ_Q with respect to the Lebesgue measure that satisfies $M^{-(d+1)} \le \rho_Q(y) \le M^{d+1}$ for all $y \in \Omega_Q$.

Note that the simplified Assumptions A1 and A2 from Section 3 follow from B1 and B2 in the case $\Omega_P = \Omega = [0,1]^d$ and $\tilde{\Omega}_P = \tilde{\Omega} = [-1,2]^d$. Additionally, the simplified assumptions restrict the class of transport maps to those such that $\Omega_Q = [0,1]^d$ and for which $\beta = \alpha$. Indeed, noting that $\|T_0\|_{W^{\alpha,2}(\tilde{\Omega})} \lesssim \|T_0\|_{C^\alpha(\tilde{\Omega})}$, we can fold the two smoothness conditions into one.

5.2 Estimator

To construct an estimator for T_0, we observe that if we had access to a Kantorovich potential f_0, then $T_0 = \nabla f_0$ by Brenier's Theorem, Theorem 1. In turn, f_0 is the minimum of the semi-dual objective (2.5). Hence, we replace population quantities with sample ones in its definition to obtain an empirical loss function. Moreover, to account for the assumed smoothness of the transport map and to ensure stability of the objective, we constrain our minimization problem to smooth and strongly convex Kantorovich potentials, restricted to a compact superset of the support of P. Then, our estimator is the gradient of the solution to this stochastic optimization problem.

More precisely, for a measurable function f, let us write

$$Pf = \int f(x)\,\mathrm{d}P(x), \quad Qf = \int f(y)\,\mathrm{d}Q(y),$$
$$\hat{P}f = \frac{1}{n}\sum_{i=1}^n f(X_i), \qquad \hat{Q}f = \frac{1}{n}\sum_{j=1}^n f(Y_i),$$

where, as in Section 3, $X_{1:n} = (X_1, \dots, X_n)$ and $Y_{1:n} = (Y_1, \dots, Y_n)$ are n *i.i.d.* samples from P and Q, respectively, that are mutually independent as well. Recall from Section 2 that the semi-dual objective is defined as $\mathcal{S}(f) = Pf + Qf^*$ for $f \in L^1(P)$, where f^* denotes the convex conjugate of f. Replacing both P and Q by their empirical counterparts, we obtain the empirical semi-dual,

$$\hat{\mathcal{S}}(f) = \hat{P}f + \hat{Q}f^*. \tag{5.2}$$

In order to incorporate smoothness regularization into the minimization of (5.2), we consider the restriction of potentials f to a wavelet expansion of finite degree, a strategy that is frequently used in non-parametric estimation [HKPT98, GN16]. For the purpose of this section, it is enough to think about wavelets as a graded orthogonal basis of $L^2(\mathbb{R}^d)$, leading to nested subspaces

$$V_0 \subseteq V_1 \subseteq \cdots \subseteq V_J \subseteq \cdots \subseteq L^2(\mathbb{R}^d),$$

that roughly correspond to increasing frequency ranges of the continuous Fourier transform of a function $f \in L^2(\mathbb{R}^d)$. Truncated wavelet decompositions yield good approximations for smooth functions and we control their approximation error in Lemma 13. We only consider the span $V_J(\tilde{\Omega}_P)$ of those basis functions of the wavelet expansion whose support has non-trivial intersection with $\tilde{\Omega}_P$. This is a finite-dimensional vector space as long as the elements of the wavelet basis have compact support. The cut-off parameter J is chosen according to the regularity of f_0 in assumption B2, see Section 5.6, or alternatively can be chosen adaptively by a straightforward but technical extension using a penalization scheme [Mas07] that we omit for readability. Alternatively, other selection methods such as Lepski's method [Lep91, Lep92, Lep93] could be used. In order to ensure the necessary regularity and the compact support of the elements of the wavelet basis, we assume throughout that the wavelet basis is given by Daubechies wavelets of sufficient order. For a more detailed treatment of wavelets, we refer the reader to Section B.

To ensure stability of the minimizer of the semi-dual with respect to perturbations of the input distributions P and Q, we further restrict the potentials f to mimic the assumptions in Proposition-Definition 7, in particular, we enforce upper and lower bounds on the Hessian $D^2 f$ on $\tilde{\Omega}_P$ by demanding $f \in \mathcal{X}(2M)$.

Combined, both wavelet regularization and strong convexity lead to the set

$$\mathcal{F}_J = \mathcal{X}(2M) \cap V_J(\tilde{\Omega}_P) \tag{5.3}$$

of candidate potentials, based on which we define the estimators

$$\hat{f}_J \in \operatorname*{argmin}_{f \in \mathcal{F}_J} \hat{\mathcal{S}}(f), \quad \hat{T}_J = \nabla \hat{f}_J, \tag{5.4}$$

for the Kantorovich potential and transport map, respectively.

Note that since we consider candidate potentials only on the compact set $\tilde{\Omega}_P$, f^* above is defined as

$$f^*(y) = \sup_{x \in \tilde{\Omega}_P} \langle x, y \rangle - f(x) = \sup_{x \in \mathbb{R}^d} \langle x, y \rangle - (f + \iota_{\tilde{\Omega}_P})(x), \quad y \in \mathbb{R}^d,$$

where $\iota_{\tilde{\Omega}_P}$ is the usual indicator function in convex analysis (see Section A).

With this, we can restate the upper bound of Theorem 2.

THEOREM 8. *Under assumptions B1 and B2, there exists $n_0 \in \mathbb{N}$ and J such that for $n \geq n_0$,*

$$\mathbb{E}_{(X_{1:n}, Y_{1:n})} \left[\int \|\hat{T}_J(x) - T_0(x)\|_2^2 \, \mathrm{d}P(x) \right] \leq C \left[n^{-\frac{2\alpha}{2\alpha - 2 + d}} (\log(n))^2 \vee \frac{1}{n} \right],$$

where n_0, C, and J may depend on $d, M, R, \Omega_P, \Omega_Q, \tilde{\Omega}_P$, n_0 may additionally depend on β, C may additionally depend on α, and J depends on n.

The cutoff J depends on α if $d \geq 3$, but in the cases $d = 1$ and $d = 2$, J can be chosen independently from α.

REMARK 9. *A few remarks are in order.*

(i) Similar upper bounds hold with high probability and can be inferred from the proof.

(ii) As written, the estimator $\hat{f}_J$ is not directly implementable since the calculation of f^* involves computing a maximum over a continuous subset of $\mathbb{R}^d$. However, this limitation can be overcome by a discretization of the space, although this is not practical even in moderate dimensions. We provide such an example implementation in Section 6, along with a more practical estimator for which we give no theoretical upper bounds.

(iii) *The numerical experiments in Section 6 suggest that restricting the optimization in (5.4) to $\mathcal{X}(2M)$, while necessary for our proofs, might not be necessary in practice.*
(iv) *The estimator employed in Theorem 8 can be made adaptive to the unknown smoothness parameter α using a standard penalization scheme, see [Mas07]. We omit this straightforward extension and instead focus on establishing minimax rates of estimation. For a more detailed account, we refer the reader to [HÏ9].*
(v) *For the sake of readability, we do not explicitly track the dependence on the parameter M. However, an inspection of the proof yields that the final rate scales like $M^{c_1 d + c_2}$ for constants $c_1, c_2 \geq 1$, i.e., there is an exponential dependence on the dimension d. Further, the dependence on R is captured in (5.15) below and amounts to $R^{\frac{2(d-2)}{2\alpha-2+d}} \log(R)$ in the case $d \geq 3$. We do not claim an optimal dependence of our rates on these parameters.*

In the rest of this section, we present the proof of Theorem 8. We begin by stating our key result, which relates the semi-dual objective to the square of our measure of performance. This result also allows us to employ a fixed-point argument when controlling the risk of our estimator using empirical process theory. Combined with approximation results for truncated wavelet expansions, these lead to a bias-variance tradeoff that achieves the minimax lower bound of Theorem 6 up to log factors.

5.3 Stability of optimal transport maps

In this section, we leverage the assumed regularity of the optimal transport map to relate the suboptimality gap in the semi-dual objective function $\mathcal{S}$ and the L^2-distance of interest.

PROPOSITION 10. *Under assumptions B1 – B2, for all $f \in \mathcal{X}(2M)$ as defined in Proposition-Definition 7, we have*

$$\frac{1}{8M}\|\nabla f(x) - \nabla f_0(x)\|^2_{L^2(P)} \leq \mathcal{S}(f) - \mathcal{S}(f_0) \leq 2M\|\nabla f(x) - \nabla f_0(x)\|^2_{L^2(P)} \tag{5.5}$$

and

$$\frac{1}{4M}\|\nabla f^*(y) - \nabla f_0^*(y)\|^2_{L^2(Q)} \leq \mathcal{S}(f) - \mathcal{S}(f_0)\,. \tag{5.6}$$

PROOF. It follows from Proposition-Definition 7(ii) and a second-order Taylor expansion that f is of *quadratic type* [Kol11] around every $x \in \Omega_P$:

$$\frac{1}{2L}\|z - x\|_2^2 \leq f(z) - f(x) - \langle \nabla f(x), z - x\rangle \leq \frac{L}{2}\|z - x\|_2^2, \quad \text{for } x \in \Omega_P, z \in \tilde{\Omega}_P, \tag{5.7}$$

for all $L \geq 2M$. It turns out that these conditions are sufficient to obtain the desired result.

The upper bound in (5.7) is of the form

$$f(z) \leq q_x(z) = f(x) + \langle \nabla f(x), z - x\rangle + \frac{L}{2}\|z - x\|_2^2 + \iota_{\tilde{\Omega}_P}(z)\,.$$

Since the convex conjugate is order reversing and because the convex conjugate q_x^* of the quadratic function q_x can be computed explicitly (Lemma 21), we have

$$\begin{aligned} f^*(\nabla f_0(x)) \geq q_x^*(\nabla f_0(x)) &= \frac{1}{2L}\|\nabla f_0(x) - \nabla f(x)\|_2^2 + \langle x, \nabla f_0(x)\rangle - f(x) \\ &\quad - \frac{L}{2} d^2\Big(\frac{\nabla f_0(x) - \nabla f(x)}{L} - x, \tilde{\Omega}_P\Big)\,. \end{aligned}$$

The squared distance term vanishes for $L = 4M$: by the triangle inequality $\|\nabla f_0(x) - \nabla f(x)\|_2 \le 4M$ and since $x \in \Omega_P$, it holds that

$$\frac{\nabla f_0(x) - \nabla f(x)}{L} - x \in \tilde{\Omega}_P.$$

Together with the fact that $Q = (\nabla f_0)_{\#}P$, this yields

$$\begin{aligned}\mathcal{S}(f) = Pf + Qf^* &= \int [f(x) + f^*(\nabla f_0(x))] \mathrm{d}P(x) \\ &\ge \frac{1}{8M}\|\nabla f - \nabla f_0\|^2_{L^2(P)} + \int \langle x, \nabla f_0(x)\rangle \mathrm{d}P(x)\,.\end{aligned}$$

Moreover, by strong duality, we have

$$\mathcal{S}(f_0) = Pf_0 + Qf_0^* = \int \langle x, \nabla f_0(x)\rangle \mathrm{d}P(x)\,.$$

The above two displays yield $\mathcal{S}(f) - \mathcal{S}(f_0) \ge (8M)^{-1}\|\nabla f - \nabla f_0\|^2_{L^2(P)}$. In the same way, using the lower bound in (5.7), we get that $\mathcal{S}(f) - \mathcal{S}(f_0) \le 2M\|\nabla f - \nabla f_0\|^2_{L^2(P)}$, which concludes the proof of (5.5).

It turns out that (5.6) is even easier to prove. Indeed, by Proposition-Definition 7(ii) and Lemma 17, we get that the upper bound in (5.7) is also true for f^* on all of $\mathbb{R}^d$. In this case, we can simply take $L = 2M$ and get similar results. □

There are many ways to leverage strong convexity in order to obtain faster rates of convergence, often known as *fixed-point arguments* [Mas07, Kol11, GN16]. In this work, we employ van de Geer's "one-shot" localization technique originally introduced in [vdG87] and stated in a form close to our needs in [vdG02].

5.4 Control of the stochastic error via empirical processes

In light of Proposition 10, the performance of our estimator $\hat{T}_J = \nabla f_J$ defined in (5.4) requires the control of $\mathcal{S}(\hat{f}_J) - \mathcal{S}(f_0)$, which can be achieved using tools from empirical process theory. To that end, for any f, define

$$\mathcal{S}_0(f) = \mathcal{S}(f) - \mathcal{S}(f_0) \qquad \text{and} \qquad \hat{\mathcal{S}}_0(f) = \hat{\mathcal{S}}(f) - \hat{\mathcal{S}}(f_0)\,,$$

and let $\bar{f}_J \in \mathcal{F}_J$. We observe that by optimality of $\hat{f}_J$ for $\hat{\mathcal{S}}$,

$$\mathcal{S}_0(\hat{f}_J) - \mathcal{S}_0(\bar{f}_J) \le \big[\mathcal{S}_0(\hat{f}_J) - \hat{\mathcal{S}}_0(\hat{f}_J)\big] + \big[\hat{\mathcal{S}}_0(\bar{f}_J) - \mathcal{S}_0(\bar{f}_J)\big].$$

To proceed, we control the localized empirical process

$$\sup_{f \in \mathcal{F}_J : \mathcal{S}_0(f) \le \tau^2} |\mathcal{S}_0(f) - \hat{\mathcal{S}}_0(f)|.$$

for some fixed $\tau^2 > 0$. More precisely, we prove the following result in Appendix F.2.

Proposition 11. *Let assumptions B1 – B2 be fulfilled and define $\mathcal{F}_J$ as in (5.3). For any $\tau > 0$, define*

$$\mathcal{F}_J(\tau^2) := \{f \in \mathcal{F}_J : \mathcal{S}_0(f) \le \tau^2\}.$$

Then, there exists $C_1 = C_1(d, \Omega_P, \tilde{\Omega}_P, \Omega_Q, M) > 0$ *such that with probability at least* $1 - \exp(-t)$*,*

$$\sup_{f \in \mathcal{F}_J(\tau^2)} |\mathcal{S}_0(f) - \hat{\mathcal{S}}_0(f)| \leq C_1 \left(\phi_J(\tau^2) + \tau\sqrt{\frac{t}{n}} + \frac{t}{n} \right),$$

where

$$\phi_J(\tau^2) = \frac{2^{J(d-2)/2}\,\tau}{\sqrt{n}} \sqrt{J \log\left(1 + \frac{C_1}{\tau}\right)} + \frac{2^{J(d-2)}J}{n} \log\left(1 + \frac{C_1}{\tau}\right) + \frac{\tau}{\sqrt{n}}. \tag{5.8}$$

Equipped with this result, we can apply van de Geer's localization technique. To simplify the presentation, assume that $\bar{f} = \bar{f}_J \in \operatorname{argmin}_{f \in \mathcal{F}_J} \mathcal{S}(f)$ exists. If not, we may repeat the proof with an ε-approximate minimizer and let $\varepsilon \to 0$. Throughout the proof, we write $\hat{f} = \hat{f}_J$ and $\|\cdot\| = \|\cdot\|_{L^2(P)}$.

Fix $\sigma > 0$ to be defined later and set

$$\hat{f}_s = s\hat{f} + (1-s)\bar{f}, \qquad s = \frac{\sigma}{\sigma + \|\nabla \hat{f} - \nabla \bar{f}\|}, \tag{5.9}$$

Note that since $s \in [0,1]$ and $\mathcal{F}_J$ is convex, we have $\hat{f}_s \in \mathcal{F}_J$.

On the one hand, $\hat{f}_s$ is localized in the sense that

$$\|\nabla \hat{f}_s - \nabla \bar{f}\| = s\|\nabla \hat{f} - \nabla \bar{f}\| = \frac{\sigma\|\nabla \hat{f} - \nabla \bar{f}\|}{\sigma + \|\nabla \hat{f} - \nabla \bar{f}\|} \leq \sigma.$$

By Proposition 10 and the triangle inequality respectively, this yields

$$\mathcal{S}_0(\hat{f}_s) \leq 2M\|\nabla \hat{f}_s - \nabla f_0\|^2 \leq 4M\left(\sigma^2 + \|\nabla \bar{f} - \nabla f_0\|^2\right) =: \tau^2.$$

Therefore, $\hat{f}_s \in \mathcal{F}_J(\tau^2)$. For the same reason, we also have that $\bar{f} \in \mathcal{F}_J(\tau^2)$.

On the other hand, $\hat{f}_s$, akin to $\hat{f}$, has empirical risk smaller than $\bar{f}$. Indeed, by convexity of $\hat{\mathcal{S}}$ and the fact that $\hat{f}$ minimizes $\hat{\mathcal{S}}$ over $\mathcal{G}$, we obtain

$$\hat{\mathcal{S}}(\hat{f}_s) \leq s\hat{\mathcal{S}}(\hat{f}) + (1-s)\hat{\mathcal{S}}(\bar{f}) \leq \hat{\mathcal{S}}(\bar{f}),$$

which yields

$$\mathcal{S}_0(\hat{f}_s) \leq \mathcal{S}_0(\bar{f}) + 2\sup_{f \in \mathcal{F}_J(\tau^2)} \big|\mathcal{S}_0(f) - \hat{\mathcal{S}}_0(f)\big|.$$

Together with Jensen's inequality and Proposition 10 respectively, the above display yields

$$\begin{aligned}\|\nabla \hat{f}_s - \nabla \bar{f}\|^2 &\leq 2\|\nabla \hat{f}_s - \nabla f_0\|^2 + 2\|\nabla f_0 - \nabla \bar{f}\|^2 \leq 16M\mathcal{S}_0(\hat{f}_s) + 16M\mathcal{S}_0(\bar{f}) \\ &\leq 32M\mathcal{S}_0(\bar{f}) + 32M \sup_{f \in \mathcal{F}_J(\tau^2)} \big|\mathcal{S}_0(f) - \hat{\mathcal{S}}_0(f)\big|.\end{aligned}$$

Next, note for s as in (5.9), we have that $\|\nabla \hat{f}_s - \nabla \bar{f}\| \geq \sigma/2$ iff $\|\nabla \hat{f} - \nabla \bar{f}\| \geq \sigma$. Hence

$$\begin{aligned}&\mathbb{P}\big(\|\nabla \hat{f} - \nabla f_0\| \geq \sigma + \|\nabla \bar{f} - \nabla f_0\|\big) \leq \mathbb{P}\big(\|\nabla \hat{f}_s - \nabla \bar{f}\|^2 \geq \sigma^2/4\big) \\ &\quad\leq \mathbb{P}\Big(\sup_{f \in \mathcal{F}_J(\tau^2)} |\mathcal{S}_0(f) - \hat{\mathcal{S}}_0(f)| \geq \frac{\sigma^2}{128M} - \mathcal{S}_0(\bar{f})\Big) \\ &\quad= \mathbb{P}\Big(\sup_{f \in \mathcal{F}_J(\tau^2)} |\mathcal{S}_0(f) - \hat{\mathcal{S}}_0(f)| \geq \frac{\tau^2}{512M^2} - \frac{1}{128M}\|\nabla \bar{f} - \nabla f_0\|^2 - \mathcal{S}_0(\bar{f})\Big).\end{aligned}$$

Recalling Proposition 11, we take σ such that

$$\frac{\tau^2}{512M^2} \geq \mathcal{S}_0(\bar{f}) + \frac{1}{128M}\|\nabla \bar{f} - \nabla f_0\|^2 + C_1\left(\phi_J(\tau^2) + \tau\sqrt{\frac{M\,t}{n}} + \frac{M^2\,t}{n}\right), \tag{5.10}$$

so that we get

$$\mathbb{P}\big(\|\nabla \hat{f} - \nabla f_0\| \geq \sigma + \|\nabla \bar{f} - \nabla f_0\|\big) \leq e^{-t}\,.$$

In particular, we can check that (5.10) is fulfilled if we choose σ such that

$$\sigma^2 \gtrsim \mathcal{S}_0(\bar{f}) + \frac{2^{J(d-2)}J}{n}\log\left(1 + C_2 n\right) + \frac{1+t}{n}\,,$$

for a suitable choice of $C_2 > 0$.

With this, and applying Theorem 10 again, we get that with probability at least $1 - e^{-t}$, it holds

$$\|\nabla \hat{f} - \nabla f_0\|^2 \lesssim \|\nabla \bar{f} - \nabla f_0\|^2 + \frac{2^{J(d-2)}J}{n}\log\left(1 + C_2 n\right) + \frac{1+t}{n}\,.$$

Moreover, integrating the tail with respect to t readily yields by Fubini's theorem that

$$\mathbb{E}\|\nabla \hat{f} - \nabla f_0\|^2 \lesssim \|\nabla \bar{f} - \nabla f_0\|^2 + \frac{2^{J(d-2)}J}{n}\log\left(1 + C_2 n\right) + \frac{1}{n}\,.$$

We have proved the following result.

PROPOSITION 12. *Let B1 – B2 hold and define $\mathcal{F}_J$ as in (5.3). Then, writing*

$$E_J := \frac{2^{J(d-2)}J}{n}\log\left(1 + C_2 n\right) + \frac{1}{n},$$

the estimator $\hat{T}_J$ defined in (5.4) satisfies

$$\mathbb{E}\|\hat{T}_J - T_0\|^2_{L^2(P)} \lesssim \inf_{f \in \mathcal{F}_J} \|\nabla f - T_0\|^2_{L^2(P)} + E_J. \tag{5.11}$$

Moreover, with probability at least $1 - \exp(-t)$,

$$\|\hat{T}_J - T_0\|^2_{L^2(P)} \lesssim \inf_{f \in \mathcal{F}_J} \|\nabla f - T_0\|^2_{L^2(P)} + E_J + \frac{t}{n}\,.$$

5.5 Control of the approximation error

Next, we control the approximation error $\inf_{f\in\mathcal{F}_J}\|\nabla f - \nabla f_0\|_{L^2(P)}$ that appears in Proposition 12. In fact, it is sufficient to control $\|\nabla \bar{f} - \nabla f_0\|_{L^2(P)}$ where $\bar{f} = \Pi_J \operatorname{ext} f_0$ is the truncation of f_0 to its first J wavelet scales after extending f_0 to all $\mathbb{R}^d$. In light of Theorem 23, we may assume that $\operatorname{ext}\bar{f}$ has the same C^β- and $W^{\alpha,2}$-norm as $\bar{f}$ up to a constant depending on $\tilde{\Omega}_P$.

To control the approximation associated with truncating a wavelet decomposition, we rely on the following lemma for Besov functions.

LEMMA 13. *Let $f \in B^s_{p,q}(\mathbb{R}^d)$ and denote by Π_J its projection onto the first scale J wavelet coefficients. That is, if*

$$f = \sum_{j=0}^{\infty} \sum_{g \in G^j} \sum_{k \in \mathbb{Z}^d} \gamma_k^{j,g} \Psi_k^{j,g}, \quad \textit{we set} \quad \Pi_J f = \sum_{j=0}^{J} \sum_{g \in G^j} \sum_{k \in \mathbb{Z}^d} \gamma_k^{j,g} \Psi_k^{j,g},$$

where $\Psi_k^{j,g}$ are multi-dimensional Daubechies wavelets and G^j the associated index sets as in Section B. Then, for all $1 \le p, q \le \infty$, $s \ge 0$,

$$\|\Pi_J f\|_{B^s_{p,q}(\mathbb{R}^d)} \le \|f\|_{B^s_{p,q}}, \tag{5.12}$$

$$\|\Pi_J f - f\|_{B^s_{p,q}(\mathbb{R}^d)} \le \|f\|_{B^s_{p,q}}. \tag{5.13}$$

Moreover, for every $q' \in [1,\infty]$, $s' > 0$, and $1 \le p \le p' \le \infty$ such that $s - d/p > s' - d/p'$,

$$\|\Pi_J f - f\|_{B^{s'}_{p',q'}} \lesssim 2^{-J(s-d/p-(s'-d/p'))} \|f\|_{B^s_{p,q}}.$$

In particular: If $f \in C^{\alpha+1}$ for $\alpha > 1$, then $\|f - \Pi_J f\|_{C^2} \lesssim 2^{-J(\alpha-1)} \|f\|_{C^{\alpha+1}}$ and if $f \in W^{\alpha+1,2}$, for $\alpha > 0$, then $\|f - \Pi_J f\|_{W^{1,2}} \lesssim 2^{-J\alpha} \|f\|_{W^{\alpha+1,2}}$.

PROOF. Write γ for the wavelet coefficients of f. The statements (5.12) and (5.13) follow immediately from the wavelet characterization of the Besov norms, (B.2).

To prove the remaining statements, note that for every j, because $\|\,.\,\|_{\ell^{p'}} \le \|\,.\,\|_{\ell^p}$ and $\|\,.\,\|_q$ and $\|\,.\,\|_{q'}$ are comparable up to a constant due to $|G^j| \le 2^d$ being finite,

$$\begin{aligned}
&2^{j(s'+\frac{d}{2}-\frac{d}{p'})} \Big[\sum_{g \in G^j} \Big(\sum_{k \in \mathbb{Z}^d} |\gamma_k^{j,g}|^{p'} \Big)^{q'/p'} \Big]^{1/q'} \\
&\quad = 2^{j(s'-\frac{d}{p'}-(s-\frac{d}{p}))}\, 2^{j(s+\frac{d}{2}-\frac{d}{p})} \Big[\sum_{g \in G^j} \Big(\sum_{k \in \mathbb{Z}^d} |\gamma_k^{j,g}|^{p'} \Big)^{q'/p'} \Big]^{1/q'} \\
&\quad \lesssim 2^{j(s'-\frac{d}{p'}-(s-\frac{d}{p}))}\, 2^{j(s+\frac{d}{2}-\frac{d}{p})} \Big[\sum_{g \in G^j} \Big(\sum_{k \in \mathbb{Z}^d} |\gamma_k^{j,g}|^{p} \Big)^{q/p} \Big]^{1/q} \\
&\quad \le 2^{j(s'-\frac{d}{p'}-(s-\frac{d}{p}))} \|f\|_{B^s_{p,q}}.
\end{aligned}$$

Then, plugging this into the wavelet expansion of $\Pi_J f - f$, we obtain

$$\begin{aligned}
\|\Pi_J f - f\|^{q'}_{B^{s'}_{p',q'}} &= \sum_{j=J+1}^{\infty} 2^{jq'(s'+\frac{d}{2}-\frac{d}{p'})} \sum_{g \in G^j} \Big(\sum_{k \in \mathbb{Z}^d} |\gamma_k^{j,g}|^{p'} \Big)^{q'/p'} \\
&\lesssim \sum_{j=J+1}^{\infty} 2^{jq'(s'-\frac{d}{p'}-(s-\frac{d}{p}))} \|f\|^{q'}_{B^s_{p,q}} \lesssim 2^{Jq'(s'-\frac{d}{p'}-(s-\frac{d}{p}))} \|f\|^{q'}_{B^s_{p,q}},
\end{aligned}$$

Finally, to obtain the special cases, note that $\|\,.\,\|_{B^s_{\infty,\infty}} \lesssim \|\,.\,\|_{C^s} \lesssim \|\,.\,\|_{B^s_{1,\infty}}$ and $\|\,.\,\|_{W^{s,2}} = \|\,.\,\|_{B^s_{2,2}}$ by Theorem 22. □

The above lemma together with Proposition-Definition 7 allows us to check that $\bar{f} \in \mathcal{F}_J$. Indeed, by Weyl's inequality, we have for any $x \in \tilde{\Omega}_P$ that

$$\begin{aligned}
\lambda_{\min}(D^2 \Pi_J \operatorname{ext} f_0(x)) &\ge \lambda_{\min}(D^2 f_0(x)) - \|D^2 \Pi_J \operatorname{ext} f_0(x) - D^2 f_0(x)\|_{\mathrm{op}} \\
&\ge M^{-1} - C \|\Pi_J \operatorname{ext} f_0 - f_0\|_{C^2(\tilde{\Omega}_P)}.
\end{aligned}$$

It follows from Lemma 13 that $\|\Pi_J \operatorname{ext} f_0 - f_0\|_{C^2(\tilde{\Omega}_P)} \lesssim 2^{-(\beta-1)J}(M^3 + R) \le 1/(2M)$, if

$$J \ge J_0 := C_3 \frac{1}{\beta - 1} \log\left(2M^4 + 2RM\right) ,$$

and $C_3 = C_3(d, \beta, \tilde{\Omega}_P)$ is large enough. This yields $\lambda_{\min}(D^2 \Pi_J \operatorname{ext} f_0(x)) \ge 1/(2M)$ and $\bar{f}$ is strongly convex. Similarly, we get $\lambda_{\max}(D^2 \Pi_J \operatorname{ext} f_0(x)) \le 2M$ and hence that $\bar{f} \in \mathcal{F}_J$. Thus,

$$\begin{aligned} \inf_{f \in \mathcal{F}_J} \|\nabla f - \nabla f_0\|^2_{L^2(P)} \le \|\nabla \bar{f} - \nabla f_0\|_{L^2(P)} &\lesssim \int_{\Omega_P} \|\nabla \bar{f} - \nabla f_0\|_2^2 \,\mathrm{d}\lambda(x) \\ &\le \|\bar{f} - f_0\|^2_{W^{1,2}(\Omega_P)} \lesssim R^2\, 2^{-2J\alpha}. \end{aligned}$$

where we used Assumption B2 and Lemma 13.

We have thus proved that

$$\inf_{f \in \mathcal{F}_J} \|\nabla f - \nabla f_0\|^2_{L^2(P)} \lesssim R^2\, 2^{-2J\alpha} \quad \text{for } J \ge J_0. \tag{5.14}$$

5.6 Bias-variance tradeoff

We are now in a position to complete the proof of Theorem 8. Combining the bounds (5.11) and (5.14), we get

$$\mathbb{E}\|\hat{T}_J - T_0\|^2_{L^2(P)} \lesssim R^2\, 2^{-2J\alpha} + 2^{J(d-2)} J \frac{\log(n)}{n} + \frac{1}{n}\,.$$

We conclude the proof by optimizing with respect to J. It yields

$$\mathbb{E}\|\hat{T}_J - T_0\|^2_{L^2(P)} \lesssim \begin{cases} n^{-1} & \text{if } d = 1, \\ \log(R)\, n^{-1} (\log(n))^2 & \text{if } d = 2, \\ R^2\, (1/R^2)^{\frac{2\alpha}{2\alpha - 2 + d}} \log(R)\, n^{\frac{-2\alpha}{2\alpha-2+d}} (\log(n))^2 & \text{if } d \ge 3. \end{cases} \tag{5.15}$$

We note that since $\alpha > 1$, in the first two cases, $d \in \{1, 2\}$, the cut-off J can be picked independently from α. Finally, high-probability bounds can be obtained in a similar manner.

6. NUMERICAL EXPERIMENTS

In this section, we provide numerical experiments with synthetic data that illustrate how leveraging the smoothness of the underlying transport map can lead to dramatically improved rates. We give two estimators exploiting smoothness: the first, $\hat{T}_{\mathrm{wav}}$ below, is an approximation to the estimator $\hat{T}_J$ in (5.4), illustrating that (5.4) can be implemented in low dimensions and that this approximation achieves favorable rates in $d = 3$. The second, $\hat{T}_{\mathrm{ker}}$ below, is a more practical heuristic two-step procedure based on smoothing the optimal matching between the empirical distributions via radial basis functions. We compare these to a baseline estimator given by the optimal transport plan between the empirical distributions.

Additional implementation details and comments on these experiments are provided in Section H of the Appendix.

6.1 Estimators

6.1.1 Baseline estimator In order to highlight the benefit of regularization, we consider the following simple estimator based on the optimal transport matrix between the empirical distributions as a baseline. Denote the empirical distributions of P and Q by

$$\hat{P} = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}, \quad \hat{Q} = \frac{1}{n}\sum_{i=1}^{n} \delta_{Y_i},$$

respectively, and calculate the optimal transport matrix $\Gamma \in \mathbb{R}^{n\times n}$

$$\hat{\Gamma} = \operatorname{argmin}\left\{ \sum_{i,j=1}^{n} \|X_i - Y_j\|^2 \Gamma_{i,j} \ : \ \begin{array}{l} \Gamma_{i,j} \geq 0 \ \forall i, j \in [n], \\ \Gamma \mathbb{1} = \mathbb{1}/n, \ \Gamma^\top \mathbb{1} = \mathbb{1}/n \end{array} \right\}$$

which can be solved exactly by linear optimization toolkits or approximated by entropic regularization [PC19]. An estimated transport function on the observations X_i is then obtained by

$$\hat{T}_{\mathrm{emp}}(X_i) = n \sum_{j=1}^{n} \hat{\Gamma}_{i,j} Y_j,$$

which corresponds to the conditional mean of the coupling given X_i. Note that since we assume that the sample sizes from P and Q are both n, the optimal $\hat{\Gamma}$ is in fact a (rescaled) permutation matrix, leading to a matching $\hat{\pi} : [n] \to [n]$, and hence to $\hat{T}_{\mathrm{emp}}(X_i) = Y_{\pi(i)}$.

Because $\hat{T}_{\mathrm{emp}}(X_i)$ above is only defined on the sample points and we do not want to introduce additional bias against the estimator, we consider the following error measure, approximating the $L^2(P)$ norm analyzed in Theorem 2:

$$\mathrm{MSE}_n(\hat{T}_{\mathrm{emp}}) = \frac{1}{n}\sum_{i=1}^{n} \left\| \hat{T}_{\mathrm{emp}}(X_i) - T_0(X_i) \right\|_2^2.$$

6.1.2 Wavelet estimator Next, we turn to an approximation of (5.4). Assume that in addition to a superset $\tilde{\Omega}_P \supseteq \Omega_P$, we are also given a superset $\tilde{\Omega}_Q \supseteq \Omega_Q$, and that both $\tilde{\Omega}_P$ and $\tilde{\Omega}_Q$ are boxes (hypercubes). We consider all functions originally defined over $\tilde{\Omega}_P$ and $\tilde{\Omega}_Q$ as given by their samples on grids with resolution $N \in \mathbb{N}$, $\boldsymbol{x} = (x_i)_{i\in[N]^d} \in (\mathbb{R}^d)^{N^d}$ and $\boldsymbol{y} = (y_i)_{i\in[N]^d} \in (\mathbb{R}^d)^{N^d}$, respectively. In particular, we write $\boldsymbol{f} = (f_i)_{i\in[N]^d} \in \mathbb{R}^{N^d}$ for the discretization of the potential f and $\boldsymbol{T} = (T_i)_{i\in[N]^d} \in (\mathbb{R}^d)^{[N]^d}$ for the discretization of the transport map T on the grid $\boldsymbol{x}$. Here, we pick $N = 65$.

We employ the following discretization/approximation schemes:

(i) The restrictions to functions up to wavelet scale $J \in \mathbb{N}$ can be obtained by parametrizing $\boldsymbol{f}$ by the inverse discrete wavelet transform up to order J, which we write as $\boldsymbol{f} = W_J^\top \gamma_J$ for wavelet coefficients $\gamma_J \in \mathbb{R}^{m_J}$ with $m_J \in \mathbb{N}$. For these experiments, we use db4 Daubechies wavelets, i.e., Daubechies wavelets with four vanishing moments.

(ii) An approximation to the (continuous) Legendre transform is given by the discrete Legendre transform,

$$\begin{aligned} \mathcal{L}(\boldsymbol{f})_j &:= \mathcal{L}(\boldsymbol{f})(y_j) := \mathcal{L}_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{f})(y_j) \\ &:= \sup_{i\in[N]^d} \{\langle x_i, y_i\rangle - f_i : i \in [N]^d\}, \end{aligned} \tag{6.1}$$

for $\boldsymbol{f} \in (\mathbb{R}^d)^{N^d}$ and $j \in [N]^d$.

(iii) The gradient of $\boldsymbol{f}$ on grid-points can be calculated by a finite-difference scheme, which we write as $\nabla_{\boldsymbol{x}}\boldsymbol{f}$.
(iv) To obtain values of these approximations at non-grid points, we appeal to linear interpolation, which we write as $\mathcal{P}_{\boldsymbol{x}\to\{X_1,\dots,X_n\}}\boldsymbol{f}$ for the interpolated value of $\boldsymbol{f}$ defined on the grid $\boldsymbol{x}$ at the point $\{X_1,\dots,X_n\}\in\mathbb{R}^d$, collected in one vector.

Given *i.i.d.* observations $X_{1:n}=\{X_1,\dots,X_n\}$ and $Y_{1:n}=\{Y_1,\dots,Y_n\}$ from P and Q, respectively, we arrive at the following optimization problem as approximation to (5.4):

$$\hat{\gamma}_J = \min_{\gamma_J\in\mathbb{R}^{m_J}} \frac{1}{n}\mathbb{1}^\top \mathcal{P}_{\boldsymbol{x}\to X_{1:n}} W_J^\top \gamma_J + \frac{1}{n}\mathbb{1}^\top \mathcal{P}_{\boldsymbol{y}\to Y_{1:n}} \mathcal{L}(W_J^\top \gamma_J). \tag{6.2}$$

Note that we dropped all boundedness and convexity constraints that were given by $\mathcal{X}(2M)$ in (5.4). In practice, this can lead to degraded estimation quality of the gradient of $W_J^\top\gamma_J$ near the boundary of Ω_P, see Section H.3 in the Appendix, which we remedy by computing the convex envelope of the resulting estimator, yielding an estimator of f_0 that is convex. This envelope can, for example, be calculated by applying the Legendre transform twice, and we set

$$\hat{\boldsymbol{T}}_J = \nabla_{\boldsymbol{x}}[\mathcal{L}_{\boldsymbol{y}\to\boldsymbol{x}}(\mathcal{L}_{\boldsymbol{x}\to\boldsymbol{y}}(W_J^\top\hat{\gamma}_J))].$$

With this, we denote by $\hat{T}_{\mathrm{wav}}^{(J)}$ the function obtained by linearly interpolating $\hat{\boldsymbol{T}}_J$,

$$\hat{T}_{\mathrm{wav}}^{(J)}(x) = \mathcal{P}_{\boldsymbol{x}\to x}\hat{\boldsymbol{T}}_J, \quad x\in\tilde{\Omega}_P.$$

For the purpose of these experiments, we select the wavelet scale $\hat{J}$ by an oracle choice, i.e., as the minimizer of an approximation to the population semi-dual (2.5), see Section H.1, while in practice, one would resort to cross-validation methods for this purpose. Finally, we set

$$\hat{T}_{\mathrm{wav}} = \hat{T}_{\mathrm{wav}}^{(\hat{J})}.$$

To obtain an error measure for $\hat{T}_{\mathrm{wav}}$ that is easily comparable to $\mathrm{MSE}_n(\hat{T}_{\mathrm{emp}})$, we consider the empirical $L^2(\hat{P})$ norm on the sample points,

$$\mathrm{MSE}_n(\hat{T}_{\mathrm{wav}}) = \frac{1}{n}\sum_{i=1}^n \left\|\hat{T}_{\mathrm{wav}}(X_i) - T_0(X_i)\right\|_2^2.$$

Note that the objective function in (6.2) can be calculated in linear time with respect to the underlying grid, that is, in $O(N^d)$, thanks to efficient algorithms for the discrete wavelet transform [Mal99] and the Linear-Time Legendre Transform algorithm [Luc97]. It can be checked that the objective is convex, rendering first-order methods provably convergent. We use the L-BFGS Quasi-Newton method to find $\hat{\gamma}_J$, even though the objective is not smooth for every γ_J due to the form of the discrete Legendre transform. In practice, we observe that it converges faster than simple (sub-)gradient descent methods.

Since $\hat{T}_{\mathrm{wav}}$ is mainly used to illustrate the practical behavior of a wavelet-based regularization of the semi-dual objective, we do not explicitly analyze the convergence of $\hat{T}_{\mathrm{wav}}$ to the estimator $\hat{T}_J$ considered in Section 5. We remark, however, that our approximations closely follow the definition of $\hat{T}_J$ and that the omitted constraints defining the set $\mathcal{X}(2M)$ could be incorporated into the approximation by means of a finite-difference discretization as well, albeit at an additional computational cost.

6.1.3 Kernel smoothing estimator While the estimator $\hat{T}_{\mathrm{wav}}$ closely follows our theoretical analysis, its applicability is limited to low dimensions by the fact that a discretization grid is needed. As a heuristic alternative, we consider smoothing the assignment obtained by $\hat{T}_{\mathrm{emp}}$. The idea of smoothing the empirical transport matrix has been previously used in practice, see for example [SST+19], where regularized linear regression was used as a post-processing step, and kernels have been applied for smoothing potential functions in [GCPB16]. Here, we obtain a smoother version of $\hat{T}_{\mathrm{emp}}$ by smoothing it via kernel-ridge regression [Mur12].

Let $\mathcal{H}$ denote a reproducing kernel Hilbert space (RKHS) with associated kernel $k(x,y)$, $x, y \in \mathbb{R}^d$, and norm $\|T\|_{\mathcal{H}}$ for $T \in \mathcal{H}$. Here, we consider the RKHS given by Gaussian radial basis functions,

$$k(x,y) = \exp(-\nu_{\mathrm{kernel}}\|x-y\|^2), \quad x, y \in \mathbb{R}^d, \quad \nu_{\mathrm{kernel}} > 0.$$

We fit an RKHS function T to the pairs $(X_i, \tilde{Y}_i = \hat{T}_{\mathrm{emp}}(X_i))$ by solving the regularized kernel regression objective

$$\hat{T}_{\mathrm{ker}}^{(\nu_{\mathrm{ridge}},\nu_{\mathrm{kernel}})} = \operatorname*{argmin}_{T \in \mathcal{H}} \sum_{i=1}^{n} \|\tilde{Y}_i - T(X_i)\|_2^2 + \nu_{\mathrm{ridge}}\|T\|_{\mathcal{H}}^2, \tag{6.3}$$

for $\nu_{\mathrm{ridge}} > 0$. By the representer theorem [Mur12], (6.3) has a solution

$$\hat{T}_{\mathrm{ker}}^{(\nu_{\mathrm{ridge}},\nu_{\mathrm{kernel}})}(x) = \sum_{i=1}^{n} \hat{w}_i k(x_i, x),$$

where

$$\hat{W} = \begin{pmatrix} \hat{w}_1^\top \\ \vdots \\ \hat{w}_n^\top \end{pmatrix} = (K + \nu_{\mathrm{ridge}} I)^{-1}\tilde{Y}, \quad \text{with } K_{i,j} = k(X_i, X_j) \text{ and } \tilde{Y} = \begin{pmatrix} \tilde{Y}_1 \\ \vdots \\ \tilde{Y}_n \end{pmatrix}.$$

We measure its performance by

$$\mathrm{MSE}_n(\hat{T}_{\mathrm{ker}}^{(\nu_{\mathrm{ridge}},\nu_{\mathrm{kernel}})}) = \frac{1}{n}\sum_{i=1}^{n} \left\| \hat{T}_{\mathrm{ker}}^{(\nu_{\mathrm{ridge}},\nu_{\mathrm{kernel}})}(X_i) - T_0(X_i) \right\|_2^2. \tag{6.4}$$

Similar to $\hat{T}_{\mathrm{wav}}$, we select the tuning parameters ν_{kernel} and ν_{ridge} by an oracle procedure, picking those parameters from a finite grid that minimize (6.4) on an independent hold-out sample $\tilde{X}_{1:n}$, and denote the resulting estimator by $\hat{T}_{\mathrm{ker}}$.

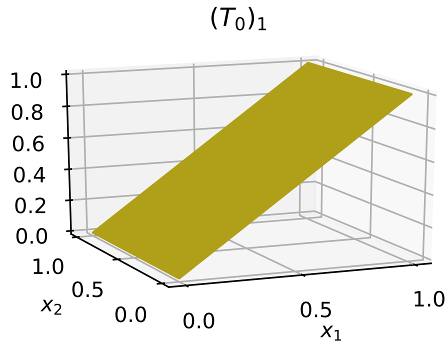
6.2 Setup

For $d \in \mathbb{N}$, we consider the following examples of smooth potentials and transport maps:

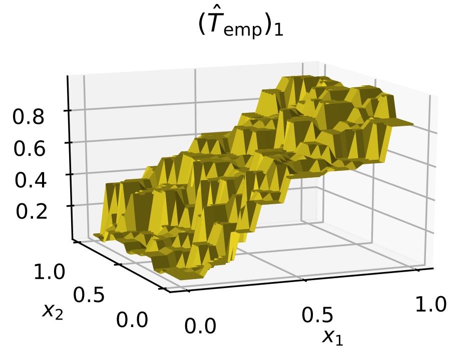
$$f_0^{(1)}(x) = \frac{1}{2}\|x\|_2^2, \qquad T_0^{(1)}(x) = x, \quad x \in \mathbb{R}^d; \tag{id}$$

$$f_0^{(2)}(x) = \sum_{i=1}^{d} \exp(x_i), \quad (T_0^{(2)}(x))_i = \exp(x_i), \quad x \in \mathbb{R}^d, \ i \in [d]; \tag{exp}$$

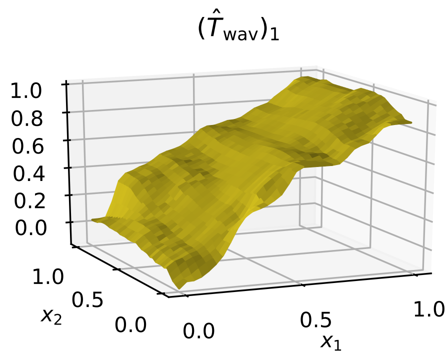
where for (id), $P^{(1)} = Q^{(1)} = \mathsf{Unif}([0,1]^d)$. For (exp), $P^{(2)} = \mathsf{Unif}([0,1]^d)$, and the target measure is defined as $Q^{(2)} = (T_0^{(2)})_\# P^{(2)}$. Note that these potentials and transport maps are C^∞ and strongly



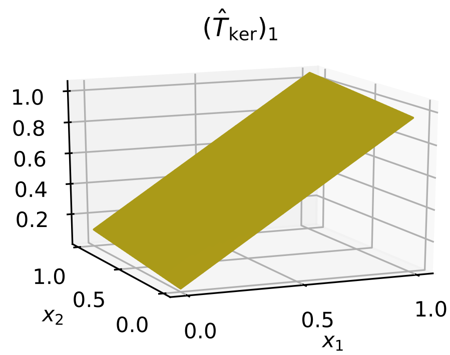
(a) Ground truth transport map



(b) Transport map estimator from empirical distributions and 1-NN interpolation



(c) Wavelet estimator, $J = 1$



(d) Kernel estimator, $\nu_{\mathrm{ridge}} = \nu_{\mathrm{kernel}} = 10^{-3}$

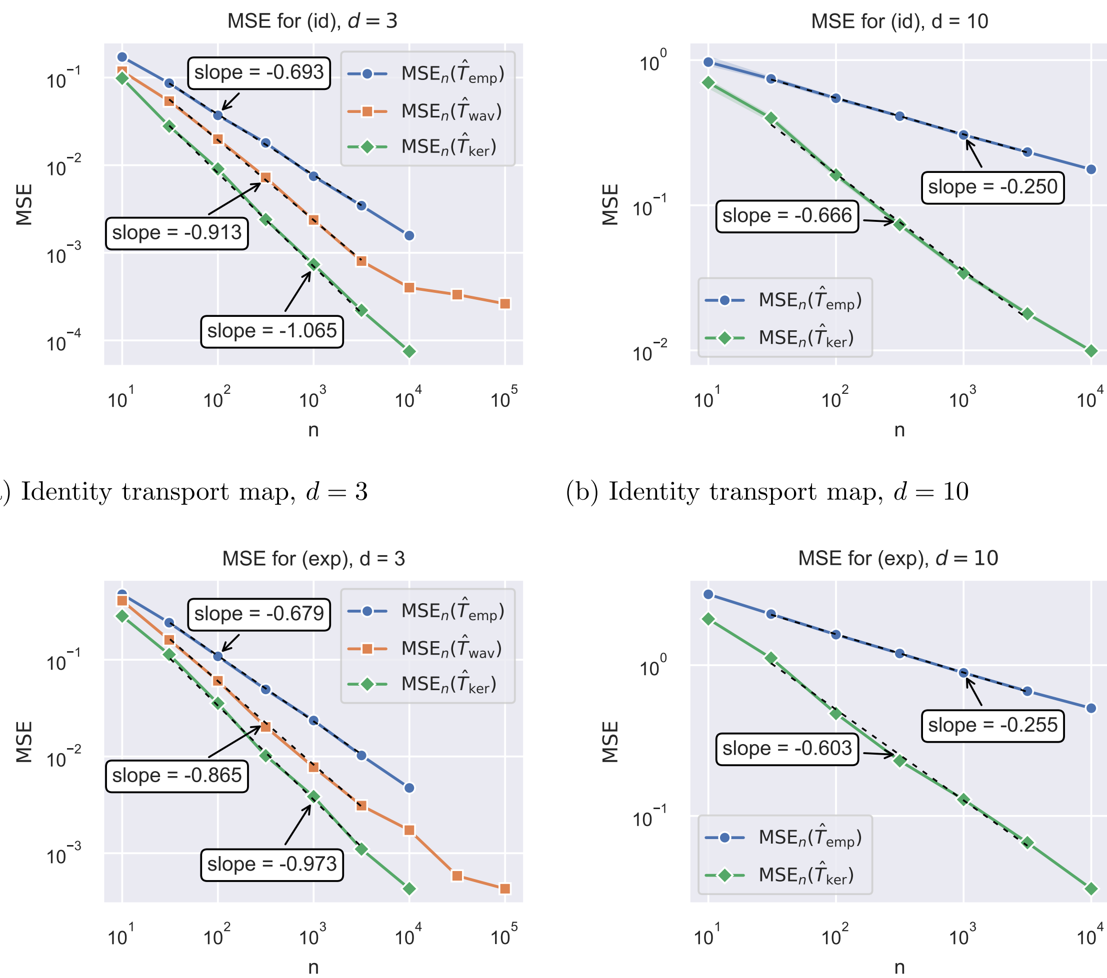
Figure 1: Qualitative comparison between $\hat{T}_{\mathrm{emp}}$, $\hat{T}_{\mathrm{wav}}$, and $\hat{T}_{\mathrm{ker}}$. Both the wavelet-based estimator $\hat{T}_{\mathrm{wav}}$ and the kernel estimator $\hat{T}_{\mathrm{ker}}$ produce a qualitatively smoother output then the optimal coupling between the empirical measures. Visualizations of the first coordinate of the transport maps.

convex on any compact convex subset of $\mathbb{R}^d$. For the purpose of qualitative comparisons, we consider case (id) in $d = 2$. In order to determine the quantitative behavior of the estimators, we study both cases for $d \in \{3, 10\}$.

The runtime of computing the optimal transport matching via linear programming scales unfavorably with the sample size, a shortcoming that could be remedied by employing recent numerical approximation techniques [ABRNW19]. Similarly, computing the kernel regression (6.3) with off-the-shelf methods scales with $O(n^3)$. For the sake of this comparison, we simply restrict our experiments on $\hat{T}_{\mathrm{emp}}$ and $\hat{T}_{\mathrm{ker}}$ to sample sizes $\leq 10^4$. Likewise, we do not compute $\hat{T}_{\mathrm{wav}}$ in $d = 10$ due to the large computational cost.

6.3 Results

6.3.1 Qualitative comparison in 2D To give a qualitative idea of the considered estimators compared to the baseline $\hat{T}_{\mathrm{emp}}$, we visualize the first coordinate of the transport map estimators for case (id) with $d = 2$ and $n = 100$ observations from $P = Q = \mathsf{Unif}([0,1]^2)$ in Figure 1, together with the ground truth transport map $(T_0^{(1)})_1$. To depict the coupling $\hat{T}_{\mathrm{emp}}$ induced by the empirical distributions, we employ 1-nearest-neighbor (1-NN) interpolation to obtain a map on $[0,1]^2$. We observe that the wavelet-based regularization in Figure 1c produces a visibly smoother map compared to the unregularized $\hat{T}_{\text{emp}}$ in Figure 1b. The kernel estimator in Figure 1d is even smoother and visually very similar to the ground truth in Figure 1a, due to the possibility of employing a large amount of regularization.



(a) Identity transport map, $d = 3$

(b) Identity transport map, $d = 10$

(c) Exponential transport map, $d = 3$

(d) Exponential transport map, $d = 10$

Figure 2: Log-log plot of MSE plotted against n, showing the median error over 32 replicates. See Section 6.2 for details.

6.3.2 Quantitative comparison in 3D and 10D To obtain a quantitative comparison, for both test cases and $d \in \{3, 10\}$, we compute $\hat{T}_{\text{emp}}$, $\hat{T}_{\text{wav}}$ (only $d = 3$), and $\hat{T}_{\text{ker}}$ over 32 replicates and a logarithmically spaced selection of sample sizes n, reporting the median error over the replicates in Figure 2. Here, the dashed lines indicate the result of linear regression on the logarithmically transformed sample sizes and error results for a selected subset of n.

In 3D, for both test cases, the error curves for the standard empirical measure-based estimator $\hat{T}_{\text{emp}}$ roughly follow a $n^{-2/3}$ rate. This corresponds to the decay of the average ℓ^2 cost of optimal matchings between samples from the uniform distribution on the cube [Tal14, Yuk06] (and Gaussian distributions [Led19]), and also matches the $n^{-2/d}$ rate for the convergence of $W_2^2(\hat{P}, P)$. We are, however, not aware of a generalization of this rate to the error measure MSE in the case of a transport map different from the identity.

The error curves for the wavelet estimator $\hat{T}_{\mathrm{wav}}$ all follow a similar trend. For low sample sizes, we obtain rates faster than $n^{-0.85}$, showing the large statistical benefit of fitting only functions that have wavelet expansions of low order. For large sample sizes, the error flattens out, which can be explained by the numerical approximation errors dominating the statistical ones. This can be readily seen from repeating the experiment with a smaller grid resolution ($N = 33$), which shows the same trend for smaller values of n, see Section H.2 in the Appendix. Moreover, for all sample sizes we considered, the error curves for $\hat{T}_{\mathrm{wav}}$ all lie below the baseline estimator. The kernel estimator $\hat{T}_{\mathrm{ker}}$ performs even better, attaining rates close to n^{-1} and yielding consistently better rates than $\hat{T}_{\mathrm{wav}}$.

We observe that the favorable behavior of $\hat{T}_{\mathrm{wav}}$ suggests that the restriction of candidate potentials to $\mathcal{X}(2M)$ in (5.4) might not be necessary and could possibly be omitted.

In 10D, for both test cases, $\hat{T}_{\mathrm{emp}}$ shows a convergence rate of about $n^{-0.25}$, which is slightly better than the expected $n^{-2/d} = n^{-0.2}$ rate. It is vastly outperformed by the kernel-based estimator that achieves rates better than $n^{-0.6}$ in both examples.

Further plots showing individual error curves for varying values of the regularization parameters can be found in Section H.4 of the appendix, illustrating the sample size-dependent performance gain achieved by $\hat{T}_{\mathrm{wav}}$ and $\hat{T}_{\mathrm{ker}}$.

To summarize, in cases where smoothness of the transport map can be assumed, its estimation greatly benefits from smoothness regularization. In particular, these experiments suggest further research on proving error bounds for the kernel estimator $\hat{T}_{\mathrm{ker}}$ under regularity assumptions on T_0, for which the minimax rates established in this work can serve as a benchmark.

Acknowledgments. The authors would like to thank Richard Nickl for pointing out relevant references on the estimation of invariant measures in the multivariate Gaussian white noise model, and Jonathan Niles-Weed for suggesting references on the rate between random matchings. The authors are also indebted to the Associate Editor and three referees for providing insightful comments that greatly improved the quality of the paper.

Philippe Rigollet was supported by NSF awards IIS-1838071, DMS-1712596, DMS-TRIPODS-1740751, and ONR grant N00014-17- 1-2147.

APPENDIX A: CONVEX ANALYSIS

In this section, we recall some useful facts from convex analysis. We refer the reader to [HL01] for a comprehensive treatment.

Recall that a set $U \subseteq \mathbb{R}^d$ is convex if for all $x, y \in U$, $t \in [0, 1]$, $tx + (1-t)y \in U$. A function $f \colon U \to \mathbb{R} \cup \{+\infty\}$ is convex if for all $x, y \in U$, $t \in [0, 1]$, it holds that

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y).$$

Moreover, we call a function μ-strongly convex if for all $x, y \in U$, $t \in [0, 1]$, we have

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) - \frac{\mu}{2} t(1-t)\|x - y\|_2.$$

For twice differentiable functions, it is often more convenient to employ the following analytic criterion for strong convexity.

LEMMA 14 ([HL01, Theorem B.4.3.1]). *Let $U \subseteq \mathbb{R}^d$ be an open convex set and $f \colon U \to \mathbb{R}$ twice differentiable. Then, f is μ-strongly convex if and only if $\lambda_{\min}(D^2 f(x)) \geq \mu$ for all $x \in U$.*

Note that even if $U \subsetneq \mathbb{R}^d$ is a proper convex subset of $\mathbb{R}^d$, we can always consider a function $f \colon U \to \mathbb{R} \cup \{+\infty\}$ to be defined on all of $\mathbb{R}^d$ by setting it to $+\infty$ outside of U. To that end, let ι_U be the *indicator* function defined by

$$\iota_U(x) = \begin{cases} 0, & x \in U \\ +\infty, & \text{otherwise.} \end{cases}$$

We define the extension of f outside U by $f + \iota_U$, which by abuse of notation we also denote by f. Note that if f is (strongly) convex on U then its extension outside U is also (strongly) convex. We call $\mathrm{dom}(f) = \{x \in \mathbb{R}^d : f(x) < +\infty\}$ the domain of a convex function.

We now recall two important notions associated with convex functions f.

First, the subdifferential of f at $x \in \mathrm{dom}(f)$ is defined as

$$\partial f(x) = \{a \in \mathbb{R}^d : f(y) \geq \langle a, y - x\rangle + f(x) \text{ for all } y \in \mathbb{R}^d\}.$$

As indicated by the following lemma, the subdifferential reduces to the gradient for differentiable functions.

LEMMA 15 ([HL01, Corollary D.2.1.4]). *Let $f \colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a convex function. If f is differentiable at $x \in \mathbb{R}^d$ with gradient $\nabla f(x)$, then $\partial f(x) = \{\nabla f(x)\}$.*

Conversely, if $\partial f(x) = \{a\}$ consists of only a single element, then f is differentiable at x with gradient $\nabla f(x) = a$.

Second, the convex conjugate, or Legendre-Fenchel conjugate, is defined for any function $f \colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ as

$$f^*(y) = \sup_{x \in \mathbb{R}^d} \langle x, y\rangle - f(x), \quad y \in \mathbb{R}^d.$$

By considering $f + \iota_U$, this definition extends to functions $f \colon U \to \mathbb{R} \cup \{+\infty\}$.

We recall the following standard facts about the convex conjugate, stated here without proof (see [HL01, Part E] for details).

LEMMA 16. *If $f \colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is convex and lower semi-continuous, then $f^{**} = f$.*

LEMMA 17. *Let U be a closed, convex set. If $f \colon U \to \mathbb{R}$ is μ-strongly convex, then $\mathrm{dom}(f^*) = \mathbb{R}^d$, and ∇f^* is μ^{-1}-Lipschitz:*

$$\|\nabla f^*(x) - \nabla f^*(y)\|_2 \leq \frac{1}{\mu}\|x - y\|_2\,, \qquad \forall\, x, y \in \mathbb{R}^d\,.$$

LEMMA 18. *Denote by $f \colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ a lower semi-continuous and convex function. Then, for $x, y \in \mathbb{R}^d$,*

$$f^*(y) = \langle x, y\rangle - f(x) \iff y \in \partial f(x) \iff x \in \partial f^*(y).$$

LEMMA 19. *If $f, g\colon U \to \mathbb{R}$, $f(x) \le g(x)$ for all $x \in U$, then $f^*(y) \ge g^*(y)$ for all $y \in \mathbb{R}$.*

LEMMA 20. *Given two functions $f, g : U \to \mathbb{R}$ for a compact set U. Abusing notation, denote the convex conjugate of the function $f + \iota_U$ by f^*. Then, $\|f^* - g^*\|_{L^\infty(\mathbb{R}^d)} \le \|f - g\|_{L^\infty(U)}$.*

The next lemma provides an explicit form for the convex conjugate of a quadratic function. It is instrumental in the stability proof for the semi-dual objective function (see Proposition 10).

LEMMA 21. *Let $a > 0$, $b, t \in \mathbb{R}^d$, $c \in \mathbb{R}$ and let $U \subseteq \mathbb{R}^d$ be a closed, convex set. Define*

$$q_t(x) = \frac{a}{2}\|x - t\|_2^2 + \langle b, x - t\rangle + c + \iota_U(x)\,, \quad x \in \mathbb{R}^d\,.$$

Then,

$$q_t^*(y) = \frac{\|y - b\|_2^2}{2a} + \langle t, y\rangle - c - \frac{a}{2} d^2\left(\frac{y - b}{a} - t, U\right),$$

where d^2 denotes the squared distance $d^2(x, U) = \inf_{y \in U} \|x - y\|_2^2$.

PROOF. Note first that for any $y \in \mathbb{R}^d$,

$$q_t^*(y) = \sup_{x \in \mathbb{R}^d} \langle x, y\rangle - q(x - t) = \sup_{x \in \mathbb{R}^d} \langle x + t, y\rangle - q(x) = q^*(y) + \langle t, y\rangle\,, \tag{A.1}$$

where

$$q(x) = \frac{a}{2}\|x\|_2^2 + b^\top x + c + \iota_{U+t}(x)$$

Moreover,

$$q^*(y) = \sup_{x \in \mathbb{R}^d} \langle x, y\rangle - q(x) = -\inf_{x \in \mathbb{R}^d} \{q(x) - \langle x, y\rangle\}. \tag{A.2}$$

Writing

$$q(x) - \langle x, y\rangle = \frac{a}{2}\left\|x - \frac{y - b}{a}\right\|_2^2 - \frac{\|y - b\|_2^2}{2a} + c + \iota_{U+t}(x)$$

we see that the infimum in (A.2) is achieved by the projection $\bar{x}$ of $(y - b)/a$ onto the closed convex set $U + t$. Moreover, the value of the objective at $\bar{x}$ is given by

$$\begin{aligned} q(\bar{x}) - \langle \bar{x}, y\rangle &= \frac{a}{2} d^2\left(\frac{y - b}{a}, U + t\right) - \frac{\|y - b\|_2^2}{2a} + c \\ &= \frac{a}{2} d^2\left(\frac{y - b}{a} - t, U\right) - \frac{\|y - b\|_2^2}{2a} + c \end{aligned}$$

Together with (A.1) and (A.2), this completes the proof of the lemma. □

APPENDIX B: WAVELETS AND FUNCTION SPACES

In this section, we give a brief overview of the different function spaces used in the paper and how their norms can be related to their wavelet coefficients.

First, we recall the definition of Hölder and Sobolev spaces. Let $\Omega \subseteq \mathbb{R}^d$ be a closed set with non-empty interior and denote by $C_u(\Omega)$ the set of uniformly continuous functions on Ω. The *Hölder* norms for any $f \in C_u(\Omega)$ are defined as follows. For any integer $k \geq 0$ and f that admits continuous derivatives up to order k, define

$$\|f\|_{C^k(\Omega)} := \sum_{|\flat| \leq k} \|\partial^\flat f\|_{L^\infty(\Omega)},$$

and for any real number $\alpha > 0$, define

$$\|f\|_{C^\alpha(\Omega)} := \|f\|_{C^{\lfloor \alpha \rfloor}(\Omega)} + \sum_{|\flat| = \lfloor \alpha \rfloor} \sup_{x \neq y, x, y \in \Omega} \frac{|\partial^\flat f(x) - \partial^\flat f(y)|}{\|x - y\|_2^{\alpha - \lfloor \alpha \rfloor}}.$$

The space $C^\alpha(\Omega)$ is then defined as the set of functions for which this norm is finite. For a vector-valued function $T \colon \Omega \to \mathbb{R}^d$, $T = (T_1, \dots, T_d)^\top$, we similarly define the norms as the sum over the individual norms,

$$\|T\|_{C^\alpha(\Omega)} := \sum_{i=1}^d \|T_i\|_{C^\alpha(\Omega)}.$$

Similarly, for an integer $k \geq 0$ and $p \in [1, \infty]$, the Sobolev norms are defined as

$$\|f\|_{W^{k,p}(\Omega)} := \sum_{|\flat| \leq k} \|\partial^\flat f\|_{L^p(\Omega)},$$

where the derivative $\partial^\flat$ is to be understood in the sense of distributions and the Sobolev space $W^{m,2}(\Omega)$ is the space of all functions for which this norm is finite. This definition can be extended to $\alpha > 0$, for example by defining $W^{\alpha,2}(\Omega)$ as the Besov space $B^\alpha_{2,2}(\Omega)$, which we define shortly.

Next, we define wavelet bases and Besov spaces, following the definitions given in [Tri06, Section 3], and we refer the reader to this reference for further details on wavelets. Denote by $\psi_{\mathfrak{M}} \in C^r(\mathbb{R})$ and $\psi_{\mathfrak{F}} \in C^r(\mathbb{R})$ a compactly supported wavelet and scaling function, respectively, for example Daubechies wavelets. This implies that

$$\psi^j_k = \begin{cases} \psi_{\mathfrak{F}}(x - k), & j = 0, k \in \mathbb{Z}, \\ 2^{(j-1)/2} \psi_{\mathfrak{M}}(2^{j-1} x - k), & j \in \mathbb{N}, k \in \mathbb{Z}, \end{cases}$$

is an orthonormal basis of $L^2(\mathbb{R})$. To obtain a basis of $L^2(\mathbb{R}^d)$, for $j \in \mathbb{N}$, set

$$G^j = \{\mathfrak{F}, \mathfrak{M}\}^d \setminus \{(\mathfrak{F}, \dots, \mathfrak{F})\}, \quad G^0 = \{(\mathfrak{F}, \dots, \mathfrak{F})\},$$

and for $g \in G^j \cup G^0$,

$$\Psi^g_k(x) = \prod_{i=1}^n \psi_{g_i}(x_i - k_i), \quad k \in \mathbb{Z}^d.$$

This gives the orthonormal basis

$$\Psi^{j,g}_k = \begin{cases} \Psi^g_k(x), & j = 0,\, g \in G^0,\, k \in \mathbb{Z}^d \\ 2^{(j-1)d/2} \Psi^g_k(2^{j-1} x), & j \in \mathbb{N},\, g \in G^j,\, k \in \mathbb{Z}^d. \end{cases}$$

Wavelet coefficients, defined as the expansion coefficients with respect to the above basis for $L^2(\mathbb{R}^d)$ functions, can be used to characterize the so-called *Besov spaces*: Let $1 \leq p, q \leq \infty$, $s \geq 0$, and let the regularity of the above wavelets satisfy

$$r > s \vee \left(\frac{2d}{p} + \frac{d}{2} - s \right).$$

With this, if f admits the wavelet representation

$$f = \sum_{j=0}^{\infty} \sum_{g \in G^j} \sum_{k \in \mathbb{Z}^d} \gamma_k^{j,g} \Psi_k^{j,g}, \tag{B.1}$$

we define the family of norms $\| \,.\, \|_{B^s_{p,q}}$ by

$$\begin{aligned} \|f\|_{B^s_{p,q}} &:= \|f\|_{B^s_{p,q}(\mathbb{R}^d)} := \|\gamma\|_{b^s_{p,q}} \\ &:= \left[\sum_{j=0}^{\infty} 2^{jq(s+\frac{d}{2}-\frac{d}{p})} \sum_{g \in G^j} \left(\sum_{k \in \mathbb{Z}^d} |\gamma_k^{j,g}|^p \right)^{q/p} \right]^{1/q}. \end{aligned} \tag{B.2}$$

In particular, by the orthonormality of the wavelets $\{\Psi_k^{j,g}\}_{j,g,k}$,

$$\|f\|_{B^0_{2,2}} = \|\gamma\|_{\ell^2} = \|f\|_{L^2(\mathbb{R}^d)}.$$

For a bounded Lipschitz domain $\Omega \subseteq \mathbb{R}^d$, we can define the Besov spaces $B^s_{p,q}(\Omega)$ by restrictions of functions on $\mathbb{R}^d$ with norm

$$\|f\|_{B^s_{p,q}(\Omega)} := \inf\{\|g\|_{B^s_{p,q}(\mathbb{R}^d)} : g|_\Omega = f\}.$$

Note that since we work with compactly supported wavelets, for a function f with compact support, only a finite number of wavelet coefficients are non-vanishing in (B.1). In particular, for non-zero wavelet coefficients are contained in a set $\Lambda(j)$ with $|\Lambda(j)| \lesssim 2^{jd}$.

The following theorem collects some basic properties of Besov spaces and their relationship to Hölder and Sobolev spaces.

THEOREM 22. *Let $\Omega \subseteq \mathbb{R}^d$ be a bounded Lipschitz domain.*

(i) [GN16, Proposition 4.3.6] Let $s, s' > 0$, $1 \le p, p', q, q' \le \infty$. Then, the following inclusions hold in the sense of continuous embeddings:
   (a) $B^s_{p,q} \subseteq B^s_{p,q'}$, if $q \le q'$,
   (b) $B^s_{p,q} \subseteq B^{s'}_{p,q'}$, if $s > s'$,

(ii) [Tri06, Theorem 1.122] Let $k \in \mathbb{N}$. Then, $W^{k,2}(\Omega) = B^k_{2,2}(\Omega)$ and we define $W^{\alpha,2}(\Omega) := B^\alpha_{2,2}(\Omega)$ for $\alpha > 0$ not an integer.

(iii) [Tri06, Theorem 1.122], [GN16, Proposition 4.3.20] Let $\alpha > 0$. If α is not integer, then

$$C^\alpha(\Omega) = B^\alpha_{\infty,\infty}(\Omega),$$

If α is integer, then

$$B^\alpha_{1,\infty}(\Omega) \subseteq C^\alpha(\Omega) \subseteq B^\alpha_{\infty,\infty}(\Omega). \tag{B.3}$$

Note that the proof in [GN16] of (B.3) is given in 1D, but extends naturally to arbitrary dimensions by summability of the wavelet coefficients.

A very useful tool to handle function spaces on domains is the availability of extension operators that preserve the norms. The following theorem guarantees the existence of such an extension operator that is the same among all Besov spaces. This allows us to characterize Besov functions on domains via the wavelet coefficients of their extensions.

THEOREM 23 (Extension operator, [Tri06, Theorem 1.105], [Ryc99]). *Let $\Omega \subseteq \mathbb{R}^d$ be a bounded Lipschitz domain. Then, there exists a linear extension operator* ext *that preserves L^2-, C^β-, and $W^{\alpha,2}$-norms. That is, there exists an extension operator* $\operatorname{ext} = \operatorname{ext}(\Omega)$ *such that for $A \in \{L^2, C^\beta, W^{\alpha,2} : \beta > 0, \alpha > 0\}$, there exist constants $C = C(\Omega, A)$ with*

$$\|\operatorname{ext} f\|_{A(\mathbb{R}^d)} \leq C\|f\|_{A(\Omega)}, \quad \textit{and} \quad \operatorname{ext} f|_\Omega = f, \quad \textit{for } f \in A.$$

We conclude this section by a lemma that provides uniform control of a function by its wavelet coefficients. It is useful to control bracketing entropy numbers.

LEMMA 24. *Let $f \in V_J(\mathbb{R}^d)$ with wavelet coefficients $\gamma_k^{j,g}$ for compactly supported mother and father wavelets. Then, $\|f\|_\infty \lesssim 2^{Jd/2}\|\gamma\|_\infty \leq 2^{Jd/2}\|\gamma\|_2$.*

PROOF. Let $x \in \mathbb{R}^d$ and write

$$\begin{aligned} |f(x)| = \Big|\sum_{j=0}^{J} \sum_{g \in G^j, k \in \mathbb{Z}^d} \gamma_k^{j,g} \Psi_k^{j,g}(x)\Big| &\leq \|\gamma\|_\infty \sum_{j=0}^{J} \sum_{g \in G^j, k \in \mathbb{Z}^d} |\Psi_k^{j,g}(x)| \\ &\lesssim \|\gamma\|_\infty \sum_{j=0}^{J} 2^{jd/2} \lesssim 2^{Jd/2}\|\gamma\|_\infty, \end{aligned}$$

where we used Hölder's inequality and the fact that only a finite number of k enters the summation for each fixed $x \in \mathbb{R}^d$. The last inequality is trivial. □

APPENDIX C: METRIC ENTROPY AND SUPREMA OF STOCHASTIC PROCESSES

Here, we collect some basic results about empirical processes that are needed in the proofs. Note that because all suprema we deal with are over subsets of finite-dimensional vector spaces, we do not consider issues of measurability in the remainder.

Denote the bracketing number of a set $\mathcal{F}$ with respect to a norm $\|.\|$ by $N_{[]}(\delta, \mathcal{F}, \|\,.\,\|)$ and define the Dudley integral as

$$\mathcal{D}_{[]}(\sigma, \mathcal{F}, \|\,.\,\|) = \int_0^\sigma \sqrt{1 + \log N_{[]}(\delta, \mathcal{F}, \|\,.\,\|)}\, \mathrm{d}\delta. \tag{C.1}$$

THEOREM 25 (Bernstein chaining, [vW07, Lemma 3.4.2]). *Let $\mathcal{F}$ be a class of measurable functions such that $\mathbb{E}[f^2] < \sigma^2$ and $\|f\|_\infty \leq M$ for all $f \in \mathcal{F}$. Then,*

$$\mathbb{E}\Big[\sup_{f \in \mathcal{F}} |\sqrt{n}(\hat{P} - P)f|\Big] \lesssim \mathcal{D}_{[]}(\sigma, \mathcal{F}, L_2(P))\Big(1 + \frac{\mathcal{D}_{[]}(\sigma, \mathcal{F}, L_2(P))}{\sigma^2\sqrt{n}} M\Big).$$

THEOREM 26 (Concentration, see [Mas07, Equation (5.50)], going back to [Bou02]). *Under the same assumptions as Theorem 25,*

$$\mathbb{P}\Big(\sup_{f \in \mathcal{F}} \sqrt{n} |(\hat{P} - P) f| \geq 2 \mathbb{E}[\sup_{f \in \mathcal{F}} \sqrt{n} |(\hat{P} - P) f|] + \sigma \sqrt{2x} + \frac{M}{\sqrt{n}} x \Big) \leq \exp(-x).$$

LEMMA 27 (Bracketing numbers from L^∞ covering numbers). *Let P be a probability measure and $A \subseteq \mathcal{F}$ be a set of functions with $N(A, L^\infty(P), \delta/2) \leq \phi(\delta)$. Then, $N_{[]}(A, L^2(P), \delta) \leq \phi(\delta)$.*

Moreover, if for every function in the original class $f \in \mathcal{F}$, $\mathbb{E}[f^2] < \sigma^2$ and $\|f\|_{L^\infty(P)} < M$, then every bracket $[f_1, f_2]$ in a δ-cover above (note that f_1 and f_2 need not be members of $\mathcal{F}$) satisfies

$$\mathbb{E}[f_j^2] < 2\sigma^2 + \frac{1}{2}\delta^2, \quad \|f_j\|_{L^\infty(\nu)} < M + \delta/2, \quad j \in \{1, 2\}.$$

PROOF. Denote by $\{f_1, \dots, f_N\}$ the centers of a minimal $\delta/2$-covering of $\mathcal{F}$ in $L^\infty(P)$. Let $i \in \{1, \dots, N\}$. Then, each $\delta/2$ ball around f_i is contained in the bracket $[f_i - \delta/2, f_i + \delta/2]$. Moreover, the $L^2(P)$ diameter of the above bracket is bounded by its $L^\infty(P)$ diameter, which is δ, so the collection of those brackets yields the desired covering with brackets.

The rest of the lemma follows from

$$\mathbb{E}[(f \pm \delta/2)^2] \leq 2\mathbb{E}[f^2] + \frac{1}{2}\delta^2, \qquad \|f \pm v\delta/2\|_{L^\infty(\nu)} \leq \|f\|_{L^\infty(\nu)} + \frac{\delta}{2}.$$

□

The following lemma can be shown by directly specifying a grid or by a volume argument such as [Ver18, Proposition 4.2.12].

LEMMA 28 (Covering numbers for norm balls). *Fix $p \in \mathbb{N}$ and denote by $B_\infty(A)$ the ℓ^∞ ball of $\mathbb{R}^d$ with radius A. Then, $N(B_\infty(A), \|\,.\,\|_\infty, \delta) \leq (3A/\delta)^d$.*

LEMMA 29 ([Rau10, Lemma 10.3]). *For $\alpha > 0$,*

$$\int_0^\alpha \sqrt{\log(1 + t^{-1})}\, \mathrm{d}t \leq \alpha \sqrt{1 + \log(1 + \alpha^{-1})}\,.$$

APPENDIX D: TOOLS FOR LOWER BOUNDS

For the convenience of the reader, in this section we restate the standard tools we use in the proof of Theorem 6 to establish lower bounds based on Fano's inequality and the Varshamov-Gilbert Lemma, taken from [Tsy09, Theorem 2.9, Lemma 2.9, Theorem 2.2].

THEOREM 30 (Lower bounds from multiple hypotheses). *[Tsy09, Theorem 2.9] Let $K \geq 2$, $\Theta = \{T_0, \dots, T_K\}$ a collection of hypotheses, and let d be a pseudometric, i.e., a bi-variate function on Θ such that*

(a) $d(T_j, T_k) \geq 0$;
(b) $d(T_j, T_j) = 0$;
(c) $d(T_j, T_k) = d(T_k, T_j)$;

(d) $d(T_j, T_\ell) \le d(T_j, T_k) + d(T_k, T_\ell)$;

for all $0 \le j, k, \ell \in \{0, \dots, K\}$.

Suppose Θ *fulfills:*

(i) $d(T_j, T_k) \ge 2s > 0$, *for all* $0 \le j < k \le K$;

(ii) $P_j \ll P_0$, *for all* $j \in [K]$, *and*

$$\frac{1}{K}\sum_{j=1}^{K} D(P_j \| P_0) \le C \log K$$

with $0 < C < 1/8$, *where* $P_j = P_{T_j}$ *denotes a probability distribution associated with every* T_j *for all* $j = 0, 1, \dots, K$.

Then,

$$\inf_{\hat{T}} \sup_{T \in \Theta} P_T(d(\hat{T}, T) \ge s) \ge \frac{\sqrt{K}}{1 + \sqrt{K}} \left(1 - 2C - \sqrt{\frac{2C}{\log K}} \right) > 0.$$

Lemma 31 (Varshamov-Gilbert lemma). *[Tsy09, Lemma 2.9] Let* $D \ge 8$. *There exists a subset* $\tau^{(0)}, \dots, \tau^{(K)}$ *of* $\{0,1\}^D$ *such that* $\tau^{(0)} = (0, \dots, 0)$,

$$\sum_{\ell=1}^{D} \mathbb{1}(\tau_l^{(j)}, \tau_l^{(k)}) \ge \frac{D}{8}, \quad \textit{for all } 0 \le j < k \le K,$$

and

$$K \ge 2^{D/8}.$$

Theorem 32 (Lower bounds from two hypotheses). *[Tsy09, Theorem 2.2] Let* T_0, T_1 *be two hypotheses with associated probability measures* $P_j = P_{T_j}$ *for* $j \in \{0, 1\}$. *Denoting* $s = d(T_0, T_1)/2$, *if*

$$D(P_1 \| P_0) \le C < \infty,$$

then

$$\inf_{\hat{T}} \max_{j \in \{0,1\}} P_j(d(\hat{T}, T_j) \ge s) \ge \frac{1}{4}\exp(-C) \vee \frac{1 - \sqrt{C/2}}{2} > 0.$$

APPENDIX E: ALTERNATIVE ASSUMPTIONS VIA CAFFARELLI'S REGULARITY THEORY

In this section, we show how to apply Theorem 8 under smoothness assumptions on the source and target distribution instead of the transport map. This is enabled by Caffarelli's regularity theory [Caf92b, Caf92a, Caf96], Theorem 33 below, which gives regularity estimates on the transport map T_0 under regularity assumptions on the source and target densities and their supports.

For simplicity, we denote an open, convex set Ω with C^2 boundary as *uniformly convex* if it can be written as the sublevel set, $\Omega = \{f < 0\}$, of a strongly convex function f. Note that uniform convexity can also be characterized by the positivity of the second fundamental form of Ω on its boundary $\partial\Omega$ [Vil09].

C1 (Supports). *Assume* Ω_P, Ω_Q *are two uniformly convex, bounded, open subsets of* $\mathbb{R}^d$ *with* $C^{\lfloor \alpha-1 \rfloor+2}$ *boundaries.*

C2 (Densities). *With* Ω_P, Ω_Q *as in Assumption C1, assume that* $\rho_P \in C^{\alpha-1}(\overline{\Omega}_P)$ *and* $\rho_Q \in C^{\alpha-1}(\overline{\Omega}_Q)$ *are two probability densities on* $\mathbb{R}^d$ *with* $\operatorname{supp}(\rho_P) = \overline{\Omega}_P$, $\operatorname{supp}(\rho_Q) = \overline{\Omega}_Q$ *that are bounded from above and below on their support. Further, denote the associated probability distributions by* P *and* Q*, respectively.*

THEOREM 33 (Caffarelli's global regularity theory).
[Vil09, Theorem 12.50(iii)], [Caf96]. Let $\alpha > 1$ *and assume that Assumptions C1 and C2 hold. Then, for the optimal transport potential* f_0*, i.e., the solution to* (2.5) *for* P *and* Q*, which is unique up to an additive constant, it holds that* $f_0 \in C^{\alpha+1}(\overline{\Omega}_P)$.

Moreover, we need the following extension lemma to smoothly extend the potential (and associated transport map) to a larger set as required by Assumption B2. Similar statements have been shown in [Gho02, Yan14, AM19].

LEMMA 34. *Let* Ω *be a convex compact subset of* $\mathbb{R}^d$*,* $f : \Omega \to \mathbb{R}$ *a convex* $C^\alpha(\Omega)$ *function for* $\alpha > 2$ *with*

$$\min_{x \in \Omega} \lambda_{\min}(D^2 f(x)) > 0.$$

Then, there exists an $\epsilon > 0$ *and an extension* $\tilde{f}$ *of* f *to* $\Omega_\epsilon = \Omega + \epsilon B(0,1)$ *such that* $f \in C^\alpha(\Omega_\epsilon)$ *and*

$$\min_{x \in \Omega_\epsilon} \lambda_{\min}(D^2 \tilde{f}(x)) > 0. \tag{E.1}$$

PROOF. The claim follows from extending f via Theorem 23 to all of $\mathbb{R}^d$ while preserving the Hölder class C^α, and observing that the strong convexity condition (E.1) on an enlargement of Ω can be ensured by uniform continuity of the extension. □

COROLLARY 35. *Let* $\alpha > 1$*. Under Assumptions C1 and C2, writing* f_0 *and* $T_0 = \nabla f_0$ *for the Kantorovich potential and optimal transport map between* P *and* Q*, respectively, there exists an estimator* $\hat{T}$ *such that*

$$\mathbb{E}_{(X_{1:n}, Y_{1:n})}\left[\int \|\hat{T}(x) - T_0(x)\|_2^2 \,\mathrm{d}P(x)\right] \lesssim \left[n^{-\frac{2\alpha}{2\alpha-2+d}}(\log(n))^2 \vee \frac{1}{n}\right],$$

where $X_{1:n}, Y_{1:n}$ *denote i.i.d. observations from* P *and* Q*, respectively.*

PROOF. The statement follows by verifying the assumptions B1 and B2 on the original transport map and concluding by Theorem 8, using the same estimator.

Assumption B1 is satisfied for $\overline{\Omega}_P$ by the (stronger) conditions imposed on it in Theorem 33.

To show Assumption B2, denote by f_0 the optimal transport potential and by T_0 the optimal transport map for the problem given by ρ_P and ρ_Q. By Theorem 33, $f_0 \in C^{\alpha+1}(\overline{\Omega}_P)$ and thus $T_0 = \nabla f_0 \in C^\alpha(\overline{\Omega}_P, \mathbb{R}^d)$. Denoting by $\tilde{f}_0$ the extension of f_0 to $\tilde{\Omega}_P = (\overline{\Omega}_P)_\epsilon$ from Lemma 34, we observe that $\tilde{\Omega}_P$ is connected and has a Lipschitz boundary. This follows from the general fact that the Minkowski sum of a convex set with C^1 boundary and another convex set has a C^1 boundary [KP91]. From the regularity of $\tilde{T}_0 = \tilde{f}_0$ and the compactness of $\tilde{\Omega}_P$, all requirements in B2 can now be easily checked. □

An inspection of the proof of our lower bound results, Theorem 6, readily yields a lower bound under smoothness assumptions on P and Q as well. In particular, after changing the domain from $[0,1]^d$ to the unit ball in d dimensions, the explicit form of the density of Q_k in (4.2) guarantees that the candidate densities fulfill assumptions C1 and C2. This indicates that the above rate is optimal up to log factors.

REMARK 36. *In Corollary 35, we do not claim any uniform bound over a regularity class, such as an explicit dependence of the constant on the $C^{\alpha-1}$-norms of ρ_P and ρ_Q alone. Although the statement of Theorem 33 strongly suggests a bound on $\|f_0\|_{C^{\alpha+1}(\overline{\Omega}_P)}$ in terms of $\|\rho_P\|_{C^{\alpha-1}(\overline{\Omega}_P)}$ and $\|\rho_Q\|_{C^{\alpha-1}(\overline{\Omega}_Q)}$, to the best of our knowledge, such bounds are not available in the literature for the "global" version of Caffarelli's regularity theory as stated Theorem 33. On the other hand, such bounds hold "locally", that is, $\|f_0\|_{C^{\alpha+1}(U)}$ can be bounded for strictly open subsets of $\overline{\Omega}_P$, see [Vil09, Theorem 12.56], [Caf92b]. This deficiency is due to the non-constructive nature of the available proofs of Theorem 33 and could possibly be remedied, but we consider it out of the scope of this paper.*

APPENDIX F: OMITTED PROOFS

F.1 Proof of Proposition-Definition 7

We proceed in order of statement of the results.

(i) It follows from B2 that there exists f_0 such that $\nabla f_0 = T_0$ and by B2(i) that $|\nabla f_0(x)| \leq M$ for all $x \in \tilde{\Omega}_P$. Moreover, since f_0 is defined up an additive constant, assume that $f_0(x_0) = 0$ for some $x_0 \in \tilde{\Omega}_P$. A first-order Taylor expansion yields that for any $x \in \tilde{\Omega}_P$,

$$|f_0(x)| = |f_0(x) - f_0(x_0)| \leq \sup_{z \in \tilde{\Omega}_P} \|\nabla f_0(z)\|_2 \|x - x_0\|_2 \leq M \operatorname{diam}(\tilde{\Omega}_P) \leq 2M^2,$$

where in the second inequality, we used B2(i), and in the third one, we used that $\operatorname{diam}(\tilde{\Omega}_P) \leq 2M$ according to B2.

(ii) follows immediately from B2(ii).

Next, as in (i), (5.1) follows from a first-order expansion: For any $x \in \tilde{\Omega}_P$, note that on the one hand

$$|f_0(x)| = |f_0(x) - f_0(x_0)| \leq 2\|T_0\|_2 M \leq 2M^2$$

so that $\|f_0\|_{L^\infty} \leq M^2$ and $\|f_0\|_{L^2} \leq 2M^3$. Together with the fact that $T_0 = \nabla f_0$, this allows us to shift the smoothness index by one speaking about potentials.

To show the statements about Q, by Lemma 14 and (ii) above, f_0 is strongly convex on $\operatorname{int}(\tilde{\Omega}_P)$ and can be extended to $+\infty$ outside of $\tilde{\Omega}_P$ and thus be also be considered a strongly convex function on $\mathbb{R}^d$. By Lemmas 17 and 18, we conclude that $T_0 = \nabla f_0$ is a bijection from Ω_P onto its image $\Omega_Q = \operatorname{supp}((\nabla f_0)_\# P)$, with both ∇f_0 and $(\nabla f_0)^{-1}$ being continuously differentiable; in other words, ∇f_0 is a C^1-diffeomorphism between Ω_P and Ω_Q. Since C^1-diffeomorphisms preserve Lipschitz domains [HMT07, Theorem 4.1] and connectedness, we can conclude that Ω_Q is a connected and bounded Lipschitz domain, and the fact that $\Omega_Q \subseteq MB_1$ follows from $T(x) \leq M$ for all $x \in \tilde{\Omega}_P$.

Finally, we turn to check the condition on the density ρ_Q. To that end, note that by the change of variables formula, Q has the density

$$\rho_Q(y) = \frac{1}{|\det D^2 f_0((\nabla f_0)^{-1}(y))|} \rho_P((\nabla f_0)^{-1}(y)) \mathbb{1}(y \in \Omega_Q).$$

This readily yields the desired bound in light of B1 which assumes boundedness of ρ_P and (ii) which gives boundedness of the Hessian $D^2 f_0$.

F.2 Proof of Proposition 11

We first bound the expectation of the supremum of interest and then obtain the high-probability bound via concentration. To begin, note that

$$\hat{\mathcal{S}}_0(f) - \mathcal{S}_0(f) = (\hat{P} - P)(f - f_0) + (\hat{Q} - Q)(f^* - f_0^*)\,, \tag{F.1}$$

which yields

$$\begin{aligned}\mathbb{E}[&\sup_{f \in \mathcal{F}_J(\tau^2)} |\hat{\mathcal{S}}_0(f) - \mathcal{S}_0(f)|]\\ &\le \mathbb{E}[\sup_{f \in \mathcal{F}_J(\tau^2)} |(\hat{P} - P)(f - f_0)|] + \mathbb{E}[\sup_{f \in \mathcal{F}_J(\tau^2)} |(\hat{Q} - Q)(f^* - f_0^*)|]\\ &=: T_1 + T_2.\end{aligned}$$

We first focus on the T_2-term. Once understood, the T_1-term can be bounded similarly.

Bound on T_2-term.
We estimate T_2 from above by two suprema, corresponding to low and high frequencies. To this end, we center all functions and consider the extension of a restriction of the functions to Ω_Q, which allows us to use wavelet expansions for harmonic analysis.

First, because $(\hat{Q} - Q)(g + c) = (\hat{Q} - Q)g$ for every function g and constant $c \in \mathbb{R}$, we may assume without loss of generality that

$$\int_{\Omega_Q} (f^*(z) - f_0^*(z))\,\mathrm{d}\lambda(z) = 0 \quad \forall f \in \mathcal{F}_J.$$

Note that f^* and f_0^* are defined over all of $\mathbb{R}^d$ but we do not control their norms over the whole space. To overcome this limitation, with a slight abuse of notation, we denote by $\operatorname{ext} f^*$ (resp. $\operatorname{ext} f_0^*$) the Lipschitz extension of the restriction of f^* (resp. f_0^*) to Ω_Q. The existence of a linear extension operator is guaranteed by Theorem 23. In particular, we control the norms of $\operatorname{ext} f^*$ and $\operatorname{ext} f_0^*$.

For a function g with wavelet expansion

$$g = \sum_{j=0}^{\infty} \sum_{g \in G^j} \sum_{k \in \mathbb{Z}^d} \gamma_k^{j,g} \Psi_k^{j,g},$$

as defined in Section B, we define the two L^2-projections

$$\Pi_J g = \sum_{j=0}^{J} \sum_{g \in G^j} \sum_{k \in \mathbb{Z}^d} \gamma_k^{j,g} \Psi_k^{j,g}\,, \qquad \Pi_{>J} g = \sum_{j=J+1}^{\infty} \sum_{g \in G^j} \sum_{k \in \mathbb{Z}^d} \gamma_k^{j,g} \Psi_k^{j,g}.$$

With this, we write $T_2 = T_{2,1} + T_{2,2}$, where

$$\begin{aligned}T_{2,1} &= \mathbb{E}[\sup_{f \in \mathcal{F}_J(\tau^2)} |(\hat{Q} - Q)\Pi_J \operatorname{ext}(f^* - f_0^*)|]\,,\\ T_{2,2} &= \mathbb{E}[\sup_{f \in \mathcal{F}_J(\tau^2)} |(\hat{Q} - Q)\Pi_{>J} \operatorname{ext}(f^* - f_0^*)|]\,.\end{aligned}$$

Note that the projected functions are again well-defined continuous functions: By assumption, all f and f_0 are strongly convex, hence their conjugates have Lipschitz-continuous gradients and therefore bounded $B^2_{\infty,\infty}(\Omega_Q)$-norm. In turn, their projections also have bounded $B^2_{\infty,\infty}(\Omega_Q)$-norm, and by Theorem 23, this implies that both $\Pi_J \operatorname{ext}(f^* - f_0^*)$ and $\Pi_{>J} \operatorname{ext}(f^* - f_0^*)$ are in $C^s(\mathbb{R}^d)$, for $s < 2$, and hence they are continuous.

Bound on $T_{2,1}$-term.

Recall that it follows from Proposition-Definition 7 that Ω_Q is a connected Lipschitz domain, so we can apply the Poincaré-Wirtinger inequality, Lemma 39, together with (5.6) from Proposition 10, to get

$$\int_{\Omega_Q} (f^* - f_0^*)^2 \,\mathrm{d}\lambda \lesssim \int_{\Omega_Q} \|\nabla f^* - \nabla f_0^*\|^2 \,\mathrm{d}\lambda \lesssim M\tau^2,$$

where we used that we assumed $f^* - f_0^*$ to be centered.

Hence, $f \in \mathcal{F}_J(\tau^2)$ implies $\|f^* - f_0^*\|_{W^{1,2}(\Omega_Q)} \lesssim \tau$, and therefore due to the properties of the extension operator ext,

$$\|\operatorname{ext}(f^* - f_0^*)\|_{W^{1,2}(\mathbb{R}^d)} \le C_4\,\tau, \tag{F.2}$$

for some constant $C_4 = C_4(\Omega_Q, M)$. Since Π_J is a non-expansive operator on Besov spaces, it follows from the above display that

$$T_{2,1} \le \sup\{|(\hat{Q} - Q)h| : h \in V_J(\mathbb{R}^d),\ \|h\|_{W^{1,2}(\mathbb{R}^d)} \le C_4\tau\}.$$

Bounding the empirical process over this standard function class can now be performed as follows. Observe first that for any function $h \in V_J(\mathbb{R}^d)$ with wavelet decomposition

$$h = \sum_{j=0}^{J} \sum_{g \in G^j} \sum_{k \in \mathbb{Z}^d} \gamma_k^{j,g} \Psi_k^{j,g}, \tag{F.3}$$

the condition $\|h\|_{W^{1,2}(\mathbb{R}^d)} \le C_4\,\tau$ is equivalent to

$$\sum_{j=0}^{J} \sum_{g \in G^j} \sum_{k \in \mathbb{Z}^d} 2^{2j} |\gamma_k^{j,g}|^2 \le C_4^2 \tau^2.$$

Next, by symmetrization, for independent copies of Rademacher random variables ε_i, $i = 1, \dots, n$,

$$\mathbb{E} \sup |(\hat{Q} - Q)h| \le \mathbb{E} \sup \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i h(Y_i),$$

where both suprema are taken over the set $\mathfrak{G}_J = \{h \in V_J(\mathbb{R}^d),\ \|h\|_{W^{1,2}(\mathbb{R}^d)} \le C_4\tau\}$. To control the Rademacher process, fix $h \in \mathfrak{G}_J$ with wavelet decomposition (F.3). By the Cauchy-Schwarz inequality,

$$\begin{aligned} \sum_{i=1}^{n} \varepsilon_i h(Y_i) &\le \Big(\sum 2^{2j} |\gamma_k^{j,g}|^2\Big)^{1/2} \Big(\sum \Big(\sum_{i=1}^{n} \frac{\varepsilon_i}{2^j} \Psi_k^{j,g}(Y_i)\Big)^2\Big)^{1/2} \\ &\le C_4\,\tau \Big(\sum \Big(\sum_{i=1}^{n} \frac{\varepsilon_i}{2^j} \Psi_k^{j,g}(Y_i)\Big)^2\Big)^{1/2}, \end{aligned}$$

where here and below, all sums without indices are over $\{0 \leq j \leq J,\, g \in G^j,\, k \in \mathbb{Z}^d\}$. Since the right-hand side in the display above does not depend on h, we get by Jensen's inequality that

$$\mathbb{E} \sup \frac{1}{n} \sum_{i=1}^n \varepsilon_i h(Y_i) \leq \frac{C_4\, \tau}{n} \Big(\mathbb{E} \sum \Big(\sum_{i=1}^n \frac{\varepsilon_i}{2^j} \Psi_k^{j,g}(Y_i) \Big)^2 \Big)^{1/2}.$$

Since $Y_i \in \Omega_Q$, a compact set by assumption, for given j and g, $\Psi_k^{j,g}(Y_i)$ is non-zero only for $k \in \Lambda(j)$ where $\Lambda(j)$ depends on the diameter of Ω_Q, and $|\Lambda(j)| \lesssim 2^{jd}$. Together with the independence of the ε_i, this yields

$$\mathbb{E} \sum \Big(\sum_{i=1}^n \frac{\varepsilon_i}{2^j} \Psi_k^{j,g}(Y_i) \Big)^2 = \sum_{\substack{0 \leq j \leq J, \\ g \in G^j, k \in \Lambda(j)}} \sum_{i=1}^n \frac{1}{2^{2j}} \mathbb{E} \left[\Psi_k^{j,g}(Y_i)^2 \right].$$

By Proposition-Definition 7 and the fact that the $\Psi_k^{j,g}$ form an orthonormal basis in $L^2(\mathbb{R}^d)$,

$$\mathbb{E} \left[\Psi_k^{j,g}(Y_i)^2 \right] \lesssim \int_{\Omega_Q} \Psi_k^{j,g}(y)^2 \,\mathrm{d}\lambda(y) \leq \int_{\mathbb{R}^d} \Psi_k^{j,g}(y)^2 \,\mathrm{d}\lambda(y) = 1.$$

Thus,

$$\sum_{\substack{0 \leq j \leq J, \\ g \in G^j, k \in \Lambda(j)}} \sum_{i=1}^n \frac{1}{2^{2j}} \mathbb{E} \left[\Psi_k^{j,g}(Y_i)^2 \right] \lesssim n \sum_{0 \leq j \leq J} \frac{2^{jd}}{2^{2j}} \lesssim \begin{cases} n, & d = 1, \\ nJ, & d = 2, \\ n 2^{J(d-2)}, & d \geq 3. \end{cases}$$

We have proved that

$$T_{2,1} \lesssim \frac{\tau}{\sqrt{n}} \mathfrak{r}_J, \qquad \mathfrak{r}_J := \begin{cases} 1, & d = 1, \\ \sqrt{J}, & d = 2, \\ 2^{J \frac{d-2}{2}}, & d \geq 3. \end{cases} \tag{F.4}$$

Bound on $T_{2,2}$-term.

To control the term $T_{2,2}$, we use a chaining bound for bracketing entropy (Theorem 25). To that end, we exhibit bounds on the L^∞-covering numbers of the corresponding function space, which in turn implies control of L^2-bracketing numbers. The idea is to bound the covering numbers in the original space $\mathcal{F}_J(\tau^2)$ and exploit continuity properties of the transformations that lead to the function class in the definition of $T_{2,2}$, in particular the operation of taking the convex conjugate.

Define the function space

$$\tilde{\mathcal{F}}_J(\tau^2) = \left\{ \Pi_{>J} \operatorname{ext}(f^* - f_0^*) : f \in \mathcal{F}_J(\tau^2) \right\},$$

and let $f_1, f_2 \in \mathcal{F}_J(\tau^2)$ have wavelet coefficients given by the sequences γ_1 and γ_2. Observe that by linearity of the projection and extension operators, we have for any cutoff $J' \geq 0$,

$$\begin{aligned} &\|\Pi_{>J} \operatorname{ext}(f_1^* - f_0^*) - \Pi_{>J} \operatorname{ext}(f_2^* - f_0^*)\|_{L^\infty(\Omega_Q)} \\ &\quad = \|\Pi_{>J} \operatorname{ext}(f_1^* - f_2^*)\|_{L^\infty(\Omega_Q)} \\ &\quad \leq \|\Pi_{>J'} \Pi_{>J} \operatorname{ext}(f_1^* - f_2^*)\|_{L^\infty(\Omega_Q)} + \|\Pi_{J'} \Pi_{>J} \operatorname{ext}(f_1^* - f_2^*)\|_{L^\infty(\Omega_Q)}. \end{aligned}$$

To control the first term in the right-hand side above, in the following lemma, we establish bounds on the potentials in $\tilde{F}_J(\tau^2)$. Its proof is deferred to the next section.

LEMMA 37. *There exists a constant $C_5 = C_5(\Omega_P, \tilde{\Omega}_P, \Omega_Q, M)$ such that for all $f \in \mathcal{X}(2M)$,*

$$\| \operatorname{ext} f^* \|_{B^2_{\infty,\infty}(\mathbb{R}^d)} \le \frac{C_5}{2}.$$

Moreover, under assumptions B1 – B2, $\| \operatorname{ext} f_0^* \|_{B^2_{\infty,\infty}(\mathbb{R}^d)} \le C_5/2$ as well.

It follows from Lemma 37 that there exists C_5 such that for any $f \in \mathcal{F}(\tau^2)$, we have

$$\|\Pi_{>J} \operatorname{ext}(f_1^* - f_2^*)\|_{B^2_{\infty,\infty}} \le C_5\,.$$

Therefore, by Lemma 13, we get

$$\|\Pi_{>J'}\Pi_{>J} \operatorname{ext}(f_1^* - f_2^*)\|_{L^\infty(\Omega_Q)} \le 2^{-2J'} \|\Pi_{>J'} \operatorname{ext}(f_1^* - f_2^*)\|_{B^2_{\infty,\infty}} \le \varepsilon$$

if we choose $J' = \lceil \log (C_5/\varepsilon)/2 \rceil$ so that $C_5\, 2^{-2J'} \le \varepsilon$.

To control the second term, we get from Lemma 24 that

$$\begin{aligned}\|\Pi_{J'}\Pi_{>J} \operatorname{ext}(f_1^* - f_2^*)\|_{L^\infty(\Omega_Q)} &\lesssim 2^{J'd/2} \|\Pi_{J'}\gamma\|_2 \le 2^{J'd/2}\|\gamma\|_2 \\ &= 2^{J'd/2} \| \operatorname{ext}(f_1^* - f_2^*)\|_{L^2(\mathbb{R}^d)}.\end{aligned}$$

Moreover, using respectively the fact that ext is a Lipschitz operator on Besov spaces, Ω_Q is bounded, and the convex conjugate is non-expansive in L^∞ by Lemma 20, we have

$$\begin{aligned}\| \operatorname{ext}(f_1^* - f_2^*)\|_{L^2(\mathbb{R}^d)} &\lesssim \|f_1^* - f_2^*\|_{L^2(\Omega_Q)} \lesssim \|f_1^* - f_2^*\|_{L^\infty(\Omega_Q)} \\ &\le \|f_1 - f_2\|_{L^\infty(\tilde{\Omega}_P)} \lesssim 2^{Jd/2}\|\gamma_1 - \gamma_2\|_\infty,\end{aligned}$$

where in the last inequality, we used Lemma 24. We have proved that

$$\|\Pi_{>J} \operatorname{ext}(f_1^* - f_0^*) - \Pi_{>J} \operatorname{ext}(f_2^* - f_0^*)\|_{L^\infty(\Omega_Q)} \le C_6\, 2^{(J+J')d/2} \|\gamma_1 - \gamma_2\|_\infty + \varepsilon\,,$$

for some constant $C_6 = C_6(\tilde{\Omega}_P, \Omega_Q)$. This inequality allows us to control the L^∞-bracketing numbers of $\tilde{\mathcal{F}}_J(\tau^2)$ using ℓ^∞-covering numbers for the wavelet coefficients. To control the latter, note that for all $f \in \mathcal{F}_J(\tau^2) \subset \mathcal{X}(2M)$, it holds $\|f\|_{B^2_{\infty,\infty}(\mathbb{R}^d)} \lesssim M^2$ so that $\|\gamma\|_{\ell^\infty} \lesssim M^2$. Moreover, these wavelet coefficients are in a space of dimension at most $C\, 2^{Jd}$, $C(\tilde{\Omega}_P, M) > 0$ because $\mathcal{F}_J(\tau^2) \subseteq V_J(\tilde{\Omega}_P)$. Hence, choosing $\varepsilon = \delta/4$, Lemmas 27, 28, and the previous display yield

$$\begin{aligned}\log N_{[\,]}(\tilde{\mathcal{F}}_J(\tau^2), \|\,.\,\|_{L^2(\Omega_Q)}, \delta) &\lesssim \log N(\tilde{\mathcal{F}}_J(\tau^2), \|\,.\,\|_{L^\infty(\Omega_Q)}, \delta/2) \\ &\lesssim 2^{Jd}J + 2^{Jd}\log\left(\frac{1}{\delta}\right).\end{aligned} \tag{F.5}$$

To apply the chaining bound of Theorem 25, note that by (F.2) and Lemma 37, respectively, combined with Lemma 13, we have

$$\|g\|_{L^2(\mathbb{R}^d)} \lesssim C_4\, 2^{-J}\,\tau, \quad \|g\|_{L^\infty(\mathbb{R}^d)} \lesssim C_5\, 2^{-2J}, \quad \text{for } g \in \tilde{\mathcal{F}}_J(\tau^2).$$

Thus,

$$T_{2,2} \lesssim \frac{1}{\sqrt{n}}\mathcal{D}_{[\,]}\big(1 + \frac{\mathcal{D}_{[\,]}}{C_4^2\, 2^{-2J}\,\tau^2\sqrt{n}}C_5\, 2^{-2J}\big) = \frac{1}{\sqrt{n}}\mathcal{D}_{[\,]}\left(1 + C_7\,\frac{\mathcal{D}_{[\,]}}{\tau^2\sqrt{n}}\right),$$

where $\mathcal{D}_{[]} = \mathcal{D}_{[]}(C_4\, 2^{-J}\tau, \tilde{\mathcal{F}}_J(\tau^2), L^2(Q))$ is the Dudley integral defined in (C.1). Moreover, by (F.5) and Lemma 29, we have

$$\begin{aligned}\mathcal{D}_{[]} &\lesssim \int_0^{C_4\, 2^{-J}\tau} \sqrt{1 + 2^{Jd}J + 2^{Jd}\log(1/\delta)}\,\mathrm{d}\delta \\ &\lesssim \tau 2^{J\frac{d-2}{2}} \sqrt{J\log(1+C_5/\tau)}\,,\end{aligned}$$

and therefore,

$$T_{2,2} \lesssim \tau \frac{2^{J\frac{d-2}{2}}}{\sqrt{n}} \sqrt{J\log(1+C_5/\tau)} + \frac{2^{J(d-2)}J}{n} \log(1+C_5/\tau)\,.$$

Together with (F.4), we have $T_2 \lesssim \phi_J(\tau^2)$ with ϕ_J defined as in (5.8) and we absorbed $\mathfrak{r}_J$ for $d \geq 2$ into the first term on the right-hand side.

Bounding T_1. T_1 can be bounded completely analogously to how we bounded T_2, with the exception that Lemma 20 is not needed. Thus, we obtain $T_1 \lesssim \phi_J(\tau^2)$.

Final bound and concentration. Collecting the above bounds on T_1 and T_2, we get

$$\mathbb{E}[\sup_{f\in\mathcal{F}_J(\tau^2)} |\hat{\mathcal{S}}_0(f) - \mathcal{S}_0(f)| \lesssim \phi_J(\tau^2)\,.$$

To obtain a bound that holds with high probability, we apply the concentration result of Theorem 26. For this, note that (F.1) can be written as

$$\hat{\mathcal{S}}_0(f) - \mathcal{S}_0(f) = ((\hat{P}\otimes\hat{Q}) - (P\otimes Q))((f-f_0)\otimes(f^*-f_0^*)),$$

where $P\otimes Q$ denotes the product measure and $(f\otimes g)(x,y) = f(x)+g(y)$. Following the same argument as in the proof of Lemma 37, we get

$$\|(f-f_0)\otimes(f^*-f_0^*)\|_{L^\infty(\Omega_P\times\Omega_Q)} \leq C_5\,.$$

Moreover, similarly to Proposition 10, we have for any $f\in\mathcal{F}_J(\tau^2)$ that

$$\|(f-f_0)\otimes(f^*-f_0^*)\|_{L^2(P\otimes Q)} \leq C_4\,\tau.$$

We can therefore apply Theorem 26 and conclude that with probability at least $1-e^{-t}$, it holds that

$$\sup_{f\in\mathcal{F}_J(\tau^2)} \left|\mathcal{S}_0(f) - \hat{\mathcal{S}}_0(f)\right| \lesssim \phi_J(\tau^2) + \tau\sqrt{\frac{t}{n}} + \frac{t}{n}\,,$$

which concludes our proof.

F.3 Proof of Lemma 37

By the boundedness conditions in the definition of $\mathcal{X}(2M)$ in Proposition-Definition 7, and the boundedness of $\tilde{\Omega}_P$ and Ω_Q, we have for $y\in\Omega_Q$ that

$$|f^*(y)| = |\sup_{x\in\tilde{\Omega}_P}\langle x,y\rangle - f(x)| \leq \frac{1}{2}C_5(\tilde{\Omega}_P,\Omega_Q,M).$$

Moreover, by Lemma 17, ∇f^* is $2M$-Lipschitz. Therefore, since Ω_Q is bounded,

$$\|\operatorname{ext} f^*\|_{B^2_{\infty,\infty}(\Omega_Q)} \lesssim_{\Omega_Q} \|f^*\|_{C^2(\Omega_Q)} \lesssim C_5.$$

We can similarly deduce the second claim by Proposition-Definition 7 since $f_0 \in \mathcal{X}(M)$, possibly choosing a larger C_5.

APPENDIX G: ADDITIONAL LEMMAS

LEMMA 38. *In the notation of the proof of Theorem 6, for any $k = 0, \ldots, K$ and $m \geq m_0$, $\nabla\phi_k$ is a bijection from $[0,1]^d$ to $[0,1]^d$.*

PROOF OF LEMMA 38. By construction, ϕ_k is strongly convex and has Lipschitz continuous derivatives, hence so does its convex conjugate ϕ_k^* by Lemma 17. In particular, ϕ_k^* is defined on all of $\mathbb{R}^d$ and for each y,

$$\phi_k^*(y) = \sup_x \langle x, y\rangle - \phi_k(x).$$

Hence, the equation $\nabla\phi_k(x) = y$ has a unique solution $x(y)$ for every $y \in \mathbb{R}^d$ which implies that $\nabla\phi_k$ is injective and that for any $y \in [0,1]^d$, there exists $x = x(y) \in \mathbb{R}^d$ such that $\nabla\phi_k(x) = y$. It remains to check that $x(y) \in [0,1]^d$ for all $y \in [0,1]^d$. To that end, note that $\phi_k(x) = \|x\|^2/2$ for $x \notin [0,1]^d$. Hence $x(y) = y$ whenever $y \notin [0,1]^d$. In particular, if $y \in [0,1]^d$, we must have $x(y) \in [0,1]^d$. This completes the proof. $\square$

LEMMA 39 (Poincaré inequality, [Eva10, Section 5.8.1], [Leo17, Theorem 13.27]). *Let $\Omega \subseteq \mathbb{R}^d$ be a bounded and connected Lipschitz domain. Then, there exists a constant $C = C(d, \Omega)$ such that for any function $f \in W^{1,2}(\Omega)$,*

$$\|f - \int_\Omega f(x)\,\mathrm{d}\lambda(x)\|_{L^2(\Omega)} \leq C\|\nabla u\|_{L^2(\Omega)}.$$

APPENDIX H: NUMERICAL EXPERIMENTS, CONTINUED

In this section, we give additional details on the numerical experiments in Section 6, using the same notation used there to define $\hat{T}_{\mathrm{emp}}$, $\hat{T}_{\mathrm{wav}}$ and $\hat{T}_{\mathrm{ker}}$.

H.1 Implementation details

All simulations are done with Python 3.8.0 and Numpy 1.17.3, where some calculations are accelerated by the just-in-time compiler of Numba 0.47.0, in particular the calculation of the Linear-Time Legendre transform [Luc97]. The discrete wavelet transform is calculated with PyWT 1.1.1 [LGW+19]. We use a second-order finite difference operator for the calculation of the numerical gradient provided by the findiff package, version 0.8.0. The optimization (6.2) is performed with the L-BFGS algorithm [LN89] as implement in Scipy 1.4.1, stopping at a relative decrease in objective function value of less than 10^{-9} and a maximum iteration number of 10000. The baseline estimator is computed with the `ot.emd` function of the Python Optimal Transport package, version 0.6.0. The kernel regression problem (6.3) is solved via scikit-learn, version 0.22.2 [PVG+11]. Plots were made using Matplotlib 3.1.2 and Seaborn 0.9.0.

The boxes in the calculation of $\hat{\boldsymbol{f}}_J$ are picked to be $\tilde{\Omega}_P = \tilde{\Omega}_Q = [-0.5, 1.5]^d$ in the case (id) and $\tilde{\Omega}_P = [-0.5, 1.5]^d$, $\tilde{\Omega}_Q = [0, 4]^d$ in the case (exp). To compute $\hat{\boldsymbol{f}}_J$ for different J, we initialize the optimization at all scales of J with the all zeros vector. An approximation to the ground truth semi-dual objective (2.5) is obtained by integrating both $W_J^\top \gamma_J$ and the interpolation $\mathcal{P}_{\boldsymbol{y} \to T_0(\boldsymbol{x})} \mathcal{L}(W_J^\top \gamma_J)$ over $[0,1]^d$ with Simpson's rule, exploiting the fact that integration over Q is equivalent to integration over P under the push-forward T_0.

The parameters ν_{kernel} and ν_{ridge} are chosen via an oracle procedure, picking the best $\nu_{\text{kernel}} \in \{10^{-9}, 10^{-8.5}, \dots, 10^{-5}\}$, $\nu_{\text{ridge}} \in \{10^{-5}, 10^{-4.5}, \dots, 10^{-1}\}$ as determined by evaluating MSE for an independent draw from P.

The data for all quantitative plots are obtained by taking the median over 32 *i.i.d.* replicates. Error bars are not shown since 95-percentile bootstrapped confidence intervals were not visible for most estimators at the present scale of the plots.

Running on one core of server processors such as an Intel® Xeon® E5-2670 v3 (2.30GHz), the calculation of all wavelet scales of $\hat{T}_{\text{wav}}$ for one replicate takes between 10 and 70 minutes, depending on the sample size and hence the conditioning of the problem. We note that the runtime and space complexity of the algorithm is determined both by the discretization size N and to a lesser extent by the sample size n, as opposed to computing the optimal transport plan between empirical distributions, whose complexity in the regimes considered here is governed entirely by the sample size n. As a comparison, computing all 100 different parameter settings for $\hat{T}_{\text{ker}}$ with $n = 10000$ takes about five hours.

H.2 Numerical error dominates for large sample size

To investigate the observed flattening out of the $\hat{T}_{\text{wav}}$ error curves in Figure 2, we repeat the experiment for (id) and $d = 3$ with a lower resolution discretization, $N = 33$. In Figure 3, we compare the resulting error to the $N = 65$ case considered in Section 6. The error bottoms out for much lower values of n, suggesting that numerical accuracy is indeed responsible for this behavior.

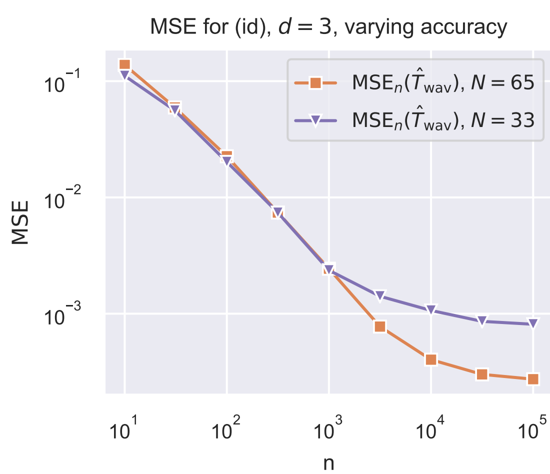


Figure 3: Estimation errors for (id), $d = 3$, low and high accuracy discretization. Median over 32 replicates. The error curve flattens out earlier for $N = 33$, suggesting that numerical approximation errors are responsible for this phenomenon.

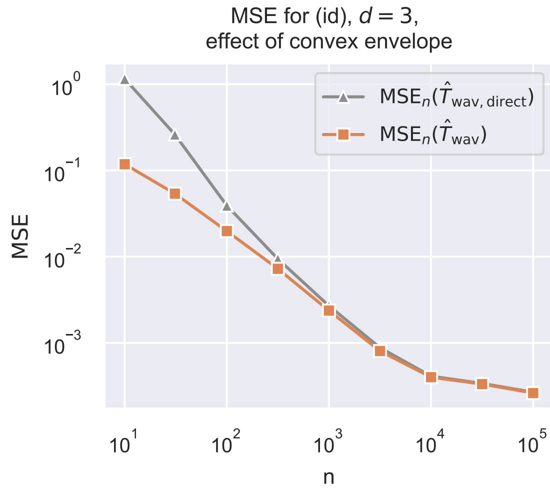


Figure 4: Estimation errors for (id), $d = 3$, direct output of (6.2), $\hat{T}_{\mathrm{wav,direct}}$, compared to its convex envelope $\hat{T}_{\mathrm{wav}}$. Median over 32 replicates. For low sample sizes, taking the convex envelope improves the estimation rate.

H.3 Improved gradient estimation by computing convex envelope

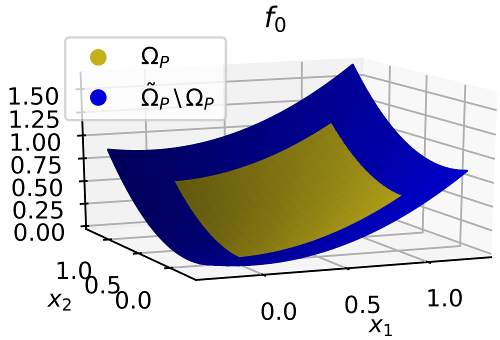
Write $\hat{\boldsymbol{f}}_J = W_J^\top \hat{\gamma}_J$. We observe that while $\hat{\boldsymbol{f}}_J$ by itself often qualitatively yields good results on Ω_P (the support of P), the lack of additional global regularity conditions in the optimization (6.2) often leads to poor approximation outside of the support, i.e., on $\tilde{\Omega}_P \setminus \Omega_P$. In particular, certain regions could simply not enter the optimization at all due to the form of the discrete Legendre transform $\mathcal{L}$ in (6.1). The numerical gradient estimate is somewhat sensitive to this, especially near the boundary, which prompts us to regularize the result further by instead considering its convex envelope, i.e., the largest convex function that lies entirely below the graph of $\hat{\boldsymbol{f}}_J$.

We give a qualitative visual example of this in Figure 5. Depicted for $d = 2$ are first the ground truth potential $f_0(x) = \frac{1}{2}\|x\|_2^2$ corresponding to the identity transport map (Figure 5a) and its derivative in the first coordinate (Figure 5b), where we set $P = \mathsf{Unif}([0,1]^2)$. The parts of those functions corresponding to the support of P are depicted in yellow, while those outside of the support are colored blue. Results of the optimization for $n = 100$ *i.i.d.* samples from P and $Q = \mathsf{Unif}([0,1]^2)$ are shown in Figures 5c (potential) and 5d (first coordinate of numerical gradient) for $J = 1$. While the estimator yields a good approximation to f_0 and T_0 on the interior of Ω_P, the outside appears very ragged. This is remedied by applying $\mathcal{L}$ twice, see Figures 5e (potential) and 5f (numerical gradient).

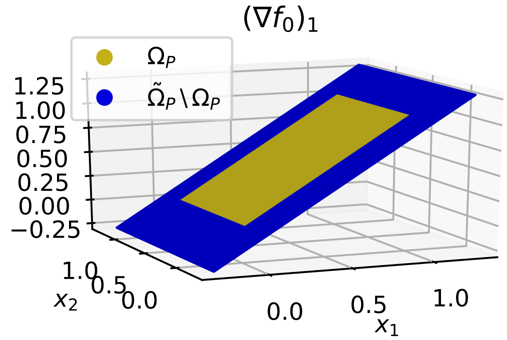
We also quantitatively compare the estimation accuracy of a linear interpolation of $\nabla_{\boldsymbol{x}} \hat{\boldsymbol{f}}_{\hat{J}}$, denoted by $\hat{T}_{\mathrm{wav,\,direct}}$ to that of $\hat{T}_{\mathrm{wav}}$ by repeating the experiment for (id) and $d = 3$, plotting the results in Figure 4. One can observe that for low sample sizes, considering the convex envelope indeed significantly improves the estimation accuracy.

H.4 Performance for varying regularization strength

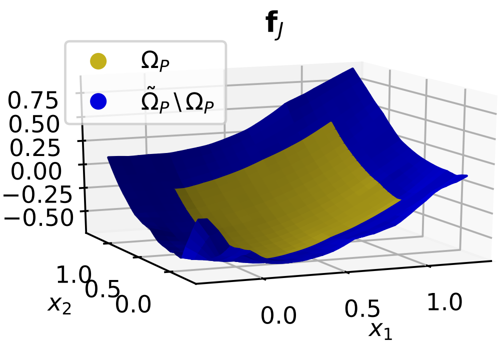
In order to further investigate the performance of the proposed regularized estimators $\hat{T}_{\mathrm{wav}}$ and $\hat{T}_{\mathrm{ker}}$, we show error plots for a selection of fixed regularization parameters in Figure 6. In the case of $\hat{T}_{\mathrm{wav}}$, we plot $\hat{T}_{\mathrm{wav}}^{(J)}$ for the whole range of $J \in \{0, \dots, 3\}$ that can be calculated by means of the



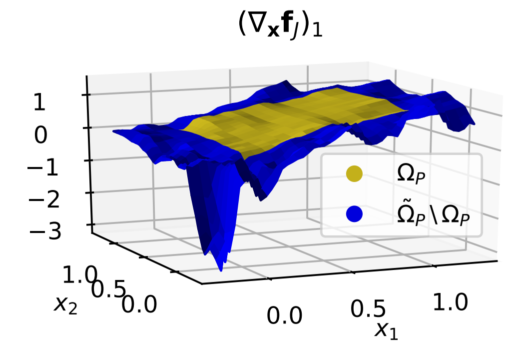
(a) Ground truth potential $f_0 = \frac{1}{2}\|x\|^2$



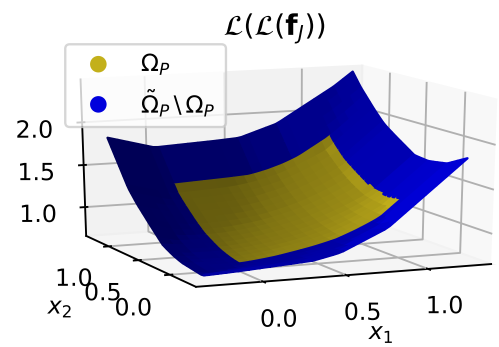
(b) Ground truth transport map, first coordinate, $(T_0(x))_1 = (\nabla f_0(x))_1 = x_1$



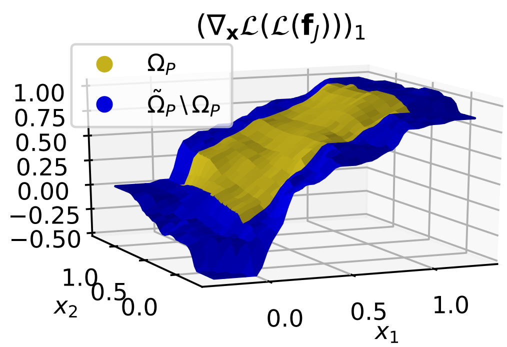
(c) Estimator $\hat{\boldsymbol{f}}_J$ with $J = 1$, see Section H.3



(d) First coordinate of numerical gradient of $\hat{\boldsymbol{f}}_J$ in Panel (c)



(e) Convex hull $\mathcal{L}(\mathcal{L}(\hat{\boldsymbol{f}}_J))$ of estimator in Panel (c)



(f) First coordinate of numerical gradient of $\mathcal{L}(\mathcal{L}(\hat{\boldsymbol{f}}_J))$ in Panel (e)

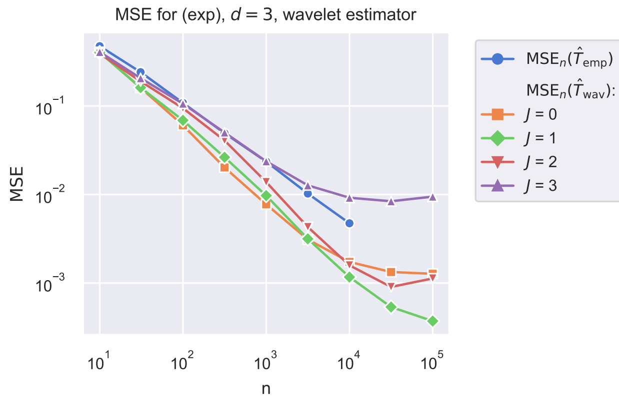
Figure 5: Numerical instability of the boundary is remedied by computing the convex envelope of $\hat{\boldsymbol{f}}$, see Section H.3. Visualization of potentials (ground truth (a), estimated potential (c), and convex envelope of estimated potential (e)), and the first coordinate of the associated gradients (b, d, f, respectively) for the identity transport map and $P = \mathsf{Unif}([0,1]^2)$ in $d = 2$, $J = 1$, $n = 100$.

discrete wavelet transform when $N = 65$. In the case of $\hat{T}_{\text{ker}}$, we plot a subset of all considered combinations $\nu_{\text{kernel}} \in \{10^{-9}, 10^{-8.5}, \ldots, 10^{-5}\}$, $\nu_{\text{ridge}} \in \{10^{-5}, 10^{-4.5}, \ldots, 10^{-1}\}$.

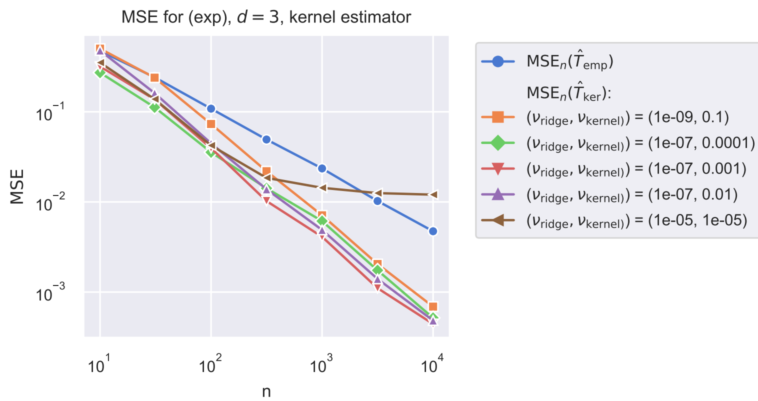
First, for the wavelet estimators $\hat{T}_{\text{wav}}^{(J)}$, we observe that the best bias-variance trade-off is achieved by $J = 0$ for smaller values of n and by $J = 1$ for $n \geq 10^{3.5}$. Moreover, the curve for $J = 3$, corresponding to no regularization, roughly follows the shape of the error curve of $\hat{T}_{\text{emp}}$, until it flattens out due to the numerical error. This is most likely due to the grid approximations involved in the definition of $\hat{T}_{\text{wav}}$.

Next, the kernel estimators $\hat{T}_{\text{ker}}^{(\nu_{\text{ridge}}, \nu_{\text{kernel}})}$ in 3D show mostly similar error curves for the range of parameters considered, although a trade-off that depends on the sample size can be observed between $(\nu_{\text{ridge}}, \nu_{\text{kernel}}) \in \{(10^{-7}, 10^{-4}), (10^{-7}, 10^{-3})\}$. Moreover, in the case of very strong regularization, $(\nu_{\text{ridge}}, \nu_{\text{kernel}}) = (10^{-5}, 10^{-5})$, the error curve flattens out for values of n as low as $10^{2.5}$. In general, for $\hat{T}_{\text{ker}}$, we observe that its performance is more sensitive to changes in ν_{kernel} than to those in ν_{ridge}.

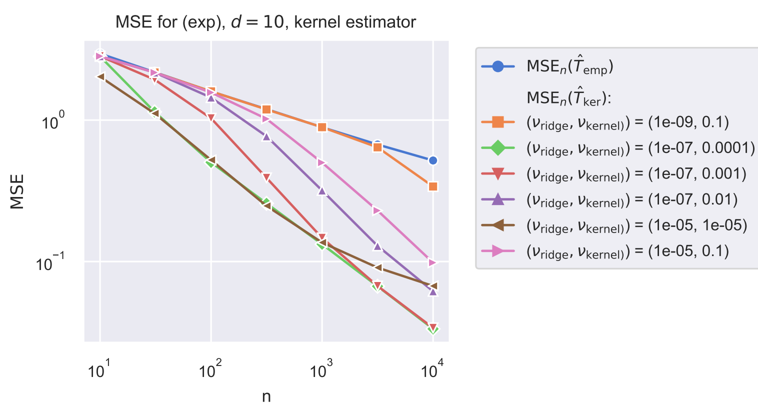
Last, In 10D, we observe a wider range of error curves according to the regularization strength, ranging from a curve that closely matches that of $\hat{T}_{\text{emp}}$ for $(\nu_{\text{ridge}}, \nu_{\text{kernel}}) = (10^{-9}, 10^{-1})$ to a curve that matches most of the error curve shown in Figure 2d for $(\nu_{\text{ridge}}, \nu_{\text{kernel}}) = (10^{-7}, 10^{-4})$.



(a) Wavelet estimator $\hat{T}_{\mathrm{wav}}$, $d = 3$



(b) Kernel estimator $\hat{T}_{\mathrm{ker}}$, $d = 3$



(c) Kernel estimator $\hat{T}_{\mathrm{ker}}$, $d = 10$

Figure 6: Log-log plot of MSE for the exponential transport map, plotted against n, showing the median error over 32 replicates. Individual curves correspond to the performance for different values of the regularization parameters, with errors for $\hat{T}_{\mathrm{emp}}$ shown for comparison.

REFERENCES

[ABM16] Ethan Anderes, Steffen Borgwardt, and Jacob Miller. Discrete wasserstein barycenters: Optimal transport for discrete data. *Mathematical Methods of Operations Research*, 84(2):389–409, 2016.

[ABRNW19] Jason Altschuler, Francis Bach, Alessandro Rudi, and Jonathan Niles-Weed. Massively scalable sinkhorn distances via the nyström method. In *Advances in Neural Information Processing Systems*, pages 4429–4439, 2019.

[ACB17] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

[AGCJ18] Jean Alaux, Edouard Grave, Marco Cuturi, and Armand Joulin. Unsupervised hyperalignment for multilingual word embeddings. *CoRR*, abs/1811.01124, 2018.

[AGP18] Adil Ahidar-Coutrix, Thibaut Le Gouic, and Quentin Paris. Convergence rates for empirical barycenters in metric spaces: Curvature, convexity and extendible geodesics. *arXiv:1806.02740 [math, stat]*, June 2018.

[AJJ18] David Alvarez-Melis, Tommi S. Jaakkola, and Stefanie Jegelka. Structured optimal transport. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, pages 1771–1780, 2018.

[AM19] Daniel Azagra and Carlos Mudarra. Smooth convex extensions of convex functions. *Calculus of Variations and Partial Differential Equations*, 58(3):84, 2019.

[AWR17] Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems*, pages 1964–1974, 2017.

[BCP17] Jérémie Bigot, Elsa Cazelles, and Nicolas Papadakis. Central limit theorems for sinkhorn divergence between probability distributions on finite spaces and statistical applications. *arXiv preprint arXiv:1711.08947*, 2017.

[BGK+17] Jérémie Bigot, Raúl Gouet, Thierry Klein, Alfredo López, et al. Geodesic pca in the wasserstein space by convex pca. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 53, pages 1–26. Institut Henri Poincaré, 2017.

[Bou02] Olivier Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500, 2002.

[Bre91] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

[BSR17] Mathieu Blondel, Vivien Seguy, and Antoine Rolet. Smooth and Sparse Optimal Transport. *arXiv:1710.06276 [cs, stat]*, October 2017.

[Caf92a] Luis A. Caffarelli. Boundary regularity of maps with convex potentials. *Communications on pure and applied mathematics*, 45(9):1141–1151, 1992.

[Caf92b] Luis A. Caffarelli. The regularity of mappings with a convex potential. *Journal of the American Mathematical Society*, 5(1):99–104, 1992.

[Caf96] Luis A. Caffarelli. Boundary Regularity of Maps with Convex Potentials–II. *Annals of mathematics*, 144(3):453–496, 1996.

[CCDM11] Claire Caillerie, Frédéric Chazal, Jérôme Dedecker, and Bertrand Michel. Deconvolution for the Wasserstein metric and geometric inference. *Electronic Journal of Statistics*, 5:1394–1423, 2011.

[CFHR17] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *Advances in Neural Information Processing Systems*, pages 3730–3739, 2017.

[CFT14] Nicolas Courty, Rémi Flamary, and Devis Tuia. Domain adaptation with regularized optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 274–289. Springer, 2014.

[CFTR17] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2017.

[CR12] Guillermo Canas and Lorenzo Rosasco. Learning probability measures with respect to optimal transport metrics. In *Advances in Neural Information Processing Systems*, pages 2492–2500, 2012.

[CS18] Sebastian Claici and Justin Solomon. Wasserstein coresets for lipschitz costs. *CoRR*, abs/1805.07412, 2018.

[CSB+18] Elsa Cazelles, Vivien Seguy, Jérémie Bigot, Marco Cuturi, and Nicolas Papadakis. Geodesic PCA versus log-pca of histograms in the wasserstein space. *SIAM J. Scientific Computing*, 40(2), 2018.

[Cut13] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

[DBGGL19] Eustasio Del Barrio, Fabrice Gamboa, Paula Gordaliza, and Jean-Michel Loubes. Obtaining fairness using optimal transport theory. *ICML*, 2019.

[dBGLL19] Eustasio del Barrio, Paula Gordaliza, Hélène Lescornel, and Jean-Michel Loubes. Central limit theorem and bootstrap procedure for wasserstein's variations with an application to structural relationships between distributions. *Journal of Multivariate Analysis*, 169:341 – 362, 2019.

[DFM15] Jérôme Dedecker, Aurélie Fischer, and Bertrand Michel. Improved rates for Wasserstein deconvolution with ordinary smooth error in dimension one. *Electronic journal of statistics*, 9(1):234–265, 2015.

[DGK18] Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn's algorithm. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1367–1376, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[DKF+18] Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IV*, pages 467–483, 2018.

[DM13] Jérôme Dedecker and Bertrand Michel. Minimax rates of convergence for Wasserstein deconvolution with supersmooth errors in any dimension. *Journal of Multivariate Analysis*, 122:278–291, 2013.

[DR07] Arnak Dalalyan and Markus Reiß. Asymptotic statistical equivalence for ergodic diffusions: The multidimensional case. *Probability theory and related fields*, 137(1-2):25–47, 2007.

[ENT11] Luca Esposito, Carlo Nitsch, and Cristina Trombetti. Best constants in Poincaré inequalities for convex domains. *arXiv preprint arXiv:1110.2960*, 2011.

[Eva10] Lawrence C. Evans. *Partial Differential Equations*. American Mathematical Soc., 2010.

[FCCR18] Rémi Flamary, Marco Cuturi, Nicolas Courty, and Alain Rakotomamonjy. Wasserstein discriminant analysis. *Machine Learning*, 107(12):1923–1945, 2018.

[FCVP17] Jean Feydy, Benjamin Charlier, François-Xavier Vialard, and Gabriel Peyré. Optimal transport for diffeomorphic registration. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017 - 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I*, pages 291–299, 2017.

[FHN$^+$19] Aden Forrow, Jan-Christian Hütter, Mor Nitzan, Geoffrey Schiebinger, Philippe Rigollet, and Jonathan Weed. Statistical optimal transport via factored couplings. *AISTATS*, 06 2019.

[FLF19] Rémi Flamary, Karim Lounici, and André Ferrari. Concentration bounds for linear Monge mapping estimation and optimal transport domain adaptation, 2019.

[GCPB16] Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems*, pages 3440–3448, 2016.

[Gho02] Mohammad Ghomi. The problem of optimal smoothing for convex functions. *Proceedings of the American Mathematical Society*, 130(8):2255–2259, 2002.

[Gig11] Nicola Gigli. On Hölder continuity-in-time of the optimal transport map towards measures along a curve. *Proceedings of the Edinburgh Mathematical Society*, 54(2):401–409, June 2011.

[GJB19] Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised alignment of embeddings with wasserstein procrustes. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 1880–1890. PMLR, 16–18 Apr 2019.

[GN16] Evarist Giné and Richard Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*, volume 40. Cambridge University Press, 2016.

[GPC18] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, pages 1608–1617, 2018.

[Gun18] Florian F Gunsilius. On the convergence rate of potentials of brenier maps. 2018.

[HÏ9] Jan-Christian Hütter. *Minimax Estimation with Structured Data: Shape Constraints, Causal Models, and Optimal Transport*. PhD thesis, Massachusetts Institute of Technology, 2019.

[HKPT98] Wolfgang Härdle, Gerard Kerkyacharian, Dominique Picard, and Alexander Tsybakov. *Wavelets, Approximation, and Statistical Applications*, volume 129. Springer Science & Business Media, 1998.

[HL01] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of Convex Analysis*. Grundlehren Text Editions. Springer-Verlag, Berlin Heidelberg, 2001.

[HMT07] Steve Hofmann, Marius Mitrea, and Michael Taylor. Geometric and transformational properties of Lipschitz domains, Semmes-Kenig-Toro domains, and other classes of finite perimeter domains. *The Journal of Geometric Analysis*, 17(4):593–647, 2007.

[IH81] I. A. Ibragimov and R. Z. Hasminskiĭ. *Statistical estimation*, volume 16 of *Applications of Mathematics*. Springer-Verlag, New York, 1981.

[JCG18] Hicham Janati, Marco Cuturi, and Alexandre Gramfort. Wasserstein regularization for sparse multi-task regression. *CoRR*, abs/1805.07833, 2018.

[Jon81] Peter W. Jones. Quasiconformal mappings and extendability of functions in Sobolev spaces. *Acta Mathematica*, 147(1):71–88, 1981.

[Kan42] Leonid Vitalievich Kantorovich. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201, 1942.

[Kan48] Leonid V. Kantorovich. On a problem of Monge. In *CR (Doklady) Acad. Sci. URSS (NS)*, volume 3, pages 225–226, 1948.

[Kol11] Vladimir Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems. École d'Été de Probabilités de Saint-Flour XXXVIII-2008.* Lecture Notes in Mathematics 2033. Berlin: Springer. ix, 254 p. EUR 48.10 , 2011.

[KP91] Steven G Krantz and Harold R Parks. On the vector sum of two convex sets in space. *Canadian Journal of Mathematics*, 43(2):347–355, 1991.

[KS84] Martin Knott and Cyril S. Smith. On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43(1):39–49, 1984.

[KSS19] Alexey Kroshnin, Vladimir Spokoiny, and Alexandra Suvorikova. Statistical inference for bures-wasserstein barycenters. *arXiv preprint arXiv:1901.00226*, 2019.

[KTM18] Marcel Klatt, Carla Tameling, and Axel Munk. Empirical regularized optimal transport: Statistical theory and applications. *arXiv preprint arXiv:1810.09880*, 2018.

[LCCS18] Hugo Lavenant, Sebastian Claici, Edward Chien, and Justin Solomon. Dynamical optimal transport on discrete surfaces. *ACM Trans. Graph.*, 37(6):250:1–250:16, 2018.

[Led19] Michel Ledoux. On optimal matching of gaussian samples. *Journal of Mathematical Sciences*, 238(4):495–522, 2019.

[Leo17] Giovanni Leoni. *A first course in Sobolev spaces*, volume 181. American Mathematical Soc., 2017.

[Lep91] OV Lepskii. On a problem of adaptive estimation in gaussian white noise. *Theory of Probability & Its Applications*, 35(3):454–466, 1991.

[Lep92] OV Lepskii. Asymptotically minimax adaptive estimation. i: Upper bounds. optimally adaptive estimates. *Theory of Probability & Its Applications*, 36(4):682–697, 1992.

[Lep93] OV Lepskii. Asymptotically minimax adaptive estimation. ii. schemes without optimal adaptation: Adaptive estimators. *Theory of Probability & Its Applications*, 37(3):433–448, 1993.

[LGW+19] Gregory Lee, Ralf Gommers, Filip Waselewski, Kai Wohlfahrt, and Aaron O'Leary. Pywavelets: A python package for wavelet analysis. *Journal of Open Source Software*, 4(36):1237, 2019.

[LN89] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

[Luc97] Yves Lucet. Faster than the fast legendre transform, the linear-time legendre transform. *Numerical Algorithms*, 16(2):171–185, 1997.

[Mal99] Stéphane Mallat. *A Wavelet Tour of Signal Processing.* Elsevier, 1999.

[Mas07] Pascal Massart. *Concentration Inequalities and Model Selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII - 2003.* École d'Été de Probabilités de Saint-Flour. Springer-Verlag, Berlin Heidelberg, 2007.

[MC98] Axel Munk and Claudia Czado. Nonparametric validation of similar distributions and assessment of goodness of fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):223–241, 1998.

[MG79] Hans-Georg Müller and Theo Gasser. Optimal convergence properties of kernel estimates of derivatives of a density function. In *Smoothing techniques for curve estimation*, pages 144–154. Springer, 1979.

[MMC16] Grégoire Montavon, Klaus-Robert Müller, and Marco Cuturi. Wasserstein training of restricted boltzmann machines. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3711–3719, 2016.

[Mon81] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

[MPZ19] Valentina Masarotto, Victor M Panaretos, and Yoav Zemel. Procrustes metrics on covariance operators and optimal transportation of gaussian processes. *Sankhya A*, 81(1):172–213, 2019.

[Mur12] Kevin P Murphy. *Machine learning: a probabilistic perspective.* MIT press, 2012.

[NWR19] Jonathan Niles-Weed and Philippe Rigollet. Estimation of wasserstein distances in the spiked transport model. *arXiv preprint arXiv:1909.07513*, 2019.

[PC19] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

[PCFH16] Michaël Perrot, Nicolas Courty, Rémi Flamary, and Amaury Habrard. Mapping estimation for discrete optimal transport. In *Advances in Neural Information Processing Systems*, pages 4197–4205, 2016.

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[PZ19] Victor M. Panaretos and Yoav Zemel. Statistical aspects of Wasserstein distances. *Annual Review of Statistics and Its Application*, 6(1):405–431, 2019.

[Rau10] Holger Rauhut. Compressive sensing and structured random matrices. *Theoretical foundations and numerical methods for sparse recovery*, 9:1–92, 2010.

[RCP16] Antoine Rolet, Marco Cuturi, and Gabriel Peyré. Fast dictionary learning with a smoothed wasserstein loss. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016, Cadiz, Spain, May 9-11, 2016*, pages 630–638, 2016.

[RR90] Ludger Rüschendorf and Svetlozar T. Rachev. A characterization of random variables with minimum L2-distance. *Journal of Multivariate Analysis*, 32(1):48–54, 1990.

[RTC17] Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

[RW18] Philippe Rigollet and Jonathan Weed. Entropic optimal transport is maximum-likelihood deconvolution. *Comptes Rendus Mathematique*, 356(11):1228 – 1235, 2018.

[RW19] Philippe Rigollet and Jonathan Weed. Uncoupled isotonic regression via minimum Wasserstein deconvolution. *Information and Inference (to appear)*, 2019.

[Ryc99] Vyacheslav S. Rychkov. On restrictions and extensions of the Besov and Triebel–Lizorkin spaces with respect to Lipschitz domains. *Journal of the London Mathematical Society*, 60(1):237–257, 1999.

[San15] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, pages 99–102, 2015.

[SC15] Vivien Seguy and Marco Cuturi. Principal geodesic analysis for probability measures under the optimal transport metric. In *Advances in Neural Information Processing*

Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, pages 3312–3320, 2015.

[SCSJ17] Matthew Staib, Sebastian Claici, Justin M. Solomon, and Stefanie Jegelka. Parallel streaming wasserstein barycenters. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 2644–2655, 2017.

[SDF+18] Vivien Seguy, Bharath Bhushan Damodaran, Remi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. Large Scale Optimal Transport and Mapping Estimation. *arXiv preprint arXiv:1711.02283*, February 2018.

[SdGP+15] Justin Solomon, Fernando de Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Trans. Graph.*, 34(4):66:1–66:11, July 2015.

[SHB+17] Morgan A. Schmitz, Matthieu Heitz, Nicolas Bonneel, Fred Maurice Ngolè Mboula, David Coeurjolly, Marco Cuturi, Gabriel Peyré, and Jean-Luc Starck. Wasserstein dictionary learning: Optimal transport-based unsupervised non-linear dictionary learning. *CoRR*, abs/1708.01955, 2017.

[SHB+18] Morgan A. Schmitz, Matthieu Heitz, Nicolas Bonneel, Fred Maurice Ngolè Mboula, David Coeurjolly, Marco Cuturi, Gabriel Peyré, and Jean-Luc Starck. Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM J. Imaging Sciences*, 11(1):643–678, 2018.

[SP18] Shashank Singh and Barnabás Póczos. Minimax distribution estimation in wasserstein distance. *arXiv preprint arXiv:1802.08855*, 2018.

[SPKS16] Justin Solomon, Gabriel Peyré, Vladimir G. Kim, and Suvrit Sra. Entropic metric alignment for correspondence problems. *ACM Trans. Graph.*, 35(4):72:1–72:13, 2016.

[SST+19] Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, Lia Lee, Jenny Chen, Justin Brumbaugh, Philippe Rigollet, Konrad Hochedlinger, Rudolf Jaenisch, Aviv Regev, and Eric S. Lander. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928 – 943.e22, 2019.

[Str18] Claudia Strauch. Adaptive invariant density estimation for ergodic diffusions over anisotropic classes. *The Annals of Statistics*, 46(6B):3451–3480, 2018.

[Tal14] Michel Talagrand. *Upper and lower bounds for stochastic processes: modern methods and classical problems*, volume 60. Springer Science & Business Media, 2014.

[TM18] Carla Tameling and Axel Munk. Computational strategies for statistical inference based on empirical optimal transport. In *2018 IEEE Data Science Workshop, DSW 2018, Lausanne, Switzerland, June 4-6, 2018*, pages 175–179, 2018.

[TMS+13] Kazuaki Tanaka, Makoto Mizuguchi, Kouta Sekine, Akitoshi Takayasu, and Shin'ichi Oishi. Estimation of an embedding constant on Lipschitz domains using extension operators. In *JSST 2013 International Conference on Simulation Technology*, 2013.

[Tri06] Hans Triebel. *Theory of Function Spaces III.* Monographs in Mathematics. Birkhäuser Basel, 2006.

[Tsy09] Alexandre B. Tsybakov. *Introduction to nonparametric estimation.* Springer Series in Statistics. Springer, New York, 2009.

[vdG87] Sara van de Geer. A new approach to least-squares estimation, with applications.

Ann. Statist., 15(2):587–602, 1987.

[vdG02] Sara van de Geer. M-estimation using penalties or sieves. *J. Statist. Plann. Inference*, 108(1-2):55–69, 2002. C. R. Rao 80th birthday felicitation volume, Part II.

[VdV00] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press, 2018.

[Vil03] Cédric Villani. *Topics in Optimal Transportation*, volume 58. American Mathematical Soc., 2003.

[Vil09] Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009. Old and new.

[vW07] Aad W. van der Vaart and Jon A. Wellner Wellner. Empirical processes indexed by estimated functions. In *Asymptotics: Particles, Processes and Inverse Problems*, pages 234–252. Institute of Mathematical Statistics, 2007.

[WB17] Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *arXiv:1707.00087 [math, stat]*, June 2017.

[WB19] Jonathan Weed and Quentin Berthet. Estimation of smooth densities in Wasserstein distance. *COLT*, 2019.

[Yan14] Min Yan. Extension of convex function. *Journal of Convex Analysis*, 21(4):965–987, 2014.

[YDV+18] Karren D Yang, Karthik Damodaran, Saradha Venkatchalapathy, Ali C Soylemezoglu, GV Shivashankar, and Caroline Uhler. Autoencoder and optimal transport to infer single-cell trajectories of biological processes. *bioRxiv*, page 455469, 2018.

[Yuk06] Joseph E Yukich. *Probability theory of classical Euclidean optimization problems.* Springer, 2006.

[ZP19] Yoav Zemel and Victor M. Panaretos. Fréchet means and Procrustes analysis in Wasserstein space. *Bernoulli*, 25(2):932–976, 05 2019.

Jan-Christian Hütter
Broad Institute of MIT and Harvard
415 Main Street,
Cambridge, MA 02142, USA
(jhuetter@broadinstitute.org)

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