MINIMAX ESTIMATION OF SMOOTH OPTIMAL TRANSPORT MAPS

BY JAN-CHRISTIAN HÜTTER 1 AND PHILIPPE RIGOLLET 2

1Broad Institute of MIT and Harvard
jhuetter@broadinstitute.org

2Department of Mathematics
Massachusetts Institute of Technology
rigollet@math.mit.edu

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.

1. Introduction. Wasserstein distances and the associated problem of optimal transport date back to the work of Gaspard Monge [62] and have since then become important tools in pure and applied mathematics [95, 96, 76]. Tools from optimal transport have been successfully employed in machine learning [2, 70, 79, 32, 7, 36, 45, 63, 74, 78, 39, 85, 5, 7, 27, 16] computer graphics [53, 83, 84, 31], statistics [80, 1, 72, 98, 101, 68, 9, 17, 71, 18, 88, 49, 8, 28, 52], and, more recently, computational biology [77, 99].

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 $T_0 : \mathbb{R}^d \rightarrow \mathbb{R}^d$ that minimizes the objective

\[ \min_T \int_{\mathbb{R}^d} \|T(x) - x\|^2 \text{d}P(x), \quad \text{s.t. } T_\# P = Q, \]

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

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

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 [47, 48], who considered the following relaxation. Instead of looking for a map $T_0$, look for a transport plan $\Gamma_0$ in the set of all possible probability 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

\[ \min_{\Gamma} \int_{\mathbb{R}^d} \|x - y\|^2 \text{d}\Gamma(x, y) \quad \text{s.t. } \Gamma \in \Pi(P, Q), \]

AMS 2000 subject classifications: 62G05

Keywords and phrases: Optimal transport, nonparametric estimation, minimax rates, wavelet estimator
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 $\Gamma_0$ 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 $\delta$ 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 $\|x - y\|_p^p$ 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 [24, 20, 19, 69, 21, 34, 73], 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 [37], which in turn enables the generalization of PCA (principal component analysis) to spaces of probability measures [9, 60].

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 $X_1, \ldots, X_n$ from $P$ and $Y_1, \ldots, Y_n$ 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 [65]. It leads to a finite-dimensional linear problem that can be approximated very efficiently due to recent algorithmic advances [22, 4, 70, 29]. 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 $x \notin \{X_1, \ldots, X_n\}$. In contrast, we would like to obtain an estimator $\hat{T}$ with guarantees in $L^2(P)$, that is, with convergence of

\[
\|\hat{T} - T_0\|_{L^2(P)} := \int \|\hat{T}(x) - T_0(x)\|^2_2 \, dP(x).
\]

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 [6]. 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, $\mathbb{E}[W^2_2(\hat{P},\hat{Q}) - W^2_2(P,Q)] \propto n^{-1/d}$ (see [67]). 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

$$
\int_{\mathbb{R}^d} \| T(x) - x \|^2_2 dP(x) - \int_{\mathbb{R}^d} \| T_0(x) - x \|^2_2 dP(x).
$$

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 [37]. They show that the optimization problem (1.1) has quadratic growth, in the sense of a stability estimate

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

provided $T_0 = \nabla f_0$ is Lipschitz continuous on $\mathbb{R}^d$ and $T_0^* P = Q$. 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 [22, 35, 4], $\ell^2$-regularization [10], smoothness regularization [69], to regularization techniques specifically adapted to the application of domain adaptation [20, 21, 19]. Notably, [35] 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 [73] 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 [15, 26, 25] 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, [33] 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 [40]. 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 [40] 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 [37]. 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, [96, Theorem 12.50(iii)], [12, 13, 14], we can obtain a variance rate of $n^{-2\alpha/(2n-2+d)}$ (up
to log factors) under assumptions quantitatively comparable to Gunsilius’s, while his results imply a sub-optimal rate of \( n^{-(2\alpha-2)/(2\alpha-2+4d)} \), see Appendix E of the supplement [43].

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 Caffarelli’s global regularity theory prevents us from claiming (near) minimax optimality over Hölder smooth densities, see Remark E.4. 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 gathered in the appendices of the supplement [43]. Moreover, the supplement 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 \([m] := \{1, \ldots, m\}\). 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 \( C, a \leq Cb \). 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 \( a \lor b \), \( a \land b \) 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 \( \text{supp} f \) to denote the support of a function or measure \( f \), and \( \text{diam} \Omega \) 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 \( \lambda \leq A \) if all eigenvalues of \( A \) are bounded below \( \lambda \), 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 \( \ell^p \) either the space \( \mathbb{R}^d \) endowed with the usual \( \ell^p \) norms \( \| \cdot \|_p \), or, by abuse of notation, the spaces of multi-dimensional sequences \( \gamma : \mathbb{Z}^m \to \mathbb{R} \) with \( \|\gamma\|_p = \sum_{k \in \mathbb{Z}^m} |\gamma_k|^p \) 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 $\lambda$ on $\mathbb{R}^d$, whose norms we denote by $\| \cdot \|_{L^p(\mathbb{R}^d)}$ and $\| \cdot \|_{L^p(\Omega)}$, respectively. By abuse of notation, for a different measure $P$, we denote the associated Lebesgue norms by $\| \cdot \|_{L^p(P)}$. We abbreviate with “a.e.” any statement that holds “almost everywhere” with respect to the Lebesgue measure.

For a differentiable one-dimensional function $f : \mathbb{R} \ni t \rightarrow \mathbb{R}$, we denote its derivative by $\frac{df}{dt}$. For a function $f : \mathbb{R}^d \ni t \rightarrow \mathbb{R}$, we denote by $\partial_i f = \partial / (\partial x_i)$ 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 $\Omega$. For a multi-index $b \in \mathbb{N}^d$, we set

$$\partial^b f = \frac{\partial}{\partial x_1^b} \ldots \frac{\partial}{\partial x_d^b} f,$$

and $|b| = \sum_{i=1}^d b_i$.

The symbol $\partial f$ is also used to denote the sub-differential of a convex function $f$, while we use the symbols $\nabla f$ for the gradient of a function $f$ and $Dg$ for the derivative of a vector-valued function $g : \mathbb{R}^d \ni \omega \rightarrow \mathbb{R}^{d_2}$, $\nabla f = (\partial_1 f, \ldots, \partial_d f)^\top$ 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 $\Omega$ as defined in Appendix B are denoted by $C^\alpha(\Omega)$ and their associated norms by $\| \cdot \|_{C^\alpha(\Omega)}$. Similarly, the $p$-Sobolev spaces of order $\alpha$ for $p \in [1, \infty]$ are denoted by $W^{\alpha,p}(\Omega)$ with norms $\| \cdot \|_{W^{\alpha,p}(\Omega)}$, 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 [90, 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

$$\min_T \mathcal{P}_M(T) \quad \text{s.t.} \quad T_# P = Q,$$

where $\mathcal{P}_M(T) := \frac{1}{2} \int_{\mathbb{R}^d} ||T(x) - x||_2^2 dP(x)$,

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

\begin{equation}
\min_{\Gamma} \tilde{\mathcal{P}}_K(\Gamma) \quad \text{s.t.} \quad \Gamma \in \Pi(P, Q)
\end{equation}

where $\tilde{\mathcal{P}}_K(\Gamma) := \frac{1}{2} \int ||x - y||_2^2 d\Gamma(x, y)$,

and $\Pi(P, Q)$ denotes the set of couplings between $P$ and $Q$, that is, the set of probability measures $\Gamma$ 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 $A \subseteq \mathbb{R}^d$.

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

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

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

$$\tilde{\mathcal{P}}_K(\Gamma) = \frac{1}{2} \int ||x - y||_2^2 d\Gamma(x, y) = \frac{1}{2} \int ||x||_2^2 dP(x) + \frac{1}{2} \int ||y||_2^2 dQ(y) - \int \langle x, y \rangle d\Gamma(x, y),$$
Since the first two terms above do not depend on $\Gamma$, we obtain the equivalent optimization problem

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

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:

$$
\min_{f,g} \int f(x) \, dP(x) + \int g(y) \, dQ(y) \quad \text{s.t. } f(x) + g(y) \geq \langle x, y \rangle, \quad P \otimes Q \text{-a.e.},
$$

$$
f \in L^1(P), g \in L^1(Q).
$$

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 $g$ given by the Legendre-Fenchel conjugate (see Section A) of $f$:

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

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

$$
\min_{f} \mathcal{S}(f) = \int f(x) \, dP(x) + \int f^*(y) \, dQ(y) \quad \text{s.t. } f \in L^1(P),
$$

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, [50, 11, 75]). 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 $\Gamma_0 = (\text{id}, T_0)_P^P$, where $T_0 = \nabla f_0$ is the gradient of a convex function $f_0 : \mathbb{R}^d \to \mathbb{R}$. 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 $Y \sim Q = (T_0)_P^P$ 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 $T$. 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 $\rho_P$ 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 $\Omega$ by 1 in every direction. Let $T = T(M)$ be the set of all differentiable functions $T : \Omega = \tilde{\Omega} \rightarrow \mathbb{R}^d$ such that $T = \nabla f$ for some differentiable convex function $f : \Omega = \tilde{\Omega} \rightarrow \mathbb{R}^d$ and

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

For $R > 1$ and $\alpha > 1$, assume that $T_0 \in T_\alpha = T_\alpha (M, R)$, where

$$T_\alpha (M, R) = \{ T \in T(M) : T \text{ is } [\alpha] \text{-times differentiable and } \| T \|_{C^\alpha(\Omega)} \leq R \}.$$  

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

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

$$\inf_{\hat{T}} \sup_{P \in M, T_0 \in T_\alpha} \mathbb{E} \left[ \int \| \hat{T}(x) - T_0(x) \|^2 dP(x) \right] \gtrsim n^{-\frac{2\alpha}{2\alpha - 2 + \alpha}} \lor \frac{1}{n},$$

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 $P \in M$ and $T_0 \in T_\alpha$, there exists an estimator $\hat{T}$, given in Section 3, 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,

$$\sup_{P \in M, T_0 \in T_\alpha} \mathbb{E} \left[ \int \| \hat{T}(x) - T_0(x) \|^2 dP(x) \right] \lesssim n^{-\frac{2\alpha}{2\alpha - 2 + \alpha}} (\log(n))^2 \lor \frac{1}{n}.$$

**Remark 3.** Assumption A1 can be significantly relaxed with respect to the geometry of $\Omega$ 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 $\text{supp}(Q) = [0, 1]^d$ 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 $\text{supp}(P)$, $\text{supp}(Q)$ and $\tilde{\Omega}$—see for example [46, 30, 89]—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), $M^{-1} \leq DT(x)$, in particular on the support of $P$, $x \in \Omega$. This yields convergence estimates for the optimal transport map as shown in [37], 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 [94, 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 \( n^{-\frac{2\alpha}{2\alpha-2+\alpha}} \). 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 [12, 13, 14]. For completeness, we also give a version of our upper bound results under smoothness assumptions on \( P \) and \( Q \) in Theorem E.1, 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 \( Q = T\#P \), estimation rates for \( \hat{T} \) immediately translate into rates for estimating \( Q \) in the 2-Wasserstein distance [97, 82, 98]. In fact, \( W_Q^2(Q, Q) \) can be bounded by \( \mathbb{E}_P[||\hat{T} - T_0||_2] \), since \((\hat{T}, T_0)\#P\) is a candidate transport plan between \( Q \) and \( \hat{Q} \). Up to log factors, our rates obtained below match those obtained in [98] 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 \( \frac{2\alpha}{2\alpha-2+\alpha} \) in the minimax rate. We prove the following theorem.

**Theorem 6.** Fix \( \alpha \geq 1 \). It holds that

\[
\inf_{\hat{T}} \sup_{P \in \mathcal{M}, T_0 \in T_n} \mathbb{E} \left[ \int ||\hat{T}(x) - T_0(x)||_2^2 dP(x) \right] \geq n^{-\frac{2\alpha}{2\alpha-2+\alpha}} \vee \frac{1}{n},
\]

where the infimum is taken over all measurable functions of \( X_1, \ldots, X_n, Y_1, \ldots, Y_n \).

**Proof.** The proof uses standard tools to establish minimax lower bounds, [91, Theorem 2.5, Lemma 2.9, Theorem 2.2], that we restate in Appendix D for the convenience of the reader as Theorem D.1, Lemma D.2, and Theorem D.3, respectively. It relies on the following construction.

Set \( P = \text{Unif}([0,1]^d) \in \mathcal{M} \), the uniform distribution on the hypercube. For \( \alpha > 1 \), let \( \xi: \mathbb{R} \to \mathbb{R} \) 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 \( g: \mathbb{R}^d \to \mathbb{R} \) by

\[
g(x) = \prod_{i \in [d]} \xi(x_i), \quad x = (x_1, \ldots, x_d),
\]

and note that \( \nabla g(x_0, \ldots, x_0) \neq 0 \) and \( \text{supp}(g) = [0,1]^d \) by the above assumptions on \( \xi \).
Let \( m = \lceil \theta n \frac{1}{\alpha - 1} \rceil \) be a positive integer where \( \theta \) 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 \( x^{(j)}_i = (j_i - 1)/m \), \( i = 1, \ldots, d \) 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 \( \text{supp}(g_j) \subseteq x^{(j)} + [0,1/m]^d \) and hence that the supports of the functions \( \{g_j\}_{j \in [m]^d} \) are pairwise disjoint.

Next, let \( b \in \mathbb{N}^d \) be a multi-index and observe that the differential operator \( \partial^b \) applied to \( g_j \) yields \( \partial^b g_j(\cdot) = m|b|^{-\alpha - 1} \partial^b g(m\cdot - x^{(j)}) \). 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 \( \|D^2 g_j(x)\|_{\text{op}} \leq 1/2 \) 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 \( \kappa \) 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 D.2, guarantees the existence of binary vectors \( \tau^{(0)}, \tau^{(1)}, \ldots, \tau^{(K)} \in \{0,1\}^{|m|^d} \) \( \tau^{(0)} = (0, \ldots, 0) \), \( K \geq 2^m/8 \) such that \( \|\tau^{(k)} - \tau^{(k')}\|_2^2 \geq m^d/4 \) for \( 0 \leq k \neq k' \leq K \). With this, we define the following collection of Kantorovich potentials:
\[
\phi_k(x) = \frac{1}{2} \|x\|^2 + \sum_{j \in [m]^d} \tau_j^{(k)} g_j(x) , \quad k = 0, \ldots, K .
\]

It is easy to see (Lemma G.1) that for any \( k = 0, \ldots, K \) and \( m \geq m_0 \), \( \nabla \phi_k \) is a bijection from \([0,1]^d\) to \([0,1]^d\). Moreover, by Weyl’s inequality and the above bound \( \|D^2 g_j(x)\|_{\text{op}} \leq 1/2 \), 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 \( \|\nabla \phi_k\|_{L^\infty([0,1]^d)} \leq 2 \) and \( \lambda_{\max}(D^2 \phi_k(x)) \leq 2 \) for all \( x \in \mathbb{R}^d \). Hence, \( T_k \equiv \nabla \phi_k \in T_{\alpha}(M, R) \) for \( M > 2 \) and \( \kappa \) small enough. We now check the conditions of Theorem D.1, where we consider the distance measure
\[
d(T_k, T_{k'}) = \int_{[0,1]^d} \|T_k - T_{k'}\|^2 \, dx , \quad 0 \leq k, k' \leq K.
\]
First, observe that for \( 0 \leq k \neq k' \leq K \), it holds that
\[
\int_{[0,1]^d} \|\nabla \phi_k(x) - \nabla \phi_{k'}(x)\|^2 \, dx
\]
\[
= \frac{\kappa^2}{m^{2\alpha + d}} \sum_{j \in [m]^d} (\tau_j^{(k)} - \tau_j^{(k')} \|^2 \int_{\mathbb{R}^d} \|\nabla g(x)\|^2 \, dx \geq \frac{1}{m^{2\alpha}} .
\]
This yields
\[
\int_{[0,1]^d} \|\nabla \phi_k(x) - \nabla \phi_{k'}(x)\|^2 \, dx \geq n^{-\frac{2\alpha-2d}{\alpha}} ,
\]
which completes checking the separation condition (i) of Theorem D.1.

To check condition (ii) of Theorem D.1, recall the Kullback-Leibler (KL) divergence between two measures \( Q, P \) such that \( Q \) is absolutely continuous with respect to \( P \) is defined by
\[
D(Q\|P) = \mathbb{E} \log \left( \frac{dQ}{dP}(W) \right) , \quad W \sim Q.
\]
In view of Lemma G.1, for any $k = 0, \ldots, K$, 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

\[
\frac{dQ_k}{dP}(y) := \frac{1}{\det D^2 \phi_k((\nabla \phi_k)^{-1}(y))} 1((\nabla \phi_k)^{-1}(y) \in [0, 1]^d).
\]

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

\[
D(Q_k \| P) = \mathbb{E} \log \left( \frac{dQ_k}{dP}(Y) \right) = \mathbb{E} \log \left( \frac{dQ_k}{dP}(\nabla \phi_k(X)) \right)
\]

\[
= - \int_{[0,1]^d} \log \left( \det D^2 \phi_k(x) \right) \, dx.
\]

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

\[
\log \left( \det D^2 \phi_k(x) \right) = \sum_{l=1}^d \log \left( 1 + \lambda_l \left( \sum_{j \in [m]^d} \tau_j^{(k)} D^2 g_j(x) \right) \right)
\]

\[
= \sum_{l=1}^d \sum_{j \in [m]^d} \log \left( 1 + \tau_j^{(k)} \lambda_l \left( D^2 g_j(x) \right) \right),
\]

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

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

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

\[
D(Q_k \| P) \leq - \sum_{j \in [m]^d} \tau_j^{(k)} \int_{[0,1]^d} \text{Tr}(D^2 g_j(x)) \, dx
\]

\[
+ \frac{1}{2} \sum_{j \in [m]^d} \int_{[0,1]^d} \|D^2 g_j(x)\|_F^2 \, dx.
\]

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

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

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

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

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 $\theta$ large enough. This completes checking (ii) in Theorem D.1 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 $\phi_0(x) = \|x\|^2/2$ and $\phi_1(x) = \phi_0(x) + (\bar{\theta}/\sqrt{n})g(x)$ for $\bar{\theta}$ chosen to ensure $\phi_0, \phi_1 \in T_\alpha$, applying Theorem D.3 in Appendix D. The separation condition is given by

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

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

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

for $\bar{\theta}$ large enough. \hfill $\square$

Looking back at this proof, we get a better understanding of the exponent in the minimax rate $n^{-\frac{2\alpha}{4\alpha-2+\beta}}$. Given that $n^{-\frac{2\beta-1}{2\beta+d}}$ is the minimax rate of estimation of the $k$th derivative of a $\beta$-smooth density in $L^2$ [64], 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 $T$ 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 $T : \mathbb{R}^d \to \mathbb{R}^d$, the existence of an antiderivative needs to be guaranteed a priori, as in the case of transport maps by assuming $T = \nabla f$ for $f : \mathbb{R}^d \to \mathbb{R}$.

Similar rates arise in the estimation of the invariant measure of a diffusion process when smoothness is imposed on the drift [23, 86]. 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 [44].

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 $\rho_P$ 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 $\Omega_P$, let $\bar{\Omega}_P$ denote a convex set with Lipschitz boundary such that $\bar{\Omega}_P \subseteq MB_1$, and $\Omega_P + M^{-1}B_1 \subseteq \Omega_P$. Let $\bar{T} = \bar{T}(M)$ be the set of all differentiable functions $T : \bar{\Omega}_P \to \mathbb{R}^d$ such that $T = \nabla f$ for some differentiable convex function $f : \bar{\Omega}_P \to \mathbb{R}^d$ and
(i) $|T(x)| \leq M$ for all $x \in \tilde{\Omega}_P$, 
(ii) $M^{-1} \leq DT(x) \leq M$ for all $x \in \tilde{\Omega}_P$.

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

$$T_0 \in \tilde{T}_\alpha = \tilde{T}_\alpha(M, R) = \{T \in \tilde{T}(M) : \|T\|_{C^3(\tilde{\Omega}_P)} \vee \|T\|_{W^{2,2}(\tilde{\Omega}_P)} \leq R\}.$$ 

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 push-forward 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 $X = X(M)$ be the set of all twice continuously differentiable functions $f : \Omega_P \to \mathbb{R}$ such that

(i) $|f(x)| \leq 2M^2$ and $|\nabla f(x)| \leq M$ for all $x \in \tilde{\Omega}_P$,

(ii) $M^{-1} \leq D^2 f(x) \leq M$ for all $x \in \tilde{\Omega}_P$.

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

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

Further, $Q = \nabla(f_0)_#P$ has a connected and bounded Lipschitz support $\Omega_Q \subseteq MB_1$ and admits a density $\rho_Q$ with respect to the Lebesgue measure that satisfies $M^{-(d+1)} \leq \rho_Q(y) \leq M^{d+1}$ 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 $\|T_0\|_{W^{2,2}(\tilde{\Omega})} \lesssim \|T_0\|_{C^{d+1}(\tilde{\Omega})}$, 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

$$Pf = \int f(x) \, dP(x), \quad Qf = \int f(y) \, dQ(y),$$

$$\hat{P}f = \frac{1}{n} \sum_{i=1}^n f(X_i), \quad \hat{Q}f = \frac{1}{n} \sum_{j=1}^n f(Y_i),$$

where, as in Section 3, $X_{1:n} = (X_1, \ldots, X_n)$ and $Y_{1:n} = (Y_1, \ldots, Y_n)$ 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 $S(f) = Pf + Qf^*$ 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,

$$\hat{S}(f) = \hat{P}f + \hat{Q}f^*.$$
In order to incorporate smoothness regularization into the minimization of (5.1), we consider the restriction of potentials $f$ to a wavelet expansion of finite degree, a strategy that is frequently used in non-parametric estimation \[41, 38\]. 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 \[61\] that we omit for readability. Alternatively, other selection methods such as Lepski’s method \[55, 56, 57\] 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 $f \in X(2M)$.

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

$$F_J = X(2M) \cap V_J(\tilde{\Omega}_P)$$

of candidate potentials, based on which we define the estimators

$$\hat{f}_J \in \arg\min_{f \in F_J} \hat{S}(f), \quad \hat{T}_J = \nabla \hat{f}_J,$$

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 $n \geq n_0$,

$$\mathbb{E}_{(X_1:n, Y_1:n)} \left[ \int \|\hat{T}_J(x) - T_0(x)\|_2^2 \, dP(x) \right] \leq C \left[ n^{-\frac{2\alpha}{2\alpha + 1}} (\log(n))^2 \vee \frac{1}{n} \right],$$

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 $\beta$, $C$ may additionally depend on $\alpha$, and $J$ depends on $n$.

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

**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 (5.3) to \( X(2M) \), 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 \( \alpha \) using a standard penalization scheme, see [61]. We omit this straightforward extension and instead focus on establishing minimax rates of estimation. For a more detailed account, we refer the reader to [42].
(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 (5.13) below and amounts to \( R^{2(d-2)} \log(R) \) 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 \( S \) and the \( L^2 \)-distance of interest.

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

\[
\frac{1}{8M} \| \nabla f(x) - \nabla f_0(x) \|_{L^2(P)}^2 \leq S(f) - S(f_0) \leq 2M \| \nabla f(x) - \nabla f_0(x) \|_{L^2(P)}^2
\]

and

\[
\frac{1}{4M} \| \nabla f^*(y) - \nabla f^*_0(y) \|_{L^2(Q)}^2 \leq S(f) - S(f_0).
\]

**Proof.** It follows from Proposition-Definition 7(ii) and a second-order Taylor expansion that \( f \) is of quadratic type [51] around every \( x \in \Omega_P \):

\[
\frac{1}{2L} \| z - x \|_2^2 \leq f(z) - f(x) - \langle \nabla f(x), z - x \rangle \leq \frac{L}{2} \| z - x \|_2^2,
\]

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

The upper bound in (5.6) is of the form

\[
f(z) \leq q_x(z) = f(x) + \langle \nabla f(x), z - x \rangle + \frac{L}{2} \| z - x \|_2^2 + \epsilon_{\Omega_P}(z).
\]
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 A.8), we have
\[
q^*_x(\nabla f_0(x)) = \frac{1}{2L}\|\nabla f_0(x) - \nabla f(x)\|^2 + \langle x, \nabla f_0(x) \rangle - f(x)
\]
\[
- \frac{L}{2} d^2 \left( \frac{\nabla f_0(x) - \nabla f(x)}{L} - x, \tilde{\Omega}_P \right).
\]
The squared distance term vanishes for $L = 4M$: by the triangle inequality $\|\nabla f_0(x) - \nabla f(x)\|_2 \leq 4M$ 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 $Q = (\nabla f_0)_{\neq P}$, this yields
\[
\mathcal{S}(f) = Pf + Qf^* = \int [f(x) + f^*(\nabla f_0(x))] dP(x)
\]
\[
\geq \frac{1}{8M} \|\nabla f - \nabla f_0\|^2_{L^2(P)} + \int \langle x, \nabla f_0(x) \rangle dP(x).
\]
Moreover, by strong duality, we have
\[
\mathcal{S}(f_0) = Pf_0 + Qf_0^* = \int \langle x, \nabla f_0(x) \rangle dP(x).
\]
The above two displays yield $\mathcal{S}(f) - \mathcal{S}(f_0) \geq (8M)^{-1} \|\nabla f - \nabla f_0\|^2_{L^2(P)}$. In the same way, using the lower bound in (5.6), we get that $\mathcal{S}(f) - \mathcal{S}(f_0) \leq 2M \|\nabla f - \nabla f_0\|^2_{L^2(P)}$, which concludes the proof of (5.4).

It turns out that (5.5) is even easier to prove. Indeed, by Proposition-Definition 7(ii) and Lemma A.4, we get that the upper bound in (5.6) 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 [61, 51, 38]. In this work, we employ van de Geer’s “one-shot” localization technique originally introduced in [92] and stated in a form close to our needs in [93].

5.4. Control of the stochastic error via empirical processes. In light of Proposition 10, the performance of our estimator $\hat{T}_j = \nabla \hat{f}_j$ defined in (5.3) 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
\[
\mathcal{S}_0(f) = \mathcal{S}(f) - \mathcal{S}(f_0) \quad \text{and} \quad \hat{\mathcal{S}}_0(f) = \hat{\mathcal{S}}(f) - \hat{\mathcal{S}}(f_0),
\]
and let $\hat{f}_j \in \mathcal{F}_j$. We observe that by optimality of $\hat{f}_j$ for $\hat{\mathcal{S}}$,
\[
\mathcal{S}_0(\hat{f}_j) - \mathcal{S}_0(\hat{f}_j) \leq [\mathcal{S}_0(\hat{f}_j) - \hat{\mathcal{S}}_0(\hat{f}_j)] + [\hat{\mathcal{S}}_0(\hat{f}_j) - \mathcal{S}_0(\hat{f}_j)].
\]
To proceed, we control the localized empirical process
\[
\sup_{f \in \mathcal{F}_j : \mathcal{S}_0(f) \leq \tau^2} |\mathcal{S}_0(f) - \hat{\mathcal{S}}_0(f)|.
\]
for some fixed $\tau^2 > 0$. More precisely, we prove the following result in Appendix F.2 of the supplement [43].
Proposition 11. Let assumptions B1 – B2 be fulfilled and define $\mathcal{F}_J$ as in (5.2). For any $\tau > 0$, define

$$
\mathcal{F}_J(\tau^2) := \{ f \in \mathcal{F}_J : S_0(f) \leq \tau^2 \}.
$$

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

$$
\sup_{f \in \mathcal{F}_J(\tau^2)} |S_0(f) - \hat{S}_0(f)| \leq C_1 \left( \phi_J(\tau^2) + \tau \sqrt{\frac{t}{n}} + \frac{t}{n} \right),
$$

where

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

Equipped with this result, we can apply van de Geer’s localization technique. To simplify the presentation, assume that $\hat{f} = \hat{f}_J \in \text{argmin}_{f \in \mathcal{F}_J} S(f)$ exists. If not, we may repeat the proof with an $\epsilon$-approximate minimizer and let $\epsilon \to 0$. 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

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

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

$$
S_0(\hat{f}_s) \leq 2M \| \nabla \hat{f}_s - \nabla f_0 \|^2 \leq 4M \left( \sigma^2 + \| \nabla \bar{f} - \nabla f_0 \|^2 \right) =: \tau^2.
$$

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

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

which yields

$$
S_0(\hat{f}_s) \leq S_0(\bar{f}) + 2 \sup_{f \in \mathcal{F}_J(\tau^2)} |S_0(f) - \hat{S}_0(f)|.
$$

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

$$
\| \nabla \hat{f}_s - \nabla \bar{f} \|^2 \leq 2 \| \nabla \hat{f}_s - \nabla f_0 \|^2 + 2 \| \nabla f_0 - \nabla \bar{f} \|^2 \leq 16MS_0(\bar{f}_s) + 16MS_0(\bar{f})
$$

$$
\leq 32M S_0(\bar{f}) + 32M \sup_{f \in \mathcal{F}_J(\tau^2)} |S_0(f) - \hat{S}_0(f)|.
$$

Next, note for $s$ as in (5.7), we have that $\| \nabla \hat{f}_s - \nabla \bar{f} \| \geq \sigma/2$ iff $\| \nabla \hat{f} - \nabla \bar{f} \| \geq \sigma$. Hence

$$
\mathbb{P} \left( \| \nabla \hat{f} - \nabla f_0 \| \geq \sigma + \| \nabla \bar{f} - \nabla f_0 \| \right) \leq \mathbb{P} \left( \| \nabla \hat{f}_s - \nabla \bar{f} \|^2 \geq \sigma^2/4 \right)
$$

$$
\leq \mathbb{P} \left( \sup_{f \in \mathcal{F}_J(\tau^2)} |S_0(f) - \hat{S}_0(f)| \geq \frac{\sigma^2}{128M} - S_0(\bar{f}) \right)
$$

$$
= \mathbb{P} \left( \sup_{f \in \mathcal{F}_J(\tau^2)} |S_0(f) - \hat{S}_0(f)| \geq \frac{\tau^2}{512M^2} - \frac{1}{128M} \| \nabla \hat{f} - \nabla f_0 \|^2 - S_0(\hat{f}) \right).
$$
Recalling Proposition 11, we take \( \sigma \) such that
\[
\frac{\tau^2}{512 M^2} \geq S_0(\tilde{f}) + \frac{1}{128M} \| \nabla \tilde{f} - \nabla f_0 \|^2 + C_1 \left( \phi_J(\tau^2) + \tau \sqrt{\frac{Mt}{n} + \frac{M^2 t}{n}} \right),
\]
so that we get
\[
P\left( \| \nabla \hat{f} - \nabla f_0 \| \geq \sigma + \| \nabla \bar{f} - \nabla f_0 \| \right) \leq e^{-t}.
\]
In particular, we can check that (5.8) is fulfilled if we choose \( \sigma \) such that
\[
\sigma^2 \gtrsim S_0(\bar{f}) + 2 \frac{J(2d-2)}{n} \log (1 + C_2 n) + \frac{1 + t}{n},
\]
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
\[
\| \nabla \hat{f} - \nabla f_0 \| \lesssim \| \nabla \bar{f} - \nabla f_0 \| + \frac{2 J(2d-2)}{n} \log (1 + C_2 n) + \frac{1 + t}{n}.
\]
Moreover, integrating the tail with respect to \( t \) readily yields by Fubini’s theorem that
\[
\mathbb{E} \| \nabla \hat{f} - \nabla f_0 \| \lesssim \| \nabla \bar{f} - \nabla f_0 \| + \frac{2 J(2d-2)}{n} \log (1 + C_2 n) + \frac{1}{n}.
\]

We have proved the following result.

**Proposition 12.** Let \( B_1 - B_2 \) hold and define \( \mathcal{F}_J \) as in (5.2). Then, writing
\[
E_J := \frac{2 J(2d-2)}{n} \log (1 + C_2 n) + \frac{1}{n},
\]
the estimator \( \hat{T}_J \) defined in (5.3) satisfies
\[
\mathbb{E} \| \hat{T}_J - T_0 \|_{L^2(P)}^2 \lesssim \inf_{f \in \mathcal{F}_J} \| \nabla f - T_0 \|_{L^2(P)}^2 + E_J.
\]
Moreover, with probability at least \( 1 - \exp(-t) \),
\[
\| \hat{T}_J - T_0 \|_{L^2(P)}^2 \lesssim \inf_{f \in \mathcal{F}_J} \| \nabla f - T_0 \|_{L^2(P)}^2 + E_J + \frac{t}{n}.
\]

### 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 \( \bar{f} = \Pi_J \text{ext } f_0 \) 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 B.2, we may assume that \( \text{ext } \bar{f} \) has the same \( C^\beta \) 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 \( \Pi_J \) its projection onto the first scale \( J \) wavelet coefficients. That is, if
\[
f = \sum_{j=0}^{\infty} \sum_{g \in G^j} \sum_{k \in \mathbb{Z}^d} \gamma_k^j \varphi_k^j g,
\]
we set \( \Pi_J f = \sum_{j=0}^{J} \sum_{g \in G^j} \sum_{k \in \mathbb{Z}^d} \gamma_k^j \varphi_k^j g \).
where $\Psi_k^{j,g}$ are multi-dimensional Daubechies wavelets and $G^j$ the associated index sets as in Section B. Then, for all $1 \leq p, q \leq \infty, s \geq 0$,

\begin{equation}
\|\Pi_j f\|_{B_{p,q}(\mathbb{R}^d)} \leq \|f\|_{B_{p,q}}.
\end{equation}

(5.10)

\begin{equation}
\|\Pi_j f - f\|_{B_{p,q}(\mathbb{R}^d)} \leq \|f\|_{B_{p,q}}.
\end{equation}

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

$$\|\Pi_j f - f\|_{B_{p',q'}^{j'}} \lesssim 2^{-J(s-d/p-(s'-d/p'))}\|f\|_{B_{p,q}}.$$ \hspace{1cm} \text{In particular: If } f \in C^{\alpha+1} \text{ for } \alpha > 1, \text{ then } \|f - \Pi_j f\|_{C^2} \lesssim 2^{-J(\alpha-1)}\|f\|_{C^{\alpha+1}} \text{ and if } f \in W^{\alpha+1,2}, \text{ for } \alpha > 0, \text{ then } \|f - \Pi_j f\|_{W^{1,2}} \lesssim 2^{-J\alpha}\|f\|_{W^{\alpha+1,2}}.$$

\text{PROOF.} \text{ Write } \gamma \text{ for the wavelet coefficients of } f. \text{ The statements (5.10) and (5.11) follow immediately from the wavelet characterization of the Besov norms, (B.2).}

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

$$2^{j(s'+\frac{d}{2}-\frac{d}{p})} \left( \sum_{g \in G^j} \left( \sum_{k \in \mathbb{Z}^d} |\gamma_k^{j,g}(p')|^{q'/p'} \right)^{1/q'} \right) \leq 2^{j(s'-\frac{d}{p'}-(s-\frac{d}{p}))} \sum_{g \in G^j} \left( \sum_{k \in \mathbb{Z}^d} |\gamma_k^{j,g}(p')|^{q'/p'} \right)^{1/q'} \leq 2^{j(s'-\frac{d}{p'}-(s-\frac{d}{p}))} \|f\|_{B_{p,q}}.$$ \hspace{1cm} \text{Then, plugging this into the wavelet expansion of } \Pi_j f - f, \text{ we obtain}

$$\|\Pi_j f - f\|_{B_{p',q'}} \lesssim \sum_{j=J+1}^\infty 2^{j\varepsilon'(s+\frac{d}{2}-\frac{d}{p})} \sum_{g \in G^j} \left( \sum_{k \in \mathbb{Z}^d} |\gamma_k^{j,g}(p')|^{q'/p'} \right)^{1/q'} \lesssim \sum_{j=J+1}^\infty 2^{j\varepsilon'(s+\frac{d}{2}-\frac{d}{p})} \|f\|_{B_{p,q}}^{q'} \lesssim 2^{j\varepsilon'(s'-\frac{d}{p'}-(s-\frac{d}{p}))} \|f\|_{B_{p,q}}^{q'}.$$ 

Finally, to obtain the special cases, note that $\|\cdot\|_{B_{q,\infty}} \lesssim \|\cdot\|_{C^2} \lesssim \|\cdot\|_{B_{1,\infty}^2}$ and $\|\cdot\|_{W^{s,2}} = \|\cdot\|_{B_{2,2}}$ by Theorem B.1. \hspace{1cm} \Box

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

$$\lambda_{\min}(D^2 \Pi_j \text{ ext } f_0(x)) \geq \lambda_{\min}(D^2 f_0(x)) - \|D^2 \Pi_j \text{ ext } f_0(x) - D^2 f_0(x)\|_{op}$$

$$\geq M^{-1} - C\|\Pi_j \text{ ext } f_0 - f_0\|_{C^2(\mathcal{O}_p)}.$$ \hspace{1cm} \text{It follows from Lemma 13 that } \|\Pi_j \text{ ext } f_0 - f_0\|_{C^2(\mathcal{O}_p)} \lesssim 2^{-(\beta-1)J}(M^3 + R) \leq 1/(2M), \text{ if}

$$J \geq J_0 := C_3 \frac{1}{\beta - 1} \log (2M^4 + 2RM),$$
and \( C_2 = C_3(d, \beta, \tilde{Q}_P) \) is large enough. This yields \( \lambda_{\min}(D^2 \Pi_f \text{ ext } f_0(x)) \geq 1/(2M) \) and \( \bar{f} \) is strongly convex. Similarly, we get \( \lambda_{\max}(D^2 \Pi_f \text{ ext } f_0(x)) \leq 2M \) and hence that \( \bar{f} \in \mathcal{F}_J \).

Thus,

\[
\inf_{\hat{f} \in \mathcal{F}_J} \| \nabla \hat{f} - \nabla f_0 \|_{L^2(P)}^2 \leq \| \nabla \bar{f} - \nabla f_0 \|_{L^2(P)}^2 \lesssim \int_{\Omega_P} \| \nabla \bar{f} - \nabla f_0 \|_2^2 \, d\lambda(x) \leq \| \bar{f} - f_0 \|_{W^{1,2}(\Omega_P)}^2 \lesssim R^2 2^{-2J_0}.
\]

where we used Assumption B2 and Lemma 13.

We have thus proved that

\[
\inf_{\hat{f} \in \mathcal{F}_J} \| \nabla \hat{f} - \nabla f_0 \|_{L^2(P)}^2 \lesssim R^2 2^{-2J_0} \quad \text{for } J \geq J_0.
\]

### 5.6. Bias-variance tradeoff.

We are now in a position to complete the proof of Theorem 8. Combining the bounds (5.9) and (5.12), we get

\[
\mathbb{E} \| \hat{T}_J - T_0 \|_{L^2(P)}^2 \lesssim R^2 2^{-2J_0} + 2^{J(d-1)} \frac{\log(n)}{n} + \frac{1}{n}.
\]

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

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

We note that since \( \alpha > 1 \), in the first two cases, \( d \in \{1, 2\} \), the cut-off \( J \) can be picked independently from \( \alpha \). 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, \( \hat{T}_{\text{wav}} \) below, is an approximation to the estimator \( \hat{T}_J \) in (5.3), illustrating that (5.3) can be implemented in low dimensions and that this approximation achieves favorable rates in \( d = 3 \). The second, \( \hat{T}_{\text{ker}} \) 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 \( \Gamma \in \mathbb{R}^{n \times n} \)

\[
\hat{\Gamma} = \arg\min \left\{ \sum_{i,j=1}^n \| X_i - Y_j \|_2^2 \Gamma_{i,j} : \begin{array}{c} \Gamma_{i,j} \geq 0 \forall i, j \in [n], \\ \Gamma 1 = 1/n, \Gamma^\top 1 = 1/n \end{array} \right\}
\]
To obtain values of these approximations at non-grid points, we appeal to linear interpolation. An estimated transport function on the observations \( X_i \) is then obtained by

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

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 \( \hat{T}_{\text{emp}}(X_i) = Y_{\pi(i)} \).

Because \( \hat{T}_{\text{emp}}(X_i) \) 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:

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

6.1.2. Wavelet estimator. Next, we turn to an approximation of (5.3). Assume that in addition to a superset \( \hat{\Omega}_P \supseteq \Omega_P \), we are also given a superset \( \hat{\Omega}_Q \supseteq \Omega_Q \), and that both \( \hat{\Omega}_P \) and \( \hat{\Omega}_Q \) are boxes (hypercubes). We consider all functions originally defined over \( \hat{\Omega}_P \) and \( \hat{\Omega}_Q \) as given by their samples on grids with resolution \( N \in \mathbb{N} \), \( x = (x_i)_{i \in [N]^d} \in (\mathbb{R}^d)^{N^d} \) and \( y = (y_i)_{i \in [N]^d} \in (\mathbb{R}^d)^{N^d} \), respectively. In particular, we write \( f = (f_i)_{i \in [N]^d} \in \mathbb{R}^{N^d} \) for the discretization of the potential \( f \) and \( T = (T_i)_{i \in [N]^d} \in (\mathbb{R}^d)^{[N]^d} \) for the discretization of the transport map \( T \) on the grid \( x \). 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 \( f \) by the inverse discrete wavelet transform up to order \( J \), which we write as \( f = W_j^T \gamma_j \) 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,

\[
\mathcal{L}(f)_j := \mathcal{L}(f)(y_j) := \mathcal{L}_{x \to y}(f)(y_j) \quad := \sup_{i \in [N]^d} \{ (x_i, y_i) - f_i : i \in [N]^d \},
\]

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

(iii) The gradient of \( f \) on grid-points can be calculated by a finite-difference scheme, which we write as \( \nabla_x f \).

(iv) To obtain values of these approximations at non-grid points, we appeal to linear interpolation, which we write as \( \mathcal{P}_{x \to \{X_1, \ldots, X_n\}} f \) for the interpolated value of \( f \) defined on the grid \( x \) at the point \( \{X_1, \ldots, X_n\} \in \mathbb{R}^d \), collected in one vector.

Given i.i.d. observations \( X_{1:n} = \{X_1, \ldots, X_n\} \) and \( Y_{1:n} = \{Y_1, \ldots, Y_n\} \) from \( P \) and \( Q \), respectively, we arrive at the following optimization problem as approximation to (5.3):

\[
\hat{\gamma}_j = \min_{\gamma_j \in \mathbb{R}^{m_J}} \frac{1}{n} \mathcal{P}_{x \to X_{1:n}} W_j^T \hat{\gamma}_j + \frac{1}{n} \mathcal{P}_{y \to Y_{1:n}} \mathcal{L}(W_j^T \hat{\gamma}_j).
\]

Note that we dropped all boundedness and convexity constraints that were given by \( X(2M) \) in (5.3). In practice, this can lead to degraded estimation quality of the gradient of \( W_j^T \gamma_j \).
near the boundary of $\Omega_P$, 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

\[
\hat{T}_J = \nabla_x [\mathcal{L}_{y \rightarrow x}(\mathcal{L}_{x \rightarrow y}(W_J^T \hat{\gamma}_J))].
\]

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

\[
\hat{T}_{\text{wav}}^{(J)}(x) = P_{x \rightarrow y} T_J, \quad x \in \Omega_P.
\]

For the purpose of these experiments, we select the wavelet scale $J$ 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

\[
\hat{T}_{\text{wav}} = \hat{T}_{\text{wav}}^{(J)}.
\]

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

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

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 [59] and the Linear-Time Legendre Transform algorithm [58]. 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 $\gamma_J$ due to the form of the discrete Legendre transform. In practice, we observe that it converges faster than simple (sub-)gradient descent methods.

Since $\hat{T}_{\text{wav}}$ 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 $\hat{T}_{\text{wav}}$ 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 $\hat{T}_{\text{wav}}$ 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 $\hat{T}_{\text{emp}}$. The idea of smoothing the empirical transport matrix has been previously used in practice, see for example [77], where regularized linear regression was used as a post-processing step, and kernels have been applied for smoothing potential functions in [35]. Here, we obtain a smoother version of $\hat{T}_{\text{emp}}$ by smoothing it via kernel-ridge regression [66].

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,

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

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

\[
\hat{T}_{\text{ker}}^{(\nu_{\text{ridge}}, \nu_{\text{kernel}})} = \arg\min_{T \in \mathcal{H}} \sum_{i=1}^n \left\| \hat{Y}_i - T(X_i) \right\|_2^2 + \nu_{\text{ridge}} \|T\|_{\mathcal{H}}^2,
\]

\[
\text{MSE}_n(\hat{T}_{\text{ker}}^{(\nu_{\text{ridge}}, \nu_{\text{kernel}})}) = \frac{1}{n} \sum_{i=1}^n \left\| \hat{Y}_i - \hat{T}_{\text{ker}}^{(\nu_{\text{ridge}}, \nu_{\text{kernel}})}(X_i) \right\|^2_2.
\]
for $\nu_{\text{ridge}} > 0$. By the representer theorem [66], (6.3) has a solution

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

where

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

We measure its performance by

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

Similar to $\hat{T}_{\text{wav}}$, we select the tuning parameters $\nu_{\text{kernel}}$ and $\nu_{\text{ridge}}$ 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 $\hat{T}_{\text{ker}}$.

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

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

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

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

where for (id), $P^{(1)} = Q^{(1)} = \text{Unif}([0, 1]^d)$. For (exp), $P^{(2)} = \text{Unif}([0, 1]^d)$, and the target measure is defined as $Q^{(2)} = (T_0^{(2)})_# P^{(2)}$. Note that these potentials and transport maps are $C^\infty$ and strongly 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 [3]. 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 $T_{\text{emp}}$ and $T_{\text{ker}}$ to sample sizes $\leq 10^4$. Likewise, we do not compute $T_{\text{wav}}$ 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 $T_{\text{emp}}$, we visualize the first coordinate of the transport map estimators for case (id) with $d = 2$ and $n = 100$ observations from $P = Q = \text{Unif}([0, 1]^2)$ in Figure 1, together with the ground truth transport map $(T_0^{(1)})_1$. To depict the coupling $T_{\text{emp}}$ 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 $T_{\text{emp}}$ in Figure 1b. The kernel estimator in Figure 1d is even smoother and visually very similar to the ground truth in 1a, due to the possibility of employing a large amount of regularization.

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 $T_{\text{emp}}$, $T_{\text{wav}}$ (only $d = 3$), and $T_{\text{ker}}$ 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 $T_{\text{emp}}$ roughly follow a $n^{-2/3}$ rate. This corresponds to the decay of the average $\ell^2$ cost of optimal matchings between samples from the uniform distribution on the cube [87, 100] (and Gaussian distributions [54]), and also matches the $n^{-2/d}$ rate for the convergence of $W_2^2(P, \hat{P})$. We are, however, not aware of a generalization of this rate to the error measure $\text{MSE}$ in the case of a transport map different from the identity.

The error curves for the wavelet estimator $T_{\text{wav}}$ 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 $T_{\text{wav}}$ all lie below
Figure 2: Log-log plot of MSE plotted against $n$, showing the median error over 32 replicates. See Section 6.2 for details.

The baseline estimator. The kernel estimator $\hat{T}_{ker}$ performs even better, attaining rates close to $n^{-1}$ and yielding consistently better rates than $\hat{T}_{wav}$.

We observe that the favorable behavior of $\hat{T}_{wav}$ suggests that the restriction of candidate potentials to $\mathcal{X}(2M)$ in (5.3) might not be necessary and could possibly be omitted.

In 10D, for both test cases, $\hat{T}_{emp}$ 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 supplement, illustrating the sample size-dependent performance gain achieved by $\hat{T}_{wav}$ and $\hat{T}_{ker}$.

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 $\hat{T}_{ker}$ 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 setting.
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.

SUPPLEMENTARY MATERIAL

Supplement to: Minimax rates of estimation for smooth optimal transport maps (; .pdf). The supplementary materials contain more background on convex functions, wavelets, and empirical processes, as well as tools to prove lower bounds, alternative assumptions based on Caffarelli’s regularity theory, and proofs, lemmas, and details on the numerical experiments omitted in the main text.

REFERENCES


