MINIMAX ESTIMATION OF SMOOTH TRANSPORT MAPS WITH KERNEL SOS*

ADRIEN VACHER †, BORIS MUZELLEC‡, FRANCIS BACH ‡, FRANÇOIS-XAVIER VIALARD †, AND ALESSANDRO RUDI ‡

Abstract. Under smoothness conditions, it was recently shown by Vacher et. al. (2021) that the squared Wasserstein distance between two distributions could be approximately computed in polynomial time with appealing statistical error bounds. In this paper, we propose to extend their result to the problem of estimating in L^2 distance the transport map between two distributions. Also building upon the kernelized Sum-Of-Squares approach, a way to model smooth positive functions, we derive a computationally tractable estimator of the transport map. Contrary to the aforementioned work, the dual problem that we solve is closer to the so-called *semi-dual* formulation of optimal transport that is known to gain convexity with respect to the linear dual formulation. After deriving a new stability result on the semi-dual and using localization-like techniques through Gagliardo-Nirenberg inequalities, we manage to prove under the same assumptions as in Vacher et. al. (2021) that our estimator is minimax optimal up to polylog factors. Then we prove that this estimator can be computed in worst case in $\tilde{O}(n^5)$ time where n is the number of samples and show how to improve its practical computation with a Nyström approximation scheme, a classical tool in kernel methods. Finally, we showcase several numerical simulations in medium dimension where we compute our estimator on simple examples.

Key words. Optimal transport, sum of squares, kernel methods

MSC codes. 62G05, 49Q22

1. Introduction.

Optimal transport (OT) provides a principled method to compare probability distributions by finding the optimal way of coupling one to another based on a cost function defined on their supports. Formally, given two probability distributions μ and ν supported over metric spaces X and Y and a cost $c : X \times Y \to \mathbb{R}$, the OT value between μ and ν was defined by Monge as follows

$$\mathrm{OT}(\mu,\nu) = \inf_{T_{\#}(\mu)=\nu} \int c(x,T(x))d\mu(x)\,, \tag{1.1}$$

where the infimum is taken over maps T that pushforward μ onto ν, that is, for all Borel set A of Y, $\mu(T^{-1}(A)) = \nu(A)$. This non-convex problem was later relaxed by Kantorovitch into the following linear program whose primal formulation reads

$$\inf_{\pi\in\mathcal{M}_+(X\times Y)} \langle \pi, c\rangle + \iota(\pi_1 = \mu) + \iota(\pi_2 = \nu)\,, \tag{1.2}$$

where $\langle\cdot,\cdot\rangle$ is the duality pairing between measures and functions, $\iota(\cdot)$ is the convex indicator function, $\mathcal{M}_+(X\times Y)$ is the set of positive Radon measures over $X\times Y$ and π_i is the i-th marginal of π. The dual formulation of this problem reads

$$\sup_{\phi,\psi} \langle\phi,\mu\rangle + \langle\psi,\nu\rangle + \iota(\phi\oplus\psi \le c)\,, \tag{1.3}$$

* Submitted to the editors DATE.

Funding: This work was funded in part by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). We also acknowledge support from the European Research Council (grants SEQUOIA 724063 and REAL 947908), and Région Ile-de-France.

†LIGM, CNRS (adrien.vacher@u-pem.fr, francois-xavier.vialard@u-pem.fr).

‡INRIA Paris and DI ENS (boris.muzellec@inria.fr, francis.bach@inria.fr, alessandro.rudi@inria.fr).

where ϕ, ψ are continuous real valued functions over X (resp. Y), that we shall refer to as *potentials* in the rest of the paper, and where $\phi \oplus \psi$ is defined over $X \times Y$ as $\phi \oplus \psi : (x, y) \mapsto \phi(x) + \psi(y)$. In the euclidean setting $X, Y \subset \mathbb{R}^d$ and for the quadratic cost $c(x, y) = \frac{\|x-y\|^2}{2}$, Brenier [6] showed that under regularity assumptions on μ, ν, the Monge problem (1.1) and the dual Kantorovitch problem (1.3) are equal and are linked as follows: given ϕ_0 the first potential solution of (1.3), the optimal transport map T_0 is given by $T_0(x) = x - \nabla\phi_0(x)$.

These optimal transportation maps are playing an increasingly important role in data sciences. Indeed, many applications such as generative modeling [2, 32, 4, 23, 28], domain adaptation [12, 11], shape matching [35, 17], dimensionality reduction [19] or predicting cell trajectories [33, 44] can be formulated as the problem of finding a map from a reference distribution to a target distribution. Yet in the aforementioned applications, the measures μ, ν are usually supported on high dimensional spaces and they are usually only available via their n-samples empirical counterparts $\hat{\mu}_n, \hat{\nu}_n$. Hence one must design an estimator $\hat{T}_n$ of the transport map T_0 that is both robust in high-dimension and that can be computed numerically for reasonable values of n. Defining the error as $e(T) = \int_x \|T - T_0\|^2 d\mu(x)$, this problem can be informally stated as follows:

Can we design an estimator $\hat{T}_n(\hat{\mu}_n, \hat{\nu}_n)$ that can be computed in dimension-free polynomial time and such that $e(\hat{T}_n)$ scales "well" with the dimension d?

1.1. Related works. Over the past few years, numerous works have focused instead on the statistical approximation of the scalar quantity $\mathrm{OT}(\mu, \nu)$ for the quadratic cost, also known as the quadratic Wasserstein distance $W_2^2(\mu, \nu)$. An intuitive choice of estimator is the *plugin* estimator $W_2^2(\hat{\mu}_n, \hat{\nu}_n)$ which is simply the squared Wasserstein distance computed on the empirical counterparts of μ, ν. However, even though it can be computed in $O(n^3 \log(n))$ time using the network simplex algorithm [1], it suffers from the *curse of dimension* as it was shown that $\mathbb{E}[|W_2^2(\mu, \nu) - W_2^2(\hat{\mu}_n, \hat{\nu}_n)|] \sim n^{-2/d}$ [16]. Another popular choice of estimator is the Sinkhorn model [13] that regularizes the OT problem with the addition of an entropic penalty in the primal problem (1.2). Even though the computational complexity of the plugin Sinkhorn model is lowered to $O(n^2)$, the error still poorly scales in $O(n^{-2/d})$ [10]. When the measures μ, ν are assumed to have smooth densities, the statistical error can be improved as the smoothness grows [21]. For instance, when the densities of μ, ν are assumed to be m-smooth, the authors of [40] designed an estimator $\hat{w}_n$ that yields an average error $\mathbb{E}[|\hat{w}_n - W_2(\mu, \nu)|]$ scaling as $n^{-\frac{1+m}{d+2m}}$; in particular, the parametric rate $1/\sqrt{n}$ is recovered in the very smooth regime $m \to \infty$. However, their estimator $\hat{w}_n$ requires $O(n^{\frac{(1+m)(2d+2)}{d+2m}})$ time to be computed. It was only recently that this computational/statistical gap was closed by [37]. Relying of the kernelized Sum-of-Squares (SoS) tool [25] (see Section 2 for a detailed background on SoS), the authors proposed an estimator $\hat{w}_n^2$ that can be computed in $O(n^{3.5})$ time and that achieved an average error $\mathbb{E}[|\hat{w}_n^2 - W_2^2(\mu, \nu)|]$ scaling in $n^{-\min(\frac{m-d-1}{2d}, 1/2)}$ when the densities are assumed to be m-smooth.

The question on how to estimate the transport map T_0 came afterwards. Again, when no smoothness assumption is made on the transport map T_0, this tasks also suffers the curse of dimension. For instance, the authors of [24] designed a so-called plugin estimator $\hat{T}_n$ of the transport map, computed via the coupling $\hat{\pi}_n$ solution of the empirical version of the primal problem (1.2), such that the error $\|\hat{T}_n - T_0\|_{L^2(\mu)}^2 := e(\hat{T}_n)$ scaled as $n^{-2/d}$. Similarly, it was shown that the maps $\hat{T}_n$ obtained with Sinkorn

model achieved an error $e(\hat{T}_n) \sim n^{-1/d}$ [29] ; note that both these estimators can be computed in either $O(n^3)$ or $O(n^2)$ time yet, as these rates show, these models do suffer the curse of dimension. At the opposite of the spectrum, it was shown in [21] if T_0 is assumed to be C^α, one could design an estimator $\hat{T}_n$ that achieved an error $e(\hat{T}_n) \sim n^{-\frac{\alpha}{\alpha+d/2-1}}$ hence recovering a $1/n$ error when smoothness grows and that this rate was actually statistically minimax ; yet this time the estimator cannot be computed numerically as it involves solving an infinite dimensional optimization program.

1.2. Contributions. Following a similar technique as in [37], we close the statistical/computational gap for the problem of optimal transport map estimation when the optimal transport map T_0 is assumed to be C^α with $\alpha > d+2$. Our main result can be summarized in the following informal theorem:

THEOREM 1.1. *If μ, ν have densities bounded from above and below on compact, convex domains and T_0 is C^α with $\alpha > d+2$ and such that $Jac(T_0)$ is non-singular on the support of μ then, given the empirical distributions $\hat{\mu}_n, \hat{\nu}_n$ there exists an estimator $\hat{T}_n$ of the transport map T_0 that can be computed in $\tilde{O}(n^5)$ time and that verifies for n sufficiently large*

$$\mathbb{E}[\|\hat{T}_n - T_0\|^2_{L^2(\mu)}] \lesssim n^{-\min(\frac{\alpha-1-d}{2d}, \frac{\alpha}{\alpha+d/2-1}) \times \min(1, \frac{2}{1+3d/(2\alpha)})}, \tag{1.4}$$

where $\tilde{O}$ and $\lesssim$ hide constants and poly-log factors.

In particular, not only we provide a statistically efficient estimator that can be computed in polynomial time, but also, when α is sufficiently large, our estimator is minimax up to poly-log factors. We highlight the fact that even if we provide statistical and computational guarantees for our estimator, this work remains mainly theoretical as our upper-bounds involve constants that are potentially exponential with the dimension. Combined with the fact that our computational complexity scales in $\tilde{O}(n^5)$, we believe that our estimator is not suited for modern big-data applications where n is typically of order $\sim 10^6$. Nevertheless, we showcase at the end of the paper a Nyström approximation of our estimator in medium dimension with gaussian densities where a small amount of samples are drawn $n \sim 10^3$.

1.3. Outline of the paper. In Section 2, we begin by explaining the estimation strategy developed in [37] and after an introduction to the theoretical background on the kernel Sum-Of-Squares method [25], we show how it allows to model smooth, positive functions in a computationally tractable manner [31]. Finally, in order to better approximate the so called *semi-dual* formulation of OT, we propose an estimator that differs from the one introduced in [37]. In Section 3, we prove that our estimator does achieve the minimax rate $n^{-\frac{\alpha}{\alpha+1+d/2}}$ by deriving a new stability result on the *semi-dual* formulation of optimal transport and by applying localization arguments through Gagliardo-Nirenberg inequalities [7]. In Section 4, we prove that our estimator $\hat{T}_n$ can be computed up to error τ in $O(n^5 \log(1/\tau))$ time. Finally, in Section 5, we show how to improve the practicality of our estimator by incorporating a Nyström sampling strategy, for which we do not provide guarantees. Using this heuristic method, we compute our improved map estimator in the case where μ, ν are gaussian distributions in medium dimensions, *e.g.* $d = 2, 4, 8$.

2. Preliminaries and background. Throughout the rest of the paper, we shall assume that the measures μ, ν are supported on $X, Y \subset \mathbb{R}^d$ and shall add additional assumptions on μ, ν throughout the section.

2.1. Estimation of the OT cost (Vacher et. al. 2021). The starting point of the estimation strategy employed in [37] relies on the empirical dual formulation of Optimal Transport with a quadratic cost. Given the n-samples empirical distributions $\hat{\mu}_n, \hat{\nu}_n$, recall that the empirical OT cost is given by

$$\begin{aligned} \widehat{\mathrm{OT}} \;&=\; \sup_{(\phi,\psi)\in C(X)\times C(Y)} \langle \phi, \hat{\mu}_n\rangle + \langle \psi, \hat{\nu}_n\rangle \\ &\quad s.t.\;\; \phi(x)+\psi(y) \le \frac{\|x-y\|^2}{2},\; \forall (x,y)\in X\times Y\,, \end{aligned} \tag{2.1}$$

where $C(X)$ (resp. $C(Y)$) is the set of continuous functions over X (resp. Y). In order to recover estimation rates that scale well with the dimension, we need to restrict the search space on the potentials to less complex spaces than $C(X)$ and $C(Y)$ respectively. One straightforward way to reduce the complexity is to assume that (ϕ_0, ψ_0), the solutions of original dual OT problem (1.3), are both smooth. Using the regularity theory of optimal transport, the smoothness of the potentials can be inherited from the smoothness of the measures μ, ν.

THEOREM 2.1 (De Philippis and Figalli in [14]). *If μ and ν have m-times differentiable densities bounded from above and below over X (resp. Y), bounded convex domain of $\mathbb{R}^d$, then the solutions (ϕ_0, ψ_0) of problem (1.3) are $(m+2)$-times differentiable over X (resp. Y).*

Yet, instead of making a smoothness assumption on the density of the measures, we shall make the less restrictive assumption of the transport potential themselves being smooth. Denoting $H_s(\Omega)$ the 2-Sobolev space of order s, for Ω a Lipschitz bounded domain of $\mathbb{R}^d$ and s a positive real number [7], we shall assume that (ϕ_0, ψ_0), the solutions of (1.3), belong to $H_s(X)$ and $H_s(Y)$ respectively. Under this assumption, the new empirical dual problem becomes

$$\begin{aligned} \widehat{\mathrm{OT}} \;&=\; \sup_{(\phi,\psi)\in H_s(X)\times H_s(Y)} \langle \phi, \hat{\mu}_n\rangle + \langle \psi, \hat{\nu}_n\rangle \\ &\quad s.t.\;\; \phi(x)+\psi(y) \le \frac{\|x-y\|^2}{2},\; \forall (x,y)\in X\times Y\,. \end{aligned} \tag{2.2}$$

Statistically speaking, the convergence of the empirical problem (2.2) toward (1.3) is faster as s grows than the convergence of (2.1) toward (1.3). Furthermore, when $s > d/2$, the space $H_s(\Omega)$, becomes a Reproducing Kernel Hilbert Space (RHKS) [3] which is a Hilbert space with appealing computational properties.

DEFINITION 2.2. *A Hilbert space H of real valued functions defined over some space Z endowed with a scalar product $\langle\cdot,\cdot\rangle_H$ is said to be an RKHS when for all $z\in Z$, the evaluation $\delta_z : f\in H \mapsto f(z)$ is continuous for the $\|\cdot\|_H$-norm. In this case, it has a unique kernel $k : Z\times Z \mapsto \mathbb{R}$ such that for all $z\in Z$, $k(z,\cdot)\in H$ and for all $f\in H$, $f(z)$ can be expressed as $f(z) = \langle f, k(z,\cdot)\rangle_H$.*

When the kernel $k(\cdot,\cdot)$ is available in closed form, its associated RKHS has an appealing computational property known as the "kernel" trick. For any Lipschitz loss $L : H \to \mathbb{R}$ depending on data points $(x_1, \cdots, x_n)$, the solution of the potentially infinite dimensional problem

$$\inf_{f\in H}\; L(f; (x_1, \cdots, x_n))$$

is of the form $f = \sum_{i=1}^n \alpha_i k(x_i, \cdot)$, see for instance [34]. Hence, if we assume that the transport potentials are in $H_s(X)$ and $H_s(Y)$ with $s > d/2$, the kernel trick does

apply and problem (2.1) becomes

$$\begin{aligned}\widehat{\mathrm{OT}} \;=\; & \sup_{(\beta,\omega)\in\mathbb{R}^n\times\mathbb{R}^n} \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n \beta_j k_X(x_i,x_j) + \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n \omega_j k_Y(y_i,y_j) \\ & s.t.\;\; \sum_{j=1}^n \beta_j k_X(x_j,x) + \sum_{j=1}^n \omega_j k_Y(y_j,y) \le \frac{\|x-y\|^2}{2},\;\forall (x,y)\in X\times Y\,,\end{aligned} \tag{2.3}$$

where we denoted $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ (resp. $\hat{\nu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$) and where k_X is the kernel of $H_s(X)$ and k_Y is the kernel of $H_s(Y)$. Yet, even though k_X and k_Y are available in closed form as shown in [41], problem (2.3) remains intractable as the constraint $\sum_{j=1}^n \beta_j k_X(x_j,x) + \sum_{j=1}^n \omega_j k_Y(y_j,y) \le \frac{\|x-y\|^2}{2}$ must be enforced overall the whole, potentially continuous, space $X\times Y$.

A naive manner to overcome this problem is to simply discretize this inequality constraint over the support of the empirical measures $\hat{\mu}_n, \hat{\nu}_n$ and implement instead

$$\sum_{j=1}^n \beta_j k_X(x_j,x_l) + \sum_{j=1}^n \omega_j k_Y(y_j,y_l) \le \frac{\|x_l-y_l\|^2}{2},\;\forall l\in\{1,\cdots,n\}\,.$$

However, the authors of [37] showed that when one discretizes an inequality constraint, the error $|\widehat{\mathrm{OT}} - \mathrm{OT}|$ scales as $n^{-1/d}$ with high probability no matter how smooth the potentials (ϕ_0,ϕ_0) are assumed to be. To overcome this issue, they use a particular structure that they proved for the optimal potentials: the inequality constraint in OT (at the optimizers) can be reformulated as an equality constraint with a smooth positive function $\gamma : X\times Y\to\mathbb{R}$ and proved that γ could be expressed as a finite Sum-Of-Squares.

THEOREM 2.3 (Vacher et. al. 2021). *Under the assumption $(\phi_0,\psi_0)\in H_{\alpha+1}(X)\times H_{\alpha+1}(Y)$, if $\alpha > d+2$ there exist functions $(w_i)_{i=1}^d$ such that $w_i\in H_{\alpha-1}(X\times Y)$ and that verify*

$$\frac{\|x-y\|^2}{2} - \phi_0(x) - \psi_0(y) = \sum_{i=1}^d w_i(x,y)^2\,, \tag{2.4}$$

where (ψ_0,ϕ_0) are the optimal transport potentials, solutions of (1.3).

Using this result, they proposed to discretize instead the equality constrained version of the OT problem and used the following estimator

$$\begin{aligned}\widehat{\mathrm{OT}}_{SoS} \;=\; & \sup_{\substack{(w_i)_{i=1}^d\in H_{\alpha-1}(X\times Y)\\ (\phi,\psi)\in H_{\alpha+1}(X)\times H_{\alpha+1}(Y)}} \langle\phi,\hat{\mu}_n\rangle + \langle\psi,\hat{\nu}_n\rangle \\ & s.t.\;\; \frac{\|x_l-y_l\|^2}{2} - \phi(x_l) - \psi(y_l) = \sum_{i=1}^d w_i(x_l,y_l)^2,\;\forall l\in\{1,\cdots,n\}\,.\end{aligned} \tag{2.5}$$

When the estimator above is properly regularized, the authors of [37] showed that with high probability $|\widehat{\mathrm{OT}}_{SoS} - \mathrm{OT}| \sim 1\sqrt{n}$ when $\alpha\to\infty$. Yet, computationally speaking, even though the kernel trick still applies to ϕ and ψ, handling numerically the functions w_i is *a priori* less clear. In the next paragraph, we give some insight on the work of [25] and show how a similar kernel trick can be applied to function $\gamma = \sum_{i=1}^d w_i^2$.

2.2. Kernel Sum-Of-Squares. One simple way to model a positive function $f : \mathbb{R} \to \mathbb{R}$ is to take the vector of the n first monomials $e_n(x) = (1, x, \cdots, x^n)$ and A a symmetric positive definite matrix $A \in \mathbb{S}_+(\mathbb{R}^n)$, the set of p.s.d. matrices of size n, and to form $f(x) = e_n(x)^\top A e_n(x)$. Denoting $(\lambda_1, \cdots, \lambda_n)$ the (positive) eigenvalues of A and $Q \in O_n(\mathbb{R})$ an orthonormal diagonalizing basis, one has $f(x) = \sum_{i=1}^n \lambda_i q_i(x)^2$ with $q_i(x) = [Qe_n(x)]_i$; as it is a sum of squared polynomials f is indeed positive. The polynomial SoS representation has been used in global optimization for over a decade now [22], yet the idea of modeling a positive function as a quadratic form on the space of monomials was recently generalized to the RKHS case by [25]. Given H an RKHS with kernel k and A a positive, self-adjoint operator on H, one can model a positive function as

$$\gamma_A(x) = \langle k(x, \cdot), Ak(x, \cdot) \rangle_H \,. \tag{2.6}$$

Assuming that A has finite rank p, it can be diagonalized in some orthonormal basis $(e_i)_{i=1}^p \in H$ and one recovers $f_A(x) = \sum_{i=1}^p \lambda_i e_i(x)^2$ where the λ_i are the positive eigenvalues of A ; similarly, any finite sum of squared functions belonging to H can be expressed in the form $x \mapsto \gamma_A(x)$. These kernel SOS models also enjoy a kernel trick where one can reformulate learning problems over the infinite set of positive, self-adjoints operators $\mathcal{S}_+(H)$ as finite dimensional problems over the finite dimensional space of positive, symmetric matrices.

THEOREM 2.4 (Marteau-Ferey et. al. 2020). *Let $L : \mathbb{R}^n \to \mathbb{R}$ be a function which is lower semi-continuous and bounded from below. Then the problem*

$$\inf_{A \in \mathcal{S}_+(H)} L(\gamma_A(x_1), \cdots, \gamma_A(x_n)) + \lambda \mathrm{Tr}(A)$$

where $\lambda > 0$, has a solution A^* that can be expressed as $A^* = \sum_{ij} b_{ij} k(x_i \cdot) \otimes k(x_j, \cdot)$ with $B = [b_{ij}] \in \mathbb{S}_+(\mathbb{R}^n)$.

The authors of [37] applied this result to the (regularized) problem (2.5) to recover the following finite reformulation of their estimator

$$\begin{aligned} \widehat{\mathrm{OT}}_{SoS} \;=\; & \sup_{\substack{(\beta,\omega)\in\mathbb{R}^{2n} \\ B\in\mathbb{S}_+(\mathbb{R}^n)}} \langle \sum_{i=1}^n \beta_i k_X(x_i, \cdot), \hat{\mu}_n \rangle + \langle \sum_{j=1}^n \omega_j k_Y(y_j, \cdot), \hat{\nu}_n \rangle + \lambda \mathrm{Tr}(B) \\ & s.t. \;\; \frac{\|x_l - y_l\|^2}{2} - \sum_{i=1}^n \beta_i k(x_i, x_l) - \sum_{j=1}^n \omega_j k(y_j, y_l) = [KBK]_{ll} \,, \end{aligned} \tag{2.7}$$

where K is the matrix $(k_{XY}((x_i, y_i), (x_j, y_j)))_{1 \le i,j \le n}$ with k_{XY} is the kernel of $H_{\alpha-1}(X \times Y)$. Using an interior point method, this problem above can be solved with a precision δ in $O(n^{3.5} \log(\frac{n}{\delta}))$ time. Hence they managed to recover an estimator of $W_2^2(\mu, \nu)$ with a favorable statistical behavior and that can be computed in polynomial time with respect to the number of available samples.

2.3. Estimation of the transport map. Even though the problem of estimating efficiently the squared Wasserstein distance was solved by [37], our problem is more delicate as we need to estimate the transport map itself, a.k.a the (gradient of the) argmin of problem (1.3). If problem (2.5) had strictly positive curvature around the minimum, we could deduce the convergence of the minimizers $(\hat{\phi}, \hat{\psi})$ toward (ϕ_0, ψ_0) at the same rate as the convergence of $\widehat{\mathrm{OT}}$ toward OT. However, the

objective function in (2.5) is linear and in particular non strongly convex. One way to remedy this is to used the so-called *semi-dual* formulation of optimal transport. Making the change of variable $(f, g) = (\frac{\|\cdot\|^2}{2} - \phi, \frac{\|\cdot\|^2}{2} - \psi)$, the dual formulation of OT can be written up to moment terms as

$$\begin{aligned} \mathrm{OT} = &\inf_{f,g} \ \langle f, \mu\rangle + \langle g, \nu\rangle \\ &s.t.\ f(x) + g(y) \geq x^\top y,\ \forall (x, y) \in X \times Y\,, \end{aligned} \tag{2.8}$$

which is also known as the Brenier formulation of OT. With this formulation, one has at the optimum $g = \mathcal{L}_X(f)$ with $\mathcal{L}_X(f)$ the Legendre transform of f restricted to X defined as

$$\mathcal{L}_X(f) : y \in Y \mapsto \sup_{x\in X} xy^\top - f(x)\,, \tag{2.9}$$

and the problem can be re-written $\mathrm{OT} = \inf_f J_X(f) \coloneqq \langle f, \mu\rangle + \langle \mathcal{L}_X(f), \nu\rangle$. If $X = \mathbb{R}^d$ and f is convex and globally M-smooth that is $\sup_{x\in\mathbb{R}^d} \|\nabla^2 f(x)\| \leq M$, it has been shown that this new objective is strongly convex with respect to the $L^2(\mu)$ norm [21] and verifies

$$\|\nabla f - T_0\|^2_{L^2(\mu)} \leq \frac{M}{2}(J_X(f) - J_X(f_0))\,,$$

where f_0 is the ground truth OT potential that verifies $\nabla f_0 = T_0$. We show in the next section that a similar result holds when X is a Lipschitz domain and f is M-smooth over X but not necessarily convex. Hence, in order to estimate the transport map T_0 we choose instead to solve

$$\begin{aligned} \inf_{\substack{A\in\mathcal{S}_+(H_{\alpha-1}(X\times Y))\\ (f,g)\in H_{\alpha+1}(X)\times H_{\alpha+1}(Y)}} &\langle f, \hat{\mu}_n\rangle + \langle g, \hat{\nu}_n\rangle + \lambda\left(\|A\|_F^2 + \|f\|^2_{H_{\alpha+1}(X)} + \|g\|^2_{H_{\alpha+1}(Y)}\right) \\ &+ \zeta \sum_l \left(f(x_l) + g(y_l) - x_l^\top y_l - \gamma_A(x_l, y_l)\right)^2, \end{aligned} \tag{2.10}$$

where λ and ζ are positive reals that we make explicit in the next section, $\|\cdot\|_F^2$ is the Froebenius norm and γ_A corresponds to (2.6) for $k = k_{XY}$ the kernel of $H_{\alpha-1}(X\times Y)$. Note that our estimator is different than the one defined by [37] (see equation (2.7)) on two levels. First, for computational purposes that will be made clearer in Section 4, we replaced the trace operator by the Froebenius norm and we use soft penalties instead of hard equality constraints. Second, as mentioned above, we rely on the Brenier formulation of OT (2.8) instead of the Kantorovitch formulation (1.3). The rational behind this slight variation is that as n grows, denoting $(\hat{f}, \hat{g})$ the minimizers of (2.10), we expect the empirical potentials to be linked as $\hat{g} \approx \mathcal{L}_X(\hat{f})$ and *a fortiori* that $\mathcal{E}_{\hat{\mu},\hat{\nu}}(\hat{f}, \hat{g}) \approx \hat{J}_X(\hat{f})$ so we can leverage the strong convexity of the semi-dual.

2.4. Assumptions. We formally state our assumptions in this paragraph and discuss their impact. Our first assumption is our main smoothness assumption on the transport potentials. As discussed above, it allows us to both temper the curse of dimension and design tractable estimators.

Assumption 2.5. The optimal transport potentials (ϕ_0, ψ_0) belong to $H_\alpha(X) \times H_\alpha(Y)$ and with $\alpha > d + 2$.

Note that our smoothness assumption slightly differs from [21] on two different levels. First we impose that $\alpha > d+2$ which is essentially to ensure that $H_{\alpha-1}(X \times Y)$ is indeed an RKHS so we can apply the SoS trick ; the extra order of regularity required is a small technical requirement that is developed in [37]. The other difference is that we make the weaker assumption that (ϕ_0, ψ_0) are in a 2-Sobolev space of order s instead of an ∞-Sobolev space of order s. Our second assumption is a set of technical hypotheses on the measures μ, ν.

Assumption 2.6. The measures μ, ν are supported over X and Y, Lipschitz domains of $\mathbb{R}^d$. Furthermore, we assume that μ, ν have continuous densities w.r.t. the Lebesgue measure that are bounded from above and below.

The goal of Assumption 2.6 is threefold: (i) it ensures that one can apply the Brézis-Mironescu inequalities [7] to quantities of the form $\|f\|_{L^p(\mu)}$, (ii) it ensures that the spaces X and Y are "well-covered" by the empirical counterparts $\hat{\mu}, \hat{\nu}$ and (iii) it ensures that the measure of points that are r-close for the quadratic distance to the boundary of the supports vanishes as r goes to zero. Equipped with these two assumptions, we prove in the next section that for well-chosen scalars (λ, ζ), we have $\|\nabla \hat{f} - T_0\|^2_{L^2(\mu)} \lesssim n^{-\frac{\alpha}{\alpha+d/2-1}}$.

3. Statistical rates. A standard way to derive optimal error bounds in non-parametric statistics is to rely on two key ingredients: the strong convexity of the true risk and the concentration of the empirical risk toward the true risk. The combination of these two ingredients, also known as localization techniques [39], lead to optimal error bounds scaling as $1/n$ instead of the classical $1/\sqrt{n}$ error rate. However, as mentioned in the previous section, our empirical risk, which is the linear part of objective (2.10) is not strongly convex. Hence, we first show that at the optimum, our linear risk nearly upper-bounds the semi-dual. Then we proceed as in standard non-parametric statistics: we extend the strong convexity properties of the semi-dual when the Legendre transform is restricted to the support of the measures and we prove the concentration of the empirical risk toward the true risk in a localized fashion using Brézis-Mironescu inequalities. Beforehand, we show that our estimators are indeed well defined.

PROPOSITION 3.1. *If $\alpha > d+1$, there exists minimizers $(\hat{f}, \hat{g}, \hat{A}) \in H_{\alpha+1}(X) \times H_{\alpha+1}(Y) \times H_{\alpha-1}(X \times Y)$ for problem (2.10).*

The proof is left in Appendix and rely on Kakutani theorem. Equipped with this result, we can start to prove the results stated above.

3.1. Upper-bound of the semi-dual. For potentials (f, g), let us define $\mathcal{E}_{\mu,\nu}(f, g) \coloneqq \langle f, \mu \rangle + \langle g, \nu \rangle$, the (deterministic) linear part of objective (2.10). In the following proposition, we prove that at the optimum of problem (2.10), $\mathcal{E}_{\mu,\nu}$ nearly upper-bounds the semi-dual J_X.

PROPOSITION 3.2. *Under Assumptions 2.5 and 2.6, it holds for any $0 \leq \delta < 1$ that the minimizers $(\hat{f}, \hat{g}, \hat{A})$ of problem (2.10) verify with probability at least $1 - \delta$*

$$J_X(\hat{f}) - \mathcal{E}_{\mu,\nu}(\hat{f}, \hat{g}) \lesssim (n/\log(n/\delta))^{-(\alpha-1-d)/(2d)} \hat{R} + (n/\log(n/\delta))^{-1/(2d)} \frac{\sqrt{1+\lambda}}{\sqrt{\zeta}},$$

where $\hat{R}$ is defined as $\hat{R} \coloneqq 1 + \|\hat{f}\|_{H_{\alpha+1}(X)} + \|\hat{g}\|_{H_{\alpha+1}(Y)} + \|\hat{A}\|_F$ and $\lesssim$ hides constants that are independent of n, δ, ζ and λ.

Proof. Using Theorem 2.3, there exist $(w_i)_{i=1}^d \in H_{\alpha-1}(X \times Y)$ such that the optimal potentials (f_0, g_0) verify for all $(x, y) \in X \times Y$

$$f_0(x) + g_0(y) - x^\top y = \sum_{i=1}^d w_i^2(x, y)\,.$$

Defining the operator

$$A \coloneqq \sum_{i=1}^d w_i \otimes w_i\,, \tag{3.1}$$

the equality above can be re-written as $f_0(x) + g_0(y) - x^\top y = \gamma_A(x, y)$. Hence, using the optimality conditions of (2.10), it holds

$$\zeta \sum_l \left(\hat{f}(x_l) + \hat{g}(y_l) - x_l^\top y_l - \gamma_{\hat{A}}(x_l, y_l) \right)^2 \le \lambda(\|f_0\|^2_{H_{\alpha+1}(X)} + \|g_0\|^2_{H_{\alpha+1}(Y)} + \|A\|_F^2) + \mathcal{E}_{\mu,\nu}(f_0, g_0)\,.$$

Now, using the fact that $\mathcal{E}_{\mu,\nu}(f_0, g_0) \le \|f_0\|_{W_0^\infty(X)} + \|g_0\|_{W_0^\infty(Y)}$, we recover that

$$\zeta \sum_l \left(\hat{f}(x_l) + \hat{g}(y_l) - x_l^\top y_l - \gamma_{\hat{A}}(x_l, y_l) \right)^2 \lesssim 1 + \lambda\,.$$

In particular, for all indexes l we have

$$\hat{f}(x_l) + \hat{g}(y_l) - x_l^\top y_l - \gamma_{\hat{A}}(x_l, y_l) \lesssim \frac{\sqrt{1+\lambda}}{\sqrt{\zeta}}\,. \tag{3.2}$$

Let us define $\hat{\theta}(x, y) \coloneqq \hat{f}(x) + \hat{g}(y) - x^\top y - \gamma_{\hat{A}}(x, y)$ for $(x, y) \in X \times Y$. As a combination of functions in $H_{\alpha-1}(X \times Y)$, $\hat{\theta}$ is indeed in $H_{\alpha-1}(X \times Y)$ and its norm can be upper bounded as

$$\|\hat{\theta}\|_{H_{\alpha-1}(X\times Y)} \lesssim (1 + \|\hat{f}\|_{H_{\alpha+1}(X)} + \|\hat{g}\|_{H_{\alpha+1}(Y)} + \|\hat{A}\|_F)\,. \tag{3.3}$$

Using the sampling inequalities [42, Proposition 2.4] together with (3.2) and (3.3), we obtain

$$\|\hat{\theta}\|_{W^\infty_{\alpha-1}(X\times Y)} \lesssim h^{\alpha-1-d}\hat{R} + h\frac{\sqrt{1+\lambda}}{\sqrt{\zeta}}\,,$$

where h is the filling distance defined as $h \coloneqq \sup_{(x,y)\in X\times Y} \min_l \|(x, y) - (x_l, y_l)\|_2$ and $\hat{R} \coloneqq (1 + \|\hat{f}\|_{H_{\alpha+1}(X)} + \|\hat{g}\|_{H_{\alpha+1}(Y)} + \|\hat{A}\|_F)$. Now, as shown in [37, Lemma 12], for all $1 > \delta > 0$ we have under Assumption 2.6 that with probability at least $1 - \delta$, if $n \ge n_0$ where n_0 is a constant independent of n, δ, we have $h \lesssim (n/\log(n/\delta))^{-1/(2d)}$ hence with probability at least $1 - \delta$ for all $(x, y) \in X \times Y$,

$$\hat{g}(y) - (x^\top y - \hat{f}(x)) \gtrsim -(n/\log(n/\delta))^{-(\alpha-1-d)/(2d)}\hat{R} - (n/\log(n/\delta))^{-1/(2d)}\frac{\sqrt{1+\lambda}}{\sqrt{\zeta}}\,.$$

Taking the maximum of the right hand side with respect to x and integrating y over ν, we recover that with probability at least $1 - \delta$

$$\mathcal{E}_{\mu,\nu}(\hat{f}, \hat{g}) - J_X(\hat{f}) \gtrsim -(n/\log(n/\delta))^{-(\alpha-1-d)/(2d)}\hat{R} - (n/\log(n/\delta))^{-1/(2d)}\frac{\sqrt{1+\lambda}}{\sqrt{\zeta}}\,.$$

∎

3.2. Strong convexity of the semi-dual. We prove in this paragraph that the semi-dual is strongly convex around its optimum in our setting. Indeed this result is already known when the following assumptions hold:

1. either the semi-dual is evaluated on convex, smooth potentials [21, Proposition 10] or either the semi-dual is evaluated on convex potentials (non necessarily smooth) with a continuous target measure [15, Theorem 2.1].
2. the Legendre transform is taken globally over $\mathbb{R}^d$ and is defined as $f^*(y) : y \mapsto \sup_{x\in\mathbb{R}^d} x^\top y - f(x)$.

Yet none of these two requirements hold in our setting as our 1) estimators $(\hat{f}, \hat{g})$ are not guaranteed to be convex and 2) the Legendre transform is taken only locally over X. Instead, we prove that [21, Proposition 10] nearly holds when the Legendre transform is taken locally at the price of restricting the measure μ to points that are "not too close" to the boundary. In what follows, we shall denote for some compact set Ω and some point $x \in \mathbb{R}^d$ the distance from x to Ω as $d(x, \Omega) \coloneqq \inf_{y\in\Omega} \|x - y\|_2$.

PROPOSITION 3.3. *If μ is a probability measure supported on a bounded open set X and ν is a probability measure such that the OT map $T_0 = \nabla f_0$ from μ to ν exists, then for a potential f with an M-Lipschitz gradient over X and strictly positive distance $r > 0$, we have*

$$\|\nabla f - T_0\|^2_{L^2(\mu_r)} \leq 2M_r^f (J_X(f) - J_X(f_0)),$$

where μ_r is the measure μ restricted to $A_r = \{x \in X | d(x, \partial X) > r\}$ the set of points the are at least at distance r from the boundary and $M_r^f = \max\left(M, \frac{\|\nabla f - T_0\|_{W_0^\infty(X)}}{r}\right)$.

Before starting the proof, note that as X expands toward the whole space $\mathbb{R}^d$, we can take r arbitrary large and recover the result $\|\nabla f - T_0\|^2_{L^2(\mu)} \leq 2M(J(f) - J(f_0))$. Even though this result is standard in OT literature [21, 24, 38], we manage to obtain it without assuming convexity of the potential f. Unfortunately, we did not manage to prove the tightness of the upper-bound and in particular, we do not know if a similar result holds when $r = 0$; we postpone this open question for future work.

Proof. We begin by noting that, since T_0 is a bijection between X and the support of ν, the term $\langle \mathcal{L}_X(f_0), \nu\rangle$ can be re-written as $\langle f_0^* \circ T_0, \mu\rangle$ where f_0^* is the Legendre transform of f_0 taken globally. Similarly, the term $\langle \mathcal{L}_X(f), \nu\rangle$ can be re-written as $\langle \mathcal{L}_X(f) \circ T_0, \mu\rangle$. Hence the difference $J_X(f) - J_X(f_0)$ reads

$$J_X(f) - J_X(f_0) = \int f(x) + \mathcal{L}_X(f)(T_0(x)) - f_0(x) - f_0^*(T_0(x)) d\mu(x). \tag{3.4}$$

The Fenchel-Young equality gives for all x in $\mathbb{R}^d$ that $f_0^*(T_0(x)) = x^\top T_0(x) - f_0(x)$, hence the f_0 terms cancel in the formula above and we obtain

$$J_X(f) - J_X(f_0) = \int f(x) + \mathcal{L}_X(f)(T_0(x)) - x^\top T_0(x) d\mu(x). \tag{3.5}$$

Now, by definition of the Legendre transform, the integrand reads pointwise for all $u \in X$

$$f(x) + \mathcal{L}_X(f)(T_0(x)) - x^\top T_0(x) \geq f(x) + u^\top T_0(x) - f(u) - x^\top T_0(x). \tag{3.6}$$

In particular, for $u = x$, the right hand side is zero hence it proves that the integrand is pointwise positive. Hence, for $r > 0$, we can lower bound the difference $J_X(f) - J_X(f_0)$

by the integrand integrated over the restricted measure μ_r

$$J_X(f) - J_X(f_0) \geq \int f(x) + \mathcal{L}_X(f)(T_0(x)) - x^\top T_0(x) d\mu_r(x)\,. \tag{3.7}$$

Let us now use again (3.6) with $u_\alpha(x) \coloneqq x + \alpha(T_0(x) - \nabla f(x))$ where $\alpha > 0$ is to be chosen later. First, we need to chose α such that u belongs to X. Let us define α^* as

$$\alpha^* = \inf\ B_x \coloneqq \{\alpha > 0 \mid u_\alpha(x) \notin X\}\,.$$

Since X is open, α^* is indeed strictly positive. For arbitrary small $\epsilon > 0$, one has $u_{\alpha^*-\epsilon}(x) \in X$ hence $u_{\alpha^*}(x) \in \overline{X}$. Furthermore, as X is opened, we can guarantee that $u_{\alpha^*}(x) \notin X$. Hence, $u_{\alpha^*}(x) \in \overline{X} \cap (\mathbb{R}^d \setminus X) = \partial X$. Now recall that $x \in A_r$ so in particular, $\|x - u^*_\alpha(x)\| > r$, *i.e* $\alpha^*\|\nabla f(x) - T_0(x)\| > r$. Hence by definition, $\alpha^* > r/\|\nabla f - T_0\|_{W_0^\infty(X)}$ and as a consequence, it suffices to take $\alpha \leq r/\|\nabla f - T_0\|_{W_0^\infty(X)}$ to guarantee that $u_\alpha(x) \in X$. Noting that for all $t \in [0,1]$, we have $tx + (1-t)u_\alpha(x) = u_{(1-t)\alpha}(x) \in X$, we can apply the Taylor inequality to the integrand at order 2 with respect to α. Denoting $\Delta(x) = f(x) + \mathcal{L}_X(f)(T_0(x)) - x^\top T_0(x)$, we recover for all x in the support of μ_r

$$\begin{aligned}
\Delta(x) &\geq f(x) + u_\alpha(x)^\top T_0(x) - f(u_\alpha(x)) - x^\top T_0(x) \\
&\geq f(x) + \alpha(T_0(x) - \nabla f(x))^\top T_0(x) - f(x) - \alpha \nabla f(x)^\top (T_0(x) - \nabla f(x)) \\
&\quad - \frac{M\alpha^2}{2}\|\nabla f(x) - T_0(x)\|^2 \\
&= \alpha\|T_0(x) - \nabla f(x)\|^2 - \frac{M\alpha^2}{2}\|\nabla f(x) - T_0(x)\|^2\,.
\end{aligned}$$

Taking the maximum with respect to $0 \leq \alpha \leq r/\|\nabla f - T_0\|_{W_0^\infty(X)}$ is separated in two cases: either $\frac{1}{M} < r/\|\nabla f - T_0\|_{W_0^\infty(X)}$ and we get that the the maximum of the r.h.s. above is $\frac{1}{2M}\|\nabla f(x) - T_0\|^2$, either $\frac{1}{M} \geq r/\|\nabla f - T_0\|_{W_0^\infty(X)}$ and we obtain the following lower bound

$$\Delta(x) \geq \left(r - \frac{M}{2}\frac{r^2}{\|\nabla f - T_0\|_{W_0^\infty(X)}}\right)\frac{\|\nabla f(x) - T_0(x)\|^2}{\|\nabla f - T_0\|_{W_0^\infty(X)}} \tag{3.8}$$

$$\geq \frac{r\|\nabla f(x) - T_0(x)\|^2}{2\|\nabla f - T_0\|_{W_0^\infty(X)}}\,. \tag{3.9}$$

Hence we have

$$J_X(f) - J_X(f_0) \geq \min\left(\frac{1}{2M}, \frac{r}{2\|\nabla f - T_0\|_{W_0^\infty(X)}}\right)\|\nabla f - T_0\|^2_{L^2(\mu_r)}\,, \tag{3.10}$$

which gives the desired result. ∎

Now, it remains to control the gap between $\|\nabla f - T_0\|_{L^2(\mu_r)}$ and $\|\nabla f - T_0\|_{L^2(\mu)}$. To this end, the lemma below guarantees, for X a Lipschitz domain, $\mu(\mathbb{R}^d \setminus A_r)$ is $O(r)$. It allows to bound the gap as

$$\|\nabla f - T_0\|^2_{L^2(\mu)} \lesssim \frac{\|\nabla f - T_0\|^2_{L^2(\mu_r)}}{r} + r\|\nabla f - T_0\|^2_{W_0^\infty(X)}\,,$$

in particular, as $\nabla \hat{f}$ converges toward T_0 the gap can be tighten.

LEMMA 3.4. *Let μ be a probability measure supported by X a Lispchitz domain of $\mathbb{R}^d$. If μ is continuous with respect to the Lesbegue measure with a density bounded from above by ρ_0, then there exists r_0 such that for all $r \leq r_0$, it holds $\mu(\mathbb{R}^d \backslash A_r) \lesssim \rho_0 r$.*

The proof is left in Appendix.

3.3. Localized concentration. In this paragraph, we collect the last technical result to derive our statistical rates. We derive a concentration bound for the empirical process $\langle g, \mu - \hat{\mu} \rangle$ when μ is supported over X a compact subset of $\mathbb{R}^d$ with a Lipschitz boundary and g is assumed to belong to some Sobolev space $H_\beta(X)$ with $\beta > d/2$. We highlight the fact that while standard concentration results rely on (localized) metric entropy bounds, we use instead the Pinelis inequality combined with Brézis-Mironescu inequalities. Had we relied on the former, we would not have matched the minimax upper-bounds in [21] in the highly smooth regime.

PROPOSITION 3.5. *Let μ be a continuous probability measure bounded from below, supported over X a connected bounded subset of $\mathbb{R}^d$ with Lipschitz boundary and let $\hat{\mu}$ be its n-samples empirical counterpart. Let g be a potential belonging to $H_\beta(X)$ with $\beta > \max(1, d/2)$. For any $0 < \delta < 1$ and $\epsilon > 0$ such that $d/2 + \epsilon \leq \beta$, it holds with probabilty at least $1 - \delta$*

$$\langle g, \mu - \hat{\mu} \rangle \lesssim \frac{\log(2/\delta)}{\epsilon \sqrt{n}} \|\nabla g\|_{L^2(\mu)}^{\frac{\beta - d/2 - \epsilon}{\beta - 1}} \|g\|_{H_\beta(X)}^{\frac{d/2 + \epsilon - 1}{\beta - 1}} . \tag{3.11}$$

Proof. Defining the mean of g over X as $m = \int_x g(x) dx$, we first observe that $\langle g, \mu - \hat{\mu} \rangle = \langle g - m, \mu - \hat{\mu} \rangle$. Then, we define the kernel mean-embedding as $w_\mu = \mathbb{E}_{Z \sim \mu}[k(Z, \cdot)]$ where k is the reproducing kernel of $H_{d/2+\epsilon}(X)$. Using this definition and reproducing property, we have

$$\langle g, \mu - \hat{\mu} \rangle = \int_x (g(x) - m) \; d(\mu(x) - \hat{\mu}(x)) \tag{3.12}$$

$$= \int_x \langle g - m, k(x, \cdot) \rangle_{H_{d/2+\epsilon}(X)} \; d(\mu(x) - \hat{\mu}(x)) \tag{3.13}$$

$$= \langle g - m, w_\mu - w_{\hat{\mu}} \rangle_{H_{d/2+\epsilon}(X)} \tag{3.14}$$

$$\leq \|g - m\|_{H_{d/2+\epsilon}(X)} \|w_\mu - w_{\hat{\mu}}\|_{H_{d/2+\epsilon}(X)} , \tag{3.15}$$

where the last inequality is obtained using Cauchy-Schwartz. For any $0 < \delta < 1$, we can upper-bound the second term of the right-hand side: the Pinelis inequality [9] yields that with probability at least $1 - \delta$,

$$\|w_\mu - w_{\hat{\mu}}\|_{H_{d/2+\epsilon}(X)} \lesssim \frac{\log(2/\delta)}{\epsilon \sqrt{n}} .$$

The first term is upper-bounded using [7, Theorem 1] with $p = p_1 = p_2 = 2$, $s = d/2 + \epsilon$, $s_1 = 1$ and $s_2 = \beta$ and we obtain

$$\|g - m\|_{H_{d/2+\epsilon}(X)} \lesssim \|g - m\|_{H_1(X)}^{\frac{\beta - d/2 - \epsilon}{\beta - 1}} \|g - m\|_{H_\beta(X)}^{\frac{d/2 + \epsilon - 1}{\beta - 1}} .$$

The therm $\|g - m\|_{H_1(X)}$ is decomposed as $\|g - m\|_{H_1(X)} = \|\nabla g\|_{L^2(X)} + \|g - m\|_{L^2(X)}$. Since $\int_X g(x) - m \; dx = 0$ and X is a bounded connected subset of $\mathbb{R}^d$ with Lipschitz, we can use the Poincaré-Wirtinger inequality [26] to recover $\|g - m\|_{L^2(X)} \lesssim$

$\|\nabla g\|_{L^2(X)}$. Using the fact that μ has a density over X bounded from below, we get $\|\nabla g\|_{L^2(X)} \lesssim \|\nabla g\|_{L^2(\mu)}$ hence with probability at least $1-\delta$

$$\langle g, \mu - \hat{\mu}\rangle \lesssim \frac{\log(2/\delta)}{\epsilon\sqrt{n}} \|g-m\|_{H_\beta(X)}^{\frac{d/2+\epsilon-1}{\beta-1}} \|\nabla g\|_{L^2(\mu)}^{\frac{\beta-d/2-\epsilon}{\beta-1}} . \tag{3.16}$$

We conclude using the fact that $\|g-m\|_{H_\beta(X)} \leq 2\|g\|_{H_\beta(X)}$. ∎

Equipped with this result, we now have all the ingredients to derive the statistical rates of our estimator of the potentials.

3.4. Proof of the main result. This paragraph is dedicated to the analysis of the convergence of our empirical estimator $\nabla\hat{f}$ toward the OT map T_0 from μ to ν using the results from the three subsections above. Equivalently, we shall study at the same time the convergence of $\nabla\hat{g}$ toward the inverse transport map T_0^{-1} *i.e.* the OT map from ν to ν.

THEOREM 3.6. *Under Assumption 2.5 and Assumption 2.6 and the additional assumption $d \geq 2$, if we set $\lambda_n = n^{-\min(\frac{\alpha-1-d}{2d}, -\frac{\alpha}{\alpha+d/2-1})}$ and $\zeta = \lambda_n^{-2} n^{-1/d}$ then, denoting T_0 the OT map from μ to ν, the minimizers $(\hat{f}, \hat{g})$ of the empirical problem 2.10 verify*

$$\mathbb{E}[\|\nabla\hat{f} - T_0\|_{L^2(\mu)}^2 + \|\nabla\hat{g} - T_0^{-1}\|_{L^2(\nu)}^2] \lesssim \lambda_n^{\min(1, \frac{2}{1+3d/(2\alpha)})}, \tag{3.17}$$

where $\lesssim$ hides poly-log factors in n and constants that do not depend on n.

Hence, as claimed in the introduction, when the smoothness parameter α is sufficiently large, we have $\lambda_n \sim n^{-\frac{\alpha}{\alpha+d/2-1}}$ and we match exactly the minimax rate found in [21]. While the slow rate $n^{-\frac{\alpha-1-d}{2d}}$ could indeed be expected in the less smooth regime as it quantifies how much the constraint $f(x)+g(y) \geq xy^\top$ is violated, the extra $\frac{2}{1+3d/(2\alpha)}$ exponent may be an artefact of our proof. It comes from the fact in Proposition 3.2, we only managed to prove strong convexity of the semi-dual with respect to the $L^2(\mu_r)$ semi-norm, where μ_r is the measure μ restricted to the point that are at least at distance r from the boundary of its support, instead of the $L^2(\mu)$ semi-norm, with an upper-bound degrading as $r \to 0$. However, this extra exponent can be removed by slightly tweaking our estimator. The idea is to sample the cost SoS constraint on a domain larger than $X \times Y$: instead of imposing the (soft) cost constraint over the pairs (x_i, y_i) one can inject noise and impose the cost constraint as

$$f(x_i+\epsilon_i^1) + g(y_i+\epsilon_i^2) - (x_i+\epsilon_i^1)(y_i+\epsilon_i^2)^\top - \gamma_A(x_i+\epsilon_i^1, y_i+\epsilon_i^2) << 1\,,$$

where the $\epsilon_i^1, \epsilon_i^2$ are uniformly drawn in the ball $B(0,\epsilon_0)$ with ϵ_0 fixed ; this simple "over-sampling" strategy allows to get rid of the edge effect introduced by Proposition 3.2.

Proof. The proof is decomposed into the following main steps: 1) we use the results of Section 3.2 on the strong convexity of the semi-dual to upper-bound the error $\|\nabla\hat{f} - T_0\|_{L^2(\mu)}$ by $J_X(\hat{f}) - J_X(f_0)$ and a residual term 2) we use the upper-bound of Proposition 3.2 to replace the difference of the semi-duals by the difference of the non-stochastic objectives $\mathcal{E}_{\mu,\nu}$ and we use the concentration result of Proposition 3.5 to concentrate the empirical objective toward the non-stochastic objective; we obtain an upper-bound on the error that depends on the RHKS norms of the empirical potentials 3) we use the same type of arguments to bound the RKHS norm

of the empirical potentials 4) we obtain two coupled upper-bounds that relate the error and the RKHS norm of the empirical potentials and we conclude. Note that the main difficulty of the proof comes from the fact that we softly penalize the norms $\|\hat{f}\|_{H_{\alpha+1}(X)}, \|\hat{g}\|_{H_{\alpha+1}(Y)}$ and $\|\hat{A}\|_F^2$ instead of imposing a hard constraint that would require an priori knowledge of these objects and lead to a less adaptive estimator. To ease the understanding of the proof, we advise the reader to treat these quantities as $O(1)$ in the first place.

1) For a Lipschitz function h and $r > 0$, we define the error $e_{2,\mu}(h) \coloneqq (\|\nabla h\|^2_{L^2(\mu)})^{1/2}$. For any probability measure μ, the error $e_{2,\mu}(\hat{f} - f_0)$ can be upper-bounded as

$$e^2_{2,\mu}(\hat{f} - f_0) \leq e^2_{2,\mu_r}(\hat{f} - f_0) + \|\nabla \hat{f} - T_0\|^2_{W_0^\infty(X)} \mu(\mathbb{R}^d \setminus A_r)\,,$$

where A_r is defined as $A_r = \{x \in X \mid d(x, \partial X) > r\}$ and where μ_r is the measure μ restricted to A_r that is such that for all Borel set B, $\mu_r(B) = \mu(B \cap A_r)$. Using Proposition 3.3, if $r > 0$ is sufficiently small, the first term is upper-bounded as

$$e^2_{2,\mu_r}(\hat{f} - f_0) \leq 2\left(\|\hat{f}\|_{W_2^\infty(X)} + \frac{\|\nabla \hat{f} - T_0\|_{W_0^\infty(X)}}{r}\right)(J_X(\hat{f}) - J_X(f_0))\,,$$

and using Lemma 3.4 we have $\mu(\mathbb{R}^d \setminus A_r) \lesssim r$. Hence, denoting for some Lipschitz function h and some measure μ the supremum norm of the gradient over the support of μ, $e_{\infty,\mu}(h) = \|\nabla h\|_{L^\infty(\mu)}$, we get with this notation

$$e^2_{2,\mu}(\hat{f} - f_0) \lesssim 2\left(\|\hat{f}\|_{W_2^\infty(X)} + \frac{e_{\infty,\mu}(\hat{f} - f_0)}{r}\right)(J_X(\hat{f}) - J_X(f_0)) + r e^2_{\infty,\mu}(\hat{f} - f_0)\,. \tag{3.18}$$

We now upper-bound the term $e_{\infty,\mu}(\hat{f} - f_0)$ by $e_{2,\mu}(\hat{f} - f_0)$ using Gagliardo-Niremberg inequalities [27, Theorem 1]. Since both $\nabla \hat{f}$ and ∇f_0 belong to the RKHS $H_\alpha(X)$ component-wise, it holds

$$e_{\infty,\mu}(\hat{f} - f_0) \lesssim \|\hat{f} - f_0\|^{\frac{d}{2\alpha}}_{H_{\alpha+1}(X)} \|\nabla \hat{f} - T_0\|^{1-\frac{d}{2\alpha}}_{L^2(X)} + \|\nabla \hat{f} - T_0\|^{1-\frac{d}{2\alpha}}_{L^2(X)}\,. \tag{3.19}$$

Since μ has a bounded density w.r.t. Lebesgue over X, we can upper bound $\|\nabla \hat{f} - T_0\|_{L^2(X)}$ by $\|\nabla \hat{f} - T_0\|_{L^2(\mu)}$ and we recover

$$e_{\infty,\mu}(\hat{f} - f_0) \lesssim e_{2,\mu}(\hat{f} - f_0)^{1-\frac{d}{2\alpha}} (\|\hat{f}\|^{\frac{d}{2\alpha}}_{H_{\alpha+1}(X)} + 1)\,, \tag{3.20}$$

which eventually yields

$$\begin{aligned} e^2_{2,\mu}(\hat{f} - f_0) \lesssim &\left(\|\hat{f}\|_{W_2^\infty(X)} + \frac{e_{2,\mu}(\hat{f} - f_0)^{1-\frac{d}{2\alpha}} (\|\hat{f}\|^{\frac{d}{2\alpha}}_{H_{\alpha+1}(X)} + 1)}{r}\right)(J_X(\hat{f}) - J_X(f_0)) \\ &+ r e_{2,\mu}(\hat{f} - f_0)^{2-\frac{d}{\alpha}} (\|\hat{f}\|^{\frac{d}{2\alpha}}_{H_{\alpha+1}(X)} + 1)^2\,. \end{aligned} \tag{3.21}$$

Conversely, a similar result holds for the empirical potential $\hat{g}$

$$\begin{aligned} e^2_{2,\nu}(\hat{g} - f_0^*) \lesssim &\left(\|\hat{g}\|_{W_2^\infty(Y)} + \frac{e_{2,\nu}(\hat{g} - f_0^*)^{1-\frac{d}{2\alpha}} (\|\hat{g}\|^{\frac{d}{2\alpha}}_{H_{\alpha+1}(Y)} + 1)}{r}\right)(J_Y(\hat{g}) - J_Y(f_0^*)) \\ &+ r e_{2,\nu}(\hat{g} - f_0^*)^{2-\frac{d}{\alpha}} (\|\hat{g}\|^{\frac{d}{2\alpha}}_{H_{\alpha+1}(Y)} + 1)^2\,. \end{aligned} \tag{3.22}$$

2) Now let us upper bound the term $J_X(\hat{f}) - J_X(f_0)$ with high probability. This term can be re-written as

$$J_X(\hat{f}) - J_X(f_0) = (J_X(\hat{f}) - \mathcal{E}_{\mu,\nu}(\hat{f},\hat{g})) + (\mathcal{E}_{\mu,\nu}(\hat{f},\hat{g}) - \mathcal{E}_{\mu,\nu}(f_0,f_0^*))\,, \tag{3.23}$$

where we recall the notation $\mathcal{E}_{\mu,\nu}(f,g) = \langle f,\mu\rangle + \langle g,\nu\rangle$. Using Proposition 3.2, we have for all $0<\delta<1$ with probability at least $1-\delta$

$$J_X(\hat{f}) - \mathcal{E}_{\mu,\nu}(\hat{f},\hat{g}) \lesssim (n/\log(n/\delta))^{-(\alpha-1-d)/(2d)}\hat{R} + (n/\log(n/\delta))^{-1/(2d)}\frac{\sqrt{1+\lambda}}{\sqrt{\zeta}}\,, \tag{3.24}$$

where $\hat{R}$ is defined as $\hat{R} \coloneqq 1+\|\hat{f}\|_{H_{\alpha+1}(X)}+\|\hat{g}\|_{H_{\alpha+1}(Y)}+\|\hat{A}\|_F$. Now, let us introduce the empirical risk $\mathcal{E}_{\hat{\mu},\hat{\nu}}(\hat{f},\hat{g})$ in the second term as

$$\begin{aligned}\mathcal{E}_{\mu,\nu}(\hat{f},\hat{g}) - \mathcal{E}_{\mu,\nu}(f_0,f_0^*) = \mathcal{E}_{\hat{\mu},\hat{\nu}}(f_0,f_0^*) - \mathcal{E}_{\hat{\mu},\hat{\nu}}(\hat{f},\hat{g}) + \mathcal{E}_{\mu,\nu}(\hat{f},\hat{g}) - \mathcal{E}_{\mu,\nu}(f_0,f_0^*)\\ + \mathcal{E}_{\hat{\mu},\hat{\nu}}(\hat{f},\hat{g}) - \mathcal{E}_{\hat{\mu},\hat{\nu}}(f_0,f_0^*)\,.\end{aligned}$$

We upper-bound the last term difference $\mathcal{E}_{\hat{\mu},\hat{\nu}}(\hat{f},\hat{g}) - \mathcal{E}_{\hat{\mu},\hat{\nu}}(f_0,f_0^*)$ using the optimality of the empirical potentials $(\hat{f},\hat{g})$ as

$$\mathcal{E}_{\hat{\mu},\hat{\nu}}(\hat{f},\hat{g}) \le \mathcal{E}_{\hat{\mu},\hat{\nu}}(f_0,f_0^*) + \lambda(\|f_0\|^2_{H_{\alpha+1}(X)} + \|f_0^*\|^2_{H_{\alpha+1}(Y)} + \|A\|_F^2)\,, \tag{3.25}$$

where the operator A is defined in equation (3.1). The first terms can be re-written as

$$\mathcal{E}_{\hat{\mu},\hat{\nu}}(f_0,f_0^*) - \mathcal{E}_{\hat{\mu},\hat{\nu}}(\hat{f},\hat{g}) + \mathcal{E}_{\mu,\nu}(\hat{f},\hat{g}) - \mathcal{E}_{\mu,\nu}(f_0,f_0^*) = \langle f_0-\hat{f},\hat{\mu}-\mu\rangle + \langle f_0^*-\hat{g},\hat{\nu}-\nu\rangle\,.$$

Let us now denote $\Delta(\hat{f},\hat{g}) \coloneqq \langle f_0-\hat{f},\hat{\mu}-\mu\rangle + \langle f_0^*-\hat{g},\hat{\nu}-\nu\rangle$ and let us use Proposition 3.5 with $\beta=\alpha+1$ to upper bound with probability at least $1-2\delta$ the r.h.s. as

$$\begin{aligned}\Delta(\hat{f},\hat{g}) \lesssim \frac{\log(2/\delta)}{\epsilon\sqrt{n}}\Big[&e_{2,\mu}(\hat{f}-f_0)^{\frac{\alpha+1-d/2-\epsilon}{\alpha}}\|\hat{f}-f_0\|^{\frac{d/2+\epsilon-1}{\alpha}}_{H_{\alpha+1}(X)}\\ &+ e_{2,\nu}(\hat{g}-f_0^*)^{\frac{\alpha+1-d/2-\epsilon}{\alpha}}\|\hat{g}-f_0^*\|^{\frac{d/2+\epsilon-1}{\alpha}}_{H_{\alpha+1}(Y)}\Big]\,.\end{aligned} \tag{3.26}$$

Assuming that $\lambda \le 1$, we have with probability at least $1-2\delta$

$$\begin{aligned}J_X(\hat{f}) - J_X(f_0) \lesssim \lambda + \hat{R}(n/\log(n/\delta))^{-\frac{\alpha-1-d}{2d}} + (n/\log(n/\delta))^{-1/(2d)}\zeta^{-1/2}\\ + \frac{\log(2/\delta)}{\epsilon\sqrt{n}}e_{2,\mu}(\hat{f}-f_0)^{\frac{\alpha+1-d/2-\epsilon}{\alpha}}\|\hat{f}-f_0\|^{\frac{d/2+\epsilon-1}{\alpha}}_{H_{\alpha+1}(X)}\\ + \frac{\log(2/\delta)}{\epsilon\sqrt{n}}e_{2,\nu}(\hat{g}-f_0^*)^{\frac{\alpha+1-d/2-\epsilon}{\alpha}}\|\hat{g}-f_0^*\|^{\frac{d/2+\epsilon-1}{\alpha}}_{H_{\alpha+1}(Y)}\,,\end{aligned} \tag{3.27}$$

and a similar result holds for $J_Y(\hat{g}) - J_Y(f_0^*)$. Let us now merge together the results from the two previous paragraphs. Denoting $\hat{Z} = e^2_{2,\mu}(\hat{f}-f_0) + e^2_{2,\nu}(\hat{g}-f_0^*)$ and picking ζ such that $n^{-1/(2d)}\zeta^{-1/2} = \lambda$ yields that with probability at least $1-4\delta$ that

for $r \leq r_0$, we have

$$\hat{Z} \lesssim \left(\hat{R} + \frac{\hat{R}^{\frac{d}{2\alpha}} \hat{Z}^{1/2-\frac{d}{4\alpha}}}{r}\right)\left(\lambda \log(n/\delta)^{\frac{1}{2d}} + \hat{R}(n/\log(n/\delta))^{-\frac{\alpha-1-d}{2d}} + \frac{\log(2/\delta)}{\epsilon\sqrt{n}} \hat{Z}^{\frac{\alpha+1-d/2-\epsilon}{2\alpha}} \hat{R}^{\frac{d/2+\epsilon-1}{\alpha}}\right) + r\hat{Z}^{1-\frac{d}{2\alpha}} \hat{R}^{\frac{d}{2\alpha}} \tag{3.28}$$

Hence, provided that the quantity $\hat{R}$ remains bounded as $n \to \infty$, we can obtain the convergence of $\hat{Z}$ as n grows.

3) To ensure that $\hat{R}$ is bounded, recall that optimality conditions of estimator (2.10) imply that

$$\lambda \hat{R}^2 \leq \lambda(\|f_0\|^2_{H_{\alpha+1}(X)} + \|f_0^*\|^2_{H_{\alpha+1}(Y)} + \|A\|_F^2) + \mathcal{E}_{\hat{\mu},\hat{\nu}}(f_0, f_0^*) - \mathcal{E}_{\hat{\mu},\hat{\nu}}(\hat{f}, \hat{g}) \tag{3.29}$$

Then, we handle the term $\mathcal{E}_{\hat{\mu},\hat{\nu}}(f_0, f_0^*) - \mathcal{E}_{\hat{\mu},\hat{\nu}}(\hat{f}, \hat{g})$ in a similar fashion as in the previous paragraph. We decompose as

$$\mathcal{E}_{\hat{\mu},\hat{\nu}}(f_0, f_0^*) - \mathcal{E}_{\hat{\mu},\hat{\nu}}(\hat{f}, \hat{g}) = \mathcal{E}_{\mu,\nu}(f_0, f_0^*) - \mathcal{E}_{\mu,\nu}(\hat{f}, \hat{g}) + \Delta(\hat{f}, \hat{g})\,, \tag{3.30}$$

where we defined $\Delta(\hat{f}, \hat{g}) \coloneqq \mathcal{E}_{\hat{\mu},\hat{\nu}}(f_0, f_0^*) - \mathcal{E}_{\hat{\mu},\hat{\nu}}(\hat{f}, \hat{g}) - (\mathcal{E}_{\mu,\nu}(f_0, f_0^*) - \mathcal{E}_{\mu,\nu}(\hat{f}, \hat{g}))$. The first term can be re-written as

$$\mathcal{E}_{\mu,\nu}(f_0, f_0^*) - \mathcal{E}_{\mu,\nu}(\hat{f}, \hat{g}) = J_X(f_0) - J_X(\hat{f}) + (J_X(\hat{f}) - \mathcal{E}_{\mu,\nu}(\hat{f}, \hat{g}))\,. \tag{3.31}$$

The term $J_X(f_0) - J_X(\hat{f})$ is negative as f_0 is the minimizer of J_X and the term $J_X(\hat{f}) - \mathcal{E}_{\mu,\nu}(\hat{f}, \hat{g})$ is upper-bounded with probability $1-\delta$ using again Proposition 3.2 as

$$J_X(\hat{f}) - \mathcal{E}_{\mu,\nu}(\hat{f}, \hat{g}) \leq (n/\log(n/\delta))^{-(\alpha-1-d)/(2d)} \hat{R} + \lambda \log(n/\delta)^{\frac{1}{2d}}\,. \tag{3.32}$$

Conversely, remark that $\Delta(\hat{f}, \hat{g}) = \langle f_0 - \hat{f}, \hat{\mu} - \mu\rangle + \langle f_0^* - \hat{g}, \hat{\nu} - \nu\rangle$ hence we can use again Proposition 3.5 to recover with probability at least $1-2\delta$

$$\Delta(\hat{f}, \hat{g}) \leq \frac{\log(2/\delta)}{\epsilon\sqrt{n}}\left[\|\nabla \hat{f} - T_0\|^{\frac{\alpha+1-d/2-\epsilon}{\alpha}}_{L^2(\mu)} \|\hat{f} - f_0\|^{\frac{d/2+\epsilon-1}{\alpha}}_{H_{\alpha+1}(X)} + \|\nabla \hat{g} - T_0^{-1}\|^{\frac{\alpha+1-d/2-\epsilon}{\alpha}}_{L^2(\nu)} \|\hat{g} - f_0^*\|^{\frac{d/2+\epsilon-1}{\alpha}}_{H_{\alpha+1}(Y)}\right]$$

Hence we recover the upper-bound with probability at least $1-3\delta$

$$\lambda \hat{R}^2 \lesssim \lambda \log(n/\delta)^{\frac{1}{2d}} + (n/\log(n/\delta))^{-\frac{\alpha-1-d}{2d}} \hat{R} + \frac{\log(2/\delta)}{\epsilon\sqrt{n}} \hat{Z}^{\frac{\alpha+1-d/2-\epsilon}{2\alpha}} \hat{R}^{\frac{d/2+\epsilon-1}{\alpha}}\,. \tag{3.33}$$

In particular, if $\hat{Z}$ converges to zero sufficiently fast, we can ensure that $\hat{R}$ is bounded up to log factors.

4) If we combine the conclusions of the paragraphs 2) and 3), we come up with the following coupled upper-bounds on $\hat{Z}$ and $\hat{R}$

$$\begin{cases} \hat{Z} \lesssim \left(\hat{R} + \frac{\hat{R}^{\frac{d}{2\alpha}} \hat{Z}^{1/2-\frac{d}{4\alpha}}}{r}\right)\left(\lambda \log(n/\delta)^{\frac{1}{2d}} + \hat{R}(n/\log(n/\delta))^{-\frac{\alpha-1-d}{2d}} \right. \\ \qquad \left. + \frac{\log(2/\delta)}{\epsilon\sqrt{n}} \hat{Z}^{\frac{\alpha+1-d/2-\epsilon}{2\alpha}} \hat{R}^{\frac{d/2+\epsilon-1}{\alpha}}\right) + r\hat{Z}^{1-\frac{d}{2\alpha}} \hat{R}^{\frac{d}{2\alpha}}\,, \\ \lambda \hat{R}^2 \lesssim \lambda \log(n/\delta)^{\frac{1}{2d}} + (n/\log(n/\delta))^{-\frac{\alpha-1-d}{2d}} \hat{R} + \frac{\log(2/\delta)}{\epsilon\sqrt{n}} \hat{Z}^{\frac{\alpha+1-d/2-\epsilon}{2\alpha}} \hat{R}^{\frac{d/2+\epsilon-1}{\alpha}}\,. \end{cases}$$

Recovering the parameters λ, ϵ and r such that $\hat{R}$ is bounded and such that $\hat{Z}$ converges as fast as possible is now a purely algebraic problem and does not present much interest on its own. Hence we conclude with the following lemma whose proof is left in appendix.

LEMMA 3.7. *Let (a_n) and (b_n) be two positive sequences such that $b_n \geq 1$ $\forall n$ that verify*

$$\begin{cases} a_n \lesssim \left(b_n + \dfrac{b_n^{\frac{d}{2\alpha}} a_n^{1/2-\frac{d}{4\alpha}}}{r}\right)\left(\lambda \log(n/\delta)^{\frac{1}{2d}} + b_n (n/\log(n/\delta))^{-\frac{\alpha-1-d}{2d}} \right. \\ \qquad \left. + \dfrac{\log(2/\delta)}{\epsilon\sqrt{n}} a_n^{\frac{\alpha+1-d/2-\epsilon}{2\alpha}} b_n^{\frac{d/2+\epsilon-1}{\alpha}}\right) + r a_n^{1-\frac{d}{2\alpha}} b_n^{\frac{d}{2\alpha}}\,, \\ \lambda b_n^2 \lesssim \lambda \log(n/\delta)^{\frac{1}{2d}} + (n/\log(n/\delta))^{-\frac{\alpha-1-d}{2d}} b_n + \dfrac{\log(2/\delta)}{\epsilon\sqrt{n}} a_n^{\frac{\alpha+1-d/2-\epsilon}{2\alpha}} b_n^{\frac{d/2+\epsilon-1}{\alpha}}\,, \end{cases}$$

for any $\epsilon, \lambda \leq 1$ and $r \leq r_0$. Choosing $\lambda_n = n^{-\min(\frac{\alpha}{\alpha+d/2-1}, -\frac{\alpha-1-d}{2d})}$, $r = r_0 \lambda_n^{\frac{3d}{2\alpha+3d}}$ and $\epsilon_n = 1/\log(n)$ then one recovers

$$a_n \lesssim \lambda_n^{\min(1, \frac{2}{1+3d/(2\alpha)})} \Big(\log(n)\log(2/\delta)\Big)^{\max(6, \frac{2\alpha+d}{d-1})} := c_n\,. \tag{3.34}$$

Applying the previous lemma to $a_n = \hat{Z}$ and $b_n = \hat{R}$ with the appropriate parameters yield that with probability at least $1-4\delta$ we have $\hat{Z} \lesssim c_n$ where c_n is defined in (3.34). There remains to obtain this result in expectation. Recall the optimality conditions that holds with probability 1

$$\lambda_n \hat{R}^2 \leq \mathcal{E}_{\mu,\nu}(f_0, f_0^*) + \lambda_n(\|f_0\|^2_{H_{\alpha+1}(X)} \|f_0^*\|^2_{H_{\alpha+1}(Y)} + \|A\|_F^2)\,. \tag{3.35}$$

In particular $\hat{R}^2 \lesssim \lambda_n^{-1}$ and, recalling that $\hat{Z} = e_\mu(\hat{f} - f_0)^2 + e_\mu(\hat{g} - f_0^*)^2 \lesssim \|\hat{f}\|^2_{H_{\alpha+1}(X)} + \|\hat{g}\|^2_{H_{\alpha+1}(Y)} + 1$, we have *a fortiori* $\hat{Z} \lesssim \lambda_n^{-1}$. Let us now upper-bound $\hat{Z}$ in expectation: for any $0 < \delta < 1$, it holds

$$\mathbb{E}[\hat{Z}] \lesssim \mathbb{E}[\hat{Z}|\hat{Z} \lesssim c_n]\mathbb{P}(\hat{Z} \lesssim c_n) + \mathbb{E}[\hat{Z}|\hat{Z} \gtrsim c_n]\mathbb{P}(\hat{Z} \gtrsim c_n)\,.$$

Using the results above, we have that $\mathbb{E}[\hat{Z}|\hat{Z} \gtrsim c_n]\mathbb{P}(\hat{Z} \gtrsim c_n) \lesssim 4\delta\lambda_n^{-1}$. Hence, it suffices to take $\delta_n = 1/n$ to recover up to poly-log factors

$$\mathbb{E}[\hat{Z}] \lesssim \lambda_n^{\min(1, \frac{2}{1+3d/(2\alpha)})}\,. \tag{3.36}$$

∎

4. Finite reformulation and associated complexity. We proved in the previous section that our estimator (2.10) was statistically optimal for well-chosen regularizers (λ, ζ) yet there remains to prove that our estimator is indeed tractable. At first glimpse, as problem (2.10) optimizes over infinite dimensional objects, it is not clear that our estimator can be computed numerically. Yet as detailed in the introduction and Section 2, standard and SoS kernel tricks apply; the following proposition shows that the objects involved in problem (2.10) admit a finite re-parametrization at the optimum.

PROPOSITION 4.1. *Denoting $(\hat{f}, \hat{g}, \hat{A})$ solutions of (2.10), there exists coefficients $(f_i)_{1\leq i\leq n}$, $(g_j)_{1\leq j\leq n}$ and a positive symmetric matrix $(a_{ij})_{1\leq i,j\leq n}$ such that*

$$\begin{cases} \hat{f}(\cdot) = \sum_{i=1}^n f_i k_X(x_i, \cdot) \\ \hat{g}(\cdot) = \sum_{j=1}^n g_j k_Y(y_j, \cdot) \\ \hat{A} = \sum_{i,j=1}^n a_{ij} k_{XY}((x_i, y_i), \cdot) \otimes k_{XY}((x_j, y_j), \cdot)\,, \end{cases} \tag{4.1}$$

where k_X (resp. k_Y) is the kernel of $H_{\alpha+1}(X)$ (resp. $H_{\alpha+1}(Y)$), k_{XY} is the kernel of $H_{\alpha-1}(X \times Y)$ whose explicit forms can be found in [37, Proposition 4] and where

$$k_{XY}((x_i, y_i), \cdot) \otimes k_{XY}((x_j, y_j), \cdot) : h \in H_{\alpha-1}(X \times Y) \mapsto h((x_j, y_j))k_{XY}((x_i, y_i), \cdot)\,.$$

Proof. Fixing the potentials (f, g), our problem (2.10) becomes

$$\inf_{A \in \mathcal{S}_+(H_{\alpha-1}(X\times Y))} \langle f, \hat{\mu}_n\rangle + \langle g, \hat{\nu}_n\rangle + \lambda\left(\|A\|_F^2 + \|f\|^2_{H_{\alpha+1}(X)} + \|g\|^2_{H_{\alpha+1}(Y)}\right) + \zeta \sum_l \left(f(x_l) + g(y_l) - x_l^\top y_l - \gamma_A(x_l, y_l)\right)^2 .$$

Using [25, Theorem 1] immediately yields that at the optimum,

$$\hat{A} = \sum_{i,j=1}^n a_{ij} k_{XY}((x_i, y_i), \cdot) \otimes k_{XY}((x_j, y_j), \cdot)\,,$$

with (a_{ij}) a positive symmetric matrix. The finite re-parametrization on $\hat{f}$ and $\hat{g}$ simply follows from the standard kernel trick. ∎

When we plug this finite re-parametrization in our estimator (2.10), we obtain a finite convex problem that we can provably solve in polynomial time. However, left under this form, the problem is ill-suited to practical resolution as it is constrained over the cone of p.s.d. matrices. Instead, we derive the dual formulation of the problem which is unconstrained.

PROPOSITION 4.2. *Problem (2.10) is equivalent to*

$$-\inf_{u \in \mathbb{R}^n} \frac{1}{4\lambda} u^T Q u + \frac{1}{4\lambda}\|(-\sum_{j=1}^n u_i \Phi_j \Phi_j^T)_+\|_F^2 + \frac{1}{4\zeta}\|u\|^2 - \frac{1}{2\lambda}\sum_{j=1}^n u_j z_j + \frac{q^2}{4\lambda}\,, \tag{4.2}$$

where $(\cdot)_+$ is the projection operator on the p.s.d. cone, $Q = (k_X(x_i, x_j) + k_Y(y_i, y_j))_{1\le i,j\le n}$, $z_j = 2\lambda x_j^\top y_j + \frac{1}{n}[Q\mathbf{1}]_j$, $q^2 = \frac{1}{n^2}\mathbf{1}^\top Q \mathbf{1}$ and where Φ_j is the j-th column of $K_{XY}^{1/2}$ with K_{XY} given by $(k_{XY}((x_i, y_i), (x_j, y_j)))_{1\le i,j\le n}$. Furthermore, the following primal-dual relations holds (at the optimum)

$$\begin{cases} \hat{f} &= \frac{1}{2\lambda}\sum_{i=1}^n (u_i - \frac{1}{n}) k_X(x_i, \cdot)\,, \\ \hat{g} &= \frac{1}{2\lambda}\sum_{i=1}^n (u_i - \frac{1}{n}) k_Y(y_i, \cdot)\,. \end{cases} \tag{4.3}$$

Proof. We start by plugging the finite dimensional re-parametrization in problem (2.10) which becomes

$$\inf_{\substack{(f,g)\in\mathbb{R}^n \\ A \in \mathbb{S}_+(\mathbb{R}^n)}} \frac{1}{n}\mathbf{1}^\top K_X f + \frac{1}{n}\mathbf{1}^\top K_Y g + \lambda\left(\|A K_{XY}\|_F^2 + f^\top K_X f + g^\top K_Y g\right) + \zeta\|K_X f + K_Y g - \mathrm{Diag}(XY^\top) - \mathrm{Diag}(K_{XY} A K_{XY})\|^2\,, \tag{4.4}$$

where $K_{XY} = (k_{XY}((x_i, y_i), (x_j, y_j)))_{i,j}$, $K_X = (k_X(x_i, x_j))_{i,j}$, $K_Y = (k_Y(y_i, y_j))_{i,j}$, $X = (x_i)_i \in \mathbb{R}^{n\times d}$, $Y = (y_i)_i \in \mathbb{R}^{n\times d}$ and where $\mathrm{Diag}(\cdot)$ stands for the diagonal

of the matrix and $\mathbf{1}$ stands for the vector of ones. Making the change of variable $B = K_{XY}^{1/2} A K_{XY}^{1/2}$ yields

$$\inf_{\substack{(f,g)\in\mathbb{R}^n \\ B\in\mathbb{S}_+(\mathbb{R}^n)}} \frac{1}{n}\mathbf{1}^\top K_X f + \frac{1}{n}\mathbf{1}^\top K_Y g + \lambda\bigg(\|B\|_F^2 + f^\top K_X f + g^\top K_Y g\bigg) + \zeta\|K_X f + K_Y g - \mathrm{Diag}(XY^\top) - \mathrm{Diag}(K_{XY}^{1/2} B K_{XY}^{1/2})\|^2 \,. \tag{4.5}$$

Defining the function $\Theta : u \in \mathbb{R}^n \mapsto \zeta\|u - \mathrm{Diag}(XY^\top)\|^2$, the operator $O : (f, g, B) \mapsto K_X f + K_Y g - \mathrm{Diag}(K_{XY}^{1/2} B K_{XY}^{1/2})$ and

$$\Omega : (f,g,B) \mapsto \frac{1}{n}\mathbf{1}^\top K_X f + \frac{1}{n}\mathbf{1}^\top K_Y g + \lambda\bigg(\|B\|_F^2 + f^\top K_X f + g^\top K_Y g\bigg) + \iota(B \in \mathbb{S}_+(\mathbb{R}^n)) \,,$$

the previous formulation reads with these notations

$$\inf_{\substack{(f,g)\in\mathbb{R}^n \\ B\in\mathbb{R}^{n\times n}}} \Omega((f,g,B)) + \Theta(O(f,g,B)) = \sup_{u\in\mathbb{R}^n} -\Omega^*(O^* u) - \Theta^*(-u) \,, \tag{4.6}$$

where the equality comes from the Fenchel-Rockafellar theorem [30] and Ω^* and Θ^* are the convex conjugate of Ω and Θ respectively and O^* is the adjoint of O. Furthermore, at the optimum we have $O^* u \in \partial\Omega((f,g,B))$ where $\partial\Omega$ is the subgradient of Ω. Let us compute explicitly the convex conjugates and the adjoint: as a quadratic function, the conjugate of Θ reads $\Theta^*(u) = \frac{\|u\|^2}{4\zeta} + u^\top \mathrm{Diag}(XY^\top)$. Let us compute the conjugate of $\eta : B \mapsto \lambda\|B\|_F^2 + \iota(B \in \mathbb{S}_+(\mathbb{R}^n))$

$$\begin{aligned}
\eta^*(B) &= \sup_{A\in\mathbb{S}_+(\mathbb{R}^n)} \langle A, B\rangle_F - \lambda\|A\|_F^2 \\
&= -\lambda \inf_{A\in\mathbb{S}_+(\mathbb{R}^n)} -\langle A, \frac{B}{\lambda}\rangle_F + \|A\|_F^2 \\
&= -\lambda \inf_{A\in\mathbb{S}_+(\mathbb{R}^n)} \|A - \frac{B}{2\lambda}\|_F^2 - \frac{\|B\|_F^2}{4\lambda^2} \\
&= \frac{\|B\|_F^2}{4\lambda} - \frac{1}{4\lambda}\inf_{A\in\mathbb{S}_+(\mathbb{R}^n)} \|2\lambda A - B\|_F^2 = \frac{1}{4\lambda}(\|B\|_F^2 - \|B_+ - B\|_F^2) \,,
\end{aligned}$$

where $(\cdot)_+$ is the projection on the p.s.d. cone with respect to the Froebenius norm. Now recall that for any $A \in \mathbb{S}_+(\mathbb{R}^n)$, we have $\langle A - B_+, B - B_+\rangle_F$ [8, Lemma 3.1] ; in particular, for $A = 0$ and $A = 2B_+$, we recover that $\langle B_+, B - B_+\rangle_F = 0$ and as a consequence $\|B\|_F^2 - \|B_+ - B\|_F^2 = \|B_+\|_F^2$. Being the independent sum of $\eta(\cdot)$ and two quadratic functions, we deduce that

$$\Omega^*((f,g,B)) = \frac{f^\top K_X^{-1} f}{4\lambda} - \frac{\mathbf{1}^\top f}{2n\lambda} + \frac{\mathbf{1}^\top K_X \mathbf{1}}{4n^2\lambda} + \frac{g^\top K_Y^{-1} g}{4\lambda} - \frac{\mathbf{1}^\top g}{2n\lambda} + \frac{\mathbf{1}^\top K_Y \mathbf{1}}{4n^2\lambda} + \frac{1}{4\lambda}\|B_+\|_F^2 \,.$$

Finally, one can check that $O^* u = (K_X u, K_Y u, -\sum_{j=1}^n u_j \Phi_j \Phi_j^\top)$ where the Φ_j are the j-column of $K_{XY}^{1/2}$. Plugging this formula in the conjugates yields

$$\begin{aligned}
-\Omega^*(O^* u) - \Theta^*(-u) = &- \frac{u^\top(K_X + K_Y)u}{4\lambda} - \frac{\|(-\sum_{j=1}^n u_j \Phi_j \Phi_j^\top)_+\|_F^2}{4\lambda} - \frac{\|u\|^2}{4\zeta} \\
&+ \frac{\mathbf{1}^\top(K_X + K_Y)u}{2n\lambda} + u^\top \mathrm{Diag}(XY^\top) - \frac{\mathbf{1}^\top(K_X + K_Y)\mathbf{1}}{n^2} \,.
\end{aligned}$$

Finally, the optimality condition $O^*u \in \partial\Omega((f,g,B))$ on the first two variables yields $K_X u = \frac{1}{n}\mathbf{1}^\top K_X + 2\lambda f^\top K_X$ and $K_Y u = \frac{1}{n}\mathbf{1}^\top K_Y + 2\lambda g^\top K_Y$ which is equivalent to $f = \frac{1}{2\lambda}\sum_{i=1}^n u_i - \frac{1}{n}$ and $g = \frac{1}{2\lambda}\sum_{i=1}^n u_i - \frac{1}{n}$ which eventually yields on the potentials themselves $\hat{f} = \frac{1}{2\lambda}\sum_{i=1}^n (u_i - \frac{1}{n})k_X(x_i,\cdot)$ and $\hat{g} = \frac{1}{2\lambda}\sum_{i=1}^n (u_i - \frac{1}{n})k_Y(y_i,\cdot)$. ∎

Now that we have a finite unconstrained convex reformulation of our estimator, we can derive its associated computational complexity. We chose an accelerated gradient descent scheme to solve (4.2), which is known to be the most efficient first order method for convex problems [8]. To recover its associated complexity, we simply need to compute the condition number of the problem which is given by the ratio of the smoothness of the objective divided by its strong-convexity constant.

PROPOSITION 4.3. *The condition number of problem (4.2) is given by* $\kappa := 1 + \frac{\zeta}{\lambda}(\xi_{max}(Q) + \|Diag(K_{XY})\|_2^2)$ *where* $\xi_{max}(Q)$ *is the squared largest singular value of* Q*. As a consequence, solving problem (4.2) with an accelerated gradient method for a* τ *precision can be done in* $O(\log(1/\tau)\sqrt{\kappa})$ *steps where each step costs* $O(n^3)$*.*

Proof. As a sum of convex terms and of $u \mapsto \frac{1}{4\zeta}\|u\|^2$, we immediately have that the objective is at least $\frac{1}{2\zeta}$-strongly convex. In order to compute the smoothness, we need to compute the smoothness of $h : u \mapsto \|(-\sum_{j=1}^n u_j \Phi_j \Phi_j^\top)_+\|_F^2$. Note that h can be decomposed as $h = \psi \circ \phi$ where $\phi(u) = -\sum_{j=1}^n u_j \Phi_j \Phi_j^\top$ and where $\psi(B) = \|B_+\|_F^2$, the squared Froebenius norm of the projection on the p.s.d. cone with respect to the Froebenius norm. By the chain rule, for any (u,v) in $\mathbb{R}^n$ it holds

$$\begin{aligned}
\|\nabla h(u) - \nabla h(v)\|_2^2 &= \|\mathrm{Jac}(\phi)(u)\nabla\psi(\phi(u)) - \mathrm{Jac}(\phi)(v)\nabla\psi(\phi(v))\|_2^2 \\
&= \sum_{j=1}^n \langle \Phi_j\Phi_j^\top, \nabla\psi(\phi(u)) - \nabla\psi(\phi(v))\rangle_F^2 \\
&\leq \sum_{j=1}^n \xi_{\max}(\nabla\psi(\phi(u)) - \nabla\psi(\phi(v)))^2 \|\Phi_j\|_2^4 \\
&\leq (\sum_{j=1}^n \|\Phi_j\|_2^4)\|\nabla\psi(\phi(u)) - \nabla\psi(\phi(v))\|_F^2 .
\end{aligned}$$

Now recall that ψ is 2-smooth with respect to the Froebenius norm [20, Equation (1.2)], so we can deduce

$$\begin{aligned}
\|\nabla h(u) - \nabla h(v)\|_2 &\leq 2(\sum_{j=1}^n \|\Phi_j\|_2^4)^{1/2}\|\phi(u) - \phi(v)\|_F \\
&= 2(\sum_{j=1}^n \|\Phi_j\|_2^4)^{1/2}\|\sum_{j=1}^n (u_j - v_j)\Phi_j\Phi_j^\top\|_F \\
&\leq 2(\sum_{j=1}^n \|\Phi_j\|_2^4)^{1/2}\sum_{j=1}^n |u_j - v_j|\|\Phi_j\Phi_j^\top\|_F \\
&\leq 2\|u-v\|_2(\sum_{j=1}^n \|\Phi_j\|_2^4)^{1/2}(\sum_{j=1}^n \|\Phi_j\Phi_j^\top\|_F^2)^{1/2} \\
&= 2\|u-v\|_2(\sum_{j=1}^n \|\Phi_j\|_2^4) .
\end{aligned}$$

By definition of Φ_j as the j-th column of $K_{XY}^{1/2}$, it reads $\|\Phi_j\|_2^4 = (K_{XY})_{jj}^2$ hence we recover that h is $2\|\mathrm{Diag}(K_{XY})\|_2^2$-smooth and by extension that the objective of (4.2) is $(\frac{1}{2\zeta} + \frac{1}{2\lambda}(\xi_{\max}(Q) + \|\mathrm{Diag}(K_{XY})\|_2^2))$-smooth. In particular, the condition number of objective (4.2) is given by $\kappa = 1 + \frac{\zeta}{\lambda}(\xi_{\max}(Q) + \|\mathrm{Diag}(K_{XY})\|_2^2)$. Using [8, Theorem 3.18], we deduce that $O(\sqrt{\kappa}\log(1/\tau))$ steps of accelerated gradient descent are required to solve (4.2) with a τ precision. Finally, each step of the accelerated gradient descent only involves computing the gradient of the objective (4.2) whose complexity is dominated by the computation of the gradient of the function h defined above. By the chain rule, recall that the gradient of h is given by $\nabla h(u) = \mathrm{Jac}(\phi)(u)\nabla\psi(\phi(u))$ where ψ and ϕ are also defined above. Using the previous computations, recall that

$$\begin{aligned}\mathrm{Jac}(\phi)(u)\nabla\psi(\phi(u)) &= (\langle \Phi_j\Phi_j^\top, \nabla\psi(\phi(u))\rangle_F)_{1\le j\le n}\\ &= (\Phi_j^\top\nabla\psi(\phi(u))\Phi_j)_{1\le j\le n}\,.\end{aligned}$$

The complexity of evaluating $\nabla\psi(\phi(u))\Phi_j$ scales as $O(n^2)$ and has to be done n-times which leads to an overall complexity of $O(n^3)$ plus the complexity to compute $\nabla\psi(\phi(u))$. Using [20, Equation (1.2)], the gradient of ψ at some matrix B is given by $\nabla\psi(B) = 2B_+$ where B_+ is the projection of B with respect to the Froebenius norm on the p.s.d. cone. If B is symmetric, this projection is obtained by computing the spectral decomposition of B as $B = \sum_{j=1}^n \lambda_j u_j u_j^\top$ and by cropping the negative eigenvalues $B_+ = \sum_{j=1}^n \max(0,\lambda_j) u_j u_j^\top$ [5, Section 8.1.1]. Since the spectral decomposition of an $\mathbb{R}^{n\times n}$ matrix scales as $O(n^3)$, we recover a total complexity of $O(n^3)$ per gradient step. ∎

Let us give a worst case bound on κ so we have a fully explicit complexity with respect to the number of samples n. First, since $k_{XY}((x,y),(x,y)) = 1$ [41], we exactly have $\|\mathrm{Diag}(K_{XY})\|_2^2 = n$. Furthermore, since $\xi_{\max}(Q) \le \mathrm{Tr}(Q)$, we have $\xi_{\max}(Q) + \|\mathrm{Diag}(K_{XY})\|_2^2 \le 3n$. There remains to bound the ratio $\frac{\zeta}{\lambda}$: if one picks the values of λ and ζ indicated in Theorem 3.6, this ratio depends on the smoothness parameter α. If $\alpha \to d+2$, we have $\lambda \sim n^{-1/2d}$ and $\zeta \sim 1$ which give a total complexity of $O(n^{3.5+1/(4d)}\log(1/\tau))$ to reach a τ-precision. On the other hand in the highly smooth regime $\alpha \to \infty$, we have $\lambda \sim 1/n$ and $\zeta \sim n^2$ which give a total complexity of $O(n^5\log(1/\tau))$ to reach a τ-precision ; in particular, we do reach a polynomial dimension-free worst case complexity. However, while the statistical rates improve with the smoothness, the computational complexity on the contrary degrades with the smoothness. We believe we could have avoided this poor dependence on the smoothness by resorting to Newton-like methods as done in [37]. However, we chose to use a first order method as it is easier to implement and scales slightly better in practice.

5. Nyström approximation and numerical simulations. In this section, we present numerical simulations of our estimator when applied to simple examples of smooth optimal transport problems in medium dimensions. Yet beforehand, we used a Nyström approximation strategy in order to reduce the $O(n^3)$ per step complexity previously found. We highlight that we do not provide any theoretical guarantees for this heuristic whose main goal was to allow for simulations to run with $n \sim 10^3$ samples ; without this heuristic, we hardly were able to run the accelerated gradient for $n \sim 10^2$.

5.1. Nyström approximation. As showcased in the proof of Proposition 4.3, the main computational bottleneck of the gradient descent is the computation of

the SVD of the matrix $-\sum_{j=1}^n u_j \Phi_j \Phi_j^\top$ followed by the computation of the scalars $\Phi_j^\top(-\sum_{j=1}^n u_j \Phi_j \Phi_j^\top)_+ \Phi_j$ for a given $u \in \mathbb{R}^n$. This complexity can be reduced by replacing the kernel matrix K_{XY} that appears in Equation (4.5) by its Nyström approximation [43] $\tilde{K} = K_{XY}^{n,r}(K_{XY}^{r,r})^{-1}(K_{XY}^{n,r})^\top$ where $r << n$ and $K_{XY}^{r,r}$ is an $r \times r$ matrix randomly extracted from K_{XY} and $K_{XY}^{n,r}$ is its corresponding $n \times r$ matrix. A square root of $\tilde{K}$ is given by $K_{XY}^{n,r}(K_{XY}^{r,r})^{-1/2} \in \mathbb{R}^{n\times r}$ which amounts in replacing the $(\Phi_j)_{1\le j\le n}$ in problem (4.2) by the $(\tilde{\Phi}_j)_{1\le j\le n}$, the n rows of the matrix $K_{XY}^{n,r}(K_{XY}^{r,r})^{-1/2}$. Hence the approximated problem to solve is now given by

$$-\inf_{u\in\mathbb{R}^n} \frac{1}{4\lambda}u^T Q u + \frac{1}{4\lambda}\|(-\sum_{j=1}^n u_i \tilde{\Phi}_j \tilde{\Phi}_j^T)_+\|_F^2 + \frac{1}{4\zeta}\|u\|^2 - \frac{1}{2\lambda}\sum_{j=1}^n u_j z_j + \frac{q^2}{4\lambda}\,. \tag{5.1}$$

The cost of forming the matrix $-\sum_{j=1}^n u_j \tilde{\Phi}_j \tilde{\Phi}_j^\top$ is now given by $O(nr^2)$, the cost to compute its SVD is $O(r^3)$ and the cost to compute the scalars $\tilde{\Phi}_j^\top(-\sum_{j=1}^n u_j \tilde{\Phi}_j \tilde{\Phi}_j^\top)_+ \tilde{\Phi}_j$ is $O(nr^2)$. Hence the total cost per gradient step is reduced to $O(n^2 + nr^2)$ where the n^2 term comes from the computation of the vector Qu.

5.2. Synthetic experiments. We describe in this paragraph the setting of our numerical experiments. We chose μ and ν to be centered Gaussians distributions whose covariance matrices C_μ, C_ν were drawn from a Wishart distribution with parameters I_d, d where I_d is the identity matrix of size d ; for this choice of distributions μ, ν the OT map T_0 has a closed form [36]. Then we drawn $n_{\text{train}}, n_{\text{test}}, n_{\text{valid}}$ samples from μ and ν respectively and we solved problem (5.1) with parameters $\zeta = 10^3$, $r = 100$ and Sobolev kernels H_s with $s = 20$. The parameter λ was taken in $\{10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$ and for each value of λ, we solved problem (5.1) on the training samples and recovered transport maps $\hat{T}_1^\lambda, \hat{T}_2^\lambda$ from μ to ν and ν to μ respectively. Then the parameter λ was chosen according to the following heuristic: we picked the value of λ that minimized the sum $\text{MMD}(\hat{T}_1^\lambda(\hat{\mu}_{\text{test}}), \hat{\nu}_{\text{test}})^2 + \text{MMD}(\hat{T}_2^\lambda(\hat{\nu}_{\text{test}}), \hat{\mu}_{\text{test}})^2$ where $\text{MMD}(\cdot)$ is the maximum mean discrepancy [18] with respect to the RKHS H_s. Finally, after selecting λ, we reported the empirical error $\text{MSE} = \|\hat{T}_1^\lambda - T_0\|^2_{L^2(\hat{\mu}_{\text{valid}})} + \|\hat{T}_2^\lambda - T_0^{-1}\|^2_{L^2(\hat{\nu}_{\text{valid}})}$. The values of $n_{\text{test}}, n_{\text{valid}}$ were both fixed to 1000 while the value of n_{train} ranged from 200 to 1000. This experiment was carried in dimensions $d = 2, 4, 8$.

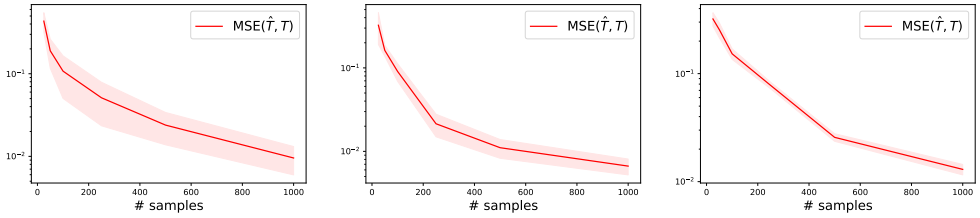

FIG. 1. *Transportation map mean square error (log scale) in dimensions* 2, 4 *and* 8 *respectively. Shaded area correspond to the standard deviation.*

The results of the experiments are reported on Figure 1. In all cases, the error does decrease with the number of samples yet it is unclear how the dimension affects the convergence rate a.k.a. the slope of the curves displayed in Figure 1. One possible explanation is that the rates of convergence only hold for n sufficiently large (see

Proposition 3.2). We plan to investigate an implementation that could run at scale for future works and hopefully observe more accurately the theoretical convergence rates stated in Theorem 3.6 as well as the effect of the dimension.

Appendix A. Additional results.

A.1. Proof of Proposition 3.1. The existence of $\hat{A}$ is ensured by [25, Proposition 7]. Let $(\hat{f}_k, \hat{g}_k)$ be a minimizing sequence of problem (2.10). Denoting (f, g) the minimizers of problem (1.3), the optimal transport potentials from μ to ν and from ν to μ respectively, and A defined in Theorem 2.3, we have for k sufficiently large that

$$\lambda \hat{R}_k^2 \leq 2(\langle f, \mu\rangle + \langle g, \nu\rangle + \lambda R^2)\,,$$

where $R^2 = \|f\|^2_{H_{\alpha+1}(X)} + \|g\|^2_{H_{\alpha+1}(Y)} + \|A\|_F^2$ and $\hat{R}_k^2 = \|\hat{f}_k\|^2_{H_{\alpha+1}(X)} + \|\hat{g}_k\|^2_{H_{\alpha+1}(Y)} + \|\hat{A}\|_F^2$. Using Kakutani theorem, we can extract from $(\hat{f}_k)$ and $(\hat{g}_k)$ weakly convergent sequences such that $\hat{f}_k \rightharpoonup \hat{f} \in H_{\alpha+1}(X)$ and $\hat{g}_k \rightharpoonup \hat{g} \in H_{\alpha+1}(Y)$. Since the integration w.r.t. μ and ν are continuous linear forms, we have in particular $\lim_k \langle \hat{f}_k, \mu\rangle \to \langle \hat{f}, \mu\rangle$ (resp. $\lim_k \langle \hat{g}_k, \mu\rangle \to \langle \hat{g}, \mu\rangle$). This implies

$$\langle \hat{f}, \mu\rangle + \langle \hat{g}, \nu\rangle + \lambda \hat{R}^2 \leq \lim_k \langle \hat{f}_k, \mu\rangle + \langle \hat{g}_k, \nu\rangle + \lambda \hat{R}_k^2\,, \tag{A.1}$$

where $\hat{R}^2 = \|\hat{f}\|^2_{H_{\alpha+1}(X)} + \|\hat{g}\|^2_{H_{\alpha+1}(Y)} + \|\hat{A}\|_F^2$.

A.2. Proof of Lemma 3.4.

Proof. By definition of a Lipschitz domain, there exists a radius $r_0 > 0$, centers $(p_i)_{i=1}^k$, radii $(r_i)_{i=1}^k$ and bi-Lipschitz bijective functions (h_i) from $B_{r_i}(p_i)$ to $B_1(0)$ that verify $h_i(X \cap B_{r_i}(p_i)) = Q_+$ and $h_i(\partial X \cap B_{r_i}(p_i)) = Q_0$ such that

$$\mathbb{R}^d \setminus A_{r_0} \subset \cup_{i=1}^k B_{r_i/3}(p_i)\,.$$

In particular, for all $r \leq r_0$, we have $\mathbb{R}^d \setminus A_r \subset \cup_{i=1}^k B_{r_i}(p_i)$. In the rest of the proof we shall denote by G_r the set $\mathbb{R}^d \setminus A_r$. Let us assume that $r \leq \min_{0\leq i\leq k} r_i$ and let $x \in G_r$. Since the boundary is compact, there exists $p \in \partial X$ that realizes the infimum and such that $d(x, \partial X) = \|x - p\| \leq r$. Furthermore there exists an index i such that $x \in B_{r_i/3}(p_i)$ and in particular, $\|p_i - p\| \leq \|p_i - x\| + \|x - p_i\| < (2r_i)/3$ which proves that p also lies in $B_{r_i}(p_i)$. Hence, we can apply the surjectivity of h_i to recover $z(x) \in Q_+$ and $z(p) \in Q_0$ such that $\|x - p\| = \|h_i^{-1}(z(x)) - h_i^{-1}(z(p))\|$. Since h_i^{-1} is also bi-Lipschitz, there exists L_i independent of r such that $\|z(x) - z(p)\| \leq rL_i$. This proves that $x \in h_i^{-1}(\tilde{G}_{L_i r})$ where $\tilde{G}_r$ is defined as

$$\tilde{G}_r = \{z \in Q_+ \;|d(z, Q_0) \leq r\}\,. \tag{A.2}$$

Using this result, we can now upper bound $\mu(G_r)$ using an union bound and the change of variable theorem.

$$\begin{aligned}
\mu(G_r) &= \int_{G_r} d\mu(x) \\
&\leq \rho_0 \sum_{i=1}^k \int_{h_i^{-1}(\tilde{G}_{L_i r})} dx \\
&\leq \rho_0 \sum_{i=1}^k \int_{\tilde{G}_{L_i r}} \|\det(Dh_i)\|_{W_0^\infty(B_{r_i}(p_i))} dx\,.
\end{aligned}$$

There remains to upper-bound $\int_{\tilde{G}_{L_i r}} dx = \mathrm{Vol}(\tilde{G}_{L_i r})$. The volume of $\tilde{G}_{L_i r}$ can be computed explicitly: it is the volume of the northern hemisphere of a ball of radius 1 minus the volume of the northern hemisphere of a ball of radius $1 - L_i r$. Hence we get

$$\mathrm{Vol}(\tilde{G}_{L_i r}) \lesssim 1 - (1 - L_i r)^d \tag{A.3}$$
$$\lesssim L_i d r\,, \tag{A.4}$$

and we recover $\mu(G_r) \lesssim r$. ∎

A.3. Proof of Lemma 3.7. Recall our two coupled upper-bounds

$$\begin{cases} a_n \lesssim \left(b_n + \dfrac{b_n^{\frac{d}{2\alpha}} a_n^{1/2 - \frac{d}{4\alpha}}}{r}\right)\left(\lambda \log(n/\delta)^{\frac{1}{2d}} + b_n (n/\log(n/\delta))^{-\frac{\alpha-1-d}{2d}} \right. \\ \qquad \left. + \dfrac{\log(2/\delta)}{\epsilon\sqrt{n}} a_n^{\frac{\alpha+1-d/2-\epsilon}{2\alpha}} b_n^{\frac{d/2+\epsilon-1}{\alpha}}\right) + r a_n^{1-\frac{d}{2\alpha}} b_n^{\frac{d}{2\alpha}}\,, \\ \lambda b_n^2 \lesssim \lambda \log(n/\delta)^{\frac{1}{2d}} + (n/\log(n/\delta))^{-\frac{\alpha-1-d}{2d}} b_n + \dfrac{\log(2/\delta)}{\epsilon\sqrt{n}} a_n^{\frac{\alpha+1-d/2-\epsilon}{2\alpha}} b_n^{\frac{d/2+\epsilon-1}{\alpha}}\,, \end{cases}$$

Defining $c_n \coloneqq \lambda \log(n/\delta)^{\frac{1}{2d}} + (n/\log(n/\delta))^{-\frac{\alpha-1-d}{2d}} b_n + \frac{\log(2/\delta)}{\epsilon\sqrt{n}} a_n^{\frac{\alpha+1-d/2-\epsilon}{2\alpha}} b_n^{\frac{d/2+\epsilon-1}{\alpha}}$, we notice that our upper-bounds can be re written as

$$\begin{cases} a_n \lesssim c_n b_n + c_n \dfrac{b_n^{\frac{d}{2\alpha}} a_n^{1/2-\frac{d}{4\alpha}}}{r} + r a_n^{1-\frac{d}{2\alpha}} b_n^{\frac{d}{2\alpha}}\,, \\ \lambda b_n^2 \lesssim c_n\,. \end{cases} \tag{A.5}$$

We observe that the upper-bound on a_n is composed of three terms. Hence we shall split our analysis in three cases: the case where the term $c_n b_n$ dominates, the case where $c_n \frac{b_n^{\frac{d}{2\alpha}} a_n^{1/2-\frac{d}{4\alpha}}}{r}$ dominates and the case where $r a_n^{1-\frac{d}{2\alpha}} b_n^{\frac{d}{2\alpha}}$ dominates. We shall make a similar analysis on c_n: in the rest of our proof we shall assume that λ is of the form $\lambda_n = n^{-\beta}$ where $\beta \leq \frac{\alpha-1-d}{2d}$. As a result, using the fact that $b_n \geq 1$, we can upper-bound c_n as

$$c_n \lesssim \lambda_n (\log(n/\delta))^{\frac{\alpha-1-d}{2d}} b_n + \frac{\log(2/\delta)}{\epsilon\sqrt{n}} a_n^{\frac{\alpha+1-d/2-\epsilon}{2\alpha}} b_n^{\frac{d/2+\epsilon-1}{\alpha}}\,. \tag{A.6}$$

Again we must split the analysis in two cases: the case where $\lambda_n (\log(n/\delta))^{\frac{\alpha-1-d}{2d}} b_n$ dominates and the case where $\frac{\log(2/\delta)}{\epsilon\sqrt{n}} a_n^{\frac{\alpha+1-d/2-\epsilon}{2\alpha}} b_n^{\frac{d/2+\epsilon-1}{\alpha}}$ dominates. In total, we have six distinct regimes: the regime $a_n \lesssim c_n b_n$ and $c_n \lesssim \lambda_n (\log(n/\delta))^{\frac{\alpha-1-d}{2d}} b_n$ which yields

$$\begin{cases} a_n \lesssim \lambda_n (\log(n/\delta))^{\frac{\alpha-1-d}{2d}} b_n^2\,, \\ \lambda_n b_n^2 \lesssim \lambda_n (\log(n/\delta))^{\frac{\alpha-1-d}{2d}} b_n\,, \end{cases} \tag{Case 1}$$

the regime $a_n \lesssim c_n \frac{b_n^{\frac{d}{2\alpha}} a_n^{1/2-\frac{d}{4\alpha}}}{r}$ and $c_n \lesssim \lambda_n (\log(n/\delta))^{\frac{\alpha-1-d}{2d}} b_n$ which yields

$$\begin{cases} a_n \lesssim \lambda_n (\log(n/\delta))^{\frac{\alpha-1-d}{2d}} \dfrac{b_n^{\frac{2\alpha+d}{2\alpha}} a_n^{1/2-\frac{d}{4\alpha}}}{r}\,, \\ \lambda_n b_n^2 \lesssim \lambda_n (\log(n/\delta))^{\frac{\alpha-1-d}{2d}} b_n\,, \end{cases} \tag{Case 2}$$

the regime $a_n \lesssim r a_n^{1-\frac{d}{2\alpha}} b_n^{\frac{d}{2\alpha}}$ and $c_n \lesssim \lambda_n (\log(n/\delta))^{\frac{\alpha-1-d}{2d}} b_n$ which yields

(Case 3)
$$\begin{cases} a_n \lesssim r a_n^{1-\frac{d}{2\alpha}} b_n^{\frac{d}{2\alpha}} , \\ \lambda_n b_n^2 \lesssim \lambda_n (\log(n/\delta))^{\frac{\alpha-1-d}{2d}} b_n , \end{cases}$$

the regime $a_n \lesssim c_n b_n$ and $c_n \lesssim \frac{\log(2/\delta)}{\epsilon\sqrt{n}} a_n^{\frac{\alpha+1-d/2-\epsilon}{2\alpha}} b_n^{\frac{d/2+\epsilon-1}{\alpha}}$ which yields

(Case 4)
$$\begin{cases} a_n \lesssim \frac{\log(2/\delta)}{\epsilon\sqrt{n}} a_n^{\frac{\alpha+1-d/2-\epsilon}{2\alpha}} b_n^{\frac{\alpha+d/2+\epsilon-1}{\alpha}} , \\ \lambda_n b_n^2 \lesssim \frac{\log(2/\delta)}{\epsilon\sqrt{n}} a_n^{\frac{\alpha+1-d/2-\epsilon}{2\alpha}} b_n^{\frac{d/2+\epsilon-1}{\alpha}} , \end{cases}$$

the regime where $a_n \lesssim c_n \frac{b_n^{\frac{d}{2\alpha}} a_n^{1/2-\frac{d}{4\alpha}}}{r}$ and $c_n \lesssim \frac{\log(2/\delta)}{\epsilon\sqrt{n}} a_n^{\frac{\alpha+1-d/2-\epsilon}{2\alpha}} b_n^{\frac{d/2+\epsilon-1}{\alpha}}$ which yields

(Case 5)
$$\begin{cases} a_n \lesssim \frac{\log(2/\delta)}{\epsilon\sqrt{n}} a_n^{\frac{\alpha+1-d/2-\epsilon}{2\alpha}} b_n^{\frac{d/2+\epsilon-1}{\alpha}} \frac{b_n^{\frac{d}{2\alpha}} a_n^{1/2-\frac{d}{4\alpha}}}{r} , \\ \lambda_n b_n^2 \lesssim \frac{\log(2/\delta)}{\epsilon\sqrt{n}} a_n^{\frac{\alpha+1-d/2-\epsilon}{2\alpha}} b_n^{\frac{d/2+\epsilon-1}{\alpha}} , \end{cases}$$

and finally the regime where $a_n \lesssim r a_n^{1-\frac{d}{2\alpha}} b_n^{\frac{d}{2\alpha}}$ and $c_n \lesssim \frac{\log(2/\delta)}{\epsilon\sqrt{n}} a_n^{\frac{\alpha+1-d/2-\epsilon}{2\alpha}} b_n^{\frac{d/2+\epsilon-1}{\alpha}}$ which yields

(Case 6)
$$\begin{cases} a_n \lesssim r a_n^{1-\frac{d}{2\alpha}} b_n^{\frac{d}{2\alpha}} , \\ \lambda_n b_n^2 \lesssim \frac{\log(2/\delta)}{\epsilon\sqrt{n}} a_n^{\frac{\alpha+1-d/2-\epsilon}{2\alpha}} b_n^{\frac{d/2+\epsilon-1}{\alpha}} . \end{cases}$$

Our purpose is to show that for well chosen λ_n, r, ϵ, we have in any cases $a_n \lesssim \lambda_n$. We shall start with Cases 1, 2, 3 which are easier to work with as we almost have $b_n = O(1)$. Then we shall move on to the remaining cases.

Cases 1, 2, 3. In these cases, we have $\lambda_n b_n^2 \lesssim \lambda_n (\log(n/\delta))^{\frac{\alpha-1-d}{2d}} b_n$ which implies $b_n \lesssim (\log(n/\delta))^{\frac{\alpha-1-d}{2d}}$. In Case 1, we thus obtain

(Case 1)
$$a_n \lesssim \lambda_n (\log(n/\delta))^{\frac{3(\alpha-1-d)}{2d}} .$$

In case 2 we obtain $a_n \lesssim \frac{\lambda_n}{r} (\log(n/\delta))^{\frac{\alpha+d-1}{2d}+\frac{d}{2\alpha}} a_n^{1/2-\frac{d}{4\alpha}}$ which yields $a_n^{\frac{2\alpha+d}{4\alpha}} \lesssim \frac{\lambda_n}{r} (\log(n/\delta))^{\frac{\alpha+d-1}{2d}+\frac{d}{2\alpha}}$ so we obtain

$$a_n \lesssim \left(\frac{\lambda_n}{r}\right)^{\frac{4\alpha}{2\alpha+d}} (\log(n/\delta))^{\frac{4\alpha(\alpha+d-1)}{2d(2\alpha+d)}+\frac{2d}{2\alpha+d}} .$$

In case 3, we have $a_n \lesssim r^{\frac{2\alpha}{d}} (\log(n/\delta))^{\frac{(\alpha-1-d)}{2d}}$. Hence we must pick r such that the two previous upper-bounds match up to the poly-log factors *i.e.* r must satisfy

$$\begin{aligned} r^{\frac{2\alpha}{d}} &= \left(\frac{\lambda_n}{r}\right)^{\frac{4\alpha}{2\alpha+d}} \\ \iff r^{\frac{2\alpha(2\alpha+d)+4\alpha d}{d(2\alpha+d)}} &= \lambda_n^{\frac{4\alpha}{2\alpha+d}} \\ \iff r &= \lambda_n^{\frac{4\alpha d}{2\alpha(2\alpha+d)+4\alpha d}} , \end{aligned}$$

hence we obtain $r_n = \lambda_n^{\frac{2d}{2\alpha+3d}}$; recalling that r must verify $r \leq r_0$, we shall thus set instead $r_n = r_0 \lambda_n^{\frac{2d}{2\alpha+3d}}$ which yields

$$\text{(Case 2)} \qquad a_n \lesssim \lambda_n^{\frac{4\alpha}{2\alpha+3d}} (\log(n/\delta))^{\frac{4\alpha(\alpha+d-1)}{2d(2\alpha+d)} + \frac{2d}{2\alpha+d}} ,$$

and

$$\text{(Case 3)} \qquad a_n \lesssim \lambda_n^{\frac{4\alpha}{2\alpha+3d}} (\log(n/\delta))^{\frac{(\alpha-1-d)}{2d}} .$$

Cases 4, 5, 6. The main difficulty of these cases is that b_n is not *a priori* $O(1)$ if $\lambda_n << 1/\sqrt{n}$. Indeed, recall that b_n now verifies $\lambda_n b_n^2 \lesssim \frac{\log(2/\delta)}{\epsilon\sqrt{n}} a_n^{\frac{\alpha+1-d/2-\epsilon}{2\alpha}} b_n^{\frac{d/2+\epsilon-1}{\alpha}}$ which yields

$$b_n \lesssim \left(\frac{\log(2/\delta)}{\epsilon\lambda_n\sqrt{n}}\right)^{\frac{\alpha}{2\alpha+1-d/2-\epsilon}} a_n^{\frac{\alpha+1-d/2-\epsilon}{2(2\alpha+1-d/2-\epsilon)}} . \tag{A.7}$$

Using this upper-bound, let us move on to each case separately.

Case 4. Recall we have $a_n \lesssim \frac{\log(2/\delta)}{\epsilon\sqrt{n}} a_n^{\frac{\alpha+1-d/2-\epsilon}{2\alpha}} b_n^{\frac{\alpha+d/2+\epsilon-1}{\alpha}}$, hence we recover

$$\begin{aligned}
a_n &\lesssim \frac{\log(2/\delta)}{\epsilon\sqrt{n}} a_n^{\frac{\alpha+1-d/2-\epsilon}{2\alpha}} \left[\left(\frac{\log(2/\delta)}{\epsilon\lambda_n\sqrt{n}}\right)^{\frac{\alpha}{2\alpha+1-d/2-\epsilon}} a_n^{\frac{\alpha+1-d/2-\epsilon}{2(2\alpha+1-d/2-\epsilon)}}\right]^{\frac{\alpha+d/2+\epsilon-1}{\alpha}} \\
&= a_n^{\frac{\alpha+1-d/2-\epsilon}{2\alpha}(1+\frac{\alpha+d/2+\epsilon-1}{2\alpha+1-d/2-\epsilon})} \left(\frac{\log(2/\delta)}{\epsilon\sqrt{n}}\right)^{1+\frac{\alpha+d/2+\epsilon-1}{2\alpha+1-d/2-\epsilon}} \lambda_n^{-\frac{\alpha+d/2+\epsilon-1}{2\alpha+1-d/2-\epsilon}} \\
&= a_n^{\frac{3(\alpha+1-d/2-\epsilon)}{2(2\alpha+1-d/2-\epsilon)}} \left(\frac{\log(2/\delta)}{\epsilon\sqrt{n}}\right)^{\frac{3\alpha}{2\alpha+1-d/2-\epsilon}} \lambda_n^{-\frac{\alpha+d/2+\epsilon-1}{2\alpha+1-d/2-\epsilon}}
\end{aligned}$$

which yields $a_n^{\frac{\alpha+d/2+\epsilon-1}{2(2\alpha+1-d/2-\epsilon)}} \lesssim \left(\frac{\log(2/\delta)}{\epsilon\sqrt{n}}\right)^{\frac{3\alpha}{2\alpha+1-d/2-\epsilon}} \lambda_n^{-\frac{\alpha+d/2+\epsilon-1}{2\alpha+1-d/2-\epsilon}}$ so we eventually get

$$a_n \lesssim \left(\frac{\log(2/\delta)}{\epsilon\sqrt{n}}\right)^{\frac{6\alpha}{\alpha+d/2+\epsilon-1}} \lambda_n^{-2} . \tag{A.8}$$

Unlike the three previous cases, the rate of convergence of a_n degrades as λ_n accelerates toward zero. Hence, we see that there appears a trade-off on how to pick λ_n: the upper-bound above must match the weakest upper-bound of Cases 1, 2, 3. When $\alpha \to \infty$, the weakest upper-bound is λ_n hence, up to poly-log factors and the ϵ term in the exponent, we have $n^{-\frac{3\alpha}{\alpha+d/2-1}} \lambda_n^{-2} = \lambda_n$ which yields $\lambda_n = n^{-\frac{\alpha}{\alpha+d/2-1}}$. Hence in the rest of the proof, we shall assume that $\lambda_n = n^{-\beta}$ with $\beta = \min(\frac{\alpha-1-d}{2d}, \frac{\alpha}{\alpha+d/2-1})$.

Case 5. Recall that we have in this case $a_n \lesssim \frac{\log(2/\delta)}{r_n\epsilon\sqrt{n}} a_n^{\frac{2\alpha+1-d-\epsilon}{2\alpha}} b_n^{\frac{d+\epsilon-1}{\alpha}}$ so we obtain

$a_n \lesssim \left(\frac{\log(2/\delta)}{r_n\epsilon\sqrt{n}}\right)^{\frac{2\alpha}{d+\epsilon-1}} b_n^2$. Combining this upper-bound with (A.7) yields

$$a_n \lesssim \left(\frac{\log(2/\delta)}{r_n\epsilon\sqrt{n}}\right)^{\frac{2\alpha}{d+\epsilon-1}} \left[\left(\frac{\log(2/\delta)}{\epsilon\lambda_n\sqrt{n}}\right)^{\frac{\alpha}{2\alpha+1-d/2-\epsilon}} a_n^{\frac{\alpha+1-d/2-\epsilon}{2(2\alpha+1-d/2-\epsilon)}}\right]^2 \tag{A.9}$$

$$\iff a_n^{\frac{\alpha}{2\alpha+1-d/2-\epsilon}} \lesssim \left(\frac{\log(2/\delta)}{r_n\epsilon\sqrt{n}}\right)^{\frac{2\alpha}{d+\epsilon-1}} \left(\frac{\log(2/\delta)}{\epsilon\lambda_n\sqrt{n}}\right)^{\frac{2\alpha}{2\alpha+1-d/2-\epsilon}} \tag{A.10}$$

$$\iff a_n \lesssim \left(\frac{\log(2/\delta)}{r_n\epsilon\sqrt{n}}\right)^{\frac{2(2\alpha+1-d/2-\epsilon)}{d+\epsilon-1}} \left(\frac{\log(2/\delta)}{\epsilon\lambda_n\sqrt{n}}\right)^{2}, \tag{A.11}$$

which eventually yields $a_n \lesssim \left(\frac{\log(2/\delta)}{\epsilon\sqrt{n}}\right)^{\frac{4\alpha+d}{d+\epsilon-1}} \lambda_n^{-2} r_n^{\frac{d+2\epsilon-4\alpha-2}{d+\epsilon-1}}$. Recalling that we set $r_n = r_0\lambda_n^{\frac{2d}{2\alpha+3d}}$, we have $\lambda_n^{-2} r_n^{\frac{d+2\epsilon-4\alpha-2}{d+\epsilon-1}} \sim \lambda_n^{-2}\lambda_n^{\frac{2d(d+2\epsilon-4\alpha-2)}{(2\alpha+3d)(d+\epsilon-1)}}$. In what follows, we shall pick $\epsilon = \epsilon_n = 1/\log(n)$; since we assumed $d \geq 2$, we can then neglect the ϵ terms in the exponents and recover

$$a_n \lesssim n^{-\frac{4\alpha+d}{2(d-1)}} \lambda_n^{-2-\frac{2d}{2\alpha+3d}\frac{4\alpha+2-d}{d-1}} \left(\log(n)\log(2/\delta)\right)^{\frac{2\alpha+d}{d-1}}.$$

Note that the exponent $h(\alpha) = 2 + \frac{2d}{2\alpha+3d}\frac{4\alpha+2-d}{d-1} = 2 + \frac{2d}{1+3d/(2\alpha)}\frac{2+(2-d)/(2\alpha)}{d-1}$ increases with α. Hence, as α grows, we have a trade-off on our upper-bound: while the term $n^{-\frac{4\alpha+d}{2(d-1)}}$ accelerates the convergence toward 0, the term $\lambda_n^{-2-\frac{2d}{2\alpha+3d}\frac{4\alpha+2-d}{d-1}}$ degrades the convergence with a doubly negative effect. First, as mentioned above, the magnitude exponent increases and furthermore, recalling that $\lambda_n = n^{-\beta(\alpha)}$ with $\beta(\alpha) = \min(\frac{\alpha-1-d}{2d}, \frac{\alpha}{\alpha+d/2-1})$, λ_n accelerates its convergence toward 0 as α grows hence the term $\lambda_n^{-2-\frac{2d}{2\alpha+3d}\frac{4\alpha+2-d}{d-1}}$ diverges more quickly. With the goal to show that we always have in fact $a_n \lesssim \lambda_n$ up to poly-log factors, we propose to quantify this trade-off and split the analysis into four smoothness regimes: a) the regime $d < \alpha \leq 2d$, b) the regime $2d < \alpha \leq 8d/3$, c) the regime $8d/3 < \alpha \leq 13d/4$ and the regime d) $13d/4 < \alpha$.

In regime a), we have $\beta(\alpha) \leq 1/2$ and *a fortiori* $\lambda_n^{-h(\alpha)} \lesssim n^{\frac{1}{2}h(2d)}$ where $h(2d)$ reads $h(2d) = 2 + \frac{2d}{4d+3d}\frac{8d-d+2}{d-1} = 2 + \frac{2}{7}\frac{7d+2}{d-1}$. Conversely, since $\alpha > d$, the term $n^{-\frac{4\alpha+d}{2(d-1)}}$ is upper-bounded by $n^{-\frac{5d}{2(d-1)}}$ which yields the following upper-bound on a_n (up to poly-log factors)

$$\begin{aligned} a_n &\lesssim n^{-\frac{5d}{2(d-1)}+1+\frac{1}{7}\frac{7d+2}{d-1}} \\ &\lesssim n^{-\frac{d}{d-1}(5/2-(d-1)/d-1-2/(7d))} \\ &= n^{-\frac{d}{d-1}(1/2+5/(7d))} \\ &\lesssim \lambda_n\,. \end{aligned}$$

In regime b), since $\frac{8d/3-d-1}{2d} < 5/6$ we have $\beta(\alpha) < 5/6$ and *a fortiori* $\lambda_n^{-h(\alpha)} \lesssim n^{\frac{5}{6}h(8d/3)}$ where $h(8d/3)$ reads $h(8d/3) = 2 + \frac{2d}{16d/3+3d}\frac{32d/3+2-d}{d-1} = 2 + \frac{6}{25}\frac{29d/3+2}{d-1}$. Conversely, the term $n^{-\frac{4\alpha+d}{2(d-1)}}$ is upper-bounded by $n^{-\frac{9d}{2(d-1)}}$ which yields the following

upper-bound on a_n (up to poly-log factors)

$$\begin{aligned} a_n &\lesssim n^{-\frac{9d}{2(d-1)}+10/6+\frac{29d/3+2}{5(d-1)}} \\ &\lesssim n^{-\frac{d}{d-1}(9/2-10(d-1)/(6d)-29/15-2/(5d))} \\ &= n^{-\frac{d}{d-1}(9/2-10/6-29/15+10/(6d)-2/(5d))} \\ &\lesssim \lambda_n \,. \end{aligned}$$

In regime c), we coarsely upper-bound $\beta(\alpha)$ by one and we evaluate the exponent $h(\alpha)$ in $\alpha = 13d/4$. We have $h(13d/4) = 2 + \frac{2d}{13d/2+3d}\frac{13d+2-d}{d-1}$ which yields

$$a_n \lesssim n^{-\frac{35d}{6(d-1)}+2+\frac{2d}{13d/2+3d}\frac{13d-d+2}{d-1}} \tag{A.12}$$

$$\lesssim n^{-\frac{d}{d-1}(35/6-2(d-1)/d-48/19-8/(19d))} \tag{A.13}$$

$$\lesssim 1/n \lesssim \lambda_n \,. \tag{A.14}$$

In regime d), we coarsely upper-bound $\beta(\alpha)$ by one and the exponent $h(\alpha)$ by $\lim_{\alpha\to\infty} h(\alpha) = 2 + 4\frac{d}{d-1}$ which yields the upper-bound

$$\begin{aligned} a_n &\lesssim n^{-\frac{14d}{2(d-1)}} n^{2+4\frac{d}{d-1}} \\ &\lesssim 1/n \lesssim \lambda_n \,. \end{aligned}$$

Case 6. Recall that in this case

$$a_n \lesssim r_n a_n^{1-\frac{d}{2\alpha}} b_n^{\frac{d}{2\alpha}} \,,$$

which implies $a_n \lesssim r_n^{\frac{2\alpha}{d}} b_n$. Using the upper-bound (A.7), we get

$$\begin{aligned} & a_n \lesssim r_n^{\frac{2\alpha}{d}} \left(\frac{\log(2/\delta)\log(n)}{\lambda_n\sqrt{n}}\right)^{\frac{\alpha}{2\alpha+1-d/2}} a_n^{\frac{\alpha+1-d/2}{2(2\alpha+1-d/2)}} \\ \iff & a_n^{\frac{3\alpha+1-d/2}{2(2\alpha+1-d/2)}} \lesssim r_n^{\frac{2\alpha}{d}} \left(\frac{\log(2/\delta)\log(n)}{\lambda_n\sqrt{n}}\right)^{\frac{\alpha}{2\alpha+1-d/2}} \\ \iff & a_n \lesssim r_n^{\frac{4\alpha(2\alpha+1-d/2)}{d(3\alpha+1-d/2)}} \left(\frac{\log(2/\delta)\log(n)}{\lambda_n\sqrt{n}}\right)^{\frac{2\alpha}{3\alpha+1-d/2}} \\ \iff & a_n \lesssim \lambda_n^{\frac{2\alpha}{3\alpha+1-d/2}\left[\frac{4(2\alpha+1-d/2)}{3d+2\alpha}-1\right]} n^{-\frac{\alpha}{3\alpha+1-d/2}} \left(\log(2/\delta)\log(n)\right)^{\frac{2\alpha}{3\alpha+1-d/2}} . \end{aligned}$$

If we further develop the exponent on λ_n and we neglect the $1-d/2$ (negative) terms in the denominators, we recover the slightly weaker upper-bound

$$a_n \lesssim \lambda_n^{\frac{2(6\alpha+4-5d)}{3(3d+2\alpha)}} n^{-1/3} \left(\log(2/\delta)\log(n)\right)^{\frac{2\alpha}{3\alpha+1-d/2}} ;$$

let us prove that for any α, this upper-bound is lower than λ_n up to the poly-log terms. The idea is the following: when the smoothness α is low, we have $\lambda_n >> n^{-1/3}$ so we indeed have that our upper bound is lower than λ_n thanks to the $n^{-1/3}$ term alone

as the exponent on λ_n is positive whenever $\alpha \geq d$. When the smoothness is high, the exponent on λ_n is greater than one hence we get to the same conclusion ; there remains the in-between cases.

We shall split the analysis in three regimes: the regime a) where $\alpha \geq 2d$, the regime b) where $3d/2 \leq \alpha \leq 2d$ and the regime c) where $\alpha \leq 3d/2$. First note that the exponent $h(\alpha) = \frac{2(6\alpha+4-5d)}{3(3d+2\alpha)}$ increases with α: indeed, the sign of h' is given by $6(3d+2\alpha)-2(6\alpha+4-5d) = 28d-8 > 0$. Hence for case a), we have in particular that $h(\alpha) \geq \frac{2(7d+4)}{3(7d)} \geq 2/3$ which gives $a_n \lesssim \lambda_n^{2/3} n^{-1/3}$. Since we always have $\lambda_n \gtrsim 1/n$, we recover $a_n \lesssim \lambda_n$. In the regime b), the exponent h verifies $h(\alpha) \geq \frac{2(4d+4)}{3(3d+3d)} \geq 4/9$. Furthermore, recall that $\lambda_n = n^{-\min(\frac{\alpha-1-d}{2d}, \frac{\alpha}{\alpha+d/2-1})}$ hence in regime b) $\lambda_n \gtrsim 1/\sqrt{n}$. Re-injecting this fact in our upper-bound yields $a_n \lesssim \lambda_n^{4/9} \lambda_n^{2/3} \lesssim \lambda_n$. Finally, in regime c), one can check that $\min(\frac{\alpha-1-d}{2d}, \frac{\alpha}{\alpha+d/2-1}) = \frac{\alpha-1-d}{2d} \leq 1/4$ and in particular $\lambda_n \gtrsim n^{-1/3}$ so we recover in particular $a_n \lesssim \lambda_n$.

Conclusion. If we pick $\lambda_n = n^{-\min(\frac{\alpha-d-1}{2d}, \frac{\alpha}{\alpha+d/2-1})}$, $r_n = r_0 \lambda_n^{\frac{2d}{3d+2\alpha}}$ and $\epsilon_n = 1/\log(n)$, we recover at worst

$$a_n \lesssim \lambda_n^{\frac{4\alpha}{2\alpha+3d}} \left(\log(n) \log(2/\delta) \right)^{\max(6, \frac{2\alpha+d}{d-1})} . \tag{A.15}$$

REFERENCES

[1] R. K. Ahuja, J. B. Orlin, and T. L. Magnanti, *Network flows: theory, algorithms, and applications*, Prentice-Hall, 1993.

[2] M. Arjovsky, S. Chintala, and L. Bottou, *Wasserstein generative adversarial networks*, in International conference on machine learning, 2017.

[3] N. Aronszajn, *Theory of reproducing kernels*, Transactions of the American Mathematical Society, (1950).

[4] E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert, *Inference in generative models using the wasserstein distance*, arXiv preprint arXiv:1701.05146, 1 (2017), p. 9.

[5] S. P. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[6] Y. Brenier, *Décomposition polaire et réarrangement monotone des champs de vecteurs*, CR Acad. Sci. Paris Sér. I Math., 305 (1987), pp. 805–808.

[7] H. Brezis and P. Mironescu, *Gagliardo–nirenberg inequalities and non-inequalities: The full story*, Annales de l'Institut Henri Poincaré C, Analyse non linéaire, (2018).

[8] S. Bubeck, *Convex optimization: Algorithms and complexity*, Foundations and Trends in Machine Learning, (2015).

[9] A. Caponnetto and E. De Vito, *Optimal rates for the regularized least-squares algorithm*, Foundations of Computational Mathematics, 7 (2007), pp. 331–368.

[10] L. Chizat, P. Roussillon, F. Léger, F.-X. Vialard, and G. Peyré, *Faster Wasserstein distance estimation with the Sinkhorn divergence*, Advances in Neural Information Processing Systems, 33 (2020).

[11] N. Courty, R. Flamary, A. Habrard, and A. Rakotomamonjy, *Joint distribution optimal transportation for domain adaptation*, Advances in Neural Information Processing Systems, (2017), pp. 3733–3742.

[12] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy, *Optimal transport for domain adaptation*, IEEE transactions on pattern analysis and machine intelligence, 39 (2016), pp. 1853–1865.

[13] M. Cuturi, *Sinkhorn distances: Lightspeed computation of optimal transport*, Advances in Neural information processing systems, 26 (2013), pp. 2292–2300.

[14] G. De Philippis and A. Figalli, *The Monge–Ampère equation and its link to optimal transportation*, Bulletin of the American Mathematical Society, 51 (2014), pp. 527–580.

[15] A. Delalande and Q. Merigot, *Quantitative stability of optimal transport maps under variations of the target measure*, arXiv, (2021).

[16] R. M. Dudley, *The speed of mean Glivenko–Cantelli convergence*, The Annals of Mathematical Statistics, 40 (1969), pp. 40–50.
[17] J. Feydy, B. Charlier, F.-X. Vialard, and G. Peyré, *Optimal transport for diffeomorphic registration*, in International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2017, pp. 291–299.
[18] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, *A kernel two-sample test*, The Journal of Machine Learning Research, 13 (2012), pp. 723–773.
[19] F. Gunsilius and S. Schennach, *Independent nonlinear component analysis*, Journal of the American Statistical Association, (2021).
[20] R. B. Holmes, *Smoothness of certain metric projections on hilbert space*, Transactions of the American Mathematical Society, (1973).
[21] J.-C. Hütter and P. Rigollet, *Minimax estimation of smooth optimal transport maps*, The Annals of Statistics, 49 (2021), pp. 1166–1194.
[22] J. B. Lasserre, *A sum of squares approximation of nonnegative polynomials*, SIAM Review, (2007).
[23] A. Makkuva, A. Taghvaei, S. Oh, and J. Lee, *Optimal transport mapping via input convex neural networks*, in International Conference on Machine Learning, PMLR, 2020, pp. 6672–6681.
[24] T. Manole, S. Balakrishnan, J. Niles-Weed, and L. Wasserman, *Plugin estimation of smooth optimal transport maps*, arXiv preprint arXiv:2107.12364, (2021).
[25] U. Marteau-Ferey, F. Bach, and A. Rudi, *Non-parametric models for non-negative functions*, Advances in Neural Information Processing Systems, (2020).
[26] N. G. Meyers and W. P. Ziemer, *Integral inequalities of Poincaré and Wirtinger type for BV functions*, American Journal of Mathematics, 99 (1977).
[27] L. Nirenberg, *An extended interpolation inequality*, Annali della Scuola Normale Superiore di Pisa - Scienze Fisiche e Matematiche, (1966).
[28] D. Onken, S. Wu Fung, X. Li, and L. Ruthotto, *Ot-flow: Fast and accurate continuous normalizing flows via optimal transport*, in AAAI Conference on Artificial Intelligence, vol. 35, 2021.
[29] A.-A. Pooladian and J. Niles-Weed, *Entropic estimation of optimal transport maps*, arXiv preprint arXiv:2109.12004, (2021).
[30] R. T. Rockafellar, *Convex Analysis*, vol. 36, Princeton University Press, 1970.
[31] A. Rudi, U. Marteau-Ferey, and F. Bach, *Finding global minima via kernel approximations*, in Arxiv preprint arXiv:2012.11978, 2020.
[32] T. Salimans, D. Metaxas, H. Zhang, and A. Radford, *Improving GANs using optimal transport*, in International Conference on Learning Representations, 2018.
[33] G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, et al., *Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming*, Cell, 176 (2019), pp. 928–943.
[34] B. Schölkopf, R. Herbrich, and A. J. Smola, *A generalized representer theorem*, 2001.
[35] Z. Su, Y. Wang, R. Shi, W. Zeng, J. Sun, F. Luo, and X. Gu, *Optimal mass transport for shape matching and comparison*, IEEE transactions on pattern analysis and machine intelligence, 37 (2015), pp. 2246–2259.
[36] A. Takatsu, *Wasserstein geometry of Gaussian measures*, Osaka Journal of Mathematics, 48 (2011), pp. 1005 – 1026.
[37] A. Vacher, B. Muzellec, A. Rudi, F. Bach, and F.-X. Vialard, *A dimension-free computational upper-bound for smooth optimal transport estimation*, Conference on Learning Theory, (2021).
[38] A. Vacher and F.-X. Vialard, *Parameter tuning and model selection in optimal transport with semi-dual brenier formulation*, in Advances in Neural Information Processing Systems, 2022.
[39] S. van de Geer, *M-estimation using penalties or sieves*, Journal of Statistical Planning and Inference, (2002).
[40] J. Weed and Q. Berthet, *Estimation of smooth densities in Wasserstein distance*, in Conference on Learning Theory, 2019, pp. 3118–3119.
[41] H. Wendland, *Scattered Data Approximation*, vol. 17, Cambridge University Press, 2004.
[42] H. Wendland and C. Rieger, *Approximate interpolation with applications to selecting smoothing parameters*, Numerische Mathematik, (2005).
[43] C. Williams and M. Seeger, *Using the Nyström method to speed up kernel machines*, Advances in Neural Information Processing Systems, 13 (2001), pp. 682–688.
[44] K. D. Yang, K. Damodaran, S. Venkatachalapathy, A. C. Soylemezoglu, G. Shivashankar, and C. Uhler, *Predicting cell lineages using autoencoders and optimal trans-*

764 *port*, PLoS computational biology, 16 (2020), p. e1007828.