A Dimension-free Computational Upper-bound for Smooth Optimal Transport Estimation

Adrien Vacher ADRIEN.VACHER@U-PEM.FR
INRIA Paris, 2 rue Simone Iff, 75012, Paris, France
LIGM, Université Gustave Eiffel, CNRS

Boris Muzellec BORIS.MUZELLEC@INRIA.FR
INRIA Paris, 2 rue Simone Iff, 75012, Paris, France
Département d'Informatique, École Normale Supérieure, Paris, France
PSL Research University, 2 rue Simone Iff, 75012, Paris, France

Alessandro Rudi ALESSANDRO.RUDI@INRIA.FR
INRIA Paris, 2 rue Simone Iff, 75012, Paris, France
Département d'Informatique, École Normale Supérieure, Paris, France
PSL Research University, 2 rue Simone Iff, 75012, Paris, France

Francis Bach FRANCIS.BACH@INRIA.FR
INRIA Paris, 2 rue Simone Iff, 75012, Paris, France
Département d'Informatique, École Normale Supérieure, Paris, France
PSL Research University, 2 rue Simone Iff, 75012, Paris, France

François-Xavier Vialard FRANCOIS-XAVIER.VIALARD@U-PEM.FR
LIGM, Université Gustave Eiffel, CNRS

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

It is well-known that plug-in statistical estimation of optimal transport suffers from the curse of dimensionality. Despite recent efforts to improve the rate of estimation with the smoothness of the problem, the computational complexity of these recently proposed methods still degrade exponentially with the dimension. In this paper, thanks to an infinite-dimensional sum-of-squares representation, we derive a statistical estimator of smooth optimal transport which achieves a precision ε from $\tilde{O}(\varepsilon^{-2})$ independent and identically distributed samples from the distributions, for a computational cost of $\tilde{O}(\varepsilon^{-4})$ when the smoothness increases, hence yielding dimension-free statistical *and* computational rates, with potentially exponentially dimension-dependent constants.

1. Introduction

The comparison between probability distributions is a fundamental task and has been extensively used in machine learning. For this purpose, optimal transport (OT) has recently gained traction in different subfields of machine learning (ML), such as natural language processing (NLP) (Xu et al., 2018; Chen et al., 2018), generative modeling (Arjovsky et al., 2017; Tolstikhin et al., 2018; Salimans et al., 2018), multi-label classification (Frogner et al., 2015), domain adaptation (Redko et al., 2019), clustering (Ho et al., 2017), and has had an impact in other areas such as imaging sciences (Feydy et al., 2017; Bonneel et al., 2011). Indeed, OT is a tool to compare data distributions which has arguably many more geometric properties than other available divergences (Peyré and Cuturi, 2019).

In practice, the optimal transport cost is often computed for the squared distance (leading to the Wasserstein-2 distance) on sampled distributions with n observations, and it is well-known that optimal transport suffers from the curse of dimensionality (Fournier and Guillin, 2015): the plug-in strategy, which simply consists in computing the Wasserstein distance between the sampled distributions, yields an estimation of the Wasserstein squared distance between a density and its sampled version in $O(1/n^{1/d})$, which degrades rapidly in high dimensions; this can only be improved to $O(1/n^{2/d})$ in the case of two different distributions (Chizat et al., 2020).

However, high dimension is the usual setting in machine learning, such as in NLP (Grave et al., 2019), and even if the intrinsic dimensionality of data can be leveraged (Weed and Bach, 2019; Niles-Weed and Rigollet, 2019), poor theoretical rates of convergence are a recurrent feature of OT. Liang (2018, 2019) recently showed that when the measures admit smooth densities, the Wasserstein-1 distance (as part of a more general class of integral probability metrics (IPM)) those minimax sample complexity rates could be improved to almost $O(1/\sqrt{n})$ when the smoothness increases. Weed and Berthet (2019) then showed equivalent rates in the case smooth densities with geometric assumptions on their supports for the Wasserstein-p distances with $p > 1$, which are not IPMs, and proposed a corresponding estimator based on a dedicated non-polynomial-time algorithm. Matching rates were then proved for the transportation maps themselves (Hütter and Rigollet, 2019) in the Wasserstein-2 setting, under smoothness assumptions on those maps. This line of work is deeply related to the regularity theory of optimal transport, that guarantees the smoothness of the optimal map in Euclidean spaces under similar assumptions on the source and target distributions, and their supports (Caffarelli, 1992; De Philippis and Figalli, 2014). Yet, to this day no practically tractable algorithm (e.g., with polynomial time) matching the bounds of Weed and Berthet (2019) and Hütter and Rigollet (2019) is known.

State of the art. An approach that has first been advocated in the machine learning community as a way to efficiently approximate empirical OT and to make it differentiable consists in adding an entropic regularization term to the OT problem (Cuturi, 2013). Rates on the sample complexity of entropic optimal transport have then been studied by Genevay et al. (2019) and Mena and Niles-Weed (2019), and were proven to be of the order $O(\frac{1}{\varepsilon^{\lfloor d/2 \rfloor}\sqrt{n}})$, for small values of ε. Although the dependency in the number of samples is in $1/\sqrt{n}$, the constant degrades exponentially with respect to the dimension. Entropic OT and Sinkhorn divergences (Genevay et al., 2018) were then leveraged as a tool to study the sample complexity of the (unregularized) OT problem itself: the most advanced results in this direction were derived by Chizat et al. (2020), who show that with few assumptions on the regularity of the Kantorovich potentials, the squared Wasserstein distance can be estimated using $O(\varepsilon^{-d/2+2})$ samples and $\tilde{O}(\varepsilon^{-(d'+5.5)})$ operations (with $d' = 2\lfloor \frac{d}{2} \rfloor$) with high probability.

Our work can be related to a current research direction which consists in developing estimators of the Wasserstein distance for classes of smooth distributions, with smoothness parameter m, that have better performances than in the general case. Related to this trend, Liang (2018, 2019) showed minimax rates for a class of integral probability metrics (IPM) that includes the Wasserstein-1 distance, as a function of the smoothness of the distributions. However, (i) for $p > 1$, the Wasserstein-p distance W_p is not an IPM and (ii) no estimators with matching rates are proposed in those two works. So far, two main contributions

leveraging smoothness that are applicable to the W_2 distance can be found. Hütter and Rigollet (2019) derive minimax rates for the estimation of the OT maps and propose an estimator which necessitates, for an L_2 error on the maps of order ε, $O(\varepsilon^{-\frac{2m-1+d/2}{2m}})$ samples. While statistically almost optimal, this estimator is not computationally feasible as it requires to project the potentials on a space of smooth, strongly-convex functions. Instead, Weed and Berthet (2019) derive estimators for the densities requiring $O(\varepsilon^{-\frac{d+2m}{1+m}})$ samples and, under the assumption that an efficient resampler is available, derive an estimator of the OT distance that can be calculated in $\tilde{O}(\varepsilon^{-(2d+d/2)})$ time.

While the contributions above do succeed in taking advantage of the smoothness from a statistical point of view, they do not manage to take advantage of the smoothness from a computational point of view. Actually, statistical-computational gaps are known to exist for some instances of high-dimensional OT, such as the spiked transport model of Niles-Weed and Rigollet (2019).

Contributions. In this paper, we bridge the statistical-computational gap of smooth OT estimation and we provide a positive answer to the question whether smoothness of the optimal potentials can be computationally beneficial to an efficient statistical estimator. More precisely, we propose an algorithm which, for a given accuracy ε, needs $O(\varepsilon^{-2})$ samples and has a computational complexity of $\tilde{O}(\varepsilon^{-\max(4,\frac{7d}{m-d})})$. Note that the computational complexity improves with the regularity of the distributions and, when $m \geq 3d$, it is $\tilde{O}(\varepsilon^{-4})$, i.e., independent of the dimension d in the exponent (but not in the constants). We thus show that smoothness can be leveraged in the computational estimation of optimal transport.

Moreover, we consider different scenarios beyond i.i.d. sampling, such as the case where we are able to compute exact integrals or where we can evaluate the densities in given points, by representing the problem in terms of kernel mean embeddings (Muandet et al., 2017). This allows to make a unified analysis for all the cases. The total error is then the sum of the error induced by approximating via the kernel mean embedding plus the error induced by subsampling the constraints. Interestingly, in the other scenarios the computational cost to achieve an error ε can be smaller than ε^{-4}, as reported below. This is particularly interesting in the case we can evaluate the densities in given points and avoids using expensive Monte-Carlo sampling techniques to obtain i.i.d. samples (see Section 6 for more details). Our results are summarized below.

Theorem 1 *Let $\varepsilon > 0$. Let μ, ν satisfy Assumption 1 for some $m > d$. Let $\widehat{OT}$ be the proposed estimator defined in Eq. (6) and computed as in Appendix F with the same parameters as in Corollary 3. The cost to achieve $|\widehat{OT} - OT(\mu,\nu)| \leq \varepsilon$ for the three scenarios is:*

1. *(Exact integral) Time: $\tilde{O}(\varepsilon^{-\frac{7d}{m-d}})$. Space: $\tilde{O}(\varepsilon^{-\frac{4d}{m-d}})$.*
2. *(Evaluation) Time: $\tilde{O}(\varepsilon^{-\frac{7d}{m-d}})$. Space: $\tilde{O}(\varepsilon^{-\frac{4d}{m-d}})$. #evaluations of μ,ν: $\tilde{O}(\varepsilon^{-\frac{d}{m+1}})$.*
3. *(Sampling) Time: $\tilde{O}(\varepsilon^{-\max(4,\frac{7d}{m-d})})$. Space: $\tilde{O}(\varepsilon^{-\frac{4d}{m-d}})$. #samples of μ,ν: $\tilde{O}(\varepsilon^{-2})$.*

The second key contribution of this paper is to provide a new representation theorem for solutions of smooth optimal transport. The inequality constraint in the dual OT problem can be replaced with an equality constraint involving a finite sum-of-squares in a Sobolev space.

In comparison with Rudi et al. (2020), it is a non-trivial extension of their representation result to the case of a *continuous* set of global minimizers instead of a *finite* set.

2. Sketch of the result and derivation of the algorithm

In this paper, we consider the optimal transport problem for the quadratic cost on bounded subsets X, Y of the Euclidean space $\mathbb{R}^d$. The set of probability measures on X is denoted by $\mathcal{P}(X)$. The optimal transport problem with quadratic cost $c(x,y) = \frac{1}{2}\|x-y\|^2$ can be stated in its dual formulation as

$$\begin{aligned} \mathrm{OT}(\mu,\nu) = \sup_{u,v \in C(\mathbb{R}^d)} \quad & \int u(x)d\mu(x) + \int v(y)d\nu(y) \\ \text{subject to} \quad & c(x,y) \geq u(x) + v(y), \quad \forall (x,y) \in X \times Y, \end{aligned} \tag{1}$$

As a standard result in optimal transport theory, the supremum is attained and the functions $u_\star, v_\star$ are referred to as the Kantorovich potentials (see Santambrogio, 2015).

The proposed approach to approximate $\mathrm{OT}(\mu,\nu)$ is the result of two main ingredients: (1) a suitable way to represent smooth functions and to approximate their integral in μ, ν, (2) a way to enforce efficiently the dense set of constraints on u, v.

Preliminary step: Representing smooth functions and integrals. We represent smooth functions via a *reproducing Kernel Hilbert space* (RKHS) (Aronszajn, 1950; Steinwart and Christmann, 2008), for which functions can be represented as linear forms. In Section 4 we show that under smoothness assumptions on μ and ν (Assumption 1) we have $u \in \mathcal{H}_X$ and $v \in \mathcal{H}_Y$ where $\mathcal{H}_X$ and $\mathcal{H}_Y$ are two suitable RKHSs on X and Y, associated with two bounded continuous feature maps $\phi_X : X \to \mathcal{H}_X$ and $\phi_Y : Y \to \mathcal{H}_Y$. Note that RKHSs offer several advantages. First, leveraging the reproducing property, we can represent the integrals in the functional of Eq. (1) as inner products in terms of the kernel mean embeddings $w_\mu \in \mathcal{H}_X$ and $w_\nu \in \mathcal{H}_Y$ where $w_\mu = \int_X \phi_X(x)d\mu(x)$ and $w_\nu = \int_Y \phi_Y(y)d\nu(y)$. Indeed, by the reproducing property, for all $u \in \mathcal{H}_X$, we have:

$$\int_X u(x)d\mu(x) = \int_X \langle u, \phi_X(x)\rangle_{\mathcal{H}_X} d\mu(x) = \Big\langle u, \int_X \phi_X(x)d\mu(x)\Big\rangle_{\mathcal{H}_X} = \langle u, w_\mu\rangle_{\mathcal{H}_X},$$

and the same reasoning holds for the integral on ν, i.e., $\int_Y v(y)d\nu(y) = \langle v, w_\nu\rangle_{\mathcal{H}_Y}$, for all $v \in \mathcal{H}_Y$. This construction is known as *kernel mean embedding* (Muandet et al., 2017). Moreover, RKHSs allow the so-called *kernel trick* (Steinwart and Christmann, 2008), i.e., to express the resulting algorithm in terms of *kernel functions* that in our case correspond to $k_X(x,x') = \langle \phi_X(x), \phi_X(x')\rangle_{\mathcal{H}_X}$ and $k_Y(y,y') = \langle \phi_Y(y), \phi_Y(y')\rangle_{\mathcal{H}_Y}$, that are known explicitly and are easily computable in $O(d)$.

The main step: Dealing with a dense set of inequalities. Even assuming that we are able to compute integrals in closed form and restricting to m-times differentiable u, v, the main challenge is to deal with the dense set of inequalities $c(x,y) \geq u(x) + v(y)$ that u, v must satisfy, for any $(x,y) \in X \times Y$. Indeed, an intuitive approach would be to subsample the set, i.e., to take ℓ points $(\tilde{x}_1, \tilde{y}_1), \ldots, (\tilde{x}_\ell, \tilde{y}_\ell)$ in $X \times Y$ and consider only the constraints $c(\tilde{x}_j, \tilde{y}_j) \geq u(\tilde{x}_j) + v(\tilde{y}_j)$ for $j = 1, \ldots, \ell$. This approach, however, is only able to leverage

the Lipschitzianity of u, v (Aubin-Frankowski and Szabo, 2020) and leads to an error in the order of $\ell^{-1/d}$ that does not allow to break the dependence in d in the exponent, and yields rates that are equivalent to the plugin estimator.

In this paper, we leverage a more refined technique to approximate the dense set of inequalities, introduced by Rudi et al. (2020) for the problem of non-convex optimization, and that allows to break the curse of dimensionality for smooth problems. The idea behind this technique is the consideration that, while a dense set of inequalities is poorly approximated by subsampling, the situation is different in the case of a dense set of *equality* constraints, for which an optimal rate of $O(\ell^{-m/d})$ is achievable for m-times differentiable constraints (Wendland and Rieger, 2005). The construction works in two steps: first, substitute the inequality constraints with equality constraints that are equivalent, and then subsample. In the next two paragraphs we explain how to apply this approach to the problem of OT.

Removing the inequalities: positive definite operator characterization. To apply the construction recalled above to our scenario, we first consider the following problem. Let $\mathcal{H}_{XY}$ be a Hilbert space on $X \times Y$ and $\phi : X \times Y \to \mathcal{H}_{XY}$. Denote by k_{XY} the kernel $k_{XY}((x,y),(x',y')) = \langle \phi(x,y),\, \phi(x',y')\rangle_{\mathcal{H}_{XY}}$ for any $(x,y),(x',y') \in X \times Y$ and by $\mathbb{S}_+(\mathcal{H}_{XY})$ the space of positive operators on $\mathcal{H}_{XY}$. We define

$$\begin{aligned} \max_{\substack{u\in\mathcal{H}_X, v\in\mathcal{H}_Y,\\ A\in\mathbb{S}_+(\mathcal{H}_{XY})}} \quad & \langle u,\, w_\mu\rangle_{\mathcal{H}_X} + \langle v,\, w_\nu\rangle_{\mathcal{H}_Y} \\ \text{subject to} \quad & \forall (x,y) \in X \times Y, \quad c(x,y) - u(x) - v(y) = \langle \phi(x,y),\, A\phi(x,y)\rangle_{\mathcal{H}_{XY}}, \end{aligned} \tag{2}$$

where the inequality in (1) is substituted with an equality w.r.t. a positive definite operator A on $\mathcal{H}_{XY}$. Note that Problem (1) is a relaxation of Problem (2): indeed, if for a given pair $u \in \mathcal{H}_X, v \in \mathcal{H}_Y$ there exists a positive definite A satisfying the equality above, then

$$c(x,y) - u(x) - v(y) = \langle \phi(x,y),\, A\phi(x,y)\rangle_{\mathcal{H}_{XY}} \geq 0, \qquad \forall (x,y) \in X \times Y,$$

so the couple (u, v) is admissible for (1). However, even for an admissible couple in (1) satisfying $u \in \mathcal{H}_X, v \in \mathcal{H}_Y$, a positive operator A may not exist. Indeed, note that the technique of representing a positivity constraint in terms of a positive matrix has a long history in the community of polynomial optimization (Lasserre, 2001; Parrilo, 2003; Lasserre, 2009), which shows that in general the resulting problem is not equivalent to the original one, for any chosen degree of polynomial approximation. This fact leads to the so-called *sum of squares hierarchies*, also used for optimal transport (Henrion and Lasserre, 2020). Instead, using kernels, Rudi et al. (2020) showed that there exists a positive operator with finite rank that matches the constraints and makes the two problems equivalent, when the constraint is attained on a finite set of points. However, such existence results cannot be used for the problem in (2), since in the case of optimal transport the set of zeros corresponds to the graph of the optimal transport map and is a smooth manifold, when μ, ν are smooth (De Philippis and Figalli, 2014).

A crucial point of our contribution is then to prove that, with a quadratic cost $c(x,y) = \frac{1}{2}\|x-y\|^2$ and under the same assumptions on the densities and their supports, or under smoothness assumptions on the Kantorovich potentials, there exists a positive operator on a suitable Hilbert space that represents the function $c(x,y) - u(x) - v(y)$ for a pair u, v

maximizing (1), making the two problems equivalent. The result is reported in Theorem 5. The proof is derived using the Fenchel dual characterization of $u_\star, v_\star$ and gives a sharp control of the rank of A.

Subsampling the constraints and approximating the integrals. We restrict the constraint of (2) to $(\tilde{x}_1, \tilde{y}_1), \dots, (\tilde{x}_\ell, \tilde{y}_\ell) \subset X \times Y$ for $\ell \in \mathbb{N}$. However, we need to add a penalization for u, v and A to avoid overfitting, since the error induced by subsampling the constraints is proportional to the trace of A (Rudi et al., 2020) and, in our case, also to the norms of u, v, as derived in Theorem 9 in Section 5. Finally, in two of the three scenarios of interest for the paper, i.e., (i) when we can only evaluate μ, ν pointwise, or (ii) when we have only i.i.d. samples from μ, ν, we do not have access to the kernel mean embeddings $w_\mu \in \mathcal{H}_X, w_\nu \in \mathcal{H}_Y$. Therefore, we need to use some estimators $\hat{w}_\mu \in \mathcal{H}_X, \hat{w}_\nu \in \mathcal{H}_Y$ that are derived in Section 6. The resulting problem is the following, for some regularization parameters $\lambda_1, \lambda_2 > 0$:

$$\begin{aligned} \max_{\substack{u \in \mathcal{H}_X, v \in \mathcal{H}_Y, \\ A \in \mathbb{S}_+(\mathcal{H}_{XY})}} \quad & \langle u,\, \hat{w}_\mu \rangle_{\mathcal{H}_X} + \langle v,\, \hat{w}_\nu \rangle_{\mathcal{H}_Y} - \lambda_1 \mathrm{Tr}(A) - \lambda_2 (\|u\|^2_{\mathcal{H}_X} + \|v\|^2_{\mathcal{H}_Y}) \\ \text{subject to} \quad & \forall j \in [\ell], \quad c(\tilde{x}_j, \tilde{y}_j) - u(\tilde{x}_j) - v(\tilde{y}_j) = \langle \phi(\tilde{x}_j, \tilde{y}_j),\, A\phi(\tilde{x}_j, \tilde{y}_j) \rangle_{\mathcal{H}_{XY}}. \end{aligned} \tag{3}$$

Let $\hat{u}, \hat{v}$ be the maximizers of the problem above (unique since the problem is strongly concave in u, v). The estimator for OT we consider corresponds to

$$\widehat{\mathrm{OT}} = \langle \hat{u},\, \hat{w}_\mu \rangle_{\mathcal{H}_X} + \langle \hat{v},\, \hat{w}_\nu \rangle_{\mathcal{H}_Y}. \tag{4}$$

Finite-dimensional characterization. In Appendix F, following Marteau-Ferey et al. (2020), we derive the dual problem of Eq. (3). Define $\mathbf{Q} \in \mathbb{R}^{\ell \times \ell}$ as $\mathbf{Q}_{i,j} = k_X(\tilde{x}_i, \tilde{x}_j) + k_Y(\tilde{y}_i, \tilde{y}_j)$ and $z_j = \hat{w}_\mu(\tilde{x}_j) + \hat{w}_\nu(\tilde{y}_j) - 2\lambda_2 c(\tilde{x}_j, \tilde{y}_j)$ for $i, j \in [\ell]$ and $q^2 = \|\hat{w}_\mu\|^2_{\mathcal{H}_X} + \|\hat{w}_\nu\|^2_{\mathcal{H}_Y}$, and let $\mathbf{I}_\ell \in \mathbb{R}^{\ell \times \ell}$ be the identity matrix. Let $\mathbf{K} \in \mathbb{R}^{\ell \times \ell}$ be defined as $\mathbf{K}_{i,j} = k_{XY}((\tilde{x}_i, \tilde{y}_i), (\tilde{x}_j, \tilde{y}_j))$ and define $\Phi_i \in \mathbb{R}^\ell$ as the i-th column of $\mathbf{R}$, the upper triangular matrix corresponding to the Cholesky decomposition of $\mathbf{K}$ (i.e., $\mathbf{R}$ satisfies $\mathbf{K} = \mathbf{R}^\top \mathbf{R}$). The dual problem writes:

$$\min_{\gamma \in \mathbb{R}^\ell} \ \tfrac{1}{4\lambda_2} \gamma^\top \mathbf{Q} \gamma - \tfrac{1}{2\lambda_2} \textstyle\sum_{j=1}^\ell \gamma_j z_j + \tfrac{q^2}{4\lambda_2} \quad \text{such that} \quad \textstyle\sum_{j=1}^\ell \gamma_j \Phi_j \Phi_j^\top + \lambda_1 \mathbf{I}_\ell \succeq 0. \tag{5}$$

In the same section in Appendix F, we derive an explicit characterization of $\hat{u}, \hat{w}, \hat{A}$ in terms of $\hat{\gamma}$, the solution of the problem above and we characterize $\widehat{\mathrm{OT}}$ as follows:

$$\widehat{\mathrm{OT}} = \tfrac{q^2}{2\lambda_2} - \tfrac{1}{2\lambda_2} \textstyle\sum_{j=1}^\ell \hat{\gamma}_j (\hat{w}_\mu(\tilde{x}_j) + \hat{w}_\nu(\tilde{y}_j)). \tag{6}$$

As it is possible to observe from the problem above and the characterization of $\widehat{\mathrm{OT}}$, the only quantities necessary to compute $\hat{\gamma}$ and $\widehat{\mathrm{OT}}$ are the kernels k_X, k_Y, k_{XY} and the evaluation of the functions $\hat{w}_\mu \in \mathcal{H}_X, \hat{w}_\nu \in \mathcal{H}_Y$ at the points $\tilde{x}_j$ and $\tilde{y}_j$ respectively for $j \in [\ell]$. In Appendix F, we consider a Newton method with self-concordant barrier to solve the problem above (Nesterov and Nemirovskii, 1994). To illustrate that this algorithm can indeed be implemented in practice, we run simulations on toy data in Appendix G. The total cost of the procedure to achieve error ε for the computation of $\widehat{\mathrm{OT}}$ is the following (see Theorem 17, Page 27 in the appendix):

$$O\big(C + E\ell + \ell^{3.5} \log \tfrac{\ell}{\varepsilon}\big) \ \text{ time}, \qquad O(\ell^2) \ \text{ memory}, \tag{7}$$

where C is the cost for computing q^2 and E is the cost to compute one z_j. Depending on the operations that we are able to perform on μ, ν and k_X, k_Y, we have three scenarios. In Section 6 we specify how to compute the vectors $\hat{w}_\mu, \hat{w}_\nu$, in Corollary 3 we report only the conditions of applicability and the resulting cost. In the next section and then in Appendix F we quantify instead how to choose $\ell, \lambda_1, \lambda_2$ to achieve $|\widehat{\mathrm{OT}} - \mathrm{OT}(\mu, \nu)| \leq \varepsilon$ with high probability and we provide a complete computational complexity in ε.

2.1. Theoretical Guarantees

Here we quantify the convergence rate of $\widehat{\mathrm{OT}}$ to OT. To simplify the exposition, in this section we will make a classical assumption on the smoothness of the densities (De Philippis and Figalli, 2014). Note however that the results of the paper hold under a more general assumption on the smoothness of the potentials (see Theorem 5).

Assumption 1 (m-times differentiable densities) *Let $m, d \in \mathbb{N}$. Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$.*

a) μ, ν have densities. Their supports, resp. $X, Y \subset \mathbb{R}^d$ are convex, bounded and open;

b) the densities are finite, bounded away from zero, with Lipschitz derivatives up to order m.

Assumption 1 is particularly adapted to our context since it guarantees that the Kantorovich potentials have a similar order of differentiability (De Philippis and Figalli, 2014). The main result, Theorem 9, is expressed for a general set of couples $(\tilde{x}_j, \tilde{y}_j)$, $j \in [\ell]$. Here, we specify it for the case where the couples are sampled independently and uniformly at random.

Theorem 2 *Let μ, ν satisfy Assumption 1 for $m > d$ and $m \geq 3$. Let $\ell \in \mathbb{N}$ and $\delta \in (0, 1]$. Let $(\tilde{x}_j, \tilde{y}_j)$ be independently sampled from the uniform distribution on $X \times Y$. Let $\widehat{OT}$ be computed with $k_X = k_{m+1}$, $k_Y = k_{m+1}$ and $k_{XY} = k_m$ where k_s for $s > 0$ is the Sobolev kernel in Eq. (8). Then, there exists ℓ_0 depending only on d, m, X, Y and C_1, C_2 depending only on $u_\star, v_\star$ and d, such that when $\ell \geq \ell_0$ and λ_1, λ_2 are chosen to satisfy*

$$\lambda_1 \geq C_1 \ell^{-m/2d+1/2} \log \tfrac{\ell}{\delta}, \quad \lambda_2 \geq \|w_\mu - \hat{w}_\mu\|_{\mathcal{H}_X} + \|w_\nu - \hat{w}_\nu\|_{\mathcal{H}_Y} + \lambda_1,$$

then, with probability $1 - \delta$, we have

$$|\widehat{OT} - OT(\mu, \nu)| \leq C_2 \lambda_2.$$

Note that while the rate does not depend exponentially in d as we will see in the rest of the section, the constants ℓ_0, C_1, C_2 depend exponentially in d in the worst case, as Rudi et al. (2020) for the case of global optimization. From the theorem above it is clear that the approximation error of $\widehat{\mathrm{OT}}$ is the sum of the error induced by the kernel mean embeddings plus the error induced by the subsampling of the inequality. Note here that the result of the theorem above holds also if the ℓ couples are i.i.d. from $\rho = \mu \otimes \nu$, as discussed in Remark 10. This can be beneficial if we do not know X, Y or we do not know how to sample from them. In the next corollary we will specialize the result depending on the considered scenarios.

Corollary 3 *Under the same assumptions as Theorem 2, let $k_X = k_{m+1}$, $k_Y = k_{m+1}, k_{XY} = k_m$ where k_s for $s > 0$ is defined in Eq. (8) and $\lambda_1 \geq C_1 \ell^{-(m-d)/2d} \log \frac{\ell}{\delta}$. Compute $\widehat{OT}$ with $\hat{w}_\mu, \hat{w}_\nu$ chosen according to one of the three scenarios below, as in Section 6. There exist C, C', C_2', C_2'' s.t. with probability at least $1 - \delta$,*

1. *(Exact integral) When we are able to compute exactly* $\int k_X(x,x')d\mu(x')d\mu(x)$ *and also* $\int_X k_X(x,x')d\mu(x)$ *for any* $x' \in X$ *(and analogously for* ν*). Choose* $\lambda_2 = \lambda_1$*. Then,*

$$|\widehat{OT} - OT(\mu,\nu)| \leq C_2 \ell^{-(m-d)/2d} \log \tfrac{\ell}{\delta}.$$

2. *(Evaluation) When we are only able to evaluate* μ, ν *on given points and to compute* $\int_X k_X(x,z)k_X(x',z)dz$, $\int_X \int_X k_X(x,z)k_X(z,z')k_X(x',z')dzdz'$*. Evaluate* μ *in* n_μ *points sampled uniformly from* X *(and* n_ν *for* ν*). Let* $\lambda_2 = \lambda_1 + C(n_\mu + n_\nu)^{-(m+1)/d} \log \frac{n_\mu+n_\nu}{\delta}$*,*

$$|\widehat{OT} - OT(\mu,\nu)| \leq C_2'(n_\mu^{-(m+1)/d} \log \tfrac{n_\mu}{\delta} + n_\nu^{-(m+1)/d} \log \tfrac{n_\nu}{\delta} + \ell^{-(m-d)/2d} \log \tfrac{\ell}{\delta}).$$

3. *(Sampling) When we are only able to sample from* μ, ν*, by using* n_μ *i.i.d. samples from* μ *and* n_ν *from* ν*. Choose* $\lambda_2 = \lambda_1 + C'(n_\mu + n_\nu)^{-1/2} \log \frac{n_\mu+n_\nu}{\delta}$*. Then,*

$$|\widehat{OT} - OT(\mu,\nu)| \leq C_2''(n_\mu^{-1/2} \log \tfrac{n_\mu}{\delta} + n_\nu^{-1/2} \log \tfrac{n_\nu}{\delta} + \ell^{-(m-d)/2d} \log \tfrac{\ell}{\delta}).$$

3. Notations and background

Let $n \in \mathbb{N}$, we denote by $[n]$ the set $\{1, \dots, n\}$. For a set Z, and a positive definite kernel $k : Z \times Z \to \mathbb{R}$ (i.e., so that all matrices of pairwise kernel evaluations are positive semi-definite), we can define a *reproducing kernel Hilbert spaces* (Aronszajn, 1950) $\mathcal{H}$ of real functions on Z, endowed with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, and a norm $\| \cdot \|_{\mathcal{H}}$. It satisfies: (1) $k(z, \cdot) \in \mathcal{H}$ for any $z \in Z$ and (2) the *reproducing property*, i.e., for any $f \in \mathcal{H}, z \in Z$ it holds that $f(z) = \langle f, k(z, \cdot) \rangle_{\mathcal{H}}$. The *canonical feature map* associated to $\mathcal{H}$ is the map $\phi : Z \to \mathcal{H}$ corresponding to $z \mapsto k(z, \cdot)$, so that $k(z, z') = \langle \phi(z), \phi(z') \rangle_{\mathcal{H}}$ (Aronszajn, 1950).

In this paper we use Sobolev spaces, defined on $Z \subseteq \mathbb{R}^d$, with $d \in \mathbb{N}$, an open set. For $s \in \mathbb{N}$, denote by $H^s(Z)$ the *Sobolev space* of functions whose weak derivatives up to order s are square-integrable, i.e., $H^s(Z) := \{f \in L^2(Z) \mid \|f\|_{H^s(Z)} < \infty\}$ and $\|f\|_{H^s(Z)} := \sum_{|\alpha| \leq s} \|D^\alpha f\|_{L^2(Z)}$ (Adams and Fournier, 2003). A remarkable property of $H^s(Z)$ that we are going to use in the proofs is that $H^s(Z) \subset C^k(Z)$ for any $s > d/2 + k$ and $k \in \mathbb{N}$. Moreover $H^{m+1}(Z) \subset H^m(Z), \forall m \in \mathbb{N}$.

Proposition 4 (Sobolev kernel, Wendland (2004)) *Let* $Z \subset \mathbb{R}^d$*,* $d \in \mathbb{N}$*, be an open bounded set. Let* $s > d/2$*. Define*

$$k_s(z, z') = c_s \|z - z'\|^{s-d/2} \mathcal{K}_{s-d/2}(\|z - z'\|), \quad \forall z, z' \in Z, \tag{8}$$

where $\mathcal{K} : \mathbb{R} \to \mathbb{R}$ *is the Bessel function of the second kind (see, e.g., Eq. 5.10 of the same book) and* $c_s = \frac{2^{1+d/2-s}}{\Gamma(s-d/2)}$*. Then the function* k_s *is a kernel. Denoting by* $\mathcal{H}_Z$ *the associated RKHS, when* Z *has Lipschitz boundary, then* $\mathcal{H}_Z = H^s(Z)$ *and the norms are equivalent.*

In the particular case of $s = d/2 + 1/2$, we have $k_s(z, z') = \exp(-\|z - z'\|)$. Note that the constant c_s is chosen such that $k_s(z, z) = 1$ for any $z \in Z$.

4. Positive operator representation for Kantorovich potentials

We start with the following representation result on the structure of the optimal potentials, which is one of our main contributions in this paper.

Theorem 5 *Let $\mu \in \mathcal{P}(X), \nu \in \mathcal{P}(Y)$ satisfying Assumption 1a and let $(u_\star, v_\star)$ be Kantorovich potentials such that $u_\star \in H^{s+2}(X)$ and $v_\star \in H^{s+2}(Y)$ for $s > d+1$. There exist functions $w_1, ..., w_d \in H^s(X \times Y)$ such that*

$$\tfrac{1}{2}\|x-y\|^2 - u_\star(x) - v_\star(y) = \textstyle\sum_{i=1}^d w_i(x,y)^2, \quad \forall (x,y) \in X \times Y.$$

Proof Denote by h the function $h(x,y) = c(x,y) - u_\star(x) - v_\star(y)$ for all $(x,y) \in X \times Y$. Let $f(x) = \frac{1}{2}\|x\|^2 - u_\star(x), x \in X$. By Brenier's theoremfor quadratic optimal transport (Brenier, 1987; Santambrogio, 2015, Theorem 1.22),

(i) $T = \nabla f$, where T is the optimal transport map from μ to ν,

(ii) f is convex on X,

(iii) h is characterized by $h(x,y) = f(x) + f^\star(y) - x^\top y$, where $f^\star : y \in Y \mapsto \sup_{x \in X} x^\top y - f(x)$ is the Fenchel-Legendre conjugate of f. Moreover, $f^\star(y) = \frac{1}{2}\|y^2\| - v_\star(y)$.

Further, from the properties of Fenchel-Legendre conjugacy (Rockafellar, 1970, Section 26), we have $T^{-1} = \nabla f^\star$. Hence, since $u_\star \in H^{s+2}(X)$ and $v_\star \in H^{s+2}(Y)$, we have

(iv) $T = \nabla f$ (resp. $T^{-1} = \nabla f^\star$) is a H^{s+1}-diffeomorphism from $\overline{X}$ to $\overline{Y}$ (resp. $\overline{Y}$ to $\overline{X}$).

Since $f \in H^{s+2}(X)$ and $s > d/2$ and X is a bounded open set with locally Lipschitz boundary (see Lemma 11), we have $H^{s+2}(X) \subset C^2(X)$ (Adams and Fournier, 2003) and the Hessian $\mathbf{H}_f(x)$ is well defined for any $x \in X$. Since, by (iv) , $T = \nabla f$ is a diffeomorphism, then, by Fenchel-Legendre conjugacy, f is strictly convex (Rockafellar, 1970). Hence by compactness of $\overline{X}$, f has a Hessian $\mathbf{H}_f(x)$ which is bounded away from 0. This implies:

(v) There exists $\rho > 0$ such that $\mathbf{H}_f(x) \succeq \rho \,\mathrm{Id}$ for all $x \in X$.

We will now use the decomposition (iii) along with a reparameterization of h to obtain a decomposition as a sum of squares. Let

$$\tilde{h}(x,z) \stackrel{\text{def}}{=} h(x, T(z)), \quad \forall (x,z) \in X \times X.$$

The effect of this change of coordinates is illustrated in Figure 1. Since f is differentiable, by the properties of the Fenchel-Legendre conjugate it holds that $f^\star(\nabla f(z)) = \nabla f(z)^\top z - f(z)$ for any $z \in X$ (Rockafellar, 1970). Therefore, from (i) we have that

$$\tilde{h}(x,z) = f(x) - f(z) - \nabla f(z)^\top (x - z).$$

Now, since X is convex, we can characterize f in terms of its Taylor expansion:

$$f(x) = f(z) + \nabla f(z)^\top (x-z) + (x-z)^\top \mathbf{R}(x,z)(x-z), \quad \forall x, z \in X,$$

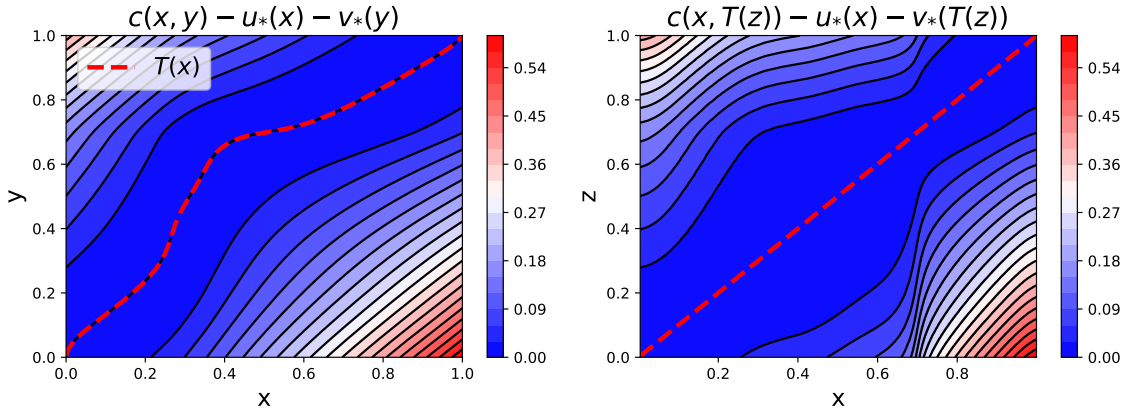

Figure 1: *Left*: dual constraint function $h(x, y)$ corresponding to the measures in Figure 2 in Appendix G. Note that h attains its minimum on the graph of transportation map T, and is elsewhere positive. *Right*: changing parameterization flattens h in the neighborhood of the graph of T. The red dotted line represent the zeros of both functions, and coincides with the graph of T in the original parameterization (*left*).

where $\mathbf{R}$ is the integral reminder $\mathbf{R}(x,z) \overset{\text{def}}{=} \int_0^1 (1-t)\mathbf{H}_f((1-t)x+tz)\mathrm{d}t$. This implies

$$\tilde{h}(x,z) = (x-z)^\top \mathbf{R}(x,z)(x-z), \quad \forall x, z \in X.$$

From (v), we get $\forall x, z, \mathbf{R}(x,z) \succeq \frac{\rho}{2}\mathrm{Id}$. In particular, for all $x, z \in X$, the matrix $\mathbf{R}(x,z)$ admits a positive square root $\sqrt{\mathbf{R}(x,z)}$. Further, since $\sqrt{\cdot}$ is C^∞ on the closed set $\{\mathbf{A} \in \mathbb{S}_+(\mathbb{R}^d) : \mathbf{A} \succeq \frac{\rho}{2}\mathrm{Id}\}$ and $\frac{\partial^2}{\partial x_i \partial x_j} f \in H^s(X)$ for all $i, j \in [d]$, the functions $r_{i,j} : (x,z) \mapsto e_i^\top \sqrt{\mathbf{R}(x,z)} e_j$ are in $H^s(X \times X)$ for all $i, j \in [d]$ (see Proposition 1 and Assumption 2b of Rudi et al., 2020), where $(e_1, ..., e_d)$ is the canonical ONB of $\mathbb{R}^d$. Define now the functions

$$\tilde{w}_i(x,z) \overset{\text{def}}{=} \sum_{j=1}^d r_{i,j}(x,z)(e_j^\top (x-z)), \quad \forall x, z \in X, i \in [d].$$

From the above arguments, it holds that $\tilde{w}_i \in H^s(X \times X), i \in [d]$, and $\tilde{h}(x,z) = \sum_{i=1}^d \tilde{w}_i^2(x,z)$. Now, since T is a H^{s+1}-diffeomorphism from X to Y, changing parameterization again we have $h(x,y) = \sum_{i=1}^d w_i^2(x,y),\ \forall (x,y) \in X \times Y$, with $w_i(x,y) = \tilde{w}_i(x, T^{-1}(y)),\ \forall (x,y) \in X \times Y$.

We conclude by proving that $w_i \in H^s(X \times Y)$ for all $i \in [d]$. Indeed, from (iv) T^{-1} is a H^{s+1}-diffeomorphism from $\bar{Y}$ to $\bar{X}$ and it is bi-Lipschitz (since T and T^{-1} are Lipschitz due to the continuity of their Hessian and the boundedness of X, Y). Define the map Q as $(x,y) \mapsto (x, T^{-1}(y))$ and note that $Q \in H^{s+1}(X \times Y, \mathbb{R}^{2d})$, by construction. Note that $\tilde{w}_i \in H^s(X \times X)$ and has bounded weak derivatives of order 1, since $s > d+1$ and $H^s(X \times X)$ is bounded (Adams and Fournier, 2003). The conditions above on $\tilde{w}_i, Q$ guarantee that $w_i = \tilde{w}_i \circ Q$ belongs to $H^s(X \times Y)$ (see, e.g., Theorem 1.2 of Campbell et al., 2015). ■

Theorem 5 implies the existence of $A_\star \in \mathbb{S}_+(\mathcal{H}_{XY})$ by effect of the reproducing property, when we consider a RKHS $\mathcal{H}_{XY}$ containing $H^s(X \times Y)$. The proof is in Appendix A, Page 17.

Corollary 6 *Let $\mathcal{H}_{XY}$ be a RKHS such that $H^s(X \times Y) \subseteq \mathcal{H}_{XY}$. Under the hypothesis of Theorem 5, there exists a positive operator $A_\star \in \mathbb{S}_+(\mathcal{H}_{XY})$ with rank at most d, such that*

$$c(x,y) - u_\star(x) - v_\star(y) = \langle \phi(x,y),\, A_* \phi(x,y) \rangle_{\mathcal{H}_{XY}}. \tag{9}$$

Finally, the following corollary shows the effect of Assumption 1 on the existence of $A_\star$. Note indeed that such an assumption implies smoothness of the Kantorovich potentials (De Philippis and Figalli, 2014). If $m > d$ they are smooth enough to apply Theorem 5 and the corollary above. The proof of the following corollary is in Appendix A, Page 17.

Corollary 7 (Effects of Asm. 1) *Let μ, ν satisfy Assumption 1 for $m > d$. Let $(u_\star, v_\star)$ be Kantorovich potentials for μ, ν. Let $\mathcal{H}_X = H^{m+3}(X), \mathcal{H}_Y = H^{m+3}(Y), \mathcal{H}_{XY} = H^m(X \times Y)$. Then $u_\star \in \mathcal{H}_X, v_\star \in \mathcal{H}_Y$ and there exists $A_\star \in \mathbb{S}_+(\mathcal{H}_{XY})$ satifying Eq. (9) and $\operatorname{rank} A_\star \leq d$.*

5. Subsampling the constraints

Given $\ell \in \mathbb{N}$ and a set of points $\tilde{Z}_\ell = \{(\tilde{x}_1, \tilde{y}_1), \dots, (\tilde{x}_\ell, \tilde{y}_\ell)\}$, define the *fill distance* (Wendland, 2004),

$$h_\ell = \sup_{x \in X, y \in Y} \min_{j \in [\ell]} \|(x,y) - (\tilde{x}_j, \tilde{y}_j)\|. \tag{10}$$

The following theorem, that is an adaptation of Theorem 4 from Rudi et al. (2020), quantifies the error of subsampling the constraints in terms of the fill distance.

Theorem 8 *Let X, Y satisfy Assumption 1a. Let $s \geq 3$ and $s > d$. Let $\mathcal{H}_X \subseteq H^s(X), \mathcal{H}_Y \subseteq H^s(Y), \mathcal{H}_{XY} = H^s(X \times Y)$. Let $\tilde{Z}_\ell \subset X \times Y$ be a set of points of cardinality ℓ and fill distance h_ℓ and let $u \in \mathcal{H}_X, v \in \mathcal{H}_Y, A \in \mathbb{S}_+(\mathcal{H}_{XY})$ satisfy*

$$c(\tilde{x}_j, \tilde{y}_j) - u(\tilde{x}_j) - v(\tilde{y}_j) = \langle \phi(\tilde{x}_j, \tilde{y}_j),\, A\phi(\tilde{x}_j, \tilde{y}_j) \rangle_{\mathcal{H}_{XY}}, \quad \forall j \in [\ell]. \tag{11}$$

There exist h_0, C_0 depending only on s, d, X, Y such that, when $h_\ell \leq h_0$, then

$$c(x,y) \geq u(x) + v(y) - \varepsilon, \quad \forall x, y \in X \times Y, \quad \text{where} \quad \varepsilon = Q h_\ell^{s-d},$$

where $Q = C_0(\|u\|_{\mathcal{H}_X} + \|v\|_{\mathcal{H}_Y} + \operatorname{Tr}(A))$. Note that h_0, C_0 depend only on d, m, X, Y.

The proof of the theorem above is reported in Appendix B, Page 17. Using the theorem above we are able to show that, given a maximizer $(\hat{u}, \hat{v}, \hat{A})$ of Problem 3 the couple $(\hat{u} - \varepsilon/2, \hat{v} - \varepsilon/2)$ is admissible for Problem 1. This is a crucial step to bound $|\widehat{\mathrm{OT}} - \mathrm{OT}|$ in terms of the fill distance h_ℓ and it is stated next.

Theorem 9 *Let μ, ν and $X, Y \subset \mathbb{R}^d$ satisfy Assumption 1a. Let $s > d$ and let $\mathcal{H}_X, \mathcal{H}_Y, \mathcal{H}_{XY}$ be RKHS on $X, Y, X \times Y$ such that $\mathcal{H}_X \subseteq H^s(X), \mathcal{H}_Y \subseteq H^s(Y), \mathcal{H}_{XY} = H^s(X \times Y)$. Assume that there exist two Kantorovich potentials $u_\star, v_\star$ such that $u_\star \in \mathcal{H}_X, v_\star \in \mathcal{H}_Y$ and there exists $A_\star \in \mathbb{S}_+(\mathcal{H}_{XY})$ that satisfies Eq. (9). Let $\widehat{OT}$ be computed according to Eq. (6) using a set of $\ell \in \mathbb{N}$ points $\tilde{Z}_\ell \subset X \times Y$ with fill distance h_ℓ. Let h_0, C_0 as in Theorem 8. Let*

$$\eta = C_0 h_\ell^{s-d}, \quad \gamma = \|w_\mu - \hat{w}_\mu\|_{\mathcal{H}_X} + \|w_\nu - \hat{w}_\nu\|_{\mathcal{H}_Y}.$$

When $h_\ell \leq h_0$, $\lambda_2 > 0$ and λ_1 is chosen such that $\lambda_1 \geq 2\eta$, then

$$|\widehat{OT} - OT| \;\leq\; 6\lambda_1 \operatorname{Tr}(A_\star) \;+\; 6\,\tfrac{(\gamma+\eta)^2}{\lambda_2} + 6\,\lambda_2\,\left(\|u_\star\|_{\mathcal{H}_X}^2 + \|v_\star\|_{\mathcal{H}_Y}^2\right).$$

The proof of the theorem above is in Appendix B, Page 17. Theorem 9 together with Theorem 5 (in particular Corollary 7) allow to prove Theorem 2. To bound h_ℓ in terms of ℓ, we used a result recalled in Lemma 12.

Proof of Theorem 2. First note that $k_X = k_{m+1}, k_Y = k_{m+1}, k_{XY} = k_m$ imply that $\mathcal{H}_X = H^{m+1}(X), \mathcal{H}_Y = H^{m+1}(X), \mathcal{H}_{XY} = H^m(X \times Y)$. Then, by Corollary 7 we have that under Assumption 1 for any Kantorovich potentials $u_\star, v_\star$ there exists a finite rank $A_\star \in \mathbb{S}_+(\mathcal{H}_{XY})$ such that $c(x,y) - u_\star(x) - v_\star(y) = \langle \phi(x,y), A_* \phi(x,y) \rangle_{\mathcal{H}_{XY}}$, and that $u_\star \in H^{m+3}(X) \subseteq H^{m+1}(X) = \mathcal{H}_X$ and analogously $v_\star \in \mathcal{H}_Y$. Among them, we select the triplet minimizing $\|u_\star\|^2_{\mathcal{H}_X} + \|v_\star\|^2_{\mathcal{H}_Y} + \mathrm{Tr}(A_\star)$ and we denote by $Q_{\mu,\nu}$ the resulting minimum. The proof is obtained by using this triplet in Theorem 9 and bounding h_ℓ as follows. First note that $X \times Y$ is a convex bounded set, since X, Y have the same property. As recalled in Lemma 11, Page 20 of the appendix, convex bounded sets have the so-called *uniform interior cone condition*. This guarantees that ℓ i.i.d. points sampled from a distribution ρ, that has a density bounded away from zero, achieve the following bound on the fill distance: $h_\ell \leq (C\ell^{-1} \log(C'\ell/\delta))^{-1/(2d)}$, with probability at least $1-\delta$, when $\ell \geq \ell_0$. Here ℓ_0, C, C' are constants that depend only on $d, X \times Y$ and the constant c_0 for which the density of ρ is bounded away from zero. The final constants C_1, C_2 depend also on $Q_{\mu,\nu}$. ■

We conclude with the following remark that is useful when we do not know the supports X, Y or we are not able to sample i.i.d. points from the uniform distribution on them.

Remark 10 (Sampling from $X \times Y$ using $\mu \otimes \nu$) *Since, under Assumption 1, we have μ and ν bounded away from 0, we can compute $\widehat{OT}$ by using ℓ i.i.d. samples from $\rho = \mu \otimes \nu$, obtaining the same guarantees as Theorem 2. Indeed, by inspecting the proof above, it is only required that ρ has support $X \times Y$, and has a density that is bounded away from 0. However, the constants ℓ_0, C_1, C_2 will depend also on how far the density of ρ is bounded away from 0.*

6. Estimators for the kernel mean embeddings of μ, ν

In this section we consider three classes of estimators $(\hat{w}_\mu, \hat{w}_\nu)$ for the kernel mean embeddings $w_\mu \in \mathcal{H}_X$ and $w_\nu \in \mathcal{H}_Y$ defined as $w_\mu = \int \phi_X(x) d\mu(x)$ and $w_\nu = \int \phi_Y(y) d\nu(y)$. As we observed in the introduction to Eq. (5), the operations we need to perform on $\hat{w}_\mu, \hat{w}_\nu$ to compute the algorithm are the evaluation of the norm $\|\hat{w}_\mu\|^2_{\mathcal{H}_X}$ and the evaluation of $\hat{w}_\mu(\tilde{x}_i)$ for $i \in [\ell]$ (and the same for $\hat{w}_\nu$). For each class we will specify such operations. Here, we assume that ϕ_X, ϕ_Y are uniformly bounded maps (as for the Sobolev kernel, where $\sup_{x \in X} \|\phi_X(x)\|^2_{\mathcal{H}_X} = \sup_{x \in X} k(x,x) = 1$, see Proposition 4). Clearly a class of estimators must only be chosen if we are able to perform the required operations.

Exact integral. Here we take $\hat{w}_\mu := w_\mu$ and $\hat{w}_\nu = w_\nu$ and we report only the operations for μ since the ones for ν are the same. This is the estimator that leads to the best rates as shown in Theorem 1 and Corollary 3. However it requires to perform the most difficult operations, i.e., (1) $\hat{w}_\mu(\tilde{x}_i) = \langle w_\mu, \phi_X(\tilde{x}_i) \rangle_{\mathcal{H}_X} = \int_X \langle \phi_X(\tilde{x}_i), \phi_X(x) \rangle d\mu(x) = \int_X k_X(\tilde{x}_i, x) d\mu(x)$ for all $i \in [\ell]$; (2) $\|\hat{w}_\mu\|^2_{\mathcal{H}_X} = \langle w_\mu, w_\mu \rangle_{\mathcal{H}_X} = \int_X \langle \phi_X(x), \phi_X(x') \rangle_{\mathcal{H}_X} d\mu(x) d\mu(x') = \int_X k_X(x,x') d\mu(x) d\mu(x')$. Moreover the costs C, E in Eq. (7) are $C = O(1), E = O(1)$.

Evaluation estimator. Here we assume we are able to evaluate $\mu(x_j)$ in a given set of points $x_1, \ldots, x_{n_\mu}$ with $n_\mu \in \mathbb{N}$ (analogously for ν on $y_1, \ldots, y_{n_\nu}$ with $n_\nu \in \mathbb{N}$). We define the estimators as $\hat{w}_\mu = \int_X \phi_X(x)\hat{g}_\mu(x)dx$ (analogously for $\hat{w}_\nu$). Here $\hat{g}_\mu \in \mathcal{H}_X$ is the *kernel least squares* estimator (Narcowich et al., 2005) of the density μ defined as $\hat{g}_\mu(x) = \sum_{j\in[n_\mu]} \alpha_j k_X(x_j, x)$ where $\alpha = \mathbf{K}_X^{-1} c_\mu$ and $\mathbf{K}_X \in \mathbb{R}^{n_\mu \times n_\mu}$, $(\mathbf{K}_X)_{i,j} = k_X(x_i, x_j)$ for all $i, j \in [n_\mu]$, while $c_\mu \in \mathbb{R}^{n_\mu}$, $c_\mu = (\mu(x_1), \ldots, \mu(x_n))$. In this case the steps are: (1) $\hat{w}_\mu(\tilde{x}_i) = \int_X \langle \phi_X(\tilde{x}_i), \phi_X(x)\rangle \hat{g}_\mu(x)dx = \sum_{j\in[n_\mu]} \alpha_j \int_X k_X(\tilde{x}_i, x)k_X(x_j, x)dx, i \in [\ell]$, and (2) $\|\hat{w}_\mu\|^2_{\mathcal{H}_X} = \sum_{i,t\in[n_\mu]} \alpha_i\alpha_t \int_X \int_X k_X(x_i, x)k(x, x')k(x_t, x')dxdx'$. Moreover the costs C, E in Eq. (7) are $C = O(n_\mu^3 + n_\nu^3), E = O(n_\mu + n_\nu)$.

Sample estimator. Here we assume we are able to sample from μ, ν. Let $x_1, \ldots, x_{n_\mu}$, $n_\mu \in \mathbb{N}$, be sampled i.i.d. from μ and $y_1, \ldots, y_{n_\nu}$, $n_\nu \in \mathbb{N}$ be sampled i.i.d. from ν. The estimators are $\hat{w}_\mu = \frac{1}{n_\mu} \sum_{j\in[n_\mu]} \phi_X(x_j)$, and analogously for $\hat{w}_\nu$. In this case the operations are: (1) $\hat{w}_\mu(\tilde{x}_i) = \frac{1}{n_\mu} \sum_{j\in[n_\mu]} k_X(x_j, \tilde{x}_i)$, and (2) $\|\hat{w}_\mu\|^2_{\mathcal{H}_X} = \frac{1}{n_\mu^2} \sum_{i,j\in[n_\mu]} k_X(x_i, x_j)$. Moreover the costs C, E in Eq. (7) are $C = O(n_\mu^2 + n_\nu^2), E = O(n_\mu + n_\nu)$.

The effects of the estimators above are studied in Theorem 1 and Corollary 3, reported in Section 1 and proven in Appendix E, Page 22, using standard tools from approximation theory and machine learning with kernel methods (Wendland, 2004; Caponnetto and De Vito, 2007; Muandet et al., 2017).

Acknowledgements. 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.

References

Robert A. Adams and John J. F. Fournier. *Sobolev Spaces*. Elsevier, 2003.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

Pierre-Cyril Aubin-Frankowski and Zoltan Szabo. Hard shape-constrained kernel machines. *Advances in Neural Information Processing Systems*, 33, 2020.

Dimitri P. Bertsekas. *Convex Optimization Theory*. Athena Scientific Belmont, 2009.

Nicolas Bonneel, Michiel Van De Panne, Sylvain Paris, and Wolfgang Heidrich. Displacement interpolation using Lagrangian mass transport. In *Proceedings of the 2011 SIGGRAPH Asia Conference*, pages 1–12, 2011.

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

Yann Brenier. Décomposition polaire et réarrangement monotone des champs de vecteurs. *CR Acad. Sci. Paris Sér. I Math.*, 305:805–808, 1987.

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

Daniel Campbell, Stanislav Hencl, and František Konopeckỳ. The weak inverse mapping theorem. *Zeitschrift für Analysis und ihre Anwendungen*, 34(3):321–342, 2015.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Liqun Chen, Shuyang Dai, Chenyang Tao, Haichao Zhang, Zhe Gan, Dinghan Shen, Yizhe Zhang, Guoyin Wang, Ruiyi Zhang, and Lawrence Carin. Adversarial text generation via feature-mover's distance. In *Advances in Neural Information Processing Systems 31*, pages 4666–4677. 2018.

Lenaic Chizat, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, and Gabriel Peyré. Faster Wasserstein distance estimation with the Sinkhorn divergence. *Advances in Neural Information Processing Systems*, 33, 2020.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, 2013.

Guido De Philippis and Alessio Figalli. The Monge–Ampère equation and its link to optimal transportation. *Bulletin of the American Mathematical Society*, 51(4):527–580, 2014.

Shai Dekel and Dany Leviatan. Whitney estimates for convex domains with applications to multivariate piecewise polynomial approximation. *Foundations of Computational Mathematics*, 4(4):345–368, 2004.

Jean Feydy, Benjamin Charlier, François-Xavier Vialard, and Gabriel Peyré. Optimal Transport for Diffeomorphic Registration. In *MICCAI 2017*, September 2017.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A. Poggio. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, pages 2053–2061, 2015.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617, 2018.

Aude Genevay, Lénaïc Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of Sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1574–1583, 2019.

Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised alignment of embeddings with Wasserstein Procrustes. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1880–1890. PMLR, 2019.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1): 723–773, 2012.

Pierre Grisvard. *Elliptic Problems in Nonsmooth Domains*. SIAM, 1985.

Didier Henrion and Jean-Bernard Lasserre. Graph recovery from incomplete moment information, 2020.

Nhat Ho, XuanLong Nguyen, Mikhail Yurochkin, Hung Hai Bui, Viet Huynh, and Dinh Phung. Multilevel clustering via Wasserstein means. In *International Conference on Machine Learning*, pages 1501–1509. PMLR, 2017.

Jan-Christian Hütter and Philippe Rigollet. Minimax rates of estimation for smooth optimal transport maps. *arXiv preprint arXiv:1905.05828*, 2019.

David Krieg and Mathias Sonnleitner. Random points are optimal for the approximation of sobolev functions. *arXiv preprint arXiv:2009.11275*, 2020.

Jean Bernard Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11(3):796–817, 2001.

Jean Bernard Lasserre. *Moments, Positive Polynomials and Their Applications*. Imperial College Press, 2009.

Tengyuan Liang. On how well generative adversarial networks learn densities: Nonparametric and parametric results. *arXiv preprint arXiv:1811.03179*, 2018.

Tengyuan Liang. Estimating certain integral probability metric (IPM) is as hard as estimating under the IPM. *arXiv preprint arXiv:1911.00730*, 2019.

Ulysse Marteau-Ferey, Francis Bach, and Alessandro Rudi. Non-parametric models for non-negative functions. *Advances in Neural Information Processing Systems*, 2020.

Gonzalo Mena and Jonathan Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In *Advances in Neural Information Processing Systems*, pages 4543–4553, 2019.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. *Kernel Mean Embedding of Distributions: A Review and Beyond*, volume 10. Now Foundations and Trends, 2017.

Francis Narcowich, Joseph Ward, and Holger Wendland. Sobolev bounds on functions with scattered zeros, with applications to radial basis function surface fitting. *Mathematics of Computation*, 74(250):743–763, 2005.

Arkadi Nemirovski. Interior point polynomial time methods in convex programming. *Lecture notes*, 2004.

Yurii Nesterov and Arkadii Nemirovskii. *Interior-Point Polynomial Algorithms in Convex Programming*. Society for Industrial and Applied Mathematics, 1994.

Jonathan Niles-Weed and Philippe Rigollet. Estimation of Wasserstein distances in the spiked transport model. *arXiv preprint 1909.07513*, 2019.

Pablo A Parrilo. Semidefinite programming relaxations for semialgebraic problems. *Mathematical programming*, 96(2):293–320, 2003.

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

Ievgen Redko, Nicolas Courty, Rémi Flamary, and Devis Tuia. Optimal transport for multi-source domain adaptation under target shift. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 849–858. PMLR, 2019.

A. Reznikov and E. B. Saff. The covering radius of randomly distributed points on a manifold. *International Mathematics Research Notices*, 2016(19):6065–6094, 2016.

R. Tyrrell Rockafellar. *Convex Analysis*, volume 36. Princeton University Press, 1970.

Lorenzo Rosasco, Mikhail Belkin, and Ernesto De Vito. On learning with integral operators. *Journal of Machine Learning Research*, 11(2), 2010.

Alessandro Rudi, Ulysse Marteau-Ferey, and Francis Bach. Finding global minima via kernel approximations. In *Arxiv preprint arXiv:2012.11978*, 2020.

Tim Salimans, Dimitris Metaxas, Han Zhang, and Alec Radford. Improving GANs using optimal transport. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Progress in Nonlinear Differential Equations and Their Applications. Springer International Publishing, 2015.

Ilya Meerovich Sobol. On the distribution of points in a cube and the approximate evaluation of integrals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 7(4):784–802, 1967.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008.

I Tolstikhin, O Bousquet, S Gelly, and B Schölkopf. Wasserstein auto-encoders. In *6th International Conference on Learning Representations (ICLR 2018)*. OpenReview. net, 2018.

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.

Jonathan Weed and Quentin Berthet. Estimation of smooth densities in Wasserstein distance. In *Conference on Learning Theory*, pages 3118–3119, 2019.

Holger Wendland. *Scattered Data Approximation*, volume 17. Cambridge University Press, 2004.

Holger Wendland and Christian Rieger. Approximate interpolation with applications to selecting smoothing parameters. *Numer. Math.*, 101(4):729–748, October 2005.

Hongteng Xu, Wenlin Wang, Wei Liu, and Lawrence Carin. Distilled Wasserstein learning for word embedding and topic modeling. In *Advances in Neural Information Processing Systems 31*, pages 1716–1725. 2018.

Appendix A. Proofs of Corollary 6 and Corollary 7

Proof of Corollary 6. Define $h(x,y) := c(x,y) - u_\star(x) - v_\star(y), \forall (x,y) \in X \times Y$. By Theorem 5, there exist $w_1, ..., w_d \in H^s(X \times Y)$ such that $h = \sum_{i=1}^d w_i^2$. Since $H^s(X \times Y) \subseteq \mathcal{H}_{XY}$, then $w_1, \dots, w_d \in \mathcal{H}_{XY}$. Hence $A_\star = \sum_{i=1}^d w_i \otimes w_i \in \mathbb{S}_+(\mathcal{H}_{XY})$. Moreover, by the reproducing property of $\mathcal{H}_{XY}$, $A_\star$ satisfies $\langle \phi(x,y),\, A_\star \phi(x,y) \rangle_{\mathcal{H}_{XY}} = \sum_{i=1}^d \langle w_j,\, \phi(x,y) \rangle^2_{\mathcal{H}_{XY}} = \sum_{i=1}^d w(x,y)^2 = h(x,y)$, for all $(x,y) \in X \times Y$. Finally, note that $\operatorname{rank} A_\star \leq d$ by construction. ■

Proof of Corollary 7. Theorem 3.3 in De Philippis and Figalli (2014) implies that Kantorovich potentials satisfy $u_\star \in C^{m+2,1}(\overline{X}), v_\star \in C^{m+2,1}(\overline{Y})$, where $C^{m+2,1}(\overline{Z})$ for an open set Z is the space of real functions over Z that are $m+2$-times differentiable, with all the derivatives of order $m+2$ that are Lipschitz continuous (Adams and Fournier, 2003, 1.29, Page 10). Since X is convex and bounded, then Lipschitz continuity of all the derivatives of order $m+2$ implies that all the weak derivatives up to order $m+3$ are in $L^\infty(X)$. Since $L^\infty(X) \subset L^2(X)$, by the boundedness of X, we have $C^{m+2,1}(\overline{X}) \subset H_2^{m+3}(X)$. Analogously $C^{m+2,1}(\overline{Y}) \subset H_2^{m+3}(Y)$. Then $u_\star \in H^{m+3}(X), v_\star \in H^{m+3}(Y)$, and we can apply Theorem 5 and Corollary 6 with $s = m+1$, when $m > d$, which guarantee the existence of $A_\star \in \mathbb{S}_+(\mathcal{H}_{XY})$, since $\mathcal{H}_{XY} = H^s(X \times Y)$. ■

Appendix B. Proofs of Theorem 8 and Theorem 9

Proof of Theorem 8. Let $f(x,y) = c(x,y) - u(x)\mathbf{1}_X(y) - v(y)\mathbf{1}_Y(x) - \langle \phi(x,y),\, A\phi(x,y) \rangle$ for any $x, y \in X \times Y$, where the function $\mathbf{1}_X(x)$ and $\mathbf{1}_Y(y)$ are respectively the constant function 1 over X and over Y. By construction $f \in H^s(X \times Y)$. Since X, Y are open bounded convex sets, then $X \times Y$ has the same property which, in turn, implies the so-called *uniform interior cone condition* (see Lemma 11, Page 20 in the appendix). Then we can apply Theorem 1 of Narcowich et al. (2005) for which there exist h_0, C_0 depending only on s, d, X, Y such that for any $h_\ell \leq h_0$ the following holds

$$\sup_{(x,y) \in X \times Y} |f(x,y)| \leq \varepsilon, \quad \varepsilon := C_1 h_\ell^{s-d} |f|_{H^s(X \times Y)},$$

where $|g|_{H^s(X\times Y)} := \sum_{|\alpha|=s} \|D^\alpha g\|_{L^2(X\times Y)} \leq \|g\|_{H^s(X\times Y)}$ for any $g \in H^s(X \times Y)$. Let $r_A(x,y) = \langle \phi(\tilde{x}_j, \tilde{y}_j),\, A\phi(\tilde{x}_j, \tilde{y}_j)\rangle_{\mathcal{H}_{XY}}$, we have

$$|f|_{H^s(X\times Y)} \leq |c|_{H^s(X\times Y)} + |u\mathbf{1}_X|_{H^s(X\times Y)} + |v\mathbf{1}_Y|_{H^s(X\times Y)} + |r_A|_{H^s(X\times Y)}.$$

Now $|c|_{H^s(X\times Y)} = 0$ since c is the quadratic cost and $s \geq 3$ and $|u\mathbf{1}_X|_{H^s(X\times Y)} = |u|_{H^s(X)} \leq \|u\|_{H^s(X)}$, $|v\mathbf{1}_Y|_{H^s(X\times Y)} = |v|_{H^s(Y)} \leq \|v\|_{H^s(Y)}$. We recall now that for any two Banach spaces satisfying $B \subseteq A$, we have $\|w\|_A \leq \|w\|_B$ for any $w \in B$. Then $\|u\|_{H^s(X)} \leq \|u\|_{\mathcal{H}_X}$, $\|v\|_{H^s(Y)} \leq \|v\|_{\mathcal{H}_Y}$. Finally $|r_A|_{H^s(X\times Y)} \leq \|r_A\|_{H^s(X\times Y)} \leq C_1 \mathrm{Tr}(A)$ via Lemma 9 page 41 from Rudi et al. (2020) and Proposition 1 of the same paper, where C_2 depends only on s, X, Y. Now since $|f(x,y)| \leq \varepsilon$ for all $(x,y) \in X \times Y$, and since $r_A(x,y) \geq 0$ for any $x, y \in X \times Y$ since A is a positive operator, we have

$$-\varepsilon \leq f(x,y) \leq f(x,y) + r_A(x,y) = c(x,y) - u(x) - v(y), \quad \forall (x,y) \in X \times Y.$$

The final result is obtained by defining $C_0 = C_1(1 + C_2)$. ■

Proof of Theorem 9. Denote by $V(u,v)$ the functional of Problem 1 and by $\hat{V}_{\lambda_1,\lambda_2}(u,v,A)$ the functional of Problem 3. Then $\mathrm{OT} = V(u_\star, v_\star)$. Denote by $\Delta(u,v)$ the quantity

$$\Delta(u,v) = |\langle u,\, \hat{w}_\mu - w_\mu\rangle_{\mathcal{H}_X} + \langle v,\, \hat{w}_\nu - w_\nu\rangle_{\mathcal{H}_Y}|.$$

Denote by $\hat{R}^2(u,v)$ the quantity $\hat{R}^2(u,v) = \|u\|^2_{\mathcal{H}_X} + \|v\|^2_{\mathcal{H}_Y}$ for any $u \in \mathcal{H}_X, v \in \mathcal{H}_Y$.

Step 0. Admissibility of $u_\star, v_\star, A_\star$ and existence of a maximizer. Note that $(u_\star, v_\star, A_\star)$ is an admissible point for Problem 3, since the triple satisfies $c_\star(x,y) - u_\star(x) - v_\star(y) = \langle \phi(x,y),\, A_\star\phi(x,y)\rangle$ $\forall (x,y) \in X \times Y$, and Problem 3 applies the same constraints but on a subset of $X \times Y$. Moreover $\lambda_1, \lambda_2 > 0$, this is enough to guarantee the existence of a maximizer for Problem 3. Indeed, as we recall in Lemma 14, Page 21, in the appendix, a form of *representer theorem* holds for Problem 3 (see Lemma 13, Page 21 in the appendix), moreover the functional is coercive on the finite-dimensional space induced by the data, it has an upper bound and the problem has an admissible point. Now denote by $(\hat{u}, \hat{v}, \hat{A})$ a maximizer of Problem 3 and define $\overline{\mathrm{OT}} := \hat{V}_{\lambda_1,\lambda_2}(\hat{u}, \hat{v}, \hat{A})$. By construction $\overline{\mathrm{OT}}$ is the maximum of $\hat{V}_{\lambda_1,\lambda_2}$. Then by definition of $\widehat{\mathrm{OT}}$ in Eq. (6),

$$\widehat{\mathrm{OT}} = \overline{\mathrm{OT}} + \lambda_1 \mathrm{Tr}(\hat{A}) + \lambda_2 R^2(\hat{u}, \hat{v}).$$

Step 1. Subsampling the inequality. The assumption on h_ℓ and the fact that the constraints of Eq. (3) satisfy Eq. (11) allow to apply Theorem 8, from which there exists a couple $(\hat{u}_\varepsilon, \hat{v}_\varepsilon)$ that is admissible for Problem 1, where $\hat{u}_\varepsilon = \hat{u} - \frac{\varepsilon}{2}$ and $\hat{v}_\varepsilon = \hat{v} - \frac{\varepsilon}{2}$, for any $\varepsilon \geq C_0 h_\ell^{s-d}(\|\hat{u}\|_{\mathcal{H}_X} + \|\hat{v}\|_{\mathcal{H}_Y} + \mathrm{Tr}(\hat{A}))$. In particular, since $\|u\|_{\mathcal{H}_X} + \|v\|_{\mathcal{H}_Y} \leq 2R(\hat{u}, \hat{v})$, we choose $\varepsilon = \eta(R(\hat{u}, \hat{v}) + \mathrm{Tr}(\hat{A}))$, with $\eta = C_0 h_\ell^{s-d}$.

Step 2. Bounding $\widehat{\mathrm{OT}} - \mathrm{OT}$. Since OT is the maximum of Problem 1 then

$$\mathrm{OT} = V(u_\star, v_\star) \geq V(\hat{u}_\varepsilon, \hat{v}_\varepsilon) = V(\hat{u}, \hat{v}) - \varepsilon \geq \widehat{\mathrm{OT}} - \Delta(\hat{u}, \hat{v}) - \varepsilon. \tag{12}$$

Analogously, since $(\hat{u}, \hat{v}, \hat{A})$ maximize Problem 3, then

$$\begin{aligned}\overline{\mathrm{OT}} = \hat{V}_{\lambda_1,\lambda_2}(\hat{u}, \hat{v}, \hat{A}) &\geq \hat{V}_{\lambda_1,\lambda_2}(u_\star, v_\star, A_\star) \\ &= V(u_\star, v_\star) - \left[V(u_\star, v_\star) - \hat{V}_{\lambda_1,\lambda_2}(u_\star, v_\star, A_\star)\right] \\ &\geq \mathrm{OT} - \Delta(u_\star, v_\star) - \lambda_1 \mathrm{Tr}(A_\star) + \lambda_2 R^2(u_\star, v_\star).\end{aligned}$$

By expressing $\overline{\mathrm{OT}}$ in terms of $\widehat{\mathrm{OT}}$ in the inequality above and combining it with Eq. (12), we have

$$\lambda_1 \mathrm{Tr}(\hat{A} - A_\star) + \lambda_2(R^2(\hat{u}, \hat{v}) - R^2(u_\star, v_\star)) - \Delta(u_\star, v_\star) \leq \widehat{\mathrm{OT}} - \mathrm{OT} \leq \Delta(\hat{u}, \hat{v}) + \varepsilon. \quad (13)$$

Step 3. Bound on Δ. Note that for two RKHS $\mathcal{H}, \mathcal{K}$ and any $u, v \in \mathcal{H}, w, z \in \mathcal{K}$, the following identity holds: $\langle u,\, v\rangle_{\mathcal{H}} + \langle w,\, z\rangle_{\mathcal{K}} = \langle (u, w),\, (v, z)\rangle_{\mathcal{H}\oplus\mathcal{K}}$, where $\mathcal{H} \oplus \mathcal{K}$ is a RKHS (Aronszajn, 1950). This implies $(\langle u,\, v\rangle_{\mathcal{H}} + \langle w,\, z\rangle_{\mathcal{K}})^2 = \|(u, w)\|^2_{\mathcal{H}\oplus\mathcal{K}} \|(v, z)\|^2_{\mathcal{H}\oplus\mathcal{K}} = (\|u\|^2_{\mathcal{H}} + \|v\|^2_{\mathcal{K}})(\|w\|^2_{\mathcal{H}} + \|z\|^2_{\mathcal{K}})$. Applying this inequality to Δ leads to the following result:

$$\Delta(a, b) \leq R(a, b)\gamma. \quad (14)$$

Step 4. Bounding $\mathrm{Tr}(\hat{A}), R(\hat{u}, \hat{v})$. The bound Eq. (13) implies

$$\lambda_1 \mathrm{Tr}(\hat{A}) + \lambda_2 R^2(\hat{u}, \hat{v}) \leq \lambda_1 \mathrm{Tr}(A_\star) + \lambda_2 R^2(u_\star, v_\star) + \Delta(u_\star, v_\star) + \Delta(\hat{u}, \hat{v}) + \varepsilon.$$

By bounding Δ via Eq. (14), expanding the definition of ε and reordering the terms in the inequality above, we obtain

$$\alpha \mathrm{Tr}(\hat{A}) + \lambda_2 R^2(\hat{u}, \hat{v}) - \beta R(\hat{u}, \hat{v}) \leq \lambda_1 \mathrm{Tr}(A_\star) + \lambda_2 R^2(u_\star, v_\star) + \gamma R(u_\star, v_\star),$$

with $\alpha = \lambda_1 - \eta$ and $\beta = \gamma + \eta$. By completing the square in $R^2(\hat{u}, \hat{v})$, the inequality above is rewritten as

$$\alpha \mathrm{Tr}(\hat{A}) + \lambda_2 (R(\hat{u}, \hat{v}) - \tfrac{\beta}{2\lambda_2})^2 \leq \lambda_1 \mathrm{Tr}(A_\star) + \lambda_2 R^2(u_\star, v_\star) + \gamma R(u_\star, v_\star) + \tfrac{\beta^2}{4\lambda_2}.$$

Since $\alpha \geq \lambda_1/2$, by the assumption $\lambda_1 \geq 2\eta$, the inequality above implies

$$\tfrac{\lambda_1}{2} \mathrm{Tr}(\hat{A}) \leq \lambda_2 S, \quad R(\hat{u}, \hat{v}) \leq \tfrac{\beta}{2\lambda_2} + \sqrt{S}. \quad (15)$$

with $S := \frac{\lambda_1}{\lambda_2} \mathrm{Tr}(A_\star) + R^2(u_\star, v_\star) + \frac{\gamma}{\lambda_2} R(u_\star, v_\star) + \frac{\beta^2}{4\lambda_2^2}$.

Conclusion. From the lower bound in Eq. (13) and the bound Eq. (14) on Δ, we have that

$$\widehat{\mathrm{OT}} - \mathrm{OT} \geq -\lambda_1 \mathrm{Tr}(A_\star) - \lambda_2 R^2(u_\star, v_\star) - \gamma R(u_\star, v_\star) \geq -\lambda_2 S.$$

From the upper bound in Eq. (13), the bound for Δ in Eq. (14), the bound for $\mathrm{Tr}(\hat{A}), R(\hat{u}, \hat{v})$ in Eq. (15), the definition of ε, and the fact that $\lambda_1 \geq 2\eta$, we have

$$\widehat{\mathrm{OT}} - \mathrm{OT} \leq \beta R(\hat{u}, \hat{v}) + \tfrac{\lambda_1}{2} \mathrm{Tr}(\hat{A}) \leq \tfrac{\beta^2}{2\lambda_2} + \beta\sqrt{S} + \lambda_2 S.$$

Then $|\widehat{\mathrm{OT}} - \mathrm{OT}| \leq \frac{\beta^2}{2\lambda_2} + \beta\sqrt{S} + \lambda_2 S$. To conclude, by noting that $S \leq \frac{\lambda_1}{\lambda_2}\mathrm{Tr}(A_\star) + (R(u_\star, v_\star) + \frac{\beta}{2\lambda_2^2})^2$, since $\gamma \leq \beta$, and that $(a+b+c)^2 \leq 3a^2 + 3b^2 + 3c^2$ for any $a, b, c \geq 0$, we have

$$\frac{\beta^2}{2\lambda_2} + \beta\sqrt{S} + \lambda_2 S \leq 2\lambda_2(\frac{\beta}{2\lambda} + \sqrt{S})^2 \leq 2\lambda_2 \left(R(u_\star, v_\star) + \frac{\beta}{\lambda_2} + \sqrt{\frac{\lambda_1}{\lambda_2}\mathrm{Tr}(A_\star)} \right)^2$$
$$\leq 6\lambda_2 (R^2(u_\star, v_\star) + \frac{\beta^2}{\lambda_2^2} + \frac{\lambda_1}{\lambda_2}\mathrm{Tr}(A_\star)).$$

■

Appendix C. Additional results on convex sets and random points

We first recall that the following property about bounded sets in Euclidean spaces.

Lemma 11 (Krieg and Sonnleitner (2020)) *Let $Z \subset \mathbb{R}^q$, $q \in \mathbb{N}$ be a non-empty open set. The following holds:*

1. *If Z is a* bounded Lipschitz domain *(Grisvard, 1985), then it satisfies the* uniform interior cone condition *(Wendland, 2004).*
2. *If Z is a convex bounded set, then it is a bounded Lipschitz domain.*

Proof For the first point, see Lemma 5 of Krieg and Sonnleitner (2020) or Theorem 1.2.2.2 of Grisvard (1985). For the second point, see Lemma 4 of Krieg and Sonnleitner (2020) or Lemma 7 in Dekel and Leviatan (2004). Also, the fact that a convex bounded open set has the uniform interior cone condition is implied by Proposition 11.26 of Wendland (2004), that proves it for the more general class of sets called *star shaped sets w.r.t. a ball.* ■

Lemma 12 (Fill distance of i.i.d points on a u.i.c. set (Reznikov and Saff, 2016)) *Let $Z \subset \mathbb{R}^q$, $q \in \mathbb{N}$ be a non-empty open set satisfying the uniform interior cone condition (see Theorem 11). Let $\tilde{Z}_\ell$ be a collection of ℓ points sampled independently and uniformly at random from a probability ρ that admits density (denote it by p) and such that $p(z) \geq c_0 > 0$ for any $z \in Z$. Let $\delta \in (0, 1]$. Then there exist ℓ_0, C_1, C_2 depending on c_0, Z, ρ, q, r_0 such that for $\ell \geq \ell_0$, the following holds with probability at least $1 - \delta$:*

$$h_\ell \leq (C_1 \ell^{-1} \log(C_2 \ell / \delta))^{1/q}.$$

Proof The uniform interior cone condition guarantees that there exists a cone C such that for any $z \in Z$ there exists a spherical cone of radius r such that $C_z \subseteq Z$ that is congruent to C and with vertex in z. Then, for any $r \leq r_0$ there exists c_1 such that

$$\rho(B(z,r) \cap Z) \geq \rho(B(z,r) \cap C_z) \geq c_0 \frac{\mathrm{vol}(B(z,r) \cap C_z)}{\mathrm{vol}(Z)} \geq \frac{c_0 c_1}{\mathrm{vol}(Z)} r^q.$$

Then we can apply Theorem 2.1 of Reznikov and Saff (2016) with $\Phi(r) = \frac{c_0 c_1}{\mathrm{vol}(Z)} r^q$ obtaining that there exists ℓ_0, C_1, C_2 depending on c_0, c_1, q, r_0, Z such that with probability at least $1 - \delta$,

$$h_\ell \leq (C_1 \ell^{-1} \log(C_2 \ell / \delta))^{1/q}.$$

■

Appendix D. Representer theorem and coercivity for Problem (3)

We first adapt the proof of the representer theorem from Marteau-Ferey et al. (2020). We use it to prove coercivity of the functional. Given the RKHS $\mathcal{H}_X, \mathcal{H}_Y, \mathcal{H}_{XY}$ and with the same notation of Problem (3) define $\hat{\mathcal{H}}_X = \text{span}\{\hat{w}_\mu, \phi_X(\tilde{x}_1), \dots, \phi_X(\tilde{x}_\ell)\}$, $\hat{\mathcal{H}}_Y = \text{span}\{\hat{w}_\nu, \phi_Y(\tilde{y}_1), \dots, \phi_Y(\tilde{y}_\ell)\}$ and $\hat{\mathcal{H}}_{XY} = \text{span}\{\phi(x_1, y_1), \dots, \phi(x_\ell, y_\ell)\}$.

Lemma 13 (Representer theorem, Marteau-Ferey et al. (2020)) *Let $\lambda_1, \lambda_2 > 0$. Denote by $V_{\lambda_1,\lambda_2}(u, v, A)$ the objective function of Problem (3). Let $u_1 \in \hat{\mathcal{H}}_X, u_2 \in \hat{\mathcal{H}}_X^\perp$, $v_1 \in \hat{\mathcal{H}}_Y, v_2 \in \hat{\mathcal{H}}_Y^\perp$ and $A_1 \in \hat{\mathcal{H}}_{XY} \otimes \hat{\mathcal{H}}_{XY}, A_2 \in \hat{\mathcal{H}}_{XY}^\perp \otimes \hat{\mathcal{H}}_{XY}, A_3 \in \hat{\mathcal{H}}_{XY}^\perp \otimes \hat{\mathcal{H}}_{XY}^\perp$. Assume that u_2 or v_2 or A_2 or A_3 are different from 0. Set $u = u_1 + u_2$, $v = v_1 + v_2$ and $A = A_1 + A_2 + A_2^* + A_3$ and assume that u, v, A is an admissible point for Problem (3) Then*

1. *(u_1, v_1, A_1) is an admissible point for Problem (3),*
2. *$V_{\lambda_1,\lambda_2}(u_1, v_1, A_1) > V_{\lambda_1,\lambda_2}(u, v, A)$.*

Proof We can decompose any $u \in \mathcal{H}_X, v \in \mathcal{H}_Y, A \in \mathbb{S}_+(\mathcal{H}_{XY})$ as $u_1 \in \hat{\mathcal{H}}_X, u_2 \in \hat{\mathcal{H}}_X^\perp$, $v_1 \in \hat{\mathcal{H}}_Y, v_2 \in \hat{\mathcal{H}}_Y^\perp$ and $A_1 \in \hat{\mathcal{H}}_{XY} \otimes \hat{\mathcal{H}}_{XY}, A_2 \in \hat{\mathcal{H}}_{XY}^\perp \otimes \hat{\mathcal{H}}_{XY}, A_3 \in \hat{\mathcal{H}}_{XY}^\perp \otimes \hat{\mathcal{H}}_{XY}^\perp$. Note that the components u_2, v_2, A_2, A_3 do not impact the constraints or the functional but are only penalized by the regularizer, indeed $\langle u_2, \phi_X(\tilde{x}_i)\rangle_{\mathcal{H}_X} = 0$ since $\phi_X(\tilde{x}_i) \in \hat{\mathcal{H}}_X$, while $u_2 \in \hat{\mathcal{H}}_X^\perp$, then

$$u(\tilde{x}_i) = \langle u, \phi_X(\tilde{x}_i)\rangle_{\mathcal{H}_X} = \langle u_1, \phi_X(\tilde{x}_i)\rangle_{\mathcal{H}_X} + \langle u_2, \phi_X(\tilde{x}_i)\rangle_{\mathcal{H}_X} = \langle u_1, \phi_X(\tilde{x}_i)\rangle_{\mathcal{H}_X} = u_1(\tilde{x}_i).$$

The same reasoning holds for $v(\tilde{x}_i)$ and for $\langle u, w_\mu\rangle_{\mathcal{H}_X}, \langle v, w_\nu\rangle_{\mathcal{H}_Y}$. For A analogously we have that $t = A_2\phi(\tilde{x}_i, \tilde{y}_i) \in \hat{\mathcal{H}}_{XY}^\perp$ so $\langle \phi(\tilde{x}_i, \tilde{y}_i), A_2\phi(\tilde{x}_i, \tilde{y}_i)\rangle_{\mathcal{H}_{XY}} = \langle \phi(\tilde{x}_i, \tilde{y}_i), t\rangle_{\mathcal{H}_{XY}} = 0$ and similarly for A_3, we have $\langle \phi(\tilde{x}_i, \tilde{y}_i), A_3\phi(\tilde{x}_i, \tilde{y}_i)\rangle_{\mathcal{H}_{XY}} = 0$. Let's see what happens to the penalization terms. For the quadratic term, we have $-\|u\|^2_{\mathcal{H}_X} = -\|u_1\|^2_{\mathcal{H}_X} - \|u_2\|^2_{\mathcal{H}_X} < -\|u_1\|^2_{\mathcal{H}_X}$ and analogously for v. For the trace term we have $\text{Tr}(A) = \text{Tr}(A_1) + \text{Tr}(A_3)$. Moreover $A \in \mathbb{S}_+(\mathcal{H}_{XY})$ implies that $A_1 \succeq 0$ and by the Schur complement property $A_3 \succeq A_2^* A_1^{-1} A_2 \succeq 0$. Then if $A_3 \succeq 0$ and different from zero we have $-\text{Tr}(A) < -\text{Tr}(A_1)$. If A_2 is different from zero this implies by the Schur complement property that A_3 is a positive definite operator different from zero and so $-\text{Tr}(A) < -\text{Tr}(A_1)$. ■

Lemma 14 (Problem (3) has a maximizer) *When $\lambda_1, \lambda_2 > 0$ and when there exists an admissible point $(\bar{u}, \bar{v}, \bar{A})$ with $\bar{u} \in \mathcal{H}_X, \bar{v} \in \mathcal{H}_Y, \bar{A} \in \mathbb{S}_+(\mathcal{H}_{XY})$, Problem (3) admits a maximizer.*

Proof First define $S = \mathcal{H}_X \times \mathcal{H}_Y \times (\mathcal{H}_{XY} \otimes \mathcal{H}_{XY})$ and $\hat{S} = \hat{\mathcal{H}}_X \times \hat{\mathcal{H}}_Y \times (\hat{\mathcal{H}}_{XY} \otimes \hat{\mathcal{H}}_{XY})$. From the lemma above, we have that for any admissible point in $S \setminus \hat{S}$ there exists another admissible point in $\hat{S}$ that has a strictly larger value. Note moreover that $\hat{S}$ is a finite dimensional Hilbert space (with dimension at most $\ell^2(\ell+1)^2$) and that (minus) the functional of Problem (3) is coercive on it. Note moreover that Eq. (3) is bounded from above since $V_{\lambda_1,\lambda_2}(u, v, A) \leq \|u\|_{\mathcal{H}_X}\|\hat{w}_\mu\|_{\mathcal{H}_X} + \|v\|_{\mathcal{H}_Y}\|\hat{w}_\nu\|_{\mathcal{H}_Y} - \lambda_2(\|u\|^2_{\mathcal{H}_X} + \|v\|^2_{\mathcal{H}_Y})$ has a maximum in u, v. Since Problem (3) has also an admissible point by assumption, then it admits at least one maximizer (see, e.g., Proposition 3.2.1, Page 119 of Bertsekas, 2009). ■

Appendix E. Proof of Corollary 3 and Theorem 1

Proof of Corollary 3. The result is obtained by plugging in Theorem 2 the bounds on $\gamma = \|w_\mu - \hat{w}_\mu\|_{\mathcal{H}_X} + \|w_\nu - \hat{w}_\nu\|_{\mathcal{H}_Y}$. We deal now with the three scenarios.

(Exact integral) In this case, since $\hat{w}_\mu := w_\mu, \hat{w}_\nu := w_\nu$, then $\gamma = 0$.

(Evaluation) We use here the construction used to analyze kernel least squares, e.g., from Caponnetto and De Vito (2007); Rosasco et al. (2010) (see also Steinwart and Christmann, 2008). We will do the construction for w_μ, since the case of w_ν is analogous. We recall that $\phi_X : X \to \mathcal{H}_X$ is uniformly bounded and continuous on X, since we are considering the Sobolev kernel (see Proposition 4). Denote by $T \in \mathbb{S}_+(\mathcal{H}_X)$ the operator defined as $T = \int \phi_X(x) \otimes \phi_X(x) dx$ where dx is the Lebesgue measure. Note that T is trace class, indeed $\mathrm{Tr}(T) = \int \mathrm{Tr}(\phi_X(x) \otimes \phi_X(x)) dx = \int \|\phi_X(x)\|^2_{\mathcal{H}_X} dx \leq \mathrm{vol}(X) \sup_{x \in X} \|\phi_X(x)\|^2_{\mathcal{H}_X} < \infty$. For any $f, g \in \mathcal{H}_X$, by the reproducing property

$$\begin{aligned}\langle f, Tg \rangle_{\mathcal{H}_X} &= \int \langle f, (\phi_X(x) \otimes \phi_X(x)) g \rangle_{\mathcal{H}_X} dx \\ &= \int \langle f, \phi_X(x) \rangle_{\mathcal{H}_X} \langle g, \phi_X(x) \rangle_{\mathcal{H}_X} dx = \int f(x) g(x) dx.\end{aligned}$$

In particular the equation above implies that $\|f\|^2_{L^2(X)} = \|T^{1/2} f\|^2_{\mathcal{H}_X}$ for any $f \in \mathcal{H}_X$. Now, note that by assumption in this Corollary, we have chosen the kernel $k_X = k_{m+1}$ so $\mathcal{H}_X$ corresponds to the Sobolev space $\mathcal{H}_X = H^{m+1}(X)$ (see Proposition 4). Now by Assumption 1, μ has a density that we denote g_μ, that is differentiable up to order m and such that all the derivatives of order m are Lipschitz continuous. Since X is convex and bounded, then Lipschitz continuity of all the derivatives of order m implies that all the weak derivatives up to order $m+1$ are in $L^\infty(X)$. Since $L^\infty(X) \subset L^2(X)$, by the boundedness of X, we have that all the weak derivatives of g_μ up to order $m+1$ belong to $L^2(X)$, i.e. $\|g_\mu\|_{H^{m+1}(X)} < \infty$ so $g_\mu \in H^{m+1}(X) = \mathcal{H}_X$. So, by the reproducing property

$$\begin{aligned}w_\mu &= \int \phi_X(x) d\mu(x) = \int \phi_X(x) g_\mu(x) dx \\ &= \int \phi_X(x) \langle \phi_X(x), g_\mu \rangle_{\mathcal{H}_X} dx = \int (\phi_X(x) \otimes \phi_X(x)) g_\mu \, dx = T g_\mu.\end{aligned}$$

Now, the estimator $\hat{w}_\mu$ is defined as $\hat{w}_\mu = \int_X \phi_X(x) \hat{g}_\mu(x)$, where $\hat{g}_\mu \in \mathcal{H}_X$ is the *Kernel Least Squares estimator* (see Narcowich et al., 2005; Caponnetto and De Vito, 2007) of the density of μ that is g_μ. With the same reasoning as above, we see that $\hat{w}_\mu = T \hat{g}_\mu$. Now note that

$$\|w_\mu - \hat{w}_\mu\|_{\mathcal{H}_X} = \|T(g_\mu - \hat{g}_\mu)\|_{\mathcal{H}_X} \leq \|T^{1/2}\|_{op} \|T^{1/2}(g_\mu - \hat{g}_\mu)\|_{\mathcal{H}_X} = \|T^{1/2}\|_{op} \|g_\mu - \hat{g}_\mu\|_{L^2(X)}.$$

Now $\|T^{1/2}\|^2_{op} = \|T^{1/2}\|^2_{op} \leq \mathrm{Tr}(T) < \infty$ as we have seen above. Moreover $\|g_\mu - \hat{g}_\mu\|_{L^2(X)}$ is controlled by classical results on approximation theory, e.g. Proposition 3.2 of Narcowich et al. (2005) (applied with $\alpha = 0$ and $q = 2$). It is possible to apply such results as the set X is convex bounded and so it satisfies the required *uniform interior cone condition* (Wendland, 2004) (see Lemma 11, Page 20). The result guarantees that there exists two constants C, h_0 depending only on X, such that $\|g_\mu - \hat{g}_\mu\|_{L^2(X)} \leq C h^{m+1} \|g_\mu\|_{\mathcal{H}_X}$, where h is the *fill distance*

(see Eq. (10)) of the sampled n_μ points, with respect to X. Now by Lemma 12, Page 20 we have that $h \leq (C' n_\mu \log(C'' n_\mu/\delta))^{1/d}$ with probability at least $1-\delta$ for some constants C', C'' depending on d, X. Then, finally for some constant C''' we have

$$\begin{aligned}\|w_\mu - \hat{w}_\mu\|_{\mathcal{H}_X} &\leq \mathrm{Tr}(T)^{1/2}\|g_\mu - \hat{g}_\mu\|_{L^2(X)} \\ &\leq \mathrm{Tr}(T)^{1/2} C \|g_\mu\|_{\mathcal{H}_X} (C' n_\mu \log(C'' n_\mu/\delta))^{1/d} \leq C''' n_\mu^{-(m+1)/d} \log(n_\mu/\delta).\end{aligned}$$

(Sampling) In this case we have $\hat{w}_\mu = \frac{1}{n_\mu}\sum_{i\in[n_\mu]} \phi_X(x_i)$ where $x_1, \dots, x_n$ are independently and identically distributed according to μ. Then $\xi_i = \phi_X(x_i)$ for $i \in [n_\nu]$ are i.i.d. random vectors and $\hat{w}_\mu = \mathbb{E}\xi_1$. Since $\phi_X : X \to \mathcal{H}_X$ is uniformly bounded on X (see Proposition 4) denote by c that bound. We have that $\|\xi_i\|_{\mathcal{H}_X} \leq c$ almost surely and $\mathbb{E}\|\xi_i - \mathbb{E}\xi_i\|^2_{\mathcal{H}_X} \leq 2c$ for all $i \in [n_\mu]$. Then we can apply the Pinelis inequality (see Proposition 2 of Caponnetto and De Vito, 2007), to control $\hat{w}_\mu = \frac{1}{n_\mu}\sum_{i=1}^{n_\mu} \xi_i$, for which the following holds with probability $1-\delta$

$$\|w_\mu - \hat{w}_\mu\|_{\mathcal{H}_X} = \|\mathbb{E}\xi_1 - \frac{1}{n_\mu}\sum_{i=1}^{n_\mu} \xi_i\|_{\mathcal{H}_X} \leq 4c n_\mu^{-1/2} \log\frac{6n_\mu}{\delta}.$$

■

Proof of Theorem 1. This theorem is a direct consequence of Corollary 3. Let $\varepsilon > 0, \ell, n_\mu, n_\nu \in \mathbb{N}$. We denote by $f(x) \asymp g(x)$ the fact that there exists two constants $0 < C_1 \leq C_2$ not depending on x such that $C_1 g(x) \leq f \leq C_2 g(x)$ (in our particular case this means that the constants will not depend on $\varepsilon, \ell, n_\mu, n_\nu$). For the rest of the proof we will fix $\ell \asymp \varepsilon^{-2d/(m-d)}$. Indeed, this choice implies that $\ell^{-(m-d)/2d} = O(\varepsilon)$. We recall, moreover, from Eq. (7) (that is proven in Appendix F) that the computational time to achieve the estimator is $\tilde{O}(C + E\ell + \ell^{3.5})$ and the required memory is $O(\ell^2)$, where E, C are specified for each scenario in Section 6. Now we will quantify the complexity for the three scenarios.

(Exact integral) In this case, by Corollary 3

$$|\widehat{\mathrm{OT}} - \mathrm{OT}| = \tilde{O}(\ell^{-(m-d)/2d}) = \tilde{O}(\varepsilon).$$

Note that, since for this scenario $C = O(1), E = O(1)$, the complexity in time is

$$\tilde{O}(C + E\ell + \ell^{3.5}) = \tilde{O}(\ell^{3.5}) = \tilde{O}(\varepsilon^{-7d/(m-d)}).$$

In space, analogously $O(\ell^2) = O(\varepsilon^{-4d/(m-d)})$.

(Evaluation) In this case we choose $n_\nu \asymp n_\mu \asymp O(\varepsilon^{-d/(m+1)})$. Indeed, with this choice $n_\mu^{-(m+1)/d} = O(\varepsilon)$ (analogously for n_ν). Then, by Corollary 3

$$|\widehat{\mathrm{OT}} - \mathrm{OT}| = \tilde{O}(n_\mu^{-(m+1)/d} + n_\nu^{-(m+1)/d} + \ell^{-(m-d)/2d}) = \tilde{O}(\varepsilon).$$

From a computational viewpoint $C = O(n_\mu^3 + n_\nu^3)$ and $E = O(n_\mu + n_\nu)$. Then the time complexity is

$$\begin{aligned}\tilde{O}(C + E\ell + \ell^{3.5}) &= \tilde{O}(n_\mu^3 + n_\nu^3 + (n_\mu + n_\nu)\ell + \ell^{3.5}) \\ &= \tilde{O}(\varepsilon^{-3d/(m+1)} + \varepsilon^{-d/(m+1)-2d/(m-d)} + \varepsilon^{-7d/(m-d)}) = O(\varepsilon^{-7d/(m-d)}),\end{aligned}$$

wile the space complexity is $O(\ell^2) = O(\varepsilon^{-4d/(m-d)})$.

(Sampling) In this case we choose $n_\nu \asymp n_\mu \asymp O(\varepsilon^{-2})$. Indeed with this choice $n_\mu^{-1/2} = O(\varepsilon)$ (analogously for n_ν). Then, by Corollary 3

$$|\widehat{\mathrm{OT}} - \mathrm{OT}| = \tilde{O}(n_\mu^{-1/2} + n_\nu^{-1/2} + \ell^{-(m-d)/2d}) = \tilde{O}(\varepsilon).$$

From a computational viewpoint $C = O(n_\mu^2 + n_\nu^2)$ and $E = O(n_\mu + n_\nu)$. Then the time complexity is

$$\begin{aligned}\tilde{O}(C + E\ell + \ell^{3.5}) &= \tilde{O}(n_\mu^2 + n_\nu^2 + (n_\mu + n_\nu)\ell + \ell^{3.5}) \\ &= \tilde{O}(\varepsilon^{-4} + \varepsilon^{-2-2d/(m-d)} + \varepsilon^{-7d/(m-d)}) = \tilde{O}(\varepsilon^{-\max(4,7d/(m-d))}).\end{aligned}$$

Also in this case the space complexity is $O(\ell^2) = O(\varepsilon^{-4d/(m-d)})$. ■

Appendix F. Dual Algorithm and Computational Bounds

In this section, we describe a dual algorithmic procedure to compute Eq. (4), and bound the computational complexity and memory footprint it requires to achieve a given precision. We start by deriving Eq. (5), the dual formulation of Eq. (3), and express Eq. (4) as a function of the dual solution $\hat{\gamma}$.

Theorem 15 *The dual problem of Eq. (3) is Eq. (5). Further, the estimator $\widehat{OT}$ can be expressed as a function of the solution $\hat{\gamma}$ of Eq. (5):*

$$\widehat{OT} = \frac{q^2}{2\lambda_2} - \frac{1}{2\lambda_2}\sum_{j=1}^{\ell} \hat{\gamma}_j(\hat{w}_\mu(\tilde{x}_j) + \hat{w}_\nu(\tilde{y}_j)).$$

Proof of Theorem 15.

Finite-dimensional representation. Let us start by formulating (3) using a finite-dimensional representation for $A \in \mathbb{S}_+(\mathcal{H}_{XY})$. Following Marteau-Ferey et al. (2020), observe that problem (3) needs only to be solved w.r.t. A in the finite-dimensional Hilbert space spanned by $\{\phi(\tilde{x}_j, \tilde{y}_j) : j \in [\ell]\}$. This result is formally proven for our setting in Appendix D. Therefore, it is sufficient to consider positive operators of the form $A = \sum_{ij=1}^{\ell} C_{ij}\phi(\tilde{x}_i, \tilde{y}_i) \otimes \phi(\tilde{x}_j, \tilde{y}_j)$, with $\mathbf{C} \in \mathbb{S}_+(\mathbb{R}^\ell)$. With this parameterization, we have $\mathrm{Tr}A = \mathrm{Tr}(\mathbf{CK})$ and $\langle \phi(\tilde{x}_i, \tilde{y}_i), \mathrm{A}\phi(\tilde{x}_j, \tilde{y}_j)\rangle = (\mathbf{KCK})_{ij}$, where $\mathbf{K} = [k_{XY}((\tilde{x}_i, \tilde{y}_i), (\tilde{x}_j, \tilde{y}_j)]_{ij=1}^{\ell}$. As Rudi et al. (2020), we can thus consider the Cholesky decomposition $\mathbf{K} = \mathbf{R}\mathbf{R}^\top$ (or alternatively take the square root $\mathbf{R} = \mathbf{K}^{1/2}$), and represent problem (3) in terms of the columns $\{\Phi_j : j \in [\ell]\}$ of $\mathbf{R}$ and solve directly for $\mathbf{B} = \mathbf{RCR}^\top$:

$$\begin{aligned} \max_{\substack{u\in\mathcal{H}_X, v\in\mathcal{H}_Y \\ \mathbf{B}\in\mathbb{S}_+(\mathbb{R}^\ell)}} \quad & \langle u, \hat{w}_\mu\rangle + \langle v, \hat{w}_\nu\rangle - \lambda_1 \mathrm{Tr}\mathbf{B} - \lambda_2(\|u\|_{\mathcal{H}_X}^2 + \|v\|_{\mathcal{H}_Y}^2) \\ \text{s.t. } \forall j \in [\ell], \quad & c(\tilde{x}_j, \tilde{y}_j) - u(\tilde{x}_j) - v(\tilde{y}_j) = \Phi_j^\top \mathbf{B}\Phi_j. \end{aligned} \tag{16}$$

Deriving the dual. Next, let us observe that problem (3) is convex, and admits a feasible point by Corollary 7 and a maximizer by Lemma 14. The same applies to (16). Therefore, strong duality holds. The Lagragian of Eq. (16) is

$$\begin{aligned}\mathcal{L}(u, v, \mathbf{B}, \gamma) = \langle u,\, \hat{w}_\mu\rangle + \langle v,\, \hat{w}_\nu\rangle - \lambda_1 \mathrm{Tr}\mathbf{B} - \lambda_2\|u\|^2_{\mathcal{H}_X} - \lambda_2\|v\|^2_{\mathcal{H}_Y} \\ + \sum_{i=1}^{\ell} \gamma_i (c(\tilde{x}_i, \tilde{y}_i) - \langle u,\, \phi_X(\tilde{x}_i)\rangle - \langle v,\, \phi_Y(\tilde{y}_i)\rangle - \Phi_i^\top \mathbf{B}\Phi_i).\end{aligned} \tag{17}$$

At the optimum, we have $\nabla_u \mathcal{L}(u, v, \mathbf{B}, \gamma) = 0$ and $\nabla_v \mathcal{L}(u, v, \mathbf{B}, \gamma) = 0$, which yields

$$\begin{aligned} u &= \frac{1}{2\lambda_2}(\hat{w}_\mu - \sum_{i=1}^{\ell} \gamma_i \phi_X(\tilde{x}_i)) \\ v &= \frac{1}{2\lambda_2}(\hat{w}_\nu - \sum_{i=1}^{\ell} \gamma_i \phi_Y(\tilde{y}_i)). \end{aligned} \tag{18}$$

Let us now derive the optimality condition on $\mathbf{B}$: we have

$$\begin{aligned}\sup_{\mathbf{B}\in\mathbb{S}_+(\mathbb{R}^\ell)} -\sum_{i=1}^{\ell} \gamma_i \Phi_i^\top \mathbf{B}\Phi_i - \lambda_1 \mathrm{Tr}\mathbf{B} &= \sup_{\mathbf{B}\in\mathbb{S}_+(\mathbb{R}^\ell)} \langle \mathbf{B},\, -(\sum_{i=1}^{\ell} \gamma_i \Phi_i \Phi_i^\top + \lambda_1 \mathbf{I}_\ell)\rangle \\ &= \begin{cases} 0 & \text{if } \sum_{i=1}^{\ell} \gamma_i \Phi_i^\top \Phi_i + \lambda_1 \mathbf{I}_\ell \succeq 0 \\ -\infty & \text{otherwise.} \end{cases}\end{aligned} \tag{19}$$

Plugging Eq. (18) and Eq. (19) in the Eq. (17), we get (5). Finally, using Eq. (18), we have

$$\widehat{\mathrm{OT}} = \langle \hat{u},\, \hat{w}_\mu\rangle_{\mathcal{H}_X} + \langle \hat{u},\, \hat{w}_\nu\rangle_{\mathcal{H}_X} = \frac{q^2}{2\lambda_2} - \frac{1}{2\lambda_2}\sum_{j=1}^{\ell} \hat{\gamma}_j (\hat{w}_\mu(\tilde{x}_j) + \hat{w}_\nu(\tilde{y}_j)).$$

■

To solve (5), it is possible to use standard software packages (Boyd and Vandenberghe, 2004). Alternatively, it can be made more scalable by adding a self-concordant barrier term to (5) and using interior point methods with Newton steps. For a given barrier penalization $\delta > 0$, we thus aim to solve

$$\begin{aligned}\min_{\gamma\in\mathbb{R}^\ell} \frac{1}{4\lambda_2}\gamma^\top \mathbf{Q}\gamma - \frac{1}{2\lambda_2}\sum_{j=1}^{\ell} \gamma_j z_j + \frac{q^2}{4\lambda_2} - \frac{\delta}{\ell}\log\det(\sum_{i=1}^{\ell} \gamma_i \Phi_i \Phi_i^\top + \lambda_1 \mathbf{I}_\ell)) \\ \text{such that} \quad \sum_{j=1}^{\ell} \gamma_j \Phi_j \Phi_j^\top + \lambda_1 \mathbf{I}_\ell \succeq 0.\end{aligned} \tag{20}$$

Starting from an initial value δ_0, the barrier method (Nemirovski, 2004) consists in iteratively solving Eq. (20) (using Newton iterations), and progressively decreasing δ. In Theorems 16 and 17, we precisely analyze the complexity of the barrier method applied to Eq. (5), and bound the number of operations required to obtain an estimator of OT with a desired accuracy.

Theorem 16 *Using a dual interior point method, a solution of problem (3) with value precision $O(\tau)$ can be obtained in $O(C + E\ell + \ell^{3.5}\log(\frac{\ell}{\tau}))$ operations and $O(\ell^2)$ memory, where E is the cost of querying $\hat{w}_\mu$ and $\hat{w}_\nu$, and C is the cost of computing q^2.*

Proof of Theorem 16. Removing terms that are constant in γ, problem (20) is equivalent to minimizing the dual functional

$$J(\gamma) \stackrel{\text{def}}{=} \frac{1}{4\lambda_2}\gamma^\top \mathbf{Q}\gamma - \frac{1}{2\lambda_2}\sum_{j=1}^{\ell}\gamma_j z_j - \frac{\delta}{\ell}\log\det(\sum_{i=1}^{\ell}\gamma_i\Phi_i\Phi_i^\top + \lambda_1\mathbf{I}_\ell).$$

Its gradient is

$$J'(\gamma)_i = \frac{1}{2\lambda_2}(\mathbf{Q}\gamma)_i - \frac{1}{2\lambda_2}z_i - \frac{\delta}{\ell}\Phi_i^\top(\Phi\,\mathrm{diag}(\gamma)\Phi^\top + \lambda_1\mathbf{I}_\ell)^{-1}\Phi_i, \quad i \in [\ell],$$

and its Hessian

$$J''(\gamma)_{ij} = \frac{1}{2\lambda_2}\mathbf{Q}_{ij} - \frac{\delta}{\ell}[\Phi_i^\top(\Phi\,\mathrm{diag}(\gamma)\Phi^\top + \lambda_1\mathbf{I}_\ell)^{-1}\Phi_j]^2, \quad i,j \in [\ell].$$

From there, we may minimize $J(\gamma)$ using damped Newton iterations

$$\gamma' = \gamma - \frac{[J''(\gamma)]^{-1}J'(\gamma)}{1 + \sqrt{\frac{\ell}{\delta}}\lambda(\gamma)},$$

where $\lambda^2(\gamma) = J'(\gamma)^\top[J''(\gamma)]^{-1}J'(\gamma)$ is the Newton decrement, or using backtracking line-search Newton iterations (Boyd and Vandenberghe, 2004).

Number of iterations. $J''(\gamma)$ can be computed and inverted in $O(\ell^3)$ operations, and assuming $\hat{w}_\mu(\tilde{x}_i), i \in [l]$ and $\hat{w}_\nu(\tilde{y}_i), i \in [l]$ are precomputed, $J'(\gamma)$ can be computed in $O(\ell^3)$ operations, hence the complexity per iteration is $O(\ell^3)$.

Let $F(\gamma) \stackrel{\text{def}}{=} \frac{1}{4\lambda_2}\gamma^\top\mathbf{Q}\gamma - \frac{1}{2\lambda_2}\sum_{j=1}^{\ell}\gamma_j z_j + \frac{q^2}{4\lambda_2}$ be the objective function of (5). Since $H(\gamma) \stackrel{\text{def}}{=} -\log\det(\sum_{i=1}^{\ell}\gamma_i\Phi_i\Phi_i^\top + \lambda_1\mathbf{I}_\ell))$ is a self-concordant barrier function of concordance parameter ℓ, standard results on barrier methods imply that δ controls the deviation (in value) to the optimum of F (Nemirovski, 2004). Moreover, a solution $\tilde{\gamma}$ to (5) of precision $\tau > 0$, i.e. satisfying $F(\tilde{\gamma}) - F(\hat{\gamma}) \leq \tau$ where $\hat{\gamma}$ is the optimum of (5), can computed in $O(\sqrt{\ell}\log\frac{\ell}{\tau})$ Newton iterations using an interior point method by progressively decreasing δ using a suitable scheme until $\delta \leq \tau$: see (Nemirovski, 2004).

Hence, taking into account the $O(E\ell)$ cost of computing $z_j, j = 1, ..., \ell$ and the $O(C)$ cost required to compute q^2, to achieve a precision $\tau > 0$ a total of $O(C + E\ell + \ell^{3.5}\log(\frac{\ell}{\tau}))$ operations and $O(\ell^2)$ memory are required.

■

In Theorem 16, we may use any of the kernel mean estimators presented in Section 6 and apply the corresponding computational costs C and E. However, the given bounds only apply to the precision in value, i.e., on $V_{\lambda_1,\lambda_2}(\hat{u},\hat{v},\hat{A}) - V_{\lambda_1,\lambda_2}(u,v,A)$, and not on the solutions (u,v,A) themselves. In Theorem 17, we derive bounds on the algorithmic approximation of the estimator in (6) obtained by minimizing Eq. (20) to a precision $\tau > 0$, as a function of its computational complexity.

Theorem 17 *Under the same notation and assumptions as Theorem 9, let $\tilde{\gamma}$ be obtained by running a barrier method on (5) to precision $\tau > 0$, i.e., by iteratively solving (20) and progressively decreasing δ until $\delta \leq \tau$, as described by Nemirovski (2004). Define $\widetilde{OT}$ as*

$$\widetilde{OT} = \frac{q^2}{2\lambda_2} - \frac{1}{2\lambda_2} \sum_{j=1}^{\ell} \tilde{\gamma}_j (\hat{w}_\mu(\tilde{x}_j) + \hat{w}_\nu(\tilde{y}_j)).$$

Then $\widetilde{OT}$ satisfies the following bound

$$|\widetilde{OT} - OT| \;\leq\; 6\lambda_2(\|u_\star\|^2_{\mathcal{H}_X} + \|v_\star\|^2_{\mathcal{H}_Y}) + 6\tfrac{\beta^2}{\lambda_2} + 6\lambda_1 \mathrm{Tr} A_\star + 6\lambda_2 \tau.$$

Further, the considered algorithm to compute $\widetilde{OT}$ has a cost of $O(C + E\ell + \ell^{3.5}\log(\frac{\ell}{\tau}))$ in time and $O(\ell^2)$ in memory.

Proof of Theorem 17. Let $\tilde{\gamma}$ be obtained by running a barrier method on Eq. (5) to precision τ, i.e. by solving (20) and decreasing δ until $\delta \leq \tau$. Then, from properties of barrier methods (Nesterov and Nemirovskii, 1994) we can associate to $\tilde{\gamma}$ the following primal feasible points, obtained by nullifying the Lagragian of (20) at $\tilde{\gamma}$:

$$\begin{aligned}
\tilde{u} &= \frac{1}{2\lambda_2}(\hat{w}_\mu - \sum_{i=1}^{\ell} \tilde{\gamma}_i \phi_X(\tilde{x}_i)) \\
\tilde{v} &= \frac{1}{2\lambda_2}(\hat{w}_\nu - \sum_{i=1}^{\ell} \tilde{\gamma}_i \phi_Y(\tilde{y}_i)) \\
\tilde{A} &= \sum_{ij=1}^{\ell} B_{ij} \phi(\tilde{x}_i, \tilde{y}_i) \otimes \phi(\tilde{x}_j, \tilde{y}_j),
\end{aligned}$$

with $\mathbf{B} = \frac{\delta}{\ell}(\sum_{i=1}^{\ell} \tilde{\gamma}_i \Phi_i \Phi_i^\top + \lambda_1 \mathbf{I}_\ell)^{-1}$. In particular, from properties of interior point methods (Nemirovski, 2004; Nesterov and Nemirovskii, 1994), we have

(i) $(\tilde{u}, \tilde{v}, \tilde{A})$ is a feasible point of (3),

(ii) The duality gap between the objective $\hat{V}_{\lambda_1,\lambda_2}$ of (3) evaluated at $(\tilde{u}, \tilde{v}, \tilde{A})$ and the objective F of (5) at $\tilde{\gamma}$ is equal to δ, i.e. $F(\tilde{\gamma}) - \hat{V}_{\lambda_1,\lambda_2}(\tilde{u}, \tilde{v}, \tilde{A}) = \delta$.

Further, note that we have

$$\begin{aligned}
\widetilde{\mathrm{OT}} &\stackrel{\mathrm{def}}{=} \frac{q^2}{2\lambda_2} - \frac{1}{2\lambda_2} \sum_{j=1}^{\ell} \tilde{\gamma}_j (\hat{w}_\mu(\tilde{x}_j) + \hat{w}_\nu(\tilde{y}_j)) \\
&= \langle \tilde{u},\, \hat{w}_\mu \rangle_{\mathcal{H}_X} + \langle \tilde{v},\, \hat{w}_\nu \rangle_{\mathcal{H}_Y}.
\end{aligned}$$

Let us now bound $\widetilde{\mathrm{OT}} - \mathrm{OT}$. We will follow similar arguments than in the proof of Theorem 9, with an additional τ precision term. As in the proof of Theorem 9, let $(\tilde{u}_\varepsilon, \tilde{v}_\varepsilon) = (\tilde{u} - \varepsilon/2, \tilde{v} - \varepsilon)$.

Since $(\tilde{u}, \tilde{v}, \tilde{A})$ satisfy the constraints of (3), from the same arguments $(\tilde{u}_\varepsilon, \tilde{v}_\varepsilon)$ defines a feasible point for (1). Hence, we have the equivalent of Eq. (12):

$$\mathrm{OT} = V(u_\star, v_\star) \geq V(\tilde{u}_\varepsilon, \tilde{v}_\varepsilon) = V(\tilde{u}, \tilde{v}) - \varepsilon \geq \widetilde{\mathrm{OT}} - \Delta(\tilde{u}, \tilde{v}) - \varepsilon. \tag{21}$$

Next, let $\hat{\gamma}$ be the optimum of (5), and F the objective function of (5). By strong duality (see Theorem 15), we have $F(\hat{\gamma}) = \hat{V}_{\lambda_1,\lambda_2}(\hat{u}, \hat{v}, \hat{A})$. Further, by optimality of $\hat{\gamma}$, we have $F(\hat{\gamma}) \leq F(\tilde{\gamma})$. Hence, using (ii) and $\delta \leq \tau$, we have

$$\begin{aligned}
\hat{V}_{\lambda_1,\lambda_2}(\tilde{u}, \tilde{v}, \tilde{A}) &= F(\tilde{\gamma}) - \delta \\
&\geq F(\hat{\gamma}) - \delta \\
&= \hat{V}_{\lambda_1,\lambda_2}(\hat{u}, \hat{v}, \hat{A}) - \delta \\
&\geq \hat{V}_{\lambda_1,\lambda_2}(u_\star, v_\star, A_\star) - \tau \\
&= V(u_\star, v_\star) - \left[V(u_\star, v_\star) - \hat{V}_{\lambda_1,\lambda_2}(u_\star, v_\star, A_\star)\right] - \tau \\
&\geq \mathrm{OT} - \Delta(u_\star, v_\star) - \lambda_1 \mathrm{Tr}(A_\star) + \lambda_2 R^2(u_\star, v_\star) - \tau.
\end{aligned}$$

From there, the rest of the proof of Theorem 9 follows identically: developing $\hat{V}_{\lambda_1,\lambda_2}(\tilde{u}, \tilde{v}, \tilde{A})$ and combining with Eq. (21), we have

$$\lambda_1 \mathrm{Tr}(\tilde{A} - A_\star) + \lambda_2 (R^2(\tilde{u}, \tilde{v}) - R^2(u_\star, v_\star)) - \Delta(u_\star, v_\star) - \tau \leq \widetilde{\mathrm{OT}} - \mathrm{OT} \leq \Delta(\tilde{u}, \tilde{v}) + \varepsilon.$$

Hence

$$\lambda_1 \mathrm{Tr}(\tilde{A}) + \lambda_2 R^2(\tilde{u}, \tilde{v}) \leq \lambda_1 \mathrm{Tr}(A_\star) + \lambda_2 R^2(u_\star, v_\star) + \Delta(u_\star, v_\star) + \Delta(\tilde{u}, \tilde{v}) + \varepsilon + \tau.$$

Therefore, replacing S from the proof of (9) with $S' := S + \tau$, and $\hat{u}, \hat{v}, \hat{A}$ with $\tilde{u}, \tilde{v}, \tilde{A}$, the rest of the proof follows and eventually yields

$$|\widetilde{\mathrm{OT}} - \mathrm{OT}| \leq 6\lambda_2 (R^2(u_\star, v_\star) + \frac{\beta^2}{\lambda_2^2} + \frac{\lambda_1}{\lambda_2} \mathrm{Tr} A_\star + \tau).$$

To conclude, note that as a consequence of Theorem 16, $\widetilde{\mathrm{OT}}$ can be computed in $O(C + E\ell + \ell^{3.5} \log(\frac{\ell}{\tau}))$ operations and $O(\ell^2)$ memory.

■

Recovering unregularized optimal transport. In Eq. (3) and in the case of empirical estimators $\hat{w}_\mu$ and $\hat{w}_\nu$ (see Section 6), we considered the case where the sample pairs $(\tilde{x}_i, \tilde{y}_i), i \in [l]$ covering $X \times Y$ are distinct from the samples $x_i \sim \mu, i \in [n_\mu]$ and $y_j \sim \nu, j \in [n_\nu]$. However, covering $X \times Y$ with the $n_\mu n_\nu$ pairs given by the μ and ν samples

$(x_i, y_j), i \in [n_\mu], j \in [n_\nu]$, we may rewrite (5) as a regularized optimal transport problem:

$$\begin{aligned} \min_{\Gamma \in \mathbb{R}^{n_\mu \times n_\nu}} \quad & \sum_{ij} \Gamma_{ij} c(x_i, y_j) + \frac{1}{2\lambda_2} r^T \mathbf{K}_X r + \frac{1}{2\lambda_2} c^T \mathbf{K}_Y c \\ \text{s.t.} \quad & \sum_{\substack{i=1,\dots,n_\mu \\ j=1,\dots,n_\nu}} \Gamma_{ij} \Phi_{ij} \Phi_{ij}^\top + \lambda_1 \mathbf{I}_{n_\mu n_\nu} \succcurlyeq 0, \\ & r_i = \frac{1}{n_\mu} - \sum_{j=1}^{n_\nu} \Gamma_{ij}, \quad i \in [n_\mu], \\ & c_j = \frac{1}{n_\nu} - \sum_{i=1}^{n_\mu} \Gamma_{ij}, \quad j \in [n_\nu], \end{aligned} \tag{22}$$

were we reindexed $\Phi_p, p \in [n_\mu n_\nu]$ as $\Phi_{ij}, i \in [n_\mu], j \in [n_\nu]$, and where $(\mathbf{K}_X)_{ij} = k_X(x_i, x_j), i, j \in [n_\mu]$, $(\mathbf{K}_Y)_{ij} = k_Y(y_i, y_j), i, j \in [n_\nu]$. Hence, (22) can be interpreted as a regularized optimal transport problem, where $\Gamma \in \mathbb{R}^{n_\mu \times n_\nu}$ plays the role of the transportation plan, and the marginal violations are penalized with maximum mean discrepancy (MMD) terms (Gretton et al., 2012). When λ_1 goes to 0, the SDP constraint becomes a $\Gamma \in \mathbb{R}_+^{n_\mu \times n_\nu}$ positivity constraint, and when λ_2 goes to 0, the MMD penalization terms enforce the hard constraints $\Gamma \mathbb{1}_{n_\nu} = \frac{\mathbb{1}_{n_\mu}}{n_\mu}$ and $\Gamma^\top \mathbb{1}_{n_\mu} = \frac{\mathbb{1}_{n_\nu}}{n_\nu}$. In particular, when $(\lambda_1, \lambda_2) \to (0, 0)$, we recover the unregularized OT problem between the empirical measures $\hat{\mu} = \frac{1}{n_\mu} \sum_{i=1}^{n_\mu} \delta_{x_i}$ and $\hat{\nu} = \frac{1}{n_\nu} \sum_{j=1}^{n_\nu} \delta_{y_j}$, which can be formally verified by deriving the dual of (3) with λ_1 and/or λ_2 equal to 0.

Appendix G. Numerical Experiments

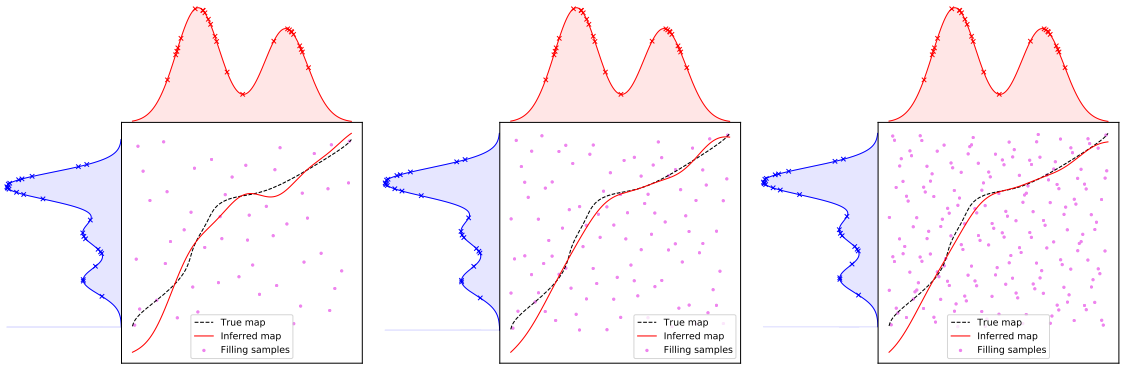

Figure 2: Effect of increasing the number of filling samples ℓ on the transportation map. *(left)*: $\ell = 50, n_\mu = n_\nu = 25$, *(middle)*: $\ell = 100, n = 25$, *(right)*: $\ell = 200, n_\mu = n_\nu = 25$.

1D transportation maps and dual constraint functions. We illustrate the algorithm described in Theorem 16 in a 1D setting in Figures 2 to 5, by representing the inferred transportation map $\hat{T}$ obtained from $\hat{u}$, defined as $\hat{T} = x - \nabla_x \hat{u}(x)$, and the corresponding constraint function $\hat{h}(x, y) = \frac{1}{2}\|x - y\|^2 - \hat{u}(x) - \hat{v}(y)$. We sample $x_1, ..., x_{n_\mu}$ *i.i.d.* from μ and $y_1, ..., y_{n_\nu}$ *i.i.d.* from ν, and use quasi-random samples $(\tilde{x}_1, \tilde{y}_1), ...(\tilde{x}_\ell, \tilde{y}_\ell)$ from a 2D

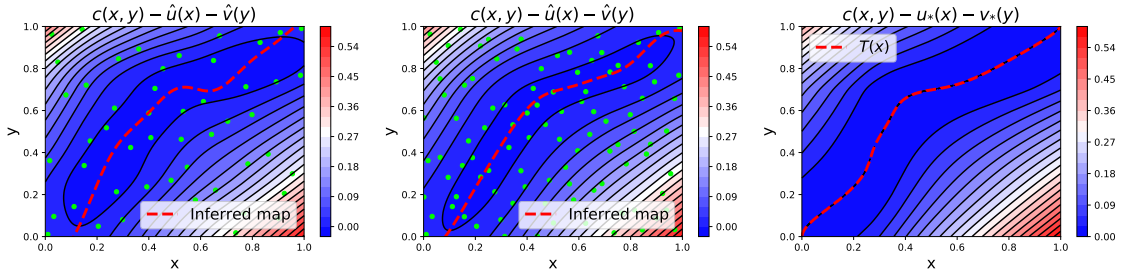

Figure 3: Effect of increasing the number of filling samples ℓ on the constraint model. *(left)*: $\ell = 50, n_\mu = n_\nu = 25$, *(middle)*: $\ell = 100, n_\mu = n_\nu = 25$, *(right)*: true function.

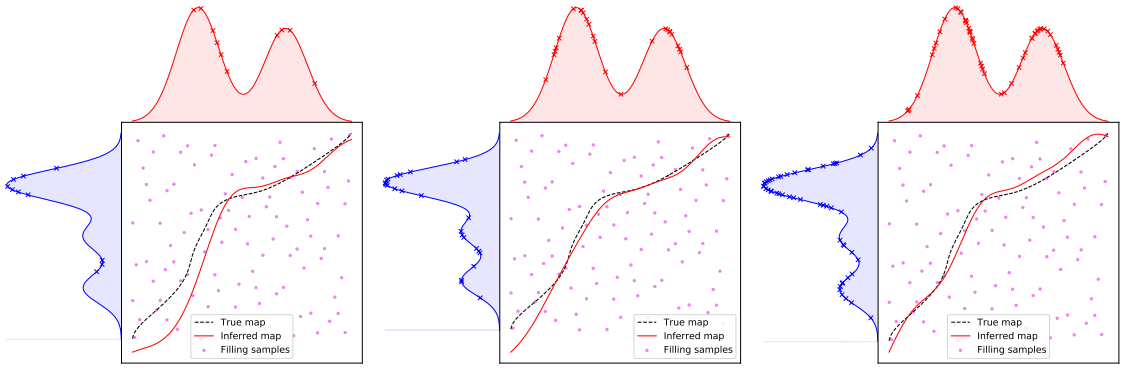

Figure 4: Effect of increasing the number of μ and ν samples on the transportation map. *(left)*: $\ell = 100, n_\mu = n_\nu = 10$, *(middle)*: $\ell = 100, n_\mu = n_\nu = 25$, *(right)*: $\ell = 100, n_\mu = n_\nu = 50$.

Sobol sequence (Sobol, 1967), and illustrate the effect of varying n, the number of μ and ν samples, and ℓ, the number of space-filling samples. For k_X, k_Y and k_{XY}, we use Gaussian kernels $k(x,y) = \exp(-\frac{\|x-y\|^2}{2\sigma^2})$ of fixed bandwidth $\sigma^2 = 0.1$, and scale the regularization parameters as $\lambda_1 = \frac{1}{\ell}$ and $\lambda_2 = \frac{1}{\sqrt{n}}$.

Convergence of $\widehat{\text{OT}}$ to OT. In Figure 6, we compare $\widehat{\text{OT}}$ to the sampled optimal transport estimator on two 4D truncated Gaussian distributions μ and ν s.t. the optimal transportation map from one to another is linear. We progressively increase the number of μ and ν samples, averaging on 20 random draws for each number of samples. The number of filling sample pairs $(\tilde{x}_i, \tilde{y}_i)$ is $\ell = 100 + n$, where $n = n_\mu = n_\nu$ is the number samples from μ and ν. We select the best estimator $\widehat{\text{OT}}$ using a grid search on (λ_1, λ_2). As such, this simulation does not provide a method for selecting those parameters, but rather illustrates that a good pair of parameters exists.

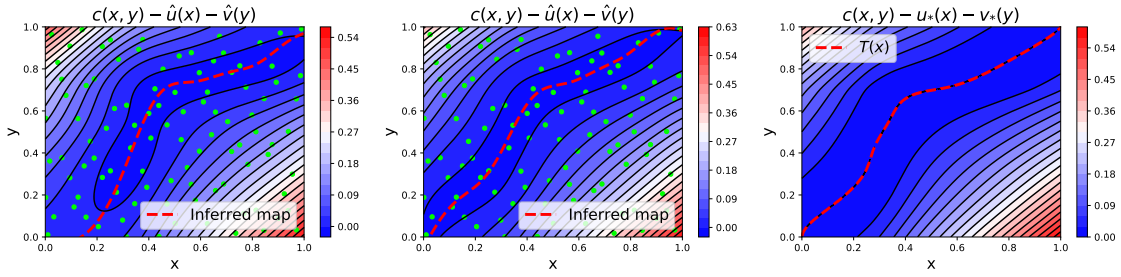

Figure 5: Effect of increasing the number of μ and ν samples on the constraint model. *(left)*: $\ell = 100, n_\mu = n_\nu = 10$, *(middle)*: $\ell = 100, n_\mu = n_\nu = 50$, *(right)*: true function.

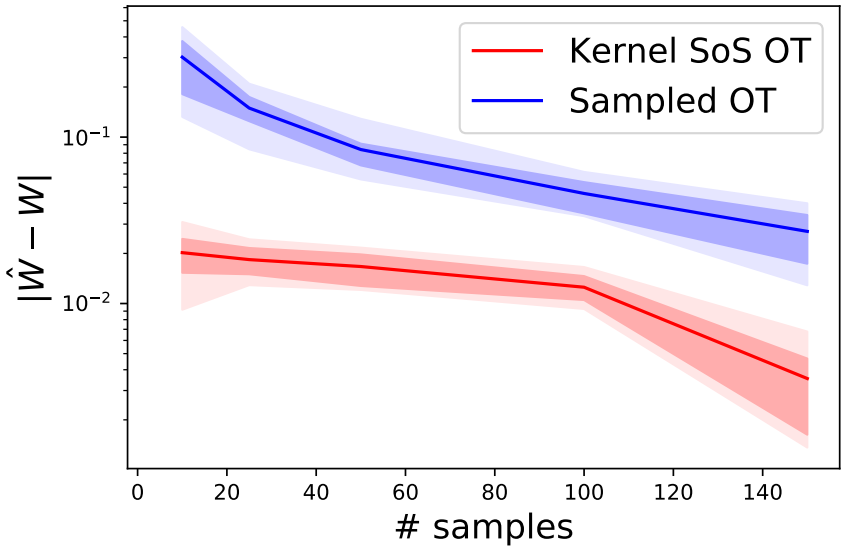

Figure 6: Convergence on 4D truncated Gaussian data with increasing number of samples (*left*). Full lines correspond to the average mean absolute error (MAE), shaded areas to $25\% - 75\%$ and $10\% - 90\%$ MAE quantiles. The parameters λ_1, λ_2 are selected via a grid search.