# Supplementary material: Statistical Optimal Transport posed as Learning Kernel Mean Embedding

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

The objective in statistical Optimal Transport (OT) is to consistently estimate the optimal transport plan/map solely using samples from the given source and target marginal distributions. This work takes the novel approach of posing statistical OT as that of learning the transport plan's kernel mean embedding from sample based estimates of marginal embeddings. A key result is that, under mild conditions, the sample complexity of the resulting estimator for the optimal transport plan as well as that for the Barycentric-projection based optimal transport map are dimension-free. Moreover, the implicit smoothing in the kernel embeddings not only improves the quality of finite sample estimation but also enables out-of-sample estimation. Also, complementary to existing $\phi$-divergence (entropy) based regularization techniques, our estimator employs a maximum mean discrepancy (MMD) based regularization to avoid over-fitting the samples. We present an appropriate representer theorem that leads to a kernelized convex formulation, which can then be potentially used to perform OT even in non-standard domains. Empirical results illustrate the efficacy of the proposed approach.

## 1 Introduction

Optimal Transport is proving to be an increasingly successful tool in solving diverse machine learning problems. Recent research shows that variants of Optimal Transport (OT) achieve state-of-the-art performance in various machine learning (ML) applications such as domain adaptation [9], NLP [2, 40, 41], robust learning [6], etc. It is also shown that OT based (Wasserstein) metrics serve as good loss functions in both supervised [15] and unsupervised [19] learning. [30] is a comprehensive monologue on the subject with focus on recent developments related to machine learning.

Given two marginal distributions over source and target domains, and a cost function between elements of the domains, the classical OT problem (Kantorovich's formulation) is that of finding the joint distribution whose marginals are equal to the given marginals, and which minimizes the expected cost with respect to this joint distribution [22]. This joint distribution is known as the (optimal) transport plan or the optimal coupling. A related object of interest for ML applications is the so-called Barycentric-projection based transport map corresponding to a transport plan (e.g., refer (11) in [34]). Though OT techniques already improve state-of-the-art in many ML applications, there are two main bottlenecks that seem to limit OT's success in ML settings:

- while continuous distributions are ubiquitous, algorithms for finding the transport plan/map over continuous domains are very scarce [18]. The situation is worse in case of non-standard domains, which are not uncommon in ML.

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

- the marginal distributions are never available, and merely samples from them are given. The variant of OT where the transport plan/map needs to be estimated merely using samples from the marginals is known as the statistical OT problem. In case of statistical OT over continuous domains, [18, 17, 14, 5] note that estimators that are free from the curse of dimensionality are not well-studied.

The concluding report from a recent workshop on OT (refer section 2 in [5]) summarizes that one of the major open problems in this area is to design estimators in context of continuous statistical OT whose sample complexity is not a strong function of the dimension (ideally dimension-free).

Our work focuses on this challenging and important problem of statistical OT over continuous domains, and seeks consistent estimators whose sample complexity is dimension-free. To this end, we take the novel approach of equivalently re-formulating the statistical OT problem solely in terms of the relevant kernel mean embeddings [25]. More specifically, our formulation finds the (characterizing) kernel mean embedding of a joint distribution with least expected cost, and whose marginal embeddings are close to the given-sample based estimates of the marginal embeddings. There are several advantages of this new approach:

1. because the samples based estimates of the kernel mean embeddings of the marginals are known to have sample complexities that are dimension-free, it is expected that the sample complexity remains dimension-free even for the proposed estimator of the transport plan embedding.
2. kernel embeddings provide implicit smoothness, as controlled by the kernel. Appropriate smoothness not only improves the quality of estimation, but also enable out-of-sample estimation.
3. while existing estimators employ $\phi$-divergence (or entropy) based regularization, our formulation employs Maximum Mean Discrepency (MMD) based regularization to avoid overfitting the samples. This is facilitated as MMD is the natural notion of distance in the kernel mean embedding space. As discussed in [38], MMD and $\phi$-divergence based regularization exhibit complementary properties and hence both are interesting to study.

To the best of our knowledge, existing works have not employed kernel mean embeddings explicitly in the context of OT.

A key result from this work is that, under mild conditions, the proposed estimator for the optimal transport plan as well as the (Barycentric-projection based) optimal transport map is statistically consistent with a sample complexity that remains dimension-free. Another key contribution is an appropriate representer theorem that guarantees finite characterization for the transport plan embedding, which leads to a fully kernelized and convex formulation. Thus the same formulation can potentially be used for obtaining estimators in all variants of OT: continuous, semi-discrete, and discrete, merely by switching the kernel between the Kronecker delta and the Gaussian kernels. More importantly, the same can be used to solve OT problems in non-standard domains using appropriate universal kernels [7]. Finally, we present an alternating direction method of multipliers (ADMM) based algorithm for efficiently solving the proposed formulation. Empirical results on synthetic and real-world datasets illustrate the efficacy of the proposed approach.

## 2 Background on Optimal Transport

Let $\mathcal{X}, \mathcal{Y}$ be any two sets that form locally compact Hausdorff topological spaces. We denote the set of all Radon probability measures over $\mathcal{X}$ by $\mathcal{M}^1(\mathcal{X})$; whereas we denote the set of strictly positive measures by $\mathcal{M}^1_+(\mathcal{X})$ . Let $c : \mathcal{X} \times \mathcal{Y}$ denote a function that evaluates the cost between elements in $\mathcal{X}, \mathcal{Y}$ and let $p_s \in \mathcal{M}^1_+(\mathcal{X}), p_t \in \mathcal{M}^1_+(\mathcal{Y})$. Then, the Kantorovich's OT formulation [22] is:

$$\begin{array}{ll} \min_{\pi \in \mathcal{M}^1(\mathcal{X},\mathcal{Y})} & \int c(x,y)\, \mathrm{d}\pi(x,y), \\ \text{s.t.} & \pi_{\mathcal{X}} = p_s, \pi_{\mathcal{Y}} = p_t, \end{array} \tag{1}$$

where $\pi_{\mathcal{X}}, \pi_{\mathcal{Y}}$ denote the marginal measures of $\pi$ over $\mathcal{X}, \mathcal{Y}$ respectively. An optimal solution of (1) is referred to as an optimal transport plan or optimal coupling.

**Statistical OT**: In the setting of statistical OT, the marginals $p_s, p_t$ are not available; however, iid samples from them are given. Let $\mathcal{D}_x = \{x_1, \ldots, x_m\}$ denote the set of $m$ iid samples from $p_s$ and let $\mathcal{D}_y = \{y_1, \ldots, y_n\}$ denote $n$ iid samples from $p_t$. The cost function is known only at the sample data points. Let $\mathcal{C} \in \mathbb{R}^{m \times n}$ denote the cost matrix with with $(i,j)^{th}$ entry as $c(x_i, y_j)$.

A popular way to estimate the optimal plan in (1) is to simply employ the sample based plug-in estimates for the marginals: $\hat{p}_s \equiv \frac{1}{m}\sum_{i=1}^{m}\delta_{x_i}$ and $\hat{p}_t \equiv \frac{1}{n}\sum_{j=1}^{n}\delta_{y_j}$, in place of the true (unknown) marginals. Here, $\delta$ denotes the Dirac delta function. In such a case, (1) simplifies as the standard discrete OT problem:

$$\begin{array}{ll} \min_{\pi\in\mathbb{R}^{m\times n}} & tr(\pi\mathcal{C}^\top), \\ \text{s.t.} & \pi\mathbf{1} = \frac{1}{m}\mathbf{1}, \pi^\top\mathbf{1} = \frac{1}{n}\mathbf{1}, \pi \geq \mathbf{0}, \end{array} \tag{2}$$

where $tr(M)$ is the trace of matrix $M$, and $\mathbf{1}, \mathbf{0}$ denote vectors/matrices with all entries as unity, zero respectively (of appropriate dimension). Since the sample complexity of (2) in estimating (1) is prohibitively high for high-dimensional domains [12], alternative estimation methods are sought after.

## 3 Proposed Methodology

We begin by re-formulating (1) solely in terms of kernel mean embeddings and operators. Let $k_1, k_2$ be characteristic kernels [16, 39] defined over $\mathcal{X}, \mathcal{Y}$ respectively. By definition, the key advantage of a characteristic kernel is that the mapping between kernel mean embeddings and $\mathcal{M}^1$ becomes one-to-one (injective). For discrete probability measures, the Kronecker delta kernel is characteristic, while for continuous measures, the Gaussian kernel is an example of a characteristic as well as a universal kernel. Let $\phi_1, \phi_2$ and $\mathcal{H}_1, \mathcal{H}_2$ denote the canonical feature maps and the reproducing kernel Hilbert spaces (RKHS) corresponding to the kernels $k_1, k_2$ respectively. Let $\langle\cdot,\cdot\rangle_{\mathcal{H}_i}, \|\cdot\|_{\mathcal{H}_i}$ denote the default inner-product, norm in the RKHS $\mathcal{H}_i$. Let $\mu_s \equiv \mathbb{E}_{X\sim p_s}\left[\phi_1(X)\right],\ \mu_t \equiv \mathbb{E}_{Y\sim p_t}\left[\phi_2(Y)\right]$ denote the kernel mean embeddings of the marginals $p_s, p_t$ respectively. Let $\Sigma_{ss} \equiv \mathbb{E}_{X\sim p_s}\left[\phi_1(X)\otimes\phi_1(X)\right]$ and $\Sigma_{tt} \equiv \mathbb{E}_{Y\sim p_t}\left[\phi_2(Y)\otimes\phi_2(Y)\right]$ denote the auto-covariance embeddings of $p_s, p_t$ respectively. Here $\otimes$ denotes tensor product. Using these embeddings one can compute expectations of functions of the respective random variables: for e.g., $\mathbb{E}[f(X)] = \mathbb{E}[\langle f, \phi(x)\rangle] = \langle f, \mathbb{E}[\phi(X)]\rangle$ etc.

Since the variable, $\pi$, is a joint measure, the cross-covariance operator, $\mathcal{U}^\pi = \mathbb{E}_{(X,Y)\sim\pi}\left[\phi_1(X)\otimes\phi_2(Y)\right]$, is the suitable kernel mean embedding to be employed. However, the constraints involve the marginals of $\pi$, whose embeddings cannot be retrieved from the cross-covariance operator alone. Hence we also employ the conditional embedding operators, $\mathcal{U}_1^\pi, \mathcal{U}_2^\pi$, which embed the conditionals $\pi_{\mathcal{Y}/\mathcal{X}}(\cdot/\cdot)$ and $\pi_{\mathcal{X}/\mathcal{Y}}(\cdot/\cdot)$ respectively. The relations between these operators and embeddings follow from the definition of conditional embedding and the kernel sum rule [37]: $\mathcal{U} = \Sigma_{ss}\mathcal{U}_1^\top = \mathcal{U}_2\Sigma_{tt}$, $\mathcal{U}_1\mu_s = \mu_t$, $\mathcal{U}_2\mu_t = \mu_s$.

In order to re-write the objective using the above operators, we assume that the cost function, $c$, can be embedded in $\mathcal{H}_2\otimes\mathcal{H}_1$. This assumption is trivially true if the domains are discrete. However, in case of continuous domains this need not be true, in general. Hence we additionally assume that the kernel corresponding to a continuous domain is universal and that the cost function is continuous in that continuous variable. It then follows from the definition of universal kernels that a continuous function like $c(\cdot,\cdot)$ can be arbitrarily closely approximated by elements in $\mathcal{H}_2\otimes\mathcal{H}_1$ [39]. Note that universal kernels are well-studied and known for non-standard domains too [7].

Now, the objective in (1) can be written as: $\mathbb{E}\left[c(X,Y)\right] = \langle c, \mathcal{U}\rangle_{\mathcal{H}_2\otimes\mathcal{H}_1}$. This leads to the following kernel embedding formulation for OT:

$$\begin{array}{ll} \min\limits_{\mathcal{U},\mathcal{U}_1,\mathcal{U}_2} & \langle c, \mathcal{U}\rangle_{\mathcal{H}_2\otimes\mathcal{H}_1} \\ \text{s.t.} & \mathcal{U}_1\mu_s = \mu_t,\ \mathcal{U}_2\mu_t = \mu_s,\ \mathcal{U} = \Sigma_{ss}\mathcal{U}_1^\top = \mathcal{U}_2\Sigma_{tt}, \\ & \mathcal{U}\in\mathcal{E}\left(\mathcal{H}_2,\mathcal{H}_1\right), \mathcal{U}_1\in\mathcal{L}\left(\mathcal{H}_1,\mathcal{H}_2\right), \mathcal{U}_2\in\mathcal{L}\left(\mathcal{H}_2,\mathcal{H}_1\right), \end{array} \tag{3}$$

where $\mathcal{L}\left(\mathcal{H}_1,\mathcal{H}_2\right)$ is the set of all linear operators from $\mathcal{H}_1 \mapsto \mathcal{H}_2$, and $\mathcal{E}\left(\mathcal{H}_2,\mathcal{H}_1\right) \equiv \left\{\mathcal{U}\in\mathcal{L}\left(\mathcal{H}_2,\mathcal{H}_1\right) \mid \exists p\in\mathcal{M}^1(\mathcal{X},\mathcal{Y}) \ni \mathcal{U} = \mathbb{E}_{(X,Y)\sim p}\left[\phi_1(X)\otimes\phi_2(Y)\right]\right\}$ is the set of all valid cross-covariance operators. The equivalence of (3) and (1) follows from the one-to-one correspondence between the measures involved and their kernel embeddings, which is guaranteed by the characteristic kernels, and from the crucial embedding characterizing constraint: $\mathcal{U}\in\mathcal{E}\left(\mathcal{H}_2,\mathcal{H}_1\right)$. Without this characterizing constraint, the formulation is not meaningful. We summarize the above re-formulation in the following theorem:

**Assumption 1.** *Both kernels $k_1, k_2$ are characteristic. Moreover, if $k_i$ is over a continuous domain, then it is universal.*

**Assumption 2.** *We assume that* $c \in \mathcal{H}_2 \otimes \mathcal{H}_1$*, where* $c$ *denotes either the exact function or the (arbitrarily) close approximation of it that can be embedded.*

**Theorem 1.** *Under Assumptions 1-2, the Kantorovich formulation of OT (1) is equivalent to (3).*

Note that unlike existing formulae for the operator embeddings [37], which eliminate two of the three operators $\mathcal{U}, \mathcal{U}_1, \mathcal{U}_2$; we critically preserve all of them in (3). This is because they facilitate efficient regularization in the statistical estimation set-up and lead to efficient algorithms (as will be shown later). Also, the characterization of embedding, $\mathcal{E}(\mathcal{H}_1, \mathcal{H}_2)$, is included only for the cross-covariance, and not explicitly included for the conditional operators. This is because the conditionals are well-defined given the cross-covariance, auto-covariance and marginal embeddings.

The main advantage of the proposed formulation (3) over (1) is that the sample based estimates for kernel mean embeddings of the marginals, which are known to have dimension-free sample complexities, can be employed directly in the statistical OT setting.

### 3.1 Proposed formulation for statistical OT

As motivated earlier, we aim to employ the standard sample based estimates for the kernel mean embeddings of the marginals in the re-formulation (3). To this end, let the estimates for the marginal kernel mean embeddings be denoted by: $\hat{\mu}_s \equiv \frac{1}{m}\sum_{i=1}^{m} \phi_1(x_i)$ and $\hat{\mu}_t \equiv \frac{1}{n}\sum_{j=1}^{n} \phi_2(y_j)$. Likewise, the estimates of the auto-covariance embeddings are given by $\hat{\Sigma}_{ss} \equiv \frac{1}{m}\sum_{i=1}^{m} \phi_1(x_i) \otimes \phi_1(x_i)$ and $\hat{\Sigma}_{tt} \equiv \frac{1}{n}\sum_{j=1}^{n} \phi_2(y_j) \otimes \phi_2(y_j)$.

In the statistical OT setting, the cost function, $c$, is only available at the given samples. In continuous domains, there will exist many functions in the RKHS that will exactly match $c$ restricted to the samples. Each such choice will lead to a valid estimator. We choose $\hat{c} \equiv \sum_{i=1}^{m}\sum_{j=1}^{n} \rho^*_{ij}\phi_1(x_i) \otimes \phi_2(y_j)$, where $\rho^* \equiv \arg\min_\rho \left\| c - \sum_{i=1}^{m}\sum_{j=1}^{n} \rho_{ij}\phi_1(x_i) \otimes \phi_2(y_j) \right\|_{\mathcal{H}_2 \otimes \mathcal{H}_1}$ and $\|\cdot\|_{\mathcal{H}_2 \otimes \mathcal{H}_1}$ is the Hilbert-Schmidt operator norm. For universal kernels, it follows that $\hat{c}$ will be equal to $c$ at the given samples, and hence the above is a valid choice for estimation. In addition, the above choice of $\hat{c}$ helps us in proving the representer theorem (Theorem 3).

Now, employing these estimates in (3) must be performed with caution as i) the equality constraints now will be in the (potentially infinite dimensional) RKHS, ii) more importantly, matching the estimates exactly will lead to over-fitting. Hence, we propose to introduce appropriate regularization by insisting that there is a close match rather than an exact match. This leads to the following kernel embedding learning formulation:

$$
\begin{aligned}
\min_{\mathcal{U}, \mathcal{U}_1, \mathcal{U}_2} \quad & \langle \hat{c}, \mathcal{U} \rangle_{\mathcal{H}_2 \otimes \mathcal{H}_1} \\
\text{s.t.} \quad & \|\mathcal{U}_1 \hat{\mu}_s - \hat{\mu}_t\|_{\mathcal{H}_2} \leq \Delta_1, \ \|\mathcal{U}_2 \hat{\mu}_t - \hat{\mu}_s\|_{\mathcal{H}_1} \leq \Delta_2, \\
& \left\|\mathcal{U} - \hat{\Sigma}_{ss}\mathcal{U}_1^\top\right\|_{\mathcal{H}_2 \otimes \mathcal{H}_1} \leq \epsilon_1, \left\|\mathcal{U} - \mathcal{U}_2\hat{\Sigma}_{tt}\right\|_{\mathcal{H}_2 \otimes \mathcal{H}_1} \leq \epsilon_2, \\
& \mathcal{U} \in \mathcal{E}(\mathcal{H}_2, \mathcal{H}_1), \mathcal{U}_1 \in \mathcal{L}(\mathcal{H}_1, \mathcal{H}_2), \mathcal{U}_2 \in \mathcal{L}(\mathcal{H}_2, \mathcal{H}_1),
\end{aligned}
\tag{4}
$$

where $\Delta_1, \Delta_2, \epsilon_1, \epsilon_2$ are regularization hyper-parameters introduced to prevent overfitting to the estimates. Setting $\Delta_i = 0 = \epsilon_i$ recovers the case where estimates of marginal mean embeddings and auto-covariances are exactly matched but it may lead to overfitting. Also, $\mathcal{U}_1, \mathcal{U}_2$ are guaranteed to be valid conditional embeddings only as $\Delta_i, \epsilon_i \rightarrow 0$. Hence, we suggest $\Delta_i, \epsilon_i = O\left(1/\sqrt{\min(m,n)}\right)$, following known sample complexities for the marginal embedding estimates, which are $O\left(1/\sqrt{m}\right), O\left(1/\sqrt{n}\right)$ respectively [36]. Since the kernel embedding estimates have sample complexities that are independent of dimension, it is expected that the statistical estimation error with the proposed formulation (4) is also independent of dimensionality. In the next theorem, we formalize the above statement:

**Assumption 3.** *Let us assume that the kernels are normalized/bounded i.e.,* $\max_{x \in \mathcal{X}} k_1(x,x) = 1, \max_{y \in \mathcal{Y}} k_2(y,y) = 1$.

**Theorem 2.** *Let* $g\left(\hat{c}, \hat{\mu}_s, \hat{\mu}_t, \hat{\Sigma}_{ss}, \hat{\Sigma}_{tt}\right)$ *denote the optimal objective of (4) in Tikhonov form. Under Assumptions 1-3, with high probability we have that,* $\left|g\left(\hat{c}, \hat{\mu}_s, \hat{\mu}_t, \hat{\Sigma}_{ss}, \hat{\Sigma}_{tt}\right) - g\left(c, \mu_s, \mu_t, \Sigma_{ss}, \Sigma_{tt}\right)\right| \leq O\left(1/\sqrt{\min(m,n)}\right)$. *The constants in the RHS of the inequality are dimension-free.*

Theorem 2 shows that with appropriate regularization one can obtain statistically consistent estimators for the embedding of the optimal transport plan by solving (4). More importantly, its proves that the sample complexity of these estimators is dimension-free. The proof of this theorem is detailed in Appendix A. The idea is to uniformly bound the difference between the population and sample versions of each of the terms in the objective. Interestingly, each of these difference terms can either be bounded by relevant estimation errors in embedding space or by approximation errors in the RKHS, both of which are known to be dimension-free.

Note that the regularization in (4) is based on the Maximum Mean Discrepancy (MMD) distances between the kernel embeddings. This characteristic of our estimators is in contrast with the popular entropic regularization [10], or $\phi$-divergence based regularization [24] in existing OT estimators. [38] argue that MMD and $\phi$-divergence based regularization have complementary properties. Hence both are interesting to study. While the dependence on dimensionality is adversely exponential with entropic regularization, if accurate solutions are desired [17], the proposed MMD based regularization for statistical OT leads to dimension-free estimation.

### 3.2 Representer theorem & Kernelization

Interestingly, (4) admits a finite parameterization facilitating it's efficient optimization. This important result is summarized in the representer theorem below:

**Theorem 3.** *Whenever (4) is solvable, there exists an optimal solution, $\mathcal{U}^*, \mathcal{U}_1^*, \mathcal{U}_2^*$, of (4) such that $\mathcal{U}^* = \sum_{i=1}^m \sum_{j=1}^n \alpha_{ij}\phi_1(x_i) \otimes \phi_2(y_j), \mathcal{U}_1^* = \sum_{i=1}^m \sum_{j=1}^n \beta_{ji}\phi_2(y_j) \otimes \phi_1(x_i), \mathcal{U}_2^* = \sum_{i=1}^m \sum_{j=1}^n \gamma_{ij}\phi_1(x_i) \otimes \phi_2(y_j)$. Here $\alpha \in \mathbb{R}^{m\times n}, \beta \in \mathbb{R}^{n\times m}, \gamma \in \mathbb{R}^{m\times n}$ that are an optimal solution for the kernelized and convex formulation (5) given below:*

$$
\begin{array}{ll}
\min\limits_{\alpha,\gamma\in\mathbb{R}^{m\times n},\beta\in\mathbb{R}^{n\times m}} & tr(\alpha\mathcal{C}^\top) \\
\text{s.t.} & \frac{1}{m^2}\mathbf{1}^\top G_1\beta^\top G_2\beta G_1\mathbf{1} - \frac{2}{mn}\mathbf{1}^\top G_2\beta G_1\mathbf{1} + \frac{1}{n^2}\mathbf{1}^\top G_2\mathbf{1} \le \Delta_1^2 \\
& \frac{1}{n^2}\mathbf{1}^\top G_2\gamma^\top G_1\gamma G_2\mathbf{1} - \frac{2}{mn}\mathbf{1}^\top G_1\gamma G_2\mathbf{1} + \frac{1}{m^2}\mathbf{1}^\top G_1\mathbf{1} \le \Delta_2^2 \\
& \left\langle G_1\alpha - \frac{1}{m}G_1^2\beta^\top, \alpha G_2 - \frac{1}{m}G_1\beta^\top G_2\right\rangle_F \le \epsilon_1^2, \\
& \left\langle \alpha G_2 - \frac{1}{n}\gamma G_2^2, G_1\alpha - \frac{1}{n}G_1\gamma G_2\right\rangle_F \le \epsilon_2^2, \\
& \alpha \ge 0, \mathbf{1}^\top\alpha\mathbf{1} = 1,
\end{array}
\tag{5}
$$

*where, $G_1$ and $G_2$ are the gram-matrices with $k_1$ and $k_2$ over $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$ respectively.*

The proof of this theorem is detailed in Appendix B. Apart from standard representer theorem-type arguments, the proof includes arguments that show that the characterizing set $\mathcal{E}(\mathcal{H}_2, \mathcal{H}_1)$ when restricted to the linear combinations of embeddings is exactly same as the convex combinations of those. This helps us replace the membership to $\mathcal{E}(\mathcal{H}_2, \mathcal{H}_1)$ constraint by a simplex constraint.

We note that (5) is jointly convex in the variables $\alpha, \beta$, and $\gamma$. This is because the constraints are either convex quadratic or linear and the objective is also linear. Hence obtaining consistent estimators using (5) is computationally tractable (refer section 3.4). It is easy to verify that (5) simplifies to the discrete OT problem (2) if both the kernels are chosen to be the Kronecker delta and all the hyper-parameters are set to zero. If one of the kernel is chosen as the Kronecker delta and the other as the Gaussian kernel, then (5) can be used for semi-discrete OT in the statistical setting. Additionally, by employing appropriate universal kernels, (5) can be used for statistical OT in non-standard domains.

We end this section with a small technical note. While the cross-covariance operator obtained by solving (5) will always be a valid one; for some hyper-parameters, which are too high, it may happen that the optimal $\beta, \gamma$ induce invalid conditional embeddings. This may make computing the transport map (6) intractable. Hence, in practice, we include additional constraints $\beta, \gamma \ge 0$.

### 3.3 Proposed Optimal Map Estimator

Once the embedding of the transport plan is obtained by solving (5), generic approaches for recovering the measure corresponding to a kernel embedding, detailed in [21, 33], can be employed to recover the corresponding transport plan. Moreover, since the recovery methods in [33] have dimension-free sample complexity, the overall sample complexity for estimating the optimal transport plan hence remains dimension-free.

We estimate the Barycentric-projection based optimal transport map, $\mathcal{T}$, at any $x \in \mathcal{X}$ as follows:

$$\begin{aligned} \mathcal{T}(x) \equiv \underset{y\in\mathcal{Y}}{\operatorname{argmin}}\, \mathbb{E}\left[c\left(y, Y\right)/x\right] &= \underset{y\in\mathcal{Y}}{\operatorname{argmin}}\, \langle c(y,\cdot), \mathcal{U}_1^* \phi_1(x)\rangle\,, \\ &= \underset{y\in\mathcal{Y}}{\operatorname{argmin}} \textstyle\sum_{j=1}^{n} \Big(c(y,y_j) \sum_{j=1}^{n}\left(\beta^*_{ji} k_1\left(x_i, x\right)\right)\Big)\,, \end{aligned} \tag{6}$$

where $\beta^*$ are obtained by solving (5) and $\mathcal{U}_1^*$ is the corresponding conditional embedding. (6) turns out to be that of finding the Karcher mean [23], whenever the cost is a squared-metric etc. Alternatively, one can directly minimize $\mathbb{E}\left[c\left(y, Y\right)/x\right]$ with respect to $y \in \mathcal{Y}$ using stochastic gradient descent (SGD). The following theorem summarizes the consistency with SGD:

**Theorem 4.** *Let the cost be a metric or it's powers greater than unity and let $\mathcal{Y}$ be compact. Then the SGD based estimator for $\mathcal{T}$ has a sample complexity that remains dimension-free.*

The proof of this theorem is detailed in Appendix C and follows from standard results in stochastic convex optimization.

An advantage with our map estimator is that it can be computed even at out-of-sample $x \in \mathcal{X}$. This is possible because of the implicit smoothing induced by the kernel.

### 3.4 Algorithms

The structure in the proposed problem (5) can be exploited to derive efficient alternating directions method of multipliers (ADMM) [4] based solvers. Further speed-up may be obtained in the special case when $\epsilon_i = 0$ in (5). This simplifies the constraints corresponding to $\epsilon_1$ and $\epsilon_2$ in (5) as $\alpha = (1/m)G_1\beta^\top$ and $\alpha = (1/n)\gamma G_2$, respectively. In addition, we re-write the regularizations corresponding to $\Delta_i^2$ in Tikhonov form. The above leads (5) to the following optimization problem:

$$\begin{aligned} \min_{\alpha\in\mathcal{A}_{mn}, \beta\in\mathbb{R}^{m\times n}\geq 0, \gamma\in\mathbb{R}^{m\times n}\geq 0} \quad & tr(\alpha\mathcal{C}^\top) + \lambda_1 \left\|\alpha\mathbf{1} - \tfrac{1}{m}\mathbf{1}\right\|_{G_1}^2 + \lambda_2 \left\|\alpha^\top\mathbf{1} - \tfrac{1}{n}\mathbf{1}\right\|_{G_2}^2 \\ \text{s.t.} \quad & \alpha = \tfrac{1}{m}G_1\beta^\top, \alpha = \tfrac{1}{n}\gamma G_2 \end{aligned} \tag{7}$$

where $\mathcal{A}_{mn} = \{x \in \mathbb{R}^{m\times n} \mid x \geq \mathbf{0}, \mathbf{1}^\top x \mathbf{1} = 1\}$ and $\lambda_i > 0$ are the regularization hyper-parameter corresponding to $\Delta_i^2$ in (5). The updates for the ADMM are summarized below:

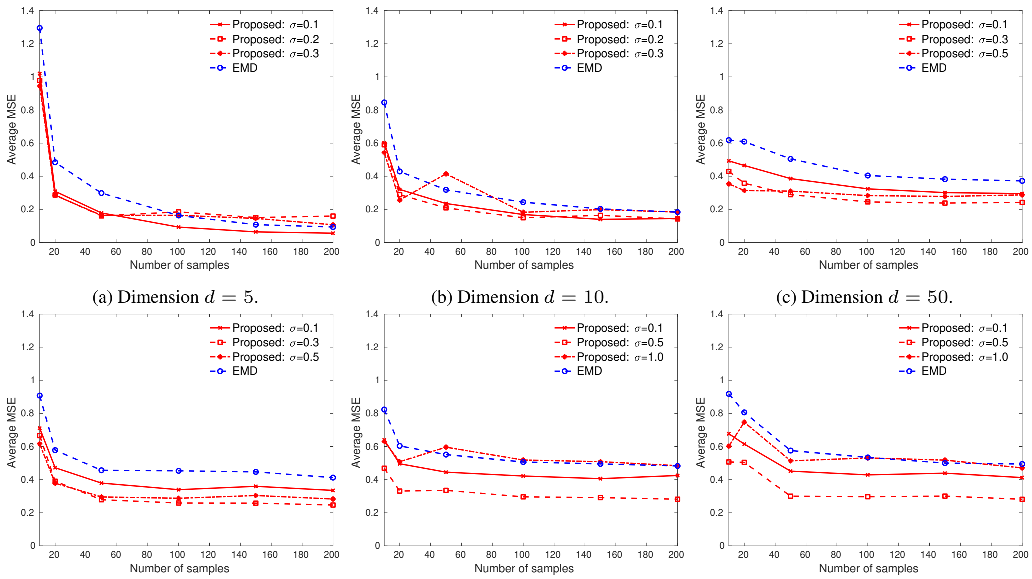
$$\begin{aligned} \alpha^{(k+1)} &:= \underset{\alpha\in\mathcal{A}_{mn}}{\operatorname{argmin}}\, \rho \left\|\alpha + \tfrac{1}{2}\left(D_1^{(k)} + D_2^{(k)} + \tfrac{\mathcal{C}}{\rho} - \tfrac{\gamma^{(k)}G_2}{n} - \tfrac{G_1\beta^{(k)\top}}{m}\right)\right\|^2 \\ &\qquad + \lambda_1 \left\|\alpha\mathbf{1} - \tfrac{\mathbf{1}}{m}\right\|_{G_1}^2 + \lambda_2 \left\|\alpha^\top\mathbf{1} - \tfrac{\mathbf{1}}{n}\right\|_{G_2}^2, \\ \beta^{(k+1)} &:= \underset{\beta\geq 0}{\operatorname{argmin}} \left\|\alpha^{(k+1)} + D_1^{(k)} - \tfrac{G_1\beta^\top}{m}\right\|^2, \\ \gamma^{(k+1)} &:= \underset{\gamma\geq 0}{\operatorname{argmin}} \left\|\alpha^{(k+1)} + D_2^{(k)} - \tfrac{\gamma G_2}{n}\right\|^2, \\ D_1^{(k+1)} &:= D_1^{(k)} + \left(\alpha^{(k+1)} - \tfrac{G_1\beta^{(k+1)\top}}{m}\right), \\ D_2^{(k+1)} &:= D_2^{(k)} + \left(\alpha^{(k+1)} - \tfrac{\gamma^{(k+1)}G_2}{n}\right), \end{aligned}$$

where $D_1$ and $D_2$ are the dual variables corresponding to the constraints $\alpha = (1/m)G_1\beta^\top$ and $\alpha = (1/n)\gamma G_2$ in (7), respectively. The optimization problems with respect to $\alpha, \beta$, and $\gamma$ can be solved efficiently using popular algorithms like conditional gradient descent, mirror descent, co-ordinate descent, conjugate gradients, etc. Since the convergence rate of these algorithms is either independent or almost independent (logarithmically dependent) on the dimensionality of the problem, the computational cost (after neglecting log factors, if any) of solving for: $\alpha$ is $O(mn)$, $\beta$ is $O(m^2n)$, and $\gamma$ is $O(mn^2)$. The updates for $D_1$ and $D_2$ have computational costs: $O(m^2n)$ and $O(mn^2)$. Without loss of generality, if we assume $m \geq n$, the per iteration cost of ADMM is $O(m^3)$.

As noted earlier, in typical cases where the hyper-parameters $\Delta_i$ are small enough, explicit constraints $\beta \geq 0, \gamma \geq 0$ are not needed:

$$\min_{\alpha\in\mathcal{A}_{mn}} \quad tr(\alpha\mathcal{C}^\top) + \lambda_1 \left\|\alpha\mathbf{1} - \tfrac{1}{m}\mathbf{1}\right\|_{G_1}^2 + \lambda_2 \left\|\alpha^\top\mathbf{1} - \tfrac{1}{n}\mathbf{1}\right\|_{G_2}^2 \tag{8}$$

Hence the computational cost in this special case is $O(mn)$, which is linear, and hence comparable to that of Sinkhorn algorithm popularly used to solve the discrete OT problem.




(a) Dimension $d = 5$. (b) Dimension $d = 10$. (c) Dimension $d = 50$.

(d) Dimension $d = 100$. (e) Dimension $d = 500$. (f) Dimension $d = 1000$.

Figure 1: Performance on the proposed estimator for the transport map (6) and the discrete OT estimator, EMD, on the problem of learning the optimal transport map between two multivariate Gaussian distributions. We observe that the proposed estimator outperforms EMD, especially in higher dimensions.

# 4 Related Work

A popular strategy for performing continuous statistical OT is to simply employ the sample based plug-in estimates for the marginals. This reduces the statistical OT problem to the classical discrete OT problem, for which efficient algorithms exist [10, 1]. However, the sample complexity of the discrete OT based estimation is plagued with the curse of dimensionality [12]. [18, 17, 14] note that estimators that are free from the curse of dimensionality are not well-studied and propose alternatives.

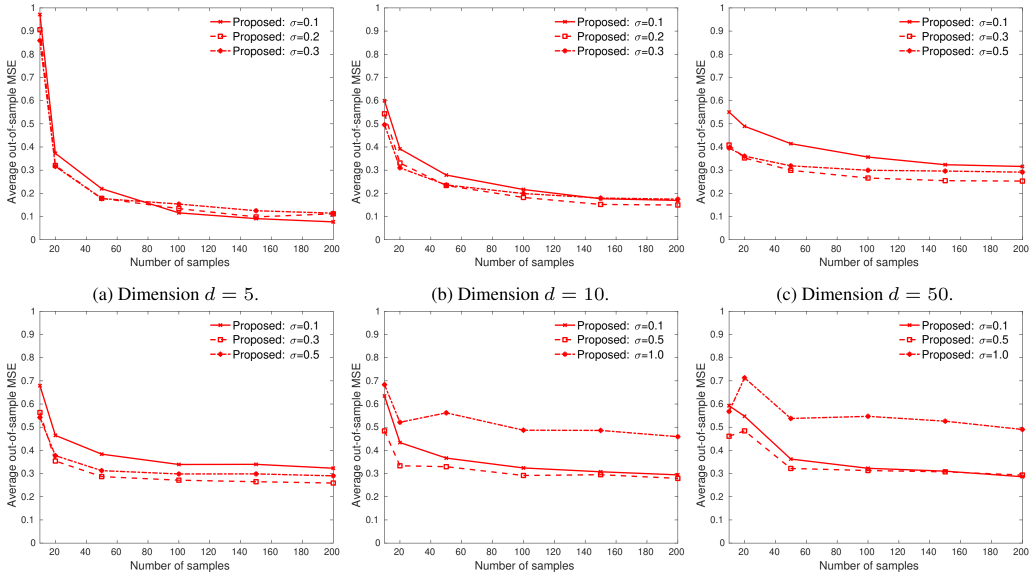
While the approach of [18] efficiently estimates the optimal dual objective, recovering the optimal transport plan from the dual's solution again requires the knowledge of the exact marginals (refer proposition 2.1 in [18]). Since estimating distributions in high-dimensional settings is known to be challenging, this alternative is not attractive for applications where the transport plan is required, e.g., domain adaptation [9] and ecological inference [26], etc.

[17] observe that continuous statistical OT is the major bottleneck for applying OT in ML problems and propose an entropic regularization based alternative. However, their results (e.g., theorem 3 in [17]) show that the curse of dimensionality is not completely removed, especially if accurate solutions are desired. Empirical results in [13, refer Figures 4 and 5] confirm that the quality of the solution degrades very quickly with entropic regularization. The alternative in [14] makes strong low-rank based assumptions, which may not be realistic in all applications. Infact, the report on a recent workshop on OT (refer section 2 in [5]) summarizes that one of the major open problems in this area is to design estimators for continuous statistical OT whose sample complexity is not a strong function of dimensionality (ideally dimension-free).

On passing we note that though there are existing works that employ kernels in context of OT [18, 29, 42, 27], none of them use the notion of kernel embedding of distributions and limit the use of kernels to either function approximation or computing MMD distance. Though relations betweeen Wasserstein and MMD distance [13] exist, none of them explore regularization with MMD distances.

# 5 Experiments

We evaluate our estimator for the transport map (6) on both synthetic and real-world datasets.




(a) Dimension $d = 5$. (b) Dimension $d = 10$. (c) Dimension $d = 50$.

(d) Dimension $d = 100$. (e) Dimension $d = 500$. (f) Dimension $d = 1000$.

Figure 2: Average out-of-sample mean square error (MSE) obtained by the proposed approach on the problem of learning the optimal transport map between two multivariate Gaussian distributions. In general, the average out-of-sample MSE decrease with increasing number of data points sampled to learn the estimator (the x-axis). We also observe that the best results obtained by proposed solution is robust to the dimensionality of the data points.

### 5.1 Learning OT map between multivariate Gaussian distributions

The optimal transport map between two Gaussian distributions $g_{source} = N(m_1, \Sigma_1)$ and $g_{target} = N(m_2, \Sigma_2)$ with squared Euclidean cost has a closed form expression [30] given by $T : x \mapsto m_2 + A(x - m_1)$, where $A = \Sigma_1^{-\frac{1}{2}}(\Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}})^{\frac{1}{2}}\Sigma_1^{-\frac{1}{2}}$. We compare the proposed estimator (6) in terms of the deviation from the optimal transport map.

**Experimental setup**: We consider mean zero Gaussian distributions with unit-trace covariances. The covariance matrices are computed as $\Sigma_1 = V_1V_1^\top/\|V_1\|_F$ and $\Sigma_2 = V_2V_2^\top/\|V_2\|_F$, where $V_1 \in \mathbb{R}^{d\times d}$ and $V_2 \in \mathbb{R}^{d\times d}$ are generated randomly from the uniform distribution. We experiment with varying dimensions and number of data-points: $d \in \{5, 10, 50, 100, 500, 1000\}$, $m \in \{10, 20, 50, 100, 150, 200\}$, and we set $n = m$ for simplicity. For each dimension $d$, we randomly generate a source-target distribution pair. Subsequently, the source and target datasets (of size $m$) are randomly generated from their respective distributions. For a every $(d, m)$, we repeat the experiments five times and report the average mean square error (MSE) results results in Figures 1 and 2. In a second set of experiments, we also study the variance in the results of a given optimal transport problem caused due to random data-points. In this setup a source-target distribution pair is randomly generated for a given $d$. From this distribution pair, source-target datasets are randomly generated five times for every $m$. Average MSE results are reported in Tables 1-3.

**Methods**: The proposed approach employs the Gaussian kernels, $k(x, z) = \exp(-\|x - z\|^2/2\sigma^2)$. We chose the same $\sigma$ values for the kernels over the source and the target domains ($k_1$ and $k_2$, respectively). Initial experiments indicate that suitable values of $\sigma$ include those that does not yield high condition number of the Gram matrices (i.e, the Gram matrices are not ill-conditioned). In our setup, in general, the condition number of the Gram matrices increase with $\sigma$ for a fixed $d$ and decrease with $d$ for a fixed $\sigma$. The $\sigma$ values used in various experiments are mentioned with the results. As a baseline, we also report the results obtained from the discrete OT estimator, henceforth referred to as EMD, learned via the discrete OT problem (2).

**Evaluation**: For a given data point $x_s$ from the source distribution, a transport map estimator maps $x_s$ to a data point $x_t$ in the target distribution. Such a mapping obtained from the optimal transport

Table 1: Average MSE on the problem of learning the optimal transport map between two given multivariate Gaussian distributions with $d = 10$. For all $m$, we randomly sample data points from a fixed randomly sampled source-target distribution. We observe that the proposed approach easily outperforms EMD.

| **Method** | $m = 10$ | $m = 20$ | $m = 50$ | $m = 100$ | $m = 150$ | $m = 200$ |
|---|---|---|---|---|---|---|
| EMD | $0.53 \pm 0.18$ | $0.42 \pm 0.11$ | $0.32 \pm 0.06$ | $0.27 \pm 0.04$ | $0.24 \pm 0.02$ | $0.21 \pm 0.01$ |
| Proposed ($\sigma = 0.1$) | $0.45 \pm 0.14$ | $\mathbf{0.23 \pm 0.05}$ | $0.23 \pm 0.03$ | $0.19 \pm 0.03$ | $\mathbf{0.17 \pm 0.01}$ | $\mathbf{0.15 \pm 0.01}$ |
| Proposed ($\sigma = 0.2$) | $0.41 \pm 0.15$ | $0.27 \pm 0.06$ | $\mathbf{0.20 \pm 0.02}$ | $\mathbf{0.17 \pm 0.02}$ | $0.23 \pm 0.10$ | $0.16 \pm 0.04$ |
| Proposed ($\sigma = 0.3$) | $\mathbf{0.37 \pm 0.14}$ | $0.25 \pm 0.06$ | $0.22 \pm 0.03$ | $0.20 \pm 0.03$ | $0.25 \pm 0.11$ | $0.19 \pm 0.01$ |

Table 2: Average MSE on the problem of learning the optimal transport map between two given multivariate Gaussian distributions with $d = 100$. For all $m$, we randomly sample data points from a fixed randomly sampled source-target distribution. We observe that the proposed approach easily outperforms EMD.

| **Method** | $m = 10$ | $m = 20$ | $m = 50$ | $m = 100$ | $m = 150$ | $m = 200$ |
|---|---|---|---|---|---|---|
| EMD | $0.69 \pm 0.11$ | $0.68 \pm 0.15$ | $0.56 \pm 0.06$ | $0.45 \pm 0.03$ | $0.42 \pm 0.01$ | $0.41 \pm 0.01$ |
| Proposed ($\sigma = 0.1$) | $0.54 \pm 0.08$ | $0.54 \pm 0.10$ | $0.41 \pm 0.03$ | $0.36 \pm 0.01$ | $0.35 \pm 0.01$ | $0.34 \pm 0.08$ |
| Proposed ($\sigma = 0.3$) | $0.47 \pm 0.06$ | $0.44 \pm 0.08$ | $\mathbf{0.31 \pm 0.02}$ | $\mathbf{0.26 \pm 0.01}$ | $\mathbf{0.25 \pm 0.01}$ | $\mathbf{0.25 \pm 0.04}$ |
| Proposed ($\sigma = 0.5$) | $\mathbf{0.40 \pm 0.06}$ | $\mathbf{0.37 \pm 0.07}$ | $0.32 \pm 0.02$ | $0.29 \pm 0.01$ | $0.30 \pm 0.02$ | $0.29 \pm 0.01$ |

map (15) is considered as the ground truth. The proposed estimator (8) and the EMD are evaluated in terms of the mean squared error (MSE) from the ground truth.

**Results**: The results of our first set of experiments are reported in Figures 1(a)-(f). We observe that the proposed estimator obtains lower average MSE (and hence better estimation of the transport map) than EMD across different number of samples $m$ and dimensions $d$. The advantage of the proposed estimator over the baseline is more pronounced at higher dimension.

In Table 1, we report the results of the second set of experiments with $d = 10$. We again observe that the proposed approach outperforms EMD. Results on the same experimental setup but with $d = 100$ and $d = 1000$ are report in Tables 2 and 3, respectively.

**Out-of-sample evaluation**: We also evaluate our estimator's ability to map out-of-sample data by sampling additional $m_{oos} = 200$ points from the source distributions in the above experiments. These source points are not used to learn the estimator and are only used for evaluation during the inference stage. The results on out-of-sample dataset, corresponding to the first set of experiments (Figure 1) are reported in Figure 2. We generate out-of-sample data points for each $(d, s)$ pair, where $d$ is the dimension of the data points and $s$ is the random seed (corresponding to five repetition discussed earlier). Hence, for a given $(d, s)$ pair, different estimators learned with varying $m$ are evaluated on the same set of out-of-sample data points.

We observe that the performance on out-of-sample data points are similar to the in-sample data points (Figures 1(a) & (b)). The average out-of-sample MSE generally decreases with increasing number of (training) samples since a better estimator is learned with more number of (training) samples. Overall,

Table 3: Average MSE on the problem of learning the optimal transport map between two given multivariate Gaussian distributions with $d = 1000$. For all $m$, we randomly sample data points from a fixed randomly sampled source-target distribution. We observe that the proposed approach easily outperforms EMD.

| **Method** | $m = 10$ | $m = 20$ | $m = 50$ | $m = 100$ | $m = 150$ | $m = 200$ |
|---|---|---|---|---|---|---|
| EMD | $0.89 \pm 0.15$ | $0.64 \pm 0.11$ | $0.59 \pm 0.06$ | $0.55 \pm 0.07$ | $0.50 \pm 0.01$ | $0.49 \pm 0.01$ |
| Proposed ($\sigma = 0.1$) | $0.69 \pm 0.12$ | $0.51 \pm 0.06$ | $0.46 \pm 0.04$ | $0.43 \pm 0.03$ | $0.41 \pm 0.01$ | $0.41 \pm 0.01$ |
| Proposed ($\sigma = 0.5$) | $\mathbf{0.53 \pm 0.16}$ | $\mathbf{0.36 \pm 0.08}$ | $\mathbf{0.33 \pm 0.05}$ | $\mathbf{0.31 \pm 0.05}$ | $\mathbf{0.29 \pm 0.02}$ | $\mathbf{0.29 \pm 0.01}$ |
| Proposed ($\sigma = 1.0$) | $0.67 \pm 0.19$ | $0.54 \pm 0.12$ | $0.55 \pm 0.13$ | $0.52 \pm 0.12$ | $0.51 \pm 0.06$ | $0.50 \pm 0.03$ |

Table 4: Accuracy obtained on the target domains of the Office-Caltech dataset. The knowledge transfer to the target domain happens via in-sample source data-points, i.e., those source data-points using which the transport plan was learned.

| **Task** | **EMD** | **OTLin [28]** | **OTKer [28]** | **Proposed** |
|---|---|---|---|---|
| $A \to C$ | $80.68 \pm 1.82$ | $82.92 \pm 1.41$ | $83.07 \pm 0.63$ | $\mathbf{86.27 \pm 1.74}$ |
| $A \to D$ | $72.66 \pm 6.58$ | $82.28 \pm 5.66$ | $82.53 \pm 3.70$ | $\mathbf{84.30 \pm 5.41}$ |
| $A \to W$ | $69.05 \pm 5.08$ | $\mathbf{77.70 \pm 3.60}$ | $76.35 \pm 4.16$ | $76.22 \pm 3.32$ |
| $C \to A$ | $82.61 \pm 3.45$ | $88.31 \pm 0.94$ | $88.09 \pm 1.50$ | $\mathbf{91.05 \pm 0.79}$ |
| $C \to D$ | $68.35 \pm 10.06$ | $79.75 \pm 6.04$ | $78.99 \pm 7.95$ | $\mathbf{82.78 \pm 3.81}$ |
| $C \to W$ | $65.54 \pm 2.74$ | $71.89 \pm 3.43$ | $70.00 \pm 3.93$ | $\mathbf{74.46 \pm 4.45}$ |
| $D \to A$ | $81.50 \pm 1.99$ | $88.57 \pm 1.86$ | $85.23 \pm 1.71$ | $\mathbf{90.92 \pm 1.23}$ |
| $D \to C$ | $76.51 \pm 2.87$ | $82.17 \pm 1.70$ | $78.22 \pm 1.80$ | $\mathbf{86.84 \pm 0.86}$ |
| $D \to W$ | $91.89 \pm 2.96$ | $\mathbf{97.57 \pm 1.09}$ | $96.35 \pm 1.10$ | $96.22 \pm 1.84$ |
| $W \to A$ | $71.22 \pm 1.54$ | $80.00 \pm 1.74$ | $76.23 \pm 2.50$ | $\mathbf{86.90 \pm 2.42}$ |
| $W \to C$ | $69.55 \pm 3.18$ | $77.58 \pm 2.34$ | $73.72 \pm 2.59$ | $\mathbf{82.28 \pm 1.31}$ |
| $W \to D$ | $80.76 \pm 4.90$ | $\mathbf{97.72 \pm 1.47}$ | $96.20 \pm 2.89$ | $96.20 \pm 3.00$ |
| Average | $75.86 \pm 1.43$ | $83.87 \pm 0.38$ | $82.08 \pm 0.95$ | $\mathbf{86.20 \pm 0.98}$ |

the results illustrate the utility of the proposed approach for out-of-sample estimation. It should be noted that the baseline EMD cannot map out-of-sample data points.

### 5.2 Domain adaptation

We experiment on the Caltech-Office dataset [20], which contains images from four domains: Amazon (online retail), the Caltech image dataset, DSLR (images taken from a high resolution DSLR camera), and Webcam (images taken from a webcam). The domains vary with respect to factors such as background, lightning conditions, noise, etc. The number of examples in each domain is: $958$ (Amazon), $1123$ (Caltech), $157$ (DSLR), and $295$ (Webcam). Each domain has images from ten classes. We perform multiclass classification in the domain adaptation setting, where each domain is in turn considered as the source or the target. Overall, there are twelve adaptation tasks (e.g., task $A \to C$ has Amazon as the source and Caltech as the target domain). We employ DeCAF6 features to represent the images [11, 28, 8].

**Experimental setup**: For learning transport plan, we randomly select ten images per class for the source domain (eight per class when DSLR is the source, due to its sample size). The remaining samples of the source domain is marked as out-of-sample source data-points. The target domain is partitioned equally into training and test sets. The transport map is learned using the source-target training sets. The 'in-sample' accuracy is then evaluated on the target's test set. We also evaluate the quality of our out-of-sample estimation as follows. Instead of projecting the source training set samples onto the target domain, we project only the out-of-sample (OOS) source data-points and compute the accuracy over the target's test set. It should be noted that the transport model has not been learned on the OOS data-points, such mappings may not be as accurate as the in-sample mapping. The OOS evaluation assesses the downstream effectiveness of OOS estimation on domain adaptation. Out-of-sample estimation is especially attractive in big data and online applications. The classification in the target domain is performed using a 1-Nearest Neighbor classifier [20, 28, 8]. The above experimentation is performed five times. The average in-sample and out-of-sample accuracy are reported in Tables 4 & 5, respectively.

**Methods**: We compare our approach with EMD, OTLin [28], and OTKer [28]. Both OTLin and OTKer aim to solve the discrete optimal transport problem and also learn a transformation approximating the corresponding transport map in a joint optimization framework. OTLin learns a linear transformation while OTKer learns a non-linear transformation (e.g., via Gaussian kernel). The learned transformation allows OTLin and OTKer to perform out-of-sample estimation as well. Both OTLin and OTKer employ two regularization parameters. As suggested by their authors [28], both the regularization parameters were chosen from the set $\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}\}$. It should be noted that best regularization parameters were selected for each task. OTKer additionally requires Gaussian kernel's hyper-parameter $\sigma$, which was chosen from the set $\{0.1, 0.5, 1, 5, 10\}$. We use the

Table 5: Accuracy obtained on the target domains of the Office-Caltech dataset. The knowledge transfer to the target domain happens via out-of-sample source data-points, i.e., those source data-points which were not used for learning the transport plan.

| Task | OTLin [28] | OTKer [28] | Proposed |
|---|---|---|---|
| $A \to C$ | $56.75 \pm 2.94$ | $79.11 \pm 2.76$ | $\mathbf{84.71 \pm 1.82}$ |
| $A \to D$ | $79.49 \pm 1.86$ | $82.79 \pm 2.06$ | $\mathbf{85.82 \pm 0.95}$ |
| $A \to W$ | $55.41 \pm 6.60$ | $76.35 \pm 1.66$ | $\mathbf{81.62 \pm 2.11}$ |
| $C \to A$ | $87.79 \pm 3.28$ | $84.54 \pm 2.78$ | $\mathbf{90.66 \pm 0.87}$ |
| $C \to D$ | $81.01 \pm 3.75$ | $74.94 \pm 3.53$ | $\mathbf{81.52 \pm 3.97}$ |
| $C \to W$ | $70.00 \pm 3.64$ | $68.11 \pm 1.16$ | $\mathbf{74.05 \pm 4.16}$ |
| $D \to A$ | $64.53 \pm 5.01$ | $81.95 \pm 2.69$ | $\mathbf{85.47 \pm 2.74}$ |
| $D \to C$ | $43.67 \pm 4.80$ | $72.79 \pm 3.04$ | $\mathbf{80.44 \pm 2.19}$ |
| $D \to W$ | $\mathbf{90.04} \pm 2.81$ | $82.02 \pm 0.72$ | $88.11 \pm 2.27$ |
| $W \to A$ | $60.09 \pm 4.77$ | $73.88 \pm 2.83$ | $\mathbf{80.04 \pm 4.11}$ |
| $W \to C$ | $49.34 \pm 8.78$ | $63.17 \pm 4.14$ | $\mathbf{76.97 \pm 1.52}$ |
| $W \to D$ | $\mathbf{95.95} \pm 1.86$ | $90.89 \pm 1.68$ | $93.42 \pm 2.45$ |
| Average | $69.51 \pm 2.70$ | $77.54 \pm 0.66$ | $\mathbf{83.60 \pm 0.22}$ |

Python Optimal Transport (POT) library (`https://github.com/PythonOT/POT`) implementations of OTLin and OTKer in our experiments. For the proposed approach, as in the previous experiments, we chose the Gaussian kernels and have same $\sigma$ values for the kernels over the source and the target domains. The $\sigma$ for our approach was also chosen from the set $\{0.1, 0.5, 1\}$.

**Results**: We observe from Tables 4 & 5 that the proposed approach outperforms the baselines, obtaining the best in-sample and out-of-sample (OOS) accuracy. As discussed, the in-sample accuracy is likely to be better than out-of-sample accuracy (for any approach). Interestingly, for a few tasks with Amazon and Caltech as the source domains, the OOS accuracy of our approach is comparable to our in-sample accuracy. In these domains, the OOS set is larger than the training set. The proposed OOS estimation is able to exploit this and provide an effective knowledge transfer. Conversely, we observe a drop in our OOS accuracy (when compared with the corresponding in-sample accuracy) in tasks with DSLR and Webcam as the source domains since the size of OOS set is quite small and hence lesser potential for knowledge transfer. On the other hand, OTLin suffers a significant drop in OOS performance, likely due the the overfitting of the learned linear transformation on the source training points. While OTKer has better OOS performance than OTLin, it has more variance between in-sample and out-of-sample performance than the proposed approach.

# 6 Conclusions

The idea of employing kernel embeddings of distributions in OT seems promising, especially in the continuous case. It not only leads to sample complexities that are dimension-free, but also provides a new regularization scheme based on MMD distances, which is complementary to existing $\phi$-divergence based regularization.

While the optimal solution of the proposed MMD regularized formulation recovers the transport plan, the objective value does not seem to have any special use. On the contrary, it has been shown that with entropic, $\phi$-divergence based regularizations the optimal objectives lead to notions of Sinkhorn divergences [13] and Hillinger-Kantorovich metrics [24]. We make an initial observation that in the special regularization, $\epsilon_i = 0, \Delta_1 = \Delta_2$, and the Tikhonov regularized form of (4), our optimal objective resembles that defining the Hillinger-Kantorovich metrics very closely. Hence, we conjecture that our optimal objective in this special case may also define a new family of metrics. However, we postpone such connections (if any) to future work.

# A Proof for Theorem 2

*Proof.* Let $\hat{h}$ denote the objective in (4), when written in Tikhonov form, as a function of variables $\mathcal{U} \in \mathcal{E}\left(\mathcal{H}_2, \mathcal{H}_1\right), \mathcal{U}_1 \in \mathcal{L}\left(\mathcal{H}_1, \mathcal{H}_2\right), \mathcal{U}_2 \in \mathcal{L}\left(\mathcal{H}_2, \mathcal{H}_1\right)$ and let $h$ denote that when the true embeddings are employed instead of their estimates. In particular, we have $\hat{h}\left(\hat{\mathcal{U}},\hat{\mathcal{U}}_1,\hat{\mathcal{U}}_2\right) = g\left(\hat{c},\hat{\mu}_s,\hat{\mu}_t,\hat{\Sigma}_{ss},\hat{\Sigma}_{tt}\right), h\left(\mathcal{U}^*,\mathcal{U}_1^*,\mathcal{U}_2^*\right) = g\left(c,\mu_s,\mu_t,\Sigma_{ss},\Sigma_{tt}\right)$, where $\hat{\mathcal{U}},\hat{\mathcal{U}}_1,\hat{\mathcal{U}}_2$ and $\mathcal{U}^*,\mathcal{U}_1^*,\mathcal{U}_2^*$ are optimal solutions to respective problems.

We begin by noting that the feasibility set of (4) is bounded. This is because: i) the set $\mathcal{E}\left(\mathcal{H}_2,\mathcal{H}_1\right)$ is bounded. This is true as $\mathcal{U} \in \mathcal{E}\left(\mathcal{H}_2,\mathcal{H}_1\right) \Rightarrow$ there exists $p \in \mathcal{M}^1(\mathcal{X}\times\mathcal{Y})$ such that $\|\mathcal{U}\| = \|\mathbb{E}_{(X,Y)\sim p}\left[\phi_1(X)\otimes\phi_2(Y)\right]\| \leq \mathbb{E}_{(X,Y)\sim p}\left[\|\phi_1(X)\otimes\phi_2(Y)\|\right] = 1$. The first inequality follows from Jensens inequality and the second equality is true for any bounded kernel like Gaussian and the Kroncker Delta. ii) By triangle inequality, $\left|\|\mathcal{U}\| - \|\hat{\Sigma}_{ss}\mathcal{U}_1\|\right| \leq \left\|\mathcal{U} - \hat{\Sigma}_{ss}\mathcal{U}_1^\top\right\| \leq \epsilon_1$. This shows that the set of all feasible $\hat{\Sigma}_{ss}\mathcal{U}_1$ is bounded, since $\mathcal{U}$ is itself bounded in the feasibility set. Now, since $maxeig(\hat{\Sigma}_{ss}) = maxeig(G_2)/n \leq tr(G_2)/n = 1$ (again true for Kronecker and Gaussian kernels), we obtain that set of all feasible $\mathcal{U}_1$ is also bounded. Similarly, set of all feasible $\mathcal{U}_2$ is bounded. Accordingly, we define $\mathcal{B}\left(\epsilon_1,\epsilon_2\right) \equiv \left\{\left(\mathcal{U}\in\mathcal{E}\left(\mathcal{H}_2,\mathcal{H}_1\right),\mathcal{U}_1\in\mathcal{L}\left(\mathcal{H}_1,\mathcal{H}_2\right),\mathcal{U}_2\in\mathcal{L}\left(\mathcal{H}_2,\mathcal{H}_1\right)\right) \mid \|\mathcal{U}\|\leq 1, \|\mathcal{U}_1\|\leq 1+\epsilon_1, \|\mathcal{U}_2\|\leq 1+\epsilon_2\right\}$. By the above argument, it is clear that there is no loss of generality in further restricting the search space to that with intersection with this bounded set, $\mathcal{B}\left(\epsilon_1,\epsilon_2\right)$, and always $\left(\mathcal{U}^*,\mathcal{U}_1^*,\mathcal{U}_2^*\right), \left(\hat{\mathcal{U}},\hat{\mathcal{U}}_1,\hat{\mathcal{U}}_2\right) \in \mathcal{B}\left(\epsilon_1,\epsilon_2\right)$ for any $m,n\in\mathbb{N}$.

The rest of the proof follows from the claim below:

**Claim 1.** *The uniform bound:*

$$\max_{\left(\mathcal{U},\mathcal{U}_1,\mathcal{U}_2\right)\in\mathcal{B}\left(\epsilon_1,\epsilon_2\right)} \left|\hat{h}\left(\mathcal{U},\mathcal{U}_1,\mathcal{U}_2\right) - h\left(\mathcal{U},\mathcal{U}_1,\mathcal{U}_2\right)\right| \leq O\left(1/\sqrt{\min(m,n)}\right)$$

*holds, where the constants in the RHS are dimension-free.*

*Proof.* Now consider the Tikhonov regularized form of (4). Then, one of the term in the objective is $\|\mathcal{U}_1\hat{\mu}_s - \hat{\mu}_t\| \leq \|\mathcal{U}_1\left(\hat{\mu}_s - \mu_s\right)\| + \|\mu_t - \hat{\mu}_t\| + \|\mathcal{U}_1\mu_s - \mu_t\|$, which is less than $\|\mathcal{U}_1\mu_s - \mu_t\| + O(\frac{1}{\sqrt{p}})$ with high probability. Here, $p = \min(m,n)$. The first inequality is by triangle inequality and the second is the crucial one that follows from sample complexity of kernel mean embeddings (see theorem 2 in [36]), and the boundedness of $\|\mathcal{U}_1\|$. Also, the constants in $O(\frac{1}{\sqrt{p}})$ are independent of samples, variables and dimensions. By symmetry, we also have with high probability that $\|\mathcal{U}_1\mu_s - \mu_t\| \leq \|\mathcal{U}_1\hat{\mu}_s - \hat{\mu}_t\| + O(\frac{1}{\sqrt{p}})$. Hence, with high probability, uniformly over the feasibility set, $|\|\mathcal{U}_1\mu_s - \mu_t\| - \|\mathcal{U}_1\hat{\mu}_s - \hat{\mu}_t\|| \leq O(\frac{1}{\sqrt{p}})$. Analogous arguments hold for the other quadratic terms too. Now, we analyze the linear objective term. By Jensen's inequality, $|\langle\hat{c},\mathcal{U}\rangle - \langle c,\mathcal{U}\rangle| \leq \sqrt{\mathbb{E}_u\left[(\hat{c}-c)^2\right]}$, where $u$ is the measure corresponding to $\mathcal{U}$. Let $\bar{u}$ denote the product measure of the given marginals. It is easy to see that $\{(x_i,y_j) \mid i\in 1,\dots,m, j\in 1,\dots,n\}$ is an iid set of samples from $\bar{u}$. By eqn. (4) in theorem 3.1 in [31] and lemma 1 in [32], we have $\sqrt{\mathbb{E}_u\left[(\hat{c}-c)^2\right]} \leq \frac{\|c\|_{\bar{u}}}{\sqrt{mn}}\left(1+\sqrt{2\log(\frac{1}{\delta})}\right)$, with probability $\delta$. Here, $\|\cdot\|_{\bar{u}}$ is same as that defined in section III of [31], and theorem 3.1 in [31] applies to our case as we assumed normalized kernels. In particular, this bound is independent of dimensions and $\mathcal{U}$. To summarize, we have, $|\langle\hat{c},\mathcal{U}\rangle - \langle c,\mathcal{U}\rangle| \leq O(\frac{1}{\sqrt{mn}})$. Finally, again by triangle inequality, $\left|\hat{h}\left(\mathcal{U},\mathcal{U}_1,\mathcal{U}_2\right) - h\left(\mathcal{U},\mathcal{U}_1,\mathcal{U}_2\right)\right|$ is less than the sum of deviations in each of the terms detailed above. Since each of these deviations is upper bounded uniformly by $O\left(\frac{1}{\sqrt{p}}\right)$, the claim is proved. □

The proof of the theorem then follows from standard arguments: $\hat{h}\left(\hat{\mathcal{U}},\hat{\mathcal{U}}_1,\hat{\mathcal{U}}_2\right) - h\left(\mathcal{U}^*,\mathcal{U}_1^*,\mathcal{U}_2^*\right) \leq h\left(\hat{\mathcal{U}},\hat{\mathcal{U}}_1,\hat{\mathcal{U}}_2\right) - h\left(\mathcal{U}^*,\mathcal{U}_1^*,\mathcal{U}_2^*\right) + \max_{\left(\mathcal{U},\mathcal{U}_1,\mathcal{U}_2\right)\in\mathcal{B}\left(\epsilon_1,\epsilon_2\right)} \hat{h}\left(\mathcal{U},\mathcal{U}_1,\mathcal{U}_2\right) - h\left(\mathcal{U},\mathcal{U}_1,\mathcal{U}_2\right) \leq h\left(\hat{\mathcal{U}},\hat{\mathcal{U}}_1,\hat{\mathcal{U}}_2\right) - h\left(\mathcal{U}^*,\mathcal{U}_1^*,\mathcal{U}_2^*\right) + O\left(1/\sqrt{\min(m,n)}\right)$ by the claim. Now, the estimation error, $h\left(\hat{\mathcal{U}},\hat{\mathcal{U}}_1,\hat{\mathcal{U}}_2\right) - h\left(\mathcal{U}^*,\mathcal{U}_1^*,\mathcal{U}_2^*\right)$, which is non-negative, is equal to $\left(\hat{h}\left(\hat{\mathcal{U}},\hat{\mathcal{U}}_1,\hat{\mathcal{U}}_2\right) - \hat{h}\left(\mathcal{U}^*,\mathcal{U}_1^*,\mathcal{U}_2^*\right)\right) + \left(h\left(\hat{\mathcal{U}},\hat{\mathcal{U}}_1,\hat{\mathcal{U}}_2\right) - \hat{h}\left(\hat{\mathcal{U}},\hat{\mathcal{U}}_1,\hat{\mathcal{U}}_2\right)\right) + \left(\hat{h}\left(\mathcal{U}^*,\mathcal{U}_1^*,\mathcal{U}_2^*\right) - h\left(\mathcal{U}^*,\mathcal{U}_1^*,\mathcal{U}_2^*\right)\right) \leq O\left(1/\sqrt{\min(m,n)}\right)$. The last inequality follows from the claim and the definition of $\left(\hat{\mathcal{U}},\hat{\mathcal{U}}_1,\hat{\mathcal{U}}_2\right)$ that it minimizes $\hat{h}$. Analogous arguments give $h\left(\mathcal{U}^*,\mathcal{U}_1^*,\mathcal{U}_2^*\right) - \hat{h}\left(\hat{\mathcal{U}},\hat{\mathcal{U}}_1,\hat{\mathcal{U}}_2\right) \leq O\left(1/\sqrt{\min(m,n)}\right)$. This not only completes the proof but also shows that the estimation error also decays with rate that is dimension-free. □

# B Proof of representer theorem

*Proof.* Without loss of generality, we consider the parameterization: $\mathcal{U}^\alpha = \sum_{i=1}^m \sum_{j=1}^n \alpha_{ij}\phi_1(x_i) \otimes \phi_2(y_j) + \mathcal{U}^\perp, \mathcal{U}_1^\beta = \sum_{i=1}^m \sum_{j=1}^n \beta_{ji}\phi_2(y_j) \otimes \phi_1(x_i) + \mathcal{U}_1^\perp, \mathcal{U}_2^\gamma = \sum_{i=1}^m \sum_{j=1}^n \gamma_{ij}\phi_1(x_i) \otimes \phi_2(y_j) + \mathcal{U}_2^\perp$, where $\mathcal{U}^\perp, \mathcal{U}_1^\perp, \mathcal{U}_2^\perp$ are the respective orthogonal complements. It is easy to see that the objective as well as the first two inequalities in (4) do not involve the orthogonal complements. Also the term $\left\|\mathcal{U} - \hat{\Sigma}_{ss}\mathcal{U}_1^\top\right\|^2_{\mathcal{H}_2\otimes\mathcal{H}_1}$ can be written as sum of a term not involving the orthogonal complements and $\|\mathcal{U}^\perp - \hat{\Sigma}_{ss}\left(\mathcal{U}_1^\perp\right)^\top\|^2_{\mathcal{H}_2\otimes\mathcal{H}_1}$. Like-wise $\left\|\mathcal{U} - \mathcal{U}_2\hat{\Sigma}_{tt}\right\|^2_{\mathcal{H}_2\otimes\mathcal{H}_1}$ can be written as sum of a term not involving the orthogonal complements as $\|\mathcal{U}^\perp - \mathcal{U}_2^\perp\hat{\Sigma}_{tt}\|^2_{\mathcal{H}_2\otimes\mathcal{H}_1}$.

Now re-writing (4), where all the norm constraints are equivalently replaced by the norm-squared constraints, in Tikhonov regularization form reads as: $\min_{f\in\mathcal{S}\subset\mathcal{H}} \hat{\mathcal{R}}[f] + \Omega[f]$, where $f = (\mathcal{U},\mathcal{U}_1,\mathcal{U}_2), \mathcal{H} = (\mathcal{H}_2 \otimes \mathcal{H}_1) \oplus (\mathcal{H}_1 \otimes \mathcal{H}_2) \oplus (\mathcal{H}_2 \otimes \mathcal{H}_1), \mathcal{S} = \mathcal{E}(\mathcal{H}_2,\mathcal{H}_1) \times \mathcal{L}(\mathcal{H}_1,\mathcal{H}_2) \times \mathcal{L}(\mathcal{H}_2,\mathcal{H}_1), \Omega[f] \equiv \|\mathcal{U}^\perp - \hat{\Sigma}_{ss}\left(\mathcal{U}_1^\perp\right)^\top\|^2_{\mathcal{H}_2\otimes\mathcal{H}_1} + \|\mathcal{U}^\perp - \mathcal{U}_2^\perp\hat{\Sigma}_{tt}\|^2_{\mathcal{H}_2\otimes\mathcal{H}_1}$ and $\hat{\mathcal{R}}[f]$ is the remaining objective that does not involve the orthogonal complements. Also, let $\hat{\mathcal{S}} \subset \mathcal{S}$ denote $\left\{f = (\mathcal{U},\mathcal{U}_1,\mathcal{U}_2) \in \mathcal{S} \mid \mathcal{U}^\perp = 0, \mathcal{U}_1^\perp = 0, \mathcal{U}_2^\perp = 0\right\}$ and let $\Pi_{\hat{\mathcal{S}}}$ denote the projection onto $\hat{\mathcal{S}}$. Now, for any $f \in \mathcal{H}$, we have that: $\hat{\mathcal{R}}[\Pi_{\hat{\mathcal{S}}}(f)] = \hat{\mathcal{R}}[f]$ and more importantly, $0 = \Omega[\Pi_{\hat{\mathcal{S}}}(f)] \leq \Omega[f]$. Consider the following argument[1]: $\min_{f\in\mathcal{S}\subset\mathcal{H}} \hat{\mathcal{R}}[f] + \Omega[f] \leq \min_{f\in\hat{\mathcal{S}}\subset\mathcal{H}} \hat{\mathcal{R}}[f] + \Omega[f] = \min_{f\in\mathcal{S}\subset\mathcal{H}} \hat{\mathcal{R}}[\Pi_{\hat{\mathcal{S}}}(f)] + \Omega[\Pi_{\hat{\mathcal{S}}}(f)] \leq \min_{f\in\mathcal{S}\subset\mathcal{H}} \hat{\mathcal{R}}[f] + \Omega[f]$. This proves that the orthogonal complements are all zero at optimality.

Now, let $\mathcal{L} \equiv \{\mathcal{U}^\alpha \mid \alpha \in \mathbb{R}^{m\times n}\}$, $\mathcal{P} \equiv \left\{\mathcal{U}^\alpha \mid \alpha \in \mathbb{R}^{m\times n}, \alpha \geq 0, \mathbf{1}^\top\alpha = 1\right\}$ and $\mathcal{A} \equiv \left\{\sum_{i=1}^{m'}\sum_{j=1}^{n'} \alpha_{ij}\phi_1(x_i') \otimes \phi_2(y_j') \mid \alpha \in \mathbb{R}^{m'\times n'}, \alpha \geq 0, \mathbf{1}^\top\alpha = 1, x_i' \in \mathcal{X}, y_j' \in \mathcal{Y}, m', n' \in \mathbb{N}\right\}$. Then, the only thing left to be shown is that $\mathcal{E}(\mathcal{H}_2,\mathcal{H}_1) \cap \mathcal{L} = \mathcal{P}$. While $\mathcal{P} \subseteq \mathcal{E}(\mathcal{H}_2,\mathcal{H}_1) \cap \mathcal{L}$ is trivial. The converse is true because of the following facts:

1. $\mathcal{E}(\mathcal{H}_2,\mathcal{H}_1) = cl(\mathcal{A})$, where $cl$ denotes the set closure. While $cl(\mathcal{A}) \subseteq \mathcal{E}(\mathcal{H}_2,\mathcal{H}_1)$ is trivial, the converse follows from the convergence of average of sample embeddings to the true embedding (see theorem 2 in [36]).
2. if $\mathcal{U} \in \mathcal{L}\backslash\mathcal{P}$, then $\mathcal{U} \notin \mathcal{A}$. This is because the expansion of embeddings in RKHS of a universal kernel are unique. Also, $\mathcal{U} \in \mathcal{L}, \mathcal{U} \notin \mathcal{A} \Rightarrow \mathcal{U} \notin cl(\mathcal{A})$.

□

# C Proof of Theorem 4

*Proof.* Firstly, by theorem 2, for low enough hyper-parameters we know that the conditional operator obtained by solving (5) are consistent with dimension-free sample complexity. Hence the Barycentric-projection problem is nothing but a stochastic optimization problem with samples as $y_j$ with likelihood $\sum_{j=1}^n \left(\beta_{ji}^* k_1(x_i, x)\right)$. Using sampling with replacement, these can be converted to $m'$ iid samples with uniform likelihood. Since the cost is assumed to be a metric or it's power greater than unity, the stochastic optimization problem is infact convex wrt $y$. Since the domain $\mathcal{Y}$ is bounded, it is

[1] See also [3] for similar a argument.

also Lipschitz continuous wrt. $y$. Hence by (7), theorem 3 in [35], the estimation error in optimal transport map when solved by SGD is $O(1/\sqrt{m'})$ and remains dimension-free. $\square$

# References


[1] Jason Altschuler, Francis Bach, Alessandro Rudi, and Jonathan Niles-Weed. Massively scalable sinkhorn distances via the nyström method. In *Advances in Neural Information Processing Systems 32*, pages 4429–4439, 2019.

[2] David Alvarez-Melis and Tommi Jaakkola. Gromov-Wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890, 2018.

[3] J. Andrew (Drew) Bagnell and Amir massoud Farahmand. Learning positive functions in a hilbert space, December 2015.

[4] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[5] Guillaume Carlier, Marco Cuturi, Brendan Pass, and Carola Schoenlieb. Optimal transport meets probability, statistics and machine learning. `https://www.birs.ca/cmo-workshops/2017/17w5093/report17w5093.pdf`, May 2017.

[6] Ruidi Chen and Ioannis Ch. Paschalidis. A robust learning approach for regression models based on distributionally robust optimization. *Journal of Machine Learning Research*, 19(1):517–564, 2018.

[7] Andreas Christmann and Ingo Steinwart. Universal kernels on non-standard input spaces. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1*, page 406–414, 2010.

[8] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2017.

[9] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *Advances in Neural Information Processing Systems*, volume 30, pages 3730–3739, 2017.

[10] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, page 2292–2300, 2013.

[11] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. DeCAF: A deep convolutional activation feature for generic visual recognition. In *ICML*, ICML'14, 2014.

[12] R. M. Dudley. The speed of mean glivenko-cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

[13] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, 2018.

[14] Aden Forrow, Jan-Christian Hütter, Mor Nitzan, Philippe Rigollet, Geoffrey Schiebinger, and Jonathan Weed. Statistical optimal transport via factored couplings. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, pages 2454–2465, 2019.

[15] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya-Polo, and Tomaso Poggio. Learning with a wasserstein loss. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, pages 2053–2061, 2015.

[16] Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, page 489–496, 2007.

[17] 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, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, pages 1574–1583, 2019.

[18] Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, page 3440–3448, 2016.

[19] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, 2017.

[20] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2012.

[21] Motonobu Kanagawa and Kenji Fukumizu. Recovering distributions from gaussian rkhs embeddings. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2014*, 2014.

[22] Leonid Kantorovich. On the transfer of masses (in russian). *Doklady Akademii Nauk*, 37(2):227–229, 1942.

[23] Hermann Karcher. Riemannian center of mass and so called karcher mean. *CORR*, 2014.

[24] Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, 2018.

[25] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

[26] Boris Muzellec, Richard Nock, Giorgio Patrini, and Frank Nielsen. Tsallis regularized optimal transport and ecological inference. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, page 2387–2393, 2017.

[27] Jung Hun Oh, Maryam Pouryahya, Aditi Iyer, Aditya P. Apte, Allen Tannenbaum, and Joseph O. Deasy. Kernel wasserstein distance, 2019.

[28] M. Perrot, N. Courty, R. Flamary, and A. Habrard. Mapping estimation for discrete optimal transport. In *NeurIPS*, 2016.

[29] Michaël Perrot, Nicolas Courty, Rémi Flamary, and Amaury Habrard. Mapping estimation for discrete optimal transport. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, page 4204–4212, 2016.

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

[31] Ali Rahimi and Benjamin Recht. Uniform approximation of functions with random bases. pages 555 – 561, 10 2008.

[32] Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, page 1313–1320, 2008.

[33] Ingmar Schuster, Mattes Mollenhauer, Stefan Klus, and Krikamol Muandet. Kernel conditional density operators. *CoRR*, abs/1905.11255, 2019.

[34] Vivien. Seguy, Bharath B. Damodaran, Remi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. Large-scale optimal transport and mapping estimation. In *International Conference on Learning Representations (ICLR)*, 2018.

[35] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *COLT*, 2009.

[36] Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A hilbert space embedding for distributions. In *Proceedings of the 18th International Conference on Algorithmic Learning Theory*, page 13–31, 2007.

[37] Le Song, Jonathan Huang, Alex Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, page 961–968, 2009.

[38] Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R. G. Lanckriet. On integral probability metrics, phi-divergences and binary classification. *arXiv preprint arXiv:0901.2698*, 2009.

[39] Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R. G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12:2389–2410, 2011.

[40] 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*, pages 1723–1732, 2018.

[41] Mikhail Yurochkin, Sebastian Claici, Edward Chien, Farzaneh Mirzazadeh, and Justin M Solomon. Hierarchical optimal transport for document representation. In *Advances in Neural Information Processing Systems*, pages 1599–1609, 2019.

[42] Z. Zhang, M. Wang, and A. Nehorai. Optimal transport in reproducing kernel hilbert spaces: Theory and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.