Cycle Consistent Probability Divergences Across Different Spaces

Zhengxin Zhang
Cornell University

Youssef Mroueh
IBM Research AI

Ziv Goldfeld
Cornell University

Bharath K. Sriperumbudur
Pennsylvania State University

Abstract

Discrepancy measures between probability distributions are at the core of statistical inference and machine learning. In many applications, distributions of interest are supported on different spaces, and yet a meaningful correspondence between data points is desired. Motivated to explicitly encode consistent bidirectional maps into the discrepancy measure, this work proposes a novel unbalanced Monge optimal transport formulation for matching, up to isometries, distributions on different spaces. Our formulation arises as a principled relaxation of the Gromov-Haussdroff distance between metric spaces, and employs two cycle-consistent maps that push forward each distribution onto the other. We study structural properties of the proposed discrepancy and, in particular, show that it captures the popular cycle-consistent generative adversarial network (GAN) framework as a special case, thereby providing the theory to explain it. Motivated by computational efficiency, we then kernelize the discrepancy and restrict the mappings to parametric function classes. The resulting kernelized version is coined the *generalized maximum mean discrepancy* (GMMD). Convergence rates for empirical estimation of GMMD are studied and experiments to support our theory are provided.

1 INTRODUCTION

Discrepancy measures between probability distributions are ubiquitous in machine learning. In practice, distributions of interests are often supported on different spaces and the goal is not only to quantify discrepancy, but also to obtain a meaningful and consistent correspondence between data points. Such problems arise, e.g., in natural language processing for unsupervised matching across different languages or ontologies (Alvarez-Melis and Jaakkola, 2018; Grave et al., 2019; Alvarez-Melis et al., 2020; Le et al., 2021b), shape matching (Bronstein et al., 2006a; Mémoli, 2011; Xu et al., 2019), heterogenous domain adaptation (Yan et al., 2018), generative modeling (Bunne et al., 2019), and many more.

Among the most popular discrepancies between distributions on incompatible spaces is the Gromov-Wasserstein distance (GW) (Mémoli, 2011) (see Séjourné et al. (2020) for an unbalanced variant). Computationally, the GW distance amounts to a quadratic assignment problem that is NP hard (Commander, 2005). To alleviate this impasse, Peyré et al. (2016) proposed an entropic regularization to the GW problem, and derived an algorithm with cubic $O(n^3)$ complexity in the number of samples (see Le et al. (2021a) for recent theoretical advances). More recently, Vayer et al. (2019) proposed slicing the GW distance, which further reduces the computational complexity to $O(n \log n)$. Despite these algorithmic advances, a common issue with GW-based discrepancies is their lack of generalization to new data points. These approaches only quantify the distance without generating a map that captures the correspondence. This requires recomputing the distance whenever one wants to account for new data points, thereby incurring an additional cost. This shortcoming motivate us to explore computationally friendly discrepancies between distributions on different spaces that explicitly encode consistent, bidirectional measure preserving mappings that capture the correspondence. This is similar in spirit to the recent interest in learning Monge optimal transport (OT) maps (Perrot et al., 2016; Makkuva et al., 2020; Paty et al., 2020; Flamary et al., 2019).

Specifically, we propose a novel unbalanced divergence between probability measures supported on different spaces that explicitly employs two cycle-consistent

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

maps that (approximately) push each distribution onto another. Cycle-consistency here is in the context of cycle generative adversarial networks (Zhu et al., 2017; Kim et al., 2017) (GANs), which requires that the two pushforward maps are roughly inverses one of another. Note that Mémoli and Needham (2021) recently proposed a quasi-metric called the Gromov-Monge distance that employs a single push forward map. A key advantage of our approach is its cycle consistency that provides a mathematical framework for the popular cycle GAN and enables a principled study thereof.

The main contributions of this paper are:

- We introduce and study in Section 3 the unbalanced bidirectional Gromov-Monge (UBGM) divergence, drawing connections between UBGM and the popular cycle GAN framework.
- Motivated by computational efficiency, we kernelize UBGM in Section 4 and restrict mappings to parametric function classes, such as neural networks (NNs). We call the resulting divergence the generalized maximum mean discrepancy (GMMD). We then derive convergence rates for two-sample empirical estimation of GMMD.
- We present numerical results in Section 5 that support our theory and demonstrate the computational efficiency, and generalization capability of the proposed framework for matching across different spaces.

2 BACKGROUND AND PRELIMINARIES

2.1 Notations

Let $(\mathcal{X}, d_\mathcal{X})$ be a compact metric space. The diameter of a set $A \subseteq \mathcal{X}$ is $\mathsf{diam}(A) := \sup_{x,x' \in A} d_\mathcal{X}(x,x')$. We use $B(x,r)$ to denote the open ball of radius $r > 0$ centered at $x \in \mathcal{X}$. For $\epsilon > 0$, a set $\mathcal{X}_\epsilon$ is called an ϵ-cover of $\mathcal{X}$ if for any $x \in \mathcal{X}$, $\inf_{x' \in \mathcal{X}_\epsilon} d_\mathcal{X}(x,x') < \epsilon$. The ϵ-covering number of $\mathcal{X}$ is $N(\mathcal{X}, d_\mathcal{X}, \epsilon) := \inf\{|\mathcal{X}_\epsilon| : \mathcal{X}_\epsilon \text{ is an } \epsilon\text{-cover of } \mathcal{X}\}$. When $\mathcal{X}$ is a subset of $\mathbb{R}^d$, we always use the metric induce by the Euclidean norm, denoted as $\|\cdot\|$. For a metric space $(\mathcal{X}, d_\mathcal{X})$, the diameter of $\mathcal{X}$ is defined as $\sup_{x,x' \in \mathcal{X}} d_\mathcal{X}(x,x')$.

For $1 \le p < \infty$, let $L^p(\mathcal{X}, \rho)$ denote the space of measurable maps $f : \mathcal{X} \to \mathbb{R}$ such that $\|f\|_{L^p(\mathcal{X},\rho)} := (\int_\mathcal{X} |f|^p \, \mathrm{d}\rho)^{1/p} < \infty$; we use the shorthand $L^p(\mathcal{X})$, $L^p(\rho)$, and L^p whenever the omitted object is clear from the context. The standard extension of these spaces to $p = \infty$ is denoted by $L^\infty(\mathcal{X})$. The Lipschitz constant of a function $f : \mathcal{X} \to \mathcal{Y}$ is $\|f\|_{\mathsf{Lip}} = \sup_{x,x' \in \mathcal{X}} \frac{d_\mathcal{Y}(f(x), f(x'))}{d_\mathcal{X}(x,x')}$, with $\mathsf{Lip}_L(\mathcal{X}, \mathcal{Y}) = \{f : \|f\|_{\mathsf{Lip}} \le L\}$ denoting the Lipschitz ball of radius $L > 0$. A mapping $f : \mathcal{X} \to \mathcal{Y}$ between metric spaces is called an isometry if $d_\mathcal{X}(x,x') = d_\mathcal{Y}\big(f(x), f(x')\big)$, for all $x, x' \in \mathcal{X}$, i.e., f preserves the metric structure. For a class of mappings from $\mathcal{X}$ to $\mathcal{Y}$, define the sup-metric on $\mathcal{F}$ as $d_\mathcal{F}\big(f_1, f_2\big) := \sup_{x \in \mathcal{X}} d_\mathcal{Y}\big(f_1(x), f_2(x)\big)$.

The probability space on which all random variables are defined is denoted by $(\Omega, \mathcal{A}, \mathbb{P})$ (assumed to be sufficiently rich), with $\mathbb{E}$ designating the corresponding expectation. The class of Borel probability measures over $\mathcal{X}$ is denoted by $\mathcal{P}(\mathcal{X})$. For $n \in \mathbb{N}$, $P^{\otimes n}$ denotes the n-fold product measure of P. Given a measurable $f : \mathcal{X} \to \mathcal{Y}$ and $P \in \mathcal{P}(\mathcal{X})$, the pushforward of P through f is $f_\sharp P(B) := P(f^{-1}(B))$, for any Borel set B. Clearly $f_\sharp P \in \mathcal{P}(\mathcal{Y})$. We also write $a \lesssim_x b$ when $a \le C_x b$, where C_x is a constant depending only on x, and write $a \lesssim b$ if the omitted constant is universal. Also denote $n \wedge m = \min\{n, m\}$.

2.2 Gromov-Haussdroff Distance

To motivate our proposed discrepancy measure, we start by recalling the Gromov-Hausdorff distance between metric spaces, which is defined as

$$d_{\mathsf{GH}}(\mathcal{X}, \mathcal{Y}) := \inf_{\mathcal{Z}, \phi_\mathcal{X}, \phi_\mathcal{Y}} d_{\mathsf{H}}^{\mathcal{Z}}\big(\phi_\mathcal{X}(\mathcal{X}), \phi_\mathcal{Y}(\mathcal{Y})\big), \tag{1}$$

where the infimum is on an ambient metric space $(\mathcal{Z}, d_\mathcal{Z})$ and isometric embeddings $\phi_\mathcal{X} : \mathcal{X} \to \mathcal{Z}$ and $\phi_\mathcal{Y} : \mathcal{Y} \to \mathcal{Z}$, with $d_{\mathsf{H}}^{\mathcal{Z}}$ as the Hausdroff distance on $\mathcal{Z}$.

Evidently, the formulation above is not computable. Nevertheless, d_{GH} has several equivalent forms (Mémoli, 2011) that, with appropriate relaxations, can be computed efficiently. Two such forms are as follows.

Correspondence set reformulation. Following Mémoli and Needham (2021), a correspondence set between $\mathcal{X}$ and $\mathcal{Y}$ is a set $R \subset \mathcal{X} \times \mathcal{Y}$ whose projections to $\mathcal{X}$ and $\mathcal{Y}$ define surjections on R. The set of all such correspondences is denoted by $\mathcal{R}(\mathcal{X}, \mathcal{Y})$. The Gromov-Hausdorff distance can be reformulated as

$$d_{\mathsf{GH}}(\mathcal{X}, \mathcal{Y}) = \frac{1}{2} \inf_{R \in \mathcal{R}(\mathcal{X}, \mathcal{Y})} \sup_{(x,y),(x',y') \in R} \Gamma_{\mathcal{X},\mathcal{Y}}(x,y,x',y'), \tag{2}$$

where $\Gamma_{\mathcal{X},\mathcal{Y}}(x,y,x',y') := \big|d_\mathcal{X}(x,x') - d_\mathcal{Y}(y,y')\big|$ is the pointwise distortion between $(x,x') \in \mathcal{X}^2$ and $(y,y') \in \mathcal{Y}^2$. Note that (2) can be written in the following compact form as an L^∞ norm:

$$d_{\mathsf{GH}}(\mathcal{X}, \mathcal{Y}) = \frac{1}{2} \inf_{R \in \mathcal{R}(\mathcal{X}, \mathcal{Y})} \|\Gamma_{\mathcal{X},\mathcal{Y}}\|_{L^\infty(R \times R)} . \tag{3}$$

Two mappings reformulation. Another important reformulation of the Gromov-Hausdroff distance in terms of mappings between spaces is:

$$d_{\mathsf{GH}}(\mathcal{X},\mathcal{Y})=\frac{1}{2}\inf_{\substack{f:\mathcal{X}\to\mathcal{Y}\\ g:\mathcal{Y}\to\mathcal{X}}}\max\{\Delta^{\infty}_{\mathcal{X}}(f),\Delta^{\infty}_{\mathcal{Y}}(g),\Delta^{\infty}_{\mathcal{X},\mathcal{Y}}(f,g)\}, \tag{4}$$

where the distortions $\Delta^{\infty}_{\mathcal{X}}$, $\Delta^{\infty}_{\mathcal{Y}}$, and $\Delta^{\infty}_{\mathcal{X},\mathcal{Y}}$ are given by[1]

$$\begin{aligned}\Delta^{\infty}_{\mathcal{X}}(f) &:= \sup_{x,x'\in\mathcal{X}}\left|d_{\mathcal{X}}(x,x')-d_{\mathcal{Y}}\big(f(x),f(x')\big)\right|\\ \Delta^{\infty}_{\mathcal{Y}}(g) &:= \sup_{y,y'\in\mathcal{Y}}\left|d_{\mathcal{X}}\big(g(y),g(y')\big)-d_{\mathcal{Y}}(y,y')\right|\\ \Delta^{\infty}_{\mathcal{X},\mathcal{Y}}(f,g) &:= \sup_{x\in\mathcal{X},y\in\mathcal{Y}}\left|d_{\mathcal{X}}\big(x,g(y)\big)-d_{\mathcal{Y}}\big(f(x),y\big)\right|.\end{aligned}$$

Formulation (4) thus measures distance by searching for low distortion maps between the two metric spaces, such that the so-called *cycle consistency* property holds, i.e., the maps are approximate inverses of one another. More specifically following Mémoli and Sapiro (2005), if $d_{\mathsf{GH}}(\mathcal{X},\mathcal{Y}) \le \epsilon$ then there exists $f:\mathcal{X}\to\mathcal{Y}$ and $g:\mathcal{Y}\to\mathcal{X}$ such that: (1) the induced metric distortions are small, i.e., $\Delta^{\infty}_{\mathcal{X}}(f)\le 2\epsilon$ and $\Delta^{\infty}_{\mathcal{Y}}(g)\le 2\epsilon$; and (2) these functions are *almost* inverses one of another, in the sense that $d_{\mathcal{X}}\big(x,g(f(x))\big)\le 2\epsilon$ and $d_{\mathcal{Y}}\big(f(g(y)),y\big)\le 2\epsilon$ (which follows from $\left|d_{\mathcal{X}}(x,g(y))-d_{\mathcal{Y}}(f(x),y)\right|\le 2\epsilon$ by taking $y=f(x)$ and $x=g(y)$, respectively). Thus, $\Delta^{\infty}_{\mathcal{X},\mathcal{Y}}$ ensures cycle consistency of the maps.

3 PROBABILITY DIVERGENCES ACROSS METRIC MEASURE SPACES

We are now ready to present the proposed discrepancy between probability measures on different spaces. In conjunction with the preceding discussion, such a discrepancy can equivalently be viewed as a distance between metric measure (mm) spaces (Mémoli, 2011).

Definition 1 (Metric measure spaces)**.** *A metric measure space is a triple $(\mathcal{X},d_{\mathcal{X}},P)$ where $(\mathcal{X},d_{\mathcal{X}})$ is a compact metric space and $P\in\mathcal{P}(\mathcal{X})$ has full support.*

3.1 The Unbalanced Bidirectional Gromov-Monge Divergence

Formulation (4) of d_{GH} is computationally appealing and has lead to many algorithmic relaxations that approximate the Gromov-Hausdroff distance (Mémoli and Sapiro, 2004, 2005; Bronstein et al., 2006b). As a stepping stone towards our unbalanced formulation, we first adapt this formulation to an extended metric between two mm spaces $(\mathcal{X},d_{\mathcal{X}},P)$ and $(\mathcal{Y},d_{\mathcal{Y}},Q)$. To that end, we first relax the distortion terms and then restrict the mappings to be measure persevering, as described next.

Let $(X,X')\sim P^{\otimes 2}$ be independent of $(Y,Y')\sim Q^{\otimes 2}$. For $1\le p<\infty$, set

$$\begin{aligned}\Delta^{(p)}_{\mathcal{X}}(f;P) &:= \Big(\mathbb{E}\Big[\left|d_{\mathcal{X}}(X,X')-d_{\mathcal{Y}}\big(f(X),f(X')\big)\right|^p\Big]\Big)^{\frac{1}{p}}\\ \Delta^{(p)}_{\mathcal{Y}}(g;Q) &:= \Big(\mathbb{E}\Big[\left|d_{\mathcal{X}}\big(g(Y),g(Y')\big)-d_{\mathcal{Y}}(Y,Y')\right|^p\Big]\Big)^{\frac{1}{p}}\\ \Delta^{(p)}_{\mathcal{X},\mathcal{Y}}(f,g;P,Q) &:= \Big(\mathbb{E}\Big[\left|d_{\mathcal{X}}\big(X,g(Y)\big)-d_{\mathcal{Y}}\big(f(X),Y\big)\right|^p\Big]\Big)^{\frac{1}{p}},\end{aligned}$$

as the L^p relaxation of the distortion terms from (4). Restricting the mappings in (4) to be 'Monge' measure preserving, i.e., $f_\sharp P=Q$ and $g_\sharp Q=P$, gives rise to pth order *bidirectional Gromov-Monge* (BGM) distance, as defined next.

Definition 2 (Bidirectional Gromov-Monge distance)**.** *Fix $1\le p<\infty$. The pth order BGM distance between $(\mathcal{X},d_{\mathcal{X}},P)$ and $(\mathcal{Y},d_{\mathcal{Y}},Q)$ is*

$$\mathsf{D}_p(P,Q) := \inf_{\substack{f:\mathcal{X}\to\mathcal{Y},\, f_\sharp P=Q\\ g:\mathcal{Y}\to\mathcal{X},\, g_\sharp Q=P}}\Delta_p(f,g;P,Q),$$

where $\Delta_p(f,g;P,Q) := \Delta^{(p)}_{\mathcal{X}}(f;P)+\Delta^{(p)}_{\mathcal{Y}}(g;Q)+\Delta^{(p)}_{\mathcal{X},\mathcal{Y}}(f,g;P,Q)$. If the set of measure preserving maps is empty, then we set $\mathsf{D}_p(P,Q)=\infty$.

This definition follows a reasoning similar to Formulation (4) of d_{GH} above: We are looking for measure preserving maps f,g, that are low metric distortion (minimizing $\Delta^{(p)}_{\mathcal{X}},\Delta^{(p)}_{\mathcal{Y}}$) and satisfy a cycle consistency propriety (minimizing $\Delta^{(p)}_{\mathcal{X}\times\mathcal{Y}}$, i.e., almost inverses one of another).

The BGM distance defines an extended metric between equivalence classes of mm spaces.

Definition 3 (Equivalence classes)**.** *Two mm spaces $(\mathcal{X},d_{\mathcal{X}},P)$ and $(\mathcal{Y},d_{\mathcal{Y}},Q)$ are called equivalent if and only if (iff) there is an invertible isometry $f:\mathcal{X}\to\mathcal{Y}$ such that $f_\sharp P=Q$ (and hence $f^{-1}_\sharp Q=P$). The set of all such equivalence classes is denoted by $\mathfrak{M}$.*

Proposition 1 (BGM distance metrizes $\mathfrak{M}$)**.** *For any $1\le p<\infty$, D_p defines an extended metric on $\mathfrak{M}$.*

Remark 1 (Finiteness of BGM)**.** *Let $(\mathcal{X},d_{\mathcal{X}},P)$ be a mm space, with $\mathcal{X}$ uncountable and P atomless. If C_∞ is the subcategory of mm spaces isomorphic to $(\mathcal{X},d_{\mathcal{X}},P)$ (Mémoli and Needham, 2021), then BGM is a finite metric on C_∞. This setting is interesting for matching isomorphic mm spaces (of same dimension), in image, shape and text matching applications.*

The proof of Proposition 1 is given in Supplement A.1. Positivity, symmetry, and the triangle inequality all

[1] The superscript ∞ indicates that $\Delta^{\infty}_{\mathcal{X}}$, $\Delta^{\infty}_{\mathcal{Y}}$, and $\Delta^{\infty}_{\mathcal{X},\mathcal{Y}}$ can be written as L^∞ norms of the appropriate pointwise distortions (e.g., $\Gamma_{\mathcal{X},\mathcal{Y}}\big(x,f(x),x',f(x')\big)$ for $\Delta^{\infty}_{\mathcal{X}}$).

follow from elementary calculations. The main challenge is in showing that if D_p nullifies then the considered spaces are isometrically isomorphic. To that end, we use the following lemma, that may be of independent interest (see Supplement A.2 for the proof).

Lemma 1 (Existence of isometries)**.** *Fix $P, Q \in \mathcal{P}(\mathcal{X})$ and let $\mathcal{F}$ and $\mathcal{G}$ be arbitrary function classes such that $\inf_{f\in\mathcal{F},\, g\in\mathcal{G}} \Delta_p(f,g;P,Q) = 0$. Then there exist minimizing sequences $(f_n)_{n\in\mathbb{N}} \subset \mathcal{F}$ and $(g_n)_{n\in\mathbb{N}} \subset \mathcal{G}$ that converge almost surely (a.s.) to isometries f and g, respectively, such that $f = g^{-1}$.*

Remark 2 (Convergent sequences)**.** *Notice that the convergence stated in the lemma is guaranteed solely by the first two terms of $\Delta_p(f,g;P,Q)$. If we drop the third term, convergent sequences still exists, but the limits are not guaranteed to be inverse of each other.*

While D_p is a valid extended metric on $\mathfrak{M}$, its evaluation is computationally challenging as it is unclear how to optimize over bidirectional Monge maps. Following the unbalanced OT framework (Chizat et al., 2018; Frogner et al., 2015), we relax the measure preserving constraint using divergences[2] $\mathsf{D}_\mathcal{X}$ and $\mathsf{D}_\mathcal{Y}$ on measures on $\mathcal{X}$ and $\mathcal{Y}$, respectively, and restrict the functions to pre-specified classes. This gives rise to the unbalanced Gromov-Monge formulation.

Definition 4 (Unbalanced bidirectional Gromov–Monge divergence)**.** *Fix $1 \le p < \infty$, let $\mathsf{D}_\mathcal{X}$ and $\mathsf{D}_\mathcal{Y}$ be divergences on $\mathcal{X}$ and $\mathcal{Y}$, respectively. Further take $\mathcal{F}$ as a class of mappings from $\mathcal{X}$ to $\mathcal{Y}$, and $\mathcal{G}$ a class of mappings from $\mathcal{Y}$ to $\mathcal{X}$. The p^{th} order UBGM divergence between $(\mathcal{X}, d_\mathcal{X}, P)$ and $(\mathcal{Y}, d_\mathcal{Y}, Q)$ is*

$$\mathsf{UD}_p^{\mathcal{F},\mathcal{G}}(P\|Q) := \inf_{\substack{f\in\mathcal{F}\\ g\in\mathcal{G}}} \Delta_p(f,g;P,Q) + \lambda_x \mathsf{D}_\mathcal{X}(g_\sharp Q, P) + \lambda_y \mathsf{D}_\mathcal{Y}(f_\sharp P, Q),$$

where $\Delta_p(f,g;P,Q)$ is given in Definition 2.

Evidently, $\mathsf{UD}_p^{\mathcal{F},\mathcal{G}}$ no longer requires optimizing over Monge maps, which alleviates the computational difficulty associated with D_p from Definition 2. The function classes $\mathcal{F}$ and $\mathcal{G}$ can also be chosen for computational convenience (e.g., NNs). The following proposition (see Supplement A.3 for the proof) shows that $\mathsf{UD}_p^{\mathcal{F},\mathcal{G}}$ is a continuous divergence.

Proposition 2 ($\mathsf{UD}_p^{\mathcal{F},\mathcal{G}}$ is a divergence)**.** *Suppose that $\mathsf{D}_\mathcal{X}$ and $\mathsf{D}_\mathcal{Y}$ are weakly continuous in their arguments, and $\mathcal{F}, \mathcal{G}$ are rich enough so that they are dense around the isometric bijections if $\mathcal{X}, \mathcal{Y}$ are isometric. Then*

1. *$\mathsf{UD}_p^{\mathcal{F},\mathcal{G}}$ is an upper semi-continuous divergence on $\mathfrak{M}$, i.e., $\mathsf{UD}_p^{\mathcal{F},\mathcal{G}}(P\|Q) \ge 0$, for all $P \in \mathcal{P}(\mathcal{X})$ and $Q \in \mathcal{P}(\mathcal{Y})$, with equality iff there exists isometries f, g, such that $f_\sharp P = Q$, $g_\sharp Q = P$, and $f = g^{-1}$.*
2. *If further $\mathcal{F}, \mathcal{G}$ are compact in the sup-metrics $d_\mathcal{F}$ and $d_\mathcal{G}$, then $\mathsf{UD}_p^{\mathcal{F},\mathcal{G}}$ is continuous with respect to (w.r.t.) weak convergence.*

3.2 Cycle GAN as Unbalanced Gromov-Monge Divergence

If $\mathsf{D}_\mathcal{X}$ and $\mathsf{D}_\mathcal{Y}$ are integral probability metrics (IPM)[3] (Zolotarev, 1984; Müller, 1997) indexed by the function classes $\mathcal{F}_\mathcal{X}$ and $\mathcal{F}_\mathcal{Y}$, respectively, then UD_1 amounts to a minimax game between the maps f, g and the witness functions ψ and ϕ of the IPMs on $\mathcal{F}_\mathcal{X}$ and $\mathcal{F}_\mathcal{Y}$ respectively, as follows:

$$\inf_{\substack{f:\mathcal{X}\to\mathcal{Y}\\ g:\mathcal{Y}\to\mathcal{X}}} \sup_{\substack{\psi\in\mathcal{F}_\mathcal{X}\\ \phi\in\mathcal{F}_\mathcal{Y}}} \Delta_1(f,g;P,Q) + \lambda_x\Big(\mathbb{E}\big[\psi\big(g(Y)\big)\big] - \mathbb{E}\big[\psi(X)\big]\Big) + \lambda_y\Big(\mathbb{E}\big[\phi\big(f(X)\big)\big] - \mathbb{E}\big[\phi(Y)\big]\Big), \quad (5)$$

where $X \sim P$ and $Y \sim Q$.

Written in this form, we see the similarity to the cycle GAN formulation (Zhu et al., 2017; Kim et al., 2017):

$$\inf_{\substack{f:\mathcal{X}\to\mathcal{Y},\\ g:\mathcal{Y}\to\mathcal{X}}} \sup_{\substack{\psi\in\mathcal{F}_\mathcal{X},\\ \phi\in\mathcal{F}_\mathcal{Y}}} \mathbb{E}\big[d_\mathcal{X}\big(X, g\circ f(X)\big)\big] + \mathbb{E}\big[d_\mathcal{Y}\big(Y, f\circ g(Y)\big)\big] + \lambda_x\Big(\mathbb{E}\big[\psi\big(g(Y)\big)\big] - \mathbb{E}\big[\psi(X)\big]\Big) + \lambda_y\Big(\mathbb{E}\big[\phi\big(f(X)\big)\big] - \mathbb{E}\big[\phi(Y)\big]\Big). \quad (6)$$

The first two terms in (6) encourage f and g to be approximate inverses of one another, similar to the role of the distortion $\Delta_1(f,g;P,Q)$ in UD_1 from (5). While the original Cycle GAN formulation did not require f and g to be isometries, this constraint was introduced in followup works (Hoshen and Wolf, 2018). We thus see that Cycle GAN, with a relaxed isometry requirement of f and g, is a particular instantiation of the UBGM divergence.

3.3 Relation to Past Works

We show briefly here the construction of two well-known discrepancies between mm spaces, namely the GW distance (Mémoli, 2011), and the Gromov-Monge (GM) distance (Mémoli and Needham, 2021). The starting point of defining these two distances is the ‘correspondence set formulation’ of d_{GH} from (3).

Gromov-Wasserstein distance. In a nutshell, the GW distance is an $L^p, p \ge 1$ relaxation of the L^∞ norm

[2] A divergence on $\mathcal{P}(\mathcal{X})$ is a functional $\mathsf{D} : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}_{\ge 0}$ such that $\mathsf{D}(P\|Q) = 0$ iff $P = Q$.

[3] An IPM indexed by a function class $\mathcal{F}$ is a pseudometric on $\mathcal{P}(\mathcal{X})$ defined as $d_\mathcal{F}(\mu,\nu) := \sup_{f\in\mathcal{F}} \int_\mathcal{X} f \,\mathrm{d}(\mu - \nu)$.

in formulation (3) of d_{GH}, along with a Kantorovich relaxation of the correspondence set using couplings.

Definition 5 (Gromov-Wasserstein distance (Mémoli, 2011))**.** *The GW distance between $(\mathcal{X}, d_{\mathcal{X}}, P)$ and $(\mathcal{Y}, d_{\mathcal{Y}}, Q)$ is*

$$\mathsf{GW}(P,Q) := \inf_{\pi\in\Pi(P,Q)} \|\Gamma_{\mathcal{X},\mathcal{Y}}\|_{L^p(\pi\otimes\pi)}$$

where $\Gamma_{\mathcal{X},\mathcal{Y}}(x,y,x',y') := \big|d_{\mathcal{X}}(x,x') - d_{\mathcal{Y}}(y,y')\big|$, and $\Pi(P,Q)$ is the set of all couplings of P,Q.

Another closely related formulation is the unbalanced GW distance from Séjourné et al. (2020). For any divergence[4] $\mathsf{D}_{\mathcal{X}}$ on $\mathcal{X}$, define its two-fold extension $\mathsf{D}_{\mathcal{X}}^{\otimes 2}(P,Q) := \mathsf{D}_{\mathcal{X}}(P\otimes P, Q\otimes Q)$.

Definition 6 (Unbalanced Gromov-Wasserstein distance (Séjourné et al., 2020))**.** *Let $\mathcal{M}_+(\mathcal{X})$ be the set of all nonnegative Borel measures on $\mathcal{X}$. The unbalanced GW distance between $(\mathcal{X}, d_{\mathcal{X}}, P)$ and $(\mathcal{Y}, d_{\mathcal{Y}}, Q)$ is*

$$\mathsf{UGW}(P,Q) := \inf_{\pi\in\mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \|\Gamma_{\mathcal{X},\mathcal{Y}}\|_{L^1(\pi\otimes\pi)} + \mathsf{D}_{\mathcal{X}}^{\otimes 2}(\pi_1\|P) + \mathsf{D}_{\mathcal{Y}}^{\otimes 2}(\pi_2\|Q)$$

where π_1, π_2 are the marginals of π on $\mathcal{X}$ and $\mathcal{Y}$.

The unbalanced relaxation of the GW distance is similar to how our UBGM distance (Definition 4) relaxes the BGM distance. A crucial difference is that both the BGM distance and its unbalanced version explicitly encode bidirectional mappings, which are important in applications as they alleviate the need to recompute the coupling matrix given new datapoints.

Gromov-Monge distance. More recently, Mémoli and Needham (2021) presented another extension of d_{GH} to a discrepancy between mm spaces. Termed the GM distance, it considers an L^p Monge relaxation of (3), as opposed to the Kanotrovich-based approach of GW. Namely, instead of using couplings, the correspondence set now comprises Monge maps, i.e., all measurable maps $f:\mathcal{X}\to\mathcal{Y}$ s.t. $f_\sharp P = Q$. Also for arbitrary f, denote $\pi_f := (\mathrm{id}, f)_\sharp P$. Clearly for Monge maps f that pushes P to Q, $\pi_f \in \Pi(P,Q)$.

Definition 7 (Gromov-Monge distance (Mémoli and Needham, 2021))**.** *The GM distance between two mm spaces $(\mathcal{X}, d_{\mathcal{X}}, P)$ and $(\mathcal{Y}, d_{\mathcal{Y}}, Q)$ is*

$$\mathsf{GM}(P,Q) := \inf_{f:\mathcal{X}\to\mathcal{Y},\, f_\sharp P = Q} \|\Gamma_{\mathcal{X},\mathcal{Y}}\|_{L^p(\pi_f\otimes\pi_f)}$$

where $\|\Gamma_{\mathcal{X},\mathcal{Y}}\|_{L^p(\pi_f\otimes\pi_f)} = \Delta_{\mathcal{X}}^{(p)}(f;P)$.

Comparing D_p to GM above, we see that while the latter uses a single low metric distortion maps (with a cost of the form $\Delta_{\mathcal{X}}^{(p)}(f;P)$), our BGM distance uses two such mappings that are approximately inverses (as enforced by $\Delta_{\mathcal{X},\mathcal{Y}}^{(p)}(f,g;P,Q)$). In a sense our definition is a symmetrized and cycle consistent version of GM.

4 KERNELIZATION: GENERALIZED MAXIMUM MEAN DISCREPANCY

Motivated by computational considerations, we now instantiate the divergences $\mathsf{D}_{\mathcal{X}}$ and $\mathsf{D}_{\mathcal{Y}}$ in $\mathsf{UD}_p^{\mathcal{F},\mathcal{G}}$ (see Definition 4) as maximum mean discrepancies (MMDs) (Gretton et al., 2012). We coin the resulting kernelized divergence as the *generalized maximum mean discrepancy* (GMMD). MMDs can be efficiently computed and offer flexibility in picking the proper kernel for each space. We start by reviewing preliminaries on MMDs (Section 4.1), after which we present the kernelized UBGM distance (Section 4.2), and explore its empirical convergence rates (Section 4.3).

4.1 Reproducing Kernel Hilbert Spaces

We define reproducing kernel Hilbert spaces (RKHS) and the associated MMD. For a separable space $\mathcal{X}$ and a continuous, positive definite, real-valued kernel $k_{\mathcal{X}}:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$, let $\mathcal{H}_{\mathcal{X}}$ denote the corresponding RKHS, in which for any $f\in\mathcal{H}_{\mathcal{X}}$, we have $f(x) = \langle f(\cdot), k(x,\cdot)\rangle_{\mathcal{H}_{\mathcal{X}}}$, for any $x\in\mathcal{X}$. See Berlinet and Thomas-Agnan (2011) for existence and uniqueness of $\mathcal{H}_{\mathcal{X}}$. There is a natural way to embed $\mathcal{P}(\mathcal{X})$ into $(\mathcal{H}_{\mathcal{X}}, k_{\mathcal{X}})$, given by the kernel mean embedding

$$\mu_{\mathcal{X}} P(x) := \int_{\mathcal{X}} k_{\mathcal{X}}(x,y)\,\mathrm{d}P(y) = \mathbb{E}\big[k_{\mathcal{X}}(x,Y)\big],$$

where $Y\sim P$. This enables defining a discrepancy measure between probability distribution as the RKHS distance between their kernel mean embeddings.

Definition 8 (Maximum mean discrepancy)**.** *Let $\mathcal{H}_{\mathcal{X}}$ be an RKHS. The MMD between $P,Q\in\mathcal{P}(\mathcal{X})$ is*

$$\begin{aligned}\mathsf{MMD}_{\mathcal{X}}(P,Q) &:= \|\mu_{\mathcal{X}}P - \mu_{\mathcal{X}}Q\|_{\mathcal{H}_{\mathcal{X}}}\\ &= \left(\int k_{\mathcal{X}}(x,y)\,\mathrm{d}(P-Q)(x)\,\mathrm{d}(P-Q)(y)\right)^{1/2}.\end{aligned}$$

When the kernel $k_{\mathcal{X}}$ is characteristic, as defined next, $\mathsf{MMD}_{\mathcal{X}}$ metrizes the space of distributions $\mathcal{P}(\mathcal{X})$.

Definition 9 (Characteristic kernel)**.** *The kernel $k_{\mathcal{X}}$ of an RKHS $\mathcal{H}_{\mathcal{X}}$ is called characteristic if the mean embedding $\mu_{\mathcal{X}}:\mathcal{P}(\mathcal{X})\to\mathcal{H}_{\mathcal{X}}$ is injective.*

[4] Séjourné et al. (2020) used f-divergences (Csiszár, 1967) for $\mathsf{D}_{\mathcal{X}}$ and $\mathsf{D}_{\mathcal{Y}}$, but we provide a general definition.

Also recall that characteristic kernels enable defining a metric on the $\mathcal{X}$ space (Sejdinovic et al., 2013). Namely, defining $\rho_{k_\mathcal{X}}(x, x') := k_\mathcal{X}(x, x) + k_\mathcal{X}(x', x') - 2k_\mathcal{X}(x, x')$, for $x, x' \in \mathcal{X}$, we have that $\left(\mathcal{X}, \sqrt{\rho_{k_\mathcal{X}}}\right)$ is a metric space. To simplify notation, we henceforth denote $\rho_{k_\mathcal{X}}$ by $\rho_\mathcal{X}$.

4.2 Generalized MMD

The GMMD is defined as follows. Throughout we assume that $\mathcal{X}$ and $\mathcal{Y}$ are compact with diameters bounded by K, and specialize to the case of $p = 1$.

Definition 10 (Generalized MMD between Metric Measure Spaces)**.** *Let $k_\mathcal{X}$ and $k_\mathcal{Y}$ be characteristic kernels on $\mathcal{X}$ and $\mathcal{Y}$, respectively. The GMMD between $(\mathcal{X}, d_\mathcal{X}, P)$ and $(\mathcal{Y}, d_\mathcal{Y}, Q)$ is*

$$\begin{aligned}\mathsf{UD}^{\mathcal{F},\mathcal{G}}(P\|Q) := \inf_{\substack{f\in\mathcal{F}\\ g\in\mathcal{G}}} & \Delta_1(f, g; P, Q)\\ & + \lambda_x \mathsf{MMD}_\mathcal{X}(P, g_\sharp Q) + \lambda_y \mathsf{MMD}_\mathcal{Y}(f_\sharp P, Q)\end{aligned}$$

where $\lambda_x, \lambda_y > 0$ are fixed regularization coefficients.

Proposition 3 (GMMD is a divergence)**.** *Consider the GMMD $\mathsf{UD}^{\mathcal{F},\mathcal{G}}$ defined above and assume that the function classes $\mathcal{F}, \mathcal{G}$ are rich enough, as defined in Proposition 2. The following hold:*

1. *$\mathsf{UD}^{\mathcal{F},\mathcal{G}}$ is a divergence on $\mathfrak{M}$, i.e., a nonnegative discrepancy measure that nullifies iff the two metric measure spaces are equivalent.*

2. *If further $\mathcal{F}, \mathcal{G}$ are compact w.r.t. their sup-metrics and $\|k_\mathcal{X}\|_{L^\infty}, \|k_\mathcal{Y}\|_{L^\infty} < \infty$, then $\mathsf{UD}^{\mathcal{F},\mathcal{G}}$ is weakly continuous.*

This proposition follows directly from Proposition 2, as MMDs are weakly continuous for bounded kernels.

Remark 3 (Kernels specify GMMD)**.** *GMMD can be fully specified by the kernels if one defines the mm spaces using the kernel induced metrics $\rho_\mathcal{X}$ and $\rho_\mathcal{Y}$.*

4.3 GMMD Empirical Estimation Rates

We now study the convergence rate of the two-sample plugin estimator of GMMD. Let $(X_i)_{i=1}^n$ and $(Y_i)_{i=1}^m$ be i.i.d. samples from P and Q, respectively. Denote by $P_n := n^{-1}\sum_{i=1}^n \delta_{X_i}$ and $Q_m := m^{-1}\sum_{i=1}^m \delta_{Y_i}$ the empirical measures associated with these samples. Since GMMD is weakly continuous (for compact $\mathcal{F}$ and $\mathcal{G}$), we immediately have $\mathsf{UD}^{\mathcal{F},\mathcal{G}}(P_n\|Q_m) \to \mathsf{UD}^{\mathcal{F},\mathcal{G}}(P\|Q)$ as $n, m \to \infty$ a.s. The focus of this section is the rate at which this convergence happens.

Theorem 1. *Suppose $k_\mathcal{X}, k_\mathcal{Y}$ are uniformly bounded by a constant C, and the diameters of $\mathcal{X}$ and $\mathcal{Y}$ are bounded by K. Further suppose that $\mathcal{F}$ and $\mathcal{G}$ are compact in $d_\mathcal{F}$ and $d_\mathcal{G}$, respectively. Then*

$$\begin{aligned}\mathbb{E}&\left[\left|\mathsf{UD}^{\mathcal{F},\mathcal{G}}(P\|Q) - \mathsf{UD}^{\mathcal{F},\mathcal{G}}(P_n\|Q_m)\right|\right]\\ &\lesssim \lambda_y \delta_n(\mathcal{F}_{k_\mathcal{Y}}) + \lambda_x \delta_m(\mathcal{G}_{k_\mathcal{X}}) + \delta_{n,m}(\mathcal{F},\mathcal{G})\\ &\qquad + \lambda_x C^{\frac{1}{2}} n^{-\frac{1}{2}} + \lambda_y C^{\frac{1}{2}} m^{-\frac{1}{2}} + K(n\wedge m)^{-1}\end{aligned}$$

where $\mathcal{F}_{k_\mathcal{Y}} := \{k_\mathcal{Y} \circ (f, f) : f \in \mathcal{F}\}$ and

$$\delta_n(\mathcal{F}_{k_\mathcal{Y}}) := \inf_{\alpha>0}\left(\alpha + \frac{1}{n}\int_\alpha^{2C} \log\left(N(\mathcal{F}_{k_\mathcal{Y}}, \|\cdot\|_\infty, \tau)\right) \mathrm{d}\tau\right)^{\frac{1}{2}},$$

with $\mathcal{G}_{k_\mathcal{X}}$ and $\delta_m(\mathcal{G}_{k_\mathcal{X}})$ defined analogously, and

$$\begin{aligned}\delta_{n,m}(\mathcal{F},\mathcal{G}) := \inf_{\alpha>0}\Bigg(\alpha + \frac{1}{\sqrt{n\wedge m}}\int_\alpha^K \Big(&\log\left(N(\mathcal{F}, d_\mathcal{F}, \tau)\right)\\ &+ \log\left(N(\mathcal{G}, d_\mathcal{G}, \tau)\right)\Big)^{\frac{1}{2}} \mathrm{d}\tau\Bigg).\end{aligned}$$

Theorem 1 bounds the estimation error for general function classes $\mathcal{F}$ and $\mathcal{G}$ in terms of the appropriate entropy integrals. The proof is given in Supplement A.4 and relies on standard chaining arguments and bounds on Rademacher chaos complexity (cf. e.g, Sriperumbudur (2016)).

In general, the above entropy integrals cannot be further simplified due to the dependence on the arbitrary classes $\mathcal{F}$ and $\mathcal{G}$. Nevertheless, the next corollary instantiates Theorem 1 to two particular function classes of interest and states explicit convergence rates.

Corollary 1 (Special cases)**.** *Under the same condition of Theorem 1, further suppose that $\mathcal{X} \subset \mathbb{R}^{d_x}, \mathcal{Y} \subset \mathbb{R}^{d_y}$ are compact, and $k_\mathcal{X}, k_\mathcal{Y}$ are L-Lipschitz in both slots.*[5]

1. *Lipschitz: For $\mathcal{F} = \mathsf{Lip}_{L_\mathcal{F}}(\mathcal{X}, \mathcal{Y})$, $\mathcal{G} = \mathsf{Lip}_{L_\mathcal{G}}(\mathcal{Y}, \mathcal{X})$, and $d_x, d_y > 2$, we have*
$$\begin{aligned}\mathbb{E}&\left[\left|\mathsf{UD}^{\mathcal{F},\mathcal{G}}_{k_\mathcal{X},k_\mathcal{Y}}(P\|Q) - \mathsf{UD}^{\mathcal{F},\mathcal{G}}_{k_\mathcal{X},k_\mathcal{Y}}(P_n\|Q_m)\right|\right]\\ &\lesssim_{\lambda_x,\lambda_y,L,L_\mathcal{F},L_\mathcal{G},C,K} \left(\frac{1}{n}\right)^{\frac{1}{2d_x}} + \left(\frac{1}{m}\right)^{\frac{1}{2d_y}}.\end{aligned}$$

2. *Parametric: Let $\Theta \subset \mathbb{R}^{k_1}, \Phi \subset \mathbb{R}^{k_2}$ be compact parameter sets with diameters bounded by K'. Take $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$, with $d_\mathcal{F}(f_{\theta_1}, f_{\theta_2}) \leq L_\Theta\|\theta_1 - \theta_2\|$, for some constant L_Θ and all $\theta_1, \theta_2 \in \Theta$. Suppose analogously for $\mathcal{G} = \{g_\phi : \phi \in \Phi\}$ and L_Φ. Then*
$$\begin{aligned}\mathbb{E}&\left[\left|\mathsf{UD}^{\mathcal{F},\mathcal{G}}_{k_\mathcal{X},k_\mathcal{Y}}(P\|Q) - \mathsf{UD}^{\mathcal{F},\mathcal{G}}_{k_\mathcal{X},k_\mathcal{Y}}(P_n\|Q_m)\right|\right]\\ &\lesssim_{\lambda_x,\lambda_y,L,L_\Theta,L_\Phi,C,K,K',k_1,k_2} \left(\frac{1}{n\wedge m}\right)^{\frac{1}{2}}.\end{aligned}$$

[5] Namely, $|k_\mathcal{X}(x_1, x_1') - k_\mathcal{X}(x_2, x_2')| \leq L(d_\mathcal{X}(x_1, x_2) + d_\mathcal{X}(x_1', x_2'))$, and similarly for $k_\mathcal{Y}$.

The proof is given in Supplement A.6. The latter bound for parametric classes is of particular practical interest, as NNs offer a convenient and trainable model for the bidirectional maps (see next section).

5 NUMERICAL EXPERIMENTS

We present applications of GMMD in shape matching. The bidirectional maps f and g are parametrized by neural networks f_θ and g_ϕ, respectively (θ and ϕ are the parameters). Algorithm 1 (Supplement B) summarizes the optimization of the GMMD objective as function of θ and ϕ. The running code for these experiments are made public on Github.[6] All experiments are run on the same machine with 4 core CPUs and a Tesla T4 GPU. The examples below highlight the qualitative and quantitative behavior of GMMD, illustrating the fact that GMMD is applicable to the same tasks as classical methods such as GW and UGW for finding correspondences. Further, GMMD amortizes the computational cost, as it results in continuous mappings that generalize to unseen datapoints drawn from the same distributions.

5.1 GMMD For Shape Matching

We consider here matching of synthetic shapes, specifically a 2-dimensional heart shape given in Figure 1(a) and its transformations through rotation (b), scaling (c) and isometrically embedding into 3-dimensional space (d). The data is generated via sampling $n = 4000$ points for each shape. The distributions for each matching experiment are the empirical measures induced by these samples, with P corresponding to 1(a) and Q_b, Q_c, and Q_d corresponding to subfigures (b), (c), and (d), respectively (the subscript is suppressed when we simultaneously refer to several experiments).

For each matching experiment, we compute GMMD using Algorithm 1 for $\lambda_x = \lambda_y = \lambda$, where $\lambda \in \{2^{-i} \times 10^3 : i = 0, \cdots, 9\}$. We use a uniform mixture of Gaussian kernels to define $\mathsf{MMD}_\mathcal{X}$ and $\mathsf{MMD}_\mathcal{Y}$ and use kernel induced metrics $\rho_\mathcal{X}$, and $\rho_\mathcal{Y}$ in the distortion Δ_1. The bandwidths used for the Gaussian kernels are median of the metric $\times\{.0001, .001, .01, .05, .25, 1, 4, 20, 100, 1000\}$. The architecture of the bidirectional maps f and g is a 3 layer ReLU NN with 200 neurons each, and an output dimension matching the target distribution dimension. We use Adam optimizer (Kingma and Ba, 2014) for 3000 epochs with a learning rate 10^{-3}.

In Figure 2, the first row corresponds to GMMD matching for $\lambda = 2^{-6} \times 10^3$. For each case, we see that the learned bidirectional maps of GMMD successfully perform the matching, i.e., $f_\sharp P \approx Q$ and $g_\sharp Q \approx P$. We also confirm that they satisfy the cycle consistency property, i.e., $f \circ g \approx \mathrm{id}_\mathcal{Y}$ and $g \circ f \approx \mathrm{id}_\mathcal{X}$. The second row shows entropic GW matchings (Peyré et al., 2016). We use the POT library (Flamary et al., 2021) to perform discrete entropic GW for an entropic regularization parameter $\varepsilon = 5e^{-4}$. Note that entropic GW results in a coupling matrix π. To obtain discrete mappings of the points we employ barycentric maps (Ferradans et al., 2014), i.e., $\tilde{f}(x_i) := (\sum_{j=1}^n \pi_{ij})^{-1} \sum_{j=1}^n \pi_{ij} y_j$ and $\tilde{g}(y_j) := (\sum_{i=1}^n \pi_{ij})^{-1} \sum_{i=1}^n \pi_{ij} x_i$.

We see from Figure 2 that the GMMD continuous maps and the discrete barycentric maps induced by the GW coupling are on par qualitatively in these matching tasks. To confirm this quantitatively, we consider the matching of the heart shape and its rotation (Figure 1(b)) since for this case an isometry exists, i.e., there are $f^\star, g^\star$ with $\Delta_1(f^\star, g^\star; P, Q_b) = 0$. Tables 1 and 2 state the values of $\mathsf{MMD}_\mathcal{Y}(f_\sharp P, Q_b)$, $\mathsf{MMD}_\mathcal{X}(P, g_\sharp Q_b)$, and $\Delta(f, g; P, Q_b)$ across different regularization parameters for the GMMD and GW-based mappings, respectively. We see that GMMD and GW indeed result in small MMD and distortions values. GMMD yields a smaller distortion than GW. Note that we also have evaluated UGW (Séjourné et al., 2020) with the code provided by the authors and found that it is sensitive to hyper-parameters choice, and did not result in an accurate matching on the considered tasks. We think that more tuning is needed for UGW. Additional results and ablation on regularization parameters and shapes are given in Supplement B.

Figure 3 presents a more complex matching of 3D shapes that consist in two different biplanes models from the Princeton Shape benchmark (Shilane et al., 2004) (for $n = 8000$). We see that GMMD and GW are also on par and that the GMMD bidirectional maps result in less outliers than barycentric GW-based maps. This robustness of GMMD is due to the use of kernel induced metrics. Quantitative evaluation is presented in Supplement B.

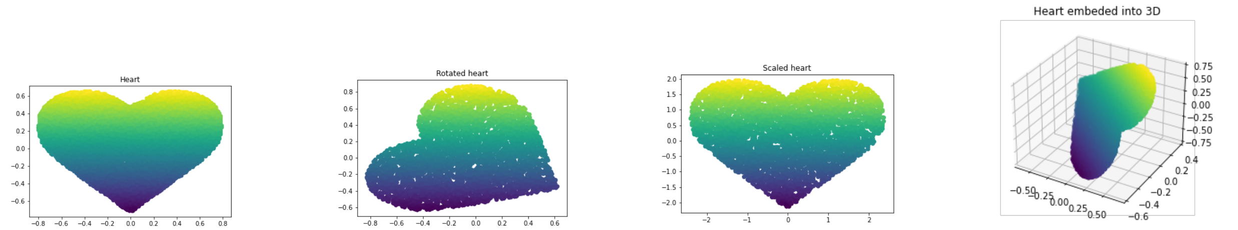
Table 1: Evaluating GMMD's mappings for P vs. Q_b.

λ	GMMD	$\mathsf{MMD}_\mathcal{X}$	$\mathsf{MMD}_\mathcal{Y}$	Δ
$2^{-8} \times 10^3$	0.310	0.0294	0.0294	0.0801
$2^{-7} \times 10^3$	0.0645	4.94e-4	4.05e-4	0.0574
$2^{-6} \times 10^3$	0.121	0.00227	0.00190	0.0560
$2^{-5} \times 10^3$	2.89	2.90e-4	0.00386	2.76

5.2 GMMD Amortization and Generalization

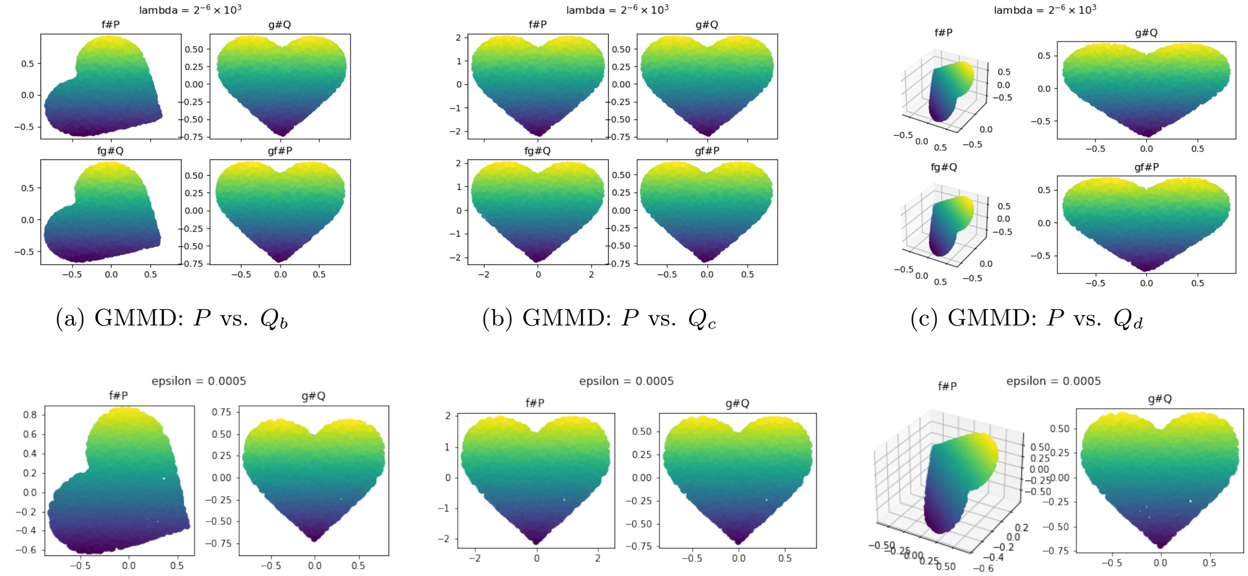
For the biplane matching experiment with sample size $n = 8000$ from each distribution, Table 3 reports the

[6] See https://github.com/ZhengxinZh/GMMD



(a) Heart shape (P) (b) Rotation (Q_b) (c) Scaling (Q_c) (d) 3D embedding (Q_d)

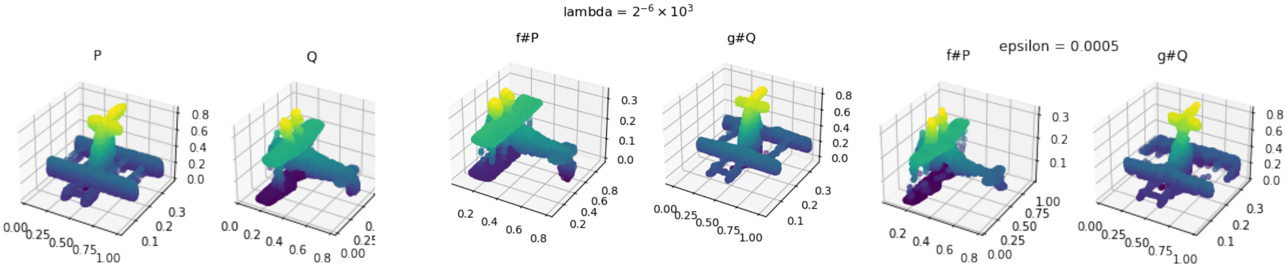
Figure 1: Heart shape and its transformations.



(a) GMMD: P vs. Q_b (b) GMMD: P vs. Q_c (c) GMMD: P vs. Q_d

(d) GW : P vs. Q_b. (e) GW : P vs. Q_c. (f) GW : P vs. Q_d.

Figure 2: First row: learned continuous GMMD Mappings and their cycle consistency in shape matching. Second row: discrete entropic GW Barycentric Mappings. The color code in the heatmaps is coordinate based.



(a) Original P and Q. (b) GMMD maps. (c) GW barycentric maps.

Figure 3: Matching 3D shapes with GMMD and entropic GW.

training time for GMMD and the runtime for GW and UGW computings. The computational complexity of training GMMD maps amounts to the complexity of gradient descent in NN training for 3000 epochs, which is $O(n)$. For entropic GW and UGW, however, the implementations are variants of the Sinkhorn algorithm, whose complexity scales as $O(n^3)$. The longer training time for GMMD is due to the large number of epochs used in gradient descent (namely, 3000), but at inference time this cost is amortized since we obtain continuous maps that generalize to unseen datapoints (see Supplement B for quantitative evaluation of the generalization). For instance, matching 8000 new datapoints sampled from P and Q each using the

Table 2: GW barycentric maps for P vs. Q_b.

ϵ	GW	$\text{MMD}_{\mathcal{X}}$	$\text{MMD}_{\mathcal{Y}}$	Δ
0.0005	0.00134	0.00420	0.00299	0.696
0.005	0.00660	0.127	0.116	1.73
0.05	0.0424	0.615	0.613	6.69
0.5	0.0686	3.99	4.12	22.9

learned mapping requires 63 ms, while with GW one would incur the cost of recomputing the coupling (26 minutes)—a three order of magnitude speedup.

Table 3: Training Time (in seconds) comparison using 8000 samples from the biplanes data.

ϵ	GW	λ	GMMD
0.0005	1566.86	$2^{-1} \times 10^3$	5048.11
ϵ	UGW	$2^{-7} \times 10^3$	5026.5
0.1	28.7508	$2^{-6} \times 10^3$	5052.89

5.3 Word Embedding Alignment

We consider the more realistic of word embedding alignment between different languages. Such tasks have been previously considered under the GW framework (Alvarez-Melis and Jaakkola, 2018), using Procrustes methods (Conneau et al., 2017), and more. We employ the GMMD to learn mappings between English and French words that are embedded into 300 dimensional spaces. A correspondence is then obtained by searching for the nearest neighbor. Our overall approach requires minimal fine tuning—see Supplement B.2 for a comprehensive description of the employed network architecture and hyperparameters values. The obtained results with comparison to existing benchmarks are stated in Table 4, where performance is measured by the percentage of correct matchings. Specifically, we compare to the entropic GW approach from Alvarez-Melis and Jaakkola (2018) with regularization parameters $\epsilon = 10^{-4}, 10^{-5}$, and to the MUSE method from Conneau et al. (2017).

Table 4: Word matching performance comparison.

	EN to FR	FR to EN
GMMD	76.1%	74.5%
GW ($\epsilon = 10^{-4}$)	79.3%	78.3%
GW ($\epsilon = 10^{-5}$)	81.3%	78.9%
MUSE	82.3%	82.1%

6 CONCLUSION

This paper introduced the UBGM divergence—a novel discrepancy measure between distributions across heterogeneous spaces, which employs bidirectional and cycle-consistent mappings. We established structural properties of the UBGM divergence and highlighted its intimate connection to the so-called cycle GAN. We also presented a kernelized variant of this divergence, termed GMMD, and analysed its statistical estimation from samples. Numerical experiments demonstrated the promise of this new divergence and compared it to other known metrics, such as the GW and UGW distances. Appealing future directions include extending the GMMD to allow optimization over kernels, sharper statistical bounds, as well as connections between the UBGM divergence and the UGW distance (in particular, under what conditions they coincide).

6.1 Societal Impact

We address potential societal impacts of our work. Though the paper is largely of theoretical nature, our empirical results demonstrate that the developed discrepancy measure and may be of practical interest. In particular, the tie between our GMMD and the popular Cycle GAN may raise issues related to inappropriate usage of GANs such as deepfakes.

Acknowledgements

We thank the anonymous reviewers and the area chair, whose feedback helped improve this paper. Z. Goldfeld is supported by the NSF CRII grant CCF-1947801, the 2020 IBM Academic Award, and the NSF CAREER Award CCF-2046018.

References

D. Alvarez-Melis and T. S. Jaakkola. Gromov-Wasserstein alignment of word embedding spaces. *arXiv preprint arXiv:1809.00013*, 2018.

D. Alvarez-Melis, Y. Mroueh, and T. Jaakkola. Unsupervised hierarchy matching with optimal transport over hyperbolic spaces. In S. Chiappa and R. Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (AISTATS-2020)*, volume 108 of *Proceedings of Machine Learning Research*, pages 1606–1617, 2020.

A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov.

Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

A. M. Bronstein, M. M. Bronstein, and R. Kimmel. Generalized multidimensional scaling: A framework for isometry-invariant partial surface matching. *Proceedings of the National Academy of Sciences*, 103 (5):1168–1172, 2006a.

A. M. Bronstein, M. M. Bronstein, and R. Kimmel. Efficient computation of isometry-invariant distances between surfaces. *SIAM Journal on Scientific Computing*, 28(5):1812–1836, 2006b.

C. Bunne, D. Alvarez-Melis, A. Krause, and S. Jegelka. Learning generative models across incomparable spaces. In *Proceedings of the International Conference on Machine Learning (ICML-2019)*, pages 851–861, June 2019.

L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. Unbalanced optimal transport: Dynamic and Kantorovich formulations. *Journal of Functional Analysis*, 274(11):3090–3123, 2018.

C. W. Commander. A survey of the quadratic assignment problem, with applications. *Morehead Electronic Journal of Applicable Mathematics*, 4: MATH–2005–01, 2005.

A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.

I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Ccientiarum Mathematicarum Hungarica*, 2: 229–318, 1967.

V. De la Pena and E. Giné. *Decoupling: from dependence to independence*. Springer Science & Business Media, 2012.

S. Ferradans, N. Papadakis, G. Peyré, and J.-F. Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014.

R. Flamary, K. Lounici, and A. Ferrari. Concentration bounds for linear monge mapping estimation and optimal transport domain adaptation. *arXiv preprint arXiv:1905.10155*, 2019.

R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, and T. Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

C. Frogner, C. Zhang, H. Mobahi, M. Araya-Polo, and T. Poggio. Learning with a Wasserstein loss. *arXiv preprint arXiv:1506.05439*, 2015.

E. Grave, A. Joulin, and Q. Berthet. Unsupervised alignment of embeddings with Wasserstein procrustes. In *Proceedings of the The 22nd International Conference on Artificial Intelligence and Statistics (AISTATS-2019)*, pages 1880–1890. PMLR, 2019.

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

Y. Hoshen and L. Wolf. Unsupervised correlation analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR-2018)*, June 2018.

T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim. Learning to discover cross-domain relations with generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML-2017)*, pages 1857–1865. PMLR, 2017.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

K. Le, D. Le, H. Nguyen, D. Do, T. Pham, and N. Ho. Entropic gromov-wasserstein between gaussian distributions. *arXiv preprint arXiv:2108.10961*, 2021a.

T. Le, N. Ho, and M. Yamada. Flow-based alignment approaches for probability measures in different spaces. In *International Conference on Artificial Intelligence and Statistics*, pages 3934–3942. PMLR, 2021b.

A. Makkuva, A. Taghvaei, S. Oh, and J. Lee. Optimal transport mapping via input convex neural networks. In *Proceedings of the International Conference on Machine Learning (ICML-2020)*, pages 6672–6681. PMLR, 2020.

F. Mémoli. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487, 2011.

F. Mémoli and T. Needham. Distance distributions and inverse problems for metric measure spaces. *arXiv preprint arXiv:1810.09646*, 2021.

F. Mémoli and G. Sapiro. Comparing point clouds. In *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing*, pages 32–40, 2004.

F. Mémoli and G. Sapiro. A theoretical and computational framework for isometry invariant recognition of point cloud data. *Foundations of Computational Mathematics*, 5(3):313–347, 2005.

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

F.-P. Paty, A. d'Aspremont, and M. Cuturi. Regularity as regularization: Smooth and strongly convex brenier potentials in optimal transport. In *International Conference on Artificial Intelligence and Statistics (AISTATS-2021)*, pages 1222–1232. PMLR, 2020.

M. Perrot, N. Courty, R. Flamary, and A. Habrard. Mapping estimation for discrete optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS-2016)*, pages 4197–4205, 2016.

G. Peyré, M. Cuturi, and J. Solomon. Gromov-Wasserstein averaging of kernel and distance matrices. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning (ICML-2016)*, volume 48 of *Proceedings of Machine Learning Research*, pages 2664–2672, New York, New York, USA, 20–22 Jun 2016. PMLR.

D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, pages 2263–2291, 2013.

T. Séjourné, F.-X. Vialard, and G. Peyré. The unbalanced Gromov Wasserstein distance: Conic formulation and relaxation. *arXiv preprint arXiv:2009.04266*, 2020.

P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser. The princeton shape benchmark. In *Proceedings of the International Conference on Shape Modeling and Applications (SMI-2004)*, pages 167–178. IEEE, 2004.

B. Sriperumbudur. On the optimal estimation of probability measures in weak and strong topologies. *Bernoulli*, 22(3):1839–1893, 2016.

A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 1996.

T. Vayer, R. Flamary, R. Tavenard, L. Chapel, and N. Courty. Sliced Gromov-Wasserstein. *arXiv preprint arXiv:1905.10124*, 2019.

M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

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

H. Xu, D. Luo, and L. Carin. Scalable Gromov-Wasserstein learning for graph partitioning and matching. *Advances in neural information processing systems (NeurIPS-19)*, 32:3052–3062, 2019.

Y. Yan, W. Li, H. Wu, H. Min, M. Tan, and Q. Wu. Semi-supervised optimal transport for heterogeneous domain adaptation. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-2018)*, volume 7, pages 2969–2975, 2018.

J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision (CCV-2017)*, pages 2223–2232, 2017.

V. M. Zolotarev. Probability metrics. *Theory of Probability & Its Applications*, 28(2):278–302, 1984.

Supplementary Material:
Cycle Consistent Probability Divergences Across Different Spaces

A PROOFS

To simplify notation we denote $\mathcal{L}_{P,Q}(f,g) := \lambda_x \mathsf{MMD}_{\mathcal{X}}(P, g_\sharp Q) + \lambda_y \mathsf{MMD}_{\mathcal{Y}}(f_\sharp P, Q) + \Delta_1(f,g;P,Q)$, which is the functional that is optimized in definition of $\mathsf{UD}^{\mathcal{F},\mathcal{G}}$.

A.1 Proof of Proposition 1

The symmetry and positivity follows directly from definition and Lemma 1, which is proven below. For the triangle inequality, fix 3 mm spaces $(\mathcal{X}, d_{\mathcal{X}}, P)$, $(\mathcal{Y}, d_{\mathcal{Y}}, Q)$, $(\mathcal{Z}, d_{\mathcal{Z}}, R)$ and functions f_1, f_2, g_1, g_2 (over the appropriate domains) with $(f_1)_\sharp P = Q$, $(f_2)_\sharp Q = R$, $(g_1)_\sharp Q = P$, $(g_2)_\sharp R = Q$. We only show the derivation for $\Delta^{(p)}_{\mathcal{X},\mathcal{Y}}$; a similar argument applies to $\Delta^{(p)}_{\mathcal{X}}, \Delta^{(p)}_{\mathcal{Y}}$. For $\Delta^{(p)}_{\mathcal{X},\mathcal{Y}}$, we have

$$\begin{aligned}
\Delta^{(p)}_{\mathcal{X},\mathcal{Y}}&(f_1, g_1; P, Q) + \Delta^{(p)}_{\mathcal{X},\mathcal{Y}}(f_2, g_2; Q, R) \\
&= \Big(\int |d_{\mathcal{X}}(x, g_1(y)) - d_{\mathcal{Y}}(f_1(x), y)|^p \, \mathrm{d}P(x) \, \mathrm{d}Q(y) \Big)^{\frac{1}{p}} \\
&\qquad + \Big(\int |d_{\mathcal{Y}}(y, g_2(z)) - d_{\mathcal{Z}}(f_2(y), z)|^p \, \mathrm{d}Q(y) \, \mathrm{d}R(z) \Big)^{\frac{1}{p}} \\
&= \Big(\int |d_{\mathcal{X}}(x, g_1(g_2(z))) - d_{\mathcal{Y}}(f_1(x), g_2(z))|^p \, \mathrm{d}P(x) \, \mathrm{d}R(z) \Big)^{\frac{1}{p}} \\
&\qquad + \Big(\int |d_{\mathcal{Y}}(f_1(x), g_2(z)) - d_{\mathcal{Z}}(f_2(f_1(x)), z)|^p \, \mathrm{d}P(x) \, \mathrm{d}R(z) \Big)^{\frac{1}{p}} \\
&\geq \Big(\int |d_{\mathcal{X}}(x, g_1(g_2(z))) - d_{\mathcal{Z}}(f_2(f_1(x)), z)|^p \, \mathrm{d}P(x) \, \mathrm{d}R(z) \Big)^{\frac{1}{p}} \\
&= \Delta^{(p)}_{\mathcal{X},\mathcal{Y}}(f_2 \circ f_1, g_1 \circ g_2; P, R).
\end{aligned}$$

Hence D_p is a metric on $\mathfrak{M}$.

A.2 Proof of Lemma 1

Suppose $\{f_n\}_{n\in\mathbb{N}}$ and $\{g_n\}_{n\in\mathbb{N}}$ are sequence such that $\Delta_p(f_n, g_n; P, Q) \to 0$. We will show that up to extracting subsequences, these sequences converge $P \otimes Q$ a.s. to isometrics, f and g, respectively, such that $f = g^{-1}$. The argument first shows that there is a countable dense $S \subseteq \mathcal{X}$ such that the distortion function ϕ_n (defined below) converges on $S \times S$ to 0. Then we take a subsequence of f_n that converges on S, and show that this subsequence also converges P-a.s. on $\mathcal{X}$, and the limit is an isometry. After applying the same to $\{g_n\}_{n\in\mathbb{N}}$, we conclude the desired convergence and demonstrate that the limits f and g satisfy $f = g^{-1}$.

We first consider the term $\Delta^{(p)}_{\mathcal{X}}(f_n; P)$. Since

$$\int \big| d_{\mathcal{X}}(x, x') - d_{\mathcal{Y}}\big(f_n(x), f_n(x')\big) \big|^p \, \mathrm{d}P(x) \, \mathrm{d}P(x') \to 0,$$

we may assume that, up to extraction of subsequences, we have

$$\phi_n(x, x') := \big| d_{\mathcal{X}}(x, x') - d_{\mathcal{Y}}\big(f_n(x), f_n(x')\big) \big| \to 0, \quad P^{\otimes 2} - a.s.$$

Set $\Omega = \{(x, x') : \phi_n(x, x') \to 0\}$ as the set of pairs for which the convergence occurs, and let $\Omega_x = \{x' : (x, x') \in$

$\Omega\}$ be the slice at $x \in \mathcal{X}$ in the first coordinate. Then $P^{\otimes 2}\big((\mathcal{X} \times \mathcal{X}) - \Omega\big) = 0$, and by Fubini's theorem

$$\int_{\mathcal{X}} P(\Omega_x)\, \mathrm{d}P(x) = P^{\otimes 2}(\Omega) = 1.$$

Denoting $A = \{x : P(\Omega_x) = 1\}$, we thus have $P(A) = 1$, and hence A is dense. We next construct S as a countable dense subset of A.

Step 1 – Separability of convergence points: We present an inductive construction of a countable dense subset $S \subset \mathcal{X}$ such that ϕ_n converges to 0 on $S \times S$. First take any $x_0 \in A$ and define

$$S_0 := \big\{x' : \phi_n(x_0, x') \text{ does not converge to } 0\big\} = \mathcal{X} - \Omega_{x_0},$$

then $P(S_0) = 0$ since $x_0 \in A$. Suppose we have points $x_0, ..., x_k \in A$, such that $\phi(x_i, x_j) \to 0$ for $i, j = 0, \dots, k$, and define $S_i = \mathcal{X} - \Omega_{x_i}$ for $i = 0, \dots, k$. Define function

$$\psi_k(x) := \min_{0 \le i \le k} \{d_{\mathcal{X}}(x, x_i)\},$$

and set $w_k = \operatorname{argmax} \psi_k(x)$. Suppose $\psi_k(w_k) > 0$, otherwise $x_0, ..., x_k$ is already dense. Since $\psi_k(x)$ is continuous, $B_k = \{\psi_k(x) > \psi_k(w_k)/2\}$ is a nonempty open set on $\mathcal{X}$. Notice that set $C_k = A - \cup_{0 \le i \le k} S_i$ still have probability 1, and any point $x' \in C_k$ satisfies that $\phi_n(x', x_i)$ converges for all $i = 0, ..., k$. Since P has full support, $B_k \cap C_k$ is not empty, hence we pick $x_{k+1} \in B_k \cap C_k$. Inductively we have sequence $\{x_k\}_{k \in \mathbb{N}}$ such that $\phi_n(x_i, x_j)$ converges for any $i, j \in \mathbb{N}$. Denote $S = \{x_k\}_{k \in \mathbb{N}}$.

Now we prove that S is dense in $\mathcal{X}$. Suppose it is not, then there is an $\epsilon > 0$ and an $\tilde{x} \in \mathcal{X}$ such that $d_{\mathcal{X}}(\tilde{x}, x_k) > \epsilon$, for all $k \in \mathbb{N}$. So $\psi_k(w_k) \ge \psi_k(\tilde{x}) > \epsilon$. By construction, $d_{\mathcal{X}}(x_{k+1}, x_i) \ge \psi_k(w_k)/2 \ge \epsilon/2$, for all $i \le k$, so $d_{\mathcal{X}}(x_i, x_j) \ge \epsilon/2$ for any $i \ne j$. This is a contradiction since $\mathcal{X}$ is compact.

Step 2 – Convergence to isometry: Next we find a subsequence of $\{f_n\}_{n \in \mathbb{N}}$ such that it converges on S to an isometry f, and extend this convergence to a.s. on $\mathcal{X}$. Now we have a countable set $S \subseteq A$ that is dense in $\mathcal{X}$ such that $\phi_n(s, t) \to 0$, $\forall s, t \in S$. We can thus take a subsequence of $\{f_n\}_{n \in \mathbb{N}}$ such that it converges on S pointwise to a mapping f. Without loss of generality (WLOG) we assume f_n converges, for any $s \in S$, as any subsequence still approaches infinum and the subsequent ϕ_n still converges to 0 on $S \times S$. Since $\lim_{n \to \infty} \big|d_{\mathcal{X}}(s, t) - d_{\mathcal{Y}}\big(f_n(s), f_n(t)\big)\big| = 0$, by continuity we have $d_{\mathcal{X}}(s, t) = d_{\mathcal{Y}}(f(s), f(t))$. For any $x \notin S$, fix a sequence $\{s_\ell\}_{\ell \in \mathbb{N}} \subseteq S$ with $s_\ell \to x$, and define

$$f(x) := \lim_{\ell \to \infty} f(s_\ell).$$

So

$$d_{\mathcal{Y}}\big(f(x), f(x')\big) = \lim_{\ell \to \infty} d\big(f(s_\ell), f(t_\ell)\big) = \lim_{\ell \to \infty} d_{\mathcal{X}}(s_\ell, t_\ell) = d_{\mathcal{X}}(x, x')$$

for $x, x' \in \mathcal{X}$, and $s_\ell \to x$, $t_\ell \to x'$. So f is extended to an isometry on $\mathcal{X}$.

Now consider any $x \in C = \cap_{s \in S} \Omega_s$, where $P(C) = 1$. Clearly for all $s \in S$,

$$\lim_{n \to \infty} d_{\mathcal{Y}}\big(f_n(x), f_n(s)\big) = d_{\mathcal{X}}(x, s).$$

We have a sequence $\{s_\ell\}_{\ell \in \mathbb{N}}$ in S such that $s_\ell \to x$, and

$$d_{\mathcal{Y}}\big(f_n(x), f(x)\big) \le d_{\mathcal{Y}}\big(f_n(x), f_n(s_\ell)\big) + d_{\mathcal{Y}}\big(f_n(s_\ell), f(s_\ell)\big) + d_{\mathcal{Y}}\big(f(s_\ell), f(x)\big),$$

which is true for all ℓ. Fix ℓ, and take upper limit in n, we have

$$\limsup_{n} d_{\mathcal{Y}}\big(f_n(x), f(x)\big) \le 2 d_{\mathcal{X}}(s_\ell, x),$$

which holds for all ℓ. Then we can take $\ell \to \infty$ which shows that $\lim_{n \to \infty} f_n(x) = f(x)$, i.e. f_n converges on C. So f_n converges to f P-a.s. Similarly, via subsequence extraction, we can find g_n that also converges Q-a.s. to an isometry g. As $\mathcal{X}, \mathcal{Y}$ are compact, the limits f and g are both surjective and have inverses.

Now consider the third term, i.e., $\int \left|d_{\mathcal{X}}\big(x, g_n(y)\big) - d_{\mathcal{Y}}\big(f_n(x), y\big)\right|^p \mathrm{d}P(x)\,\mathrm{d}Q(y) \to 0$. Since $\mathcal{X}, \mathcal{Y}$ are bounded, by dominated convergence theorem we have

$$\begin{aligned}&\int \left|d_{\mathcal{X}}\big(x, g(y)\big) - d_{\mathcal{Y}}\big(f(x), y\big)\right|^p \mathrm{d}P(x)\,\mathrm{d}Q(y)\\ &= \lim_n \int \left|d_{\mathcal{X}}\big(x, g_n(y)\big) - d_{\mathcal{Y}}\big(f_n(x), y\big)\right|^p \mathrm{d}P(x)\,\mathrm{d}Q(y) = 0.\end{aligned}$$

Thus $d_{\mathcal{X}}(x, x') = d_{\mathcal{X}}\big(g \circ f(x), x'\big)$ holds $P \otimes g_\sharp Q$-a.s., hence holds densely on $\mathcal{X} \times \mathcal{X}$. By continuity this holds for all $\mathcal{X} \times \mathcal{X}$. So $g \circ f(x) = x$, i.e. $f = g^{-1}$.

A.3 Proof of Proposition 2

Non-negativity of $\mathsf{UD}_p^{\mathcal{F},\mathcal{G}}(P\|Q)$ is immediate. The fact that it nullifies when the mm spaces are equivalent, as specified in Definition 3, is also straightforward. For the opposite implication, let $(\mathcal{X}, d_{\mathcal{X}}, P)$ and $(\mathcal{Y}, d_{\mathcal{Y}}, Q)$ be mm spaces such that $\mathsf{UD}_p^{\mathcal{F},\mathcal{G}}(P\|Q) = 0$. Since all summands in the definition of $\mathsf{UD}_p^{\mathcal{F},\mathcal{G}}(P\|Q)$ are nonnegative we have that

$$\inf_{f \in \mathcal{F}, g \in \mathcal{G}} \Delta_p(f, g; P, Q) = 0.$$

By Lemma 1, there exist infimizing sequences $\{f_n\}_{n\in\mathbb{N}} \subset \mathcal{F}$ and $\{g_n\}_{n\in\mathbb{N}} \subset \mathcal{G}$ that converge P, Q-a.s. to isometries $f : \mathcal{X} \to \mathcal{Y}$ and $g : \mathcal{Y} \to \mathcal{X}$, respectively. This further implies weak convergence of the pushforward measures, i.e., $(f_n)_\sharp P \overset{w}{\to} f_\sharp P$ and $(g_n)_\sharp Q \overset{w}{\to} g_\sharp Q$. In fact, for any bounded continuous function ϕ on $\mathcal{Y}$, $\phi \circ f_n$ converges to $\phi \circ f$ P-a.s. Consequently, we have

$$\int \phi(f_n(x))\,\mathrm{d}P(x) \to \int \phi(f(x))\,\mathrm{d}P(x),$$

and hence $(f_n)_\sharp P \overset{w}{\to} f_\sharp P$ (the argument for g_n is analogous). Since $\mathsf{D}_{\mathcal{X}}$ and $\mathsf{D}_{\mathcal{Y}}$ are weakly continuous in their arguments, we have

$$\begin{aligned}0 &= \lim_{n\to\infty} \Delta_p(f_n, g_n; P, Q) + \mathsf{D}_{\mathcal{X}}\big((g_n)_\sharp Q \big\| P\big) + \mathsf{D}_{\mathcal{Y}}\big((f_n)_\sharp P \big\| Q\big)\\ &= \mathsf{D}_{\mathcal{X}}(g_\sharp Q\|P) + \mathsf{D}_{\mathcal{Y}}(f_\sharp P\|Q),\end{aligned}$$

which further implies that $\mathsf{D}_{\mathcal{X}}(g_\sharp Q\|P) = \mathsf{D}_{\mathcal{Y}}(f_\sharp P\|Q) = 0$. We conclude that $f_\sharp P = Q$ and $g_\sharp Q = P$ for the isometries f and g that are inverses of each other, which establishes equivalence of the mm space.

We next prove continuity. Suppose $P_n \overset{w}{\to} P$ and $Q_n \overset{w}{\to} Q$. For any fixed $f \in \mathcal{F}$ and $g \in \mathcal{G}$,

$$\mathsf{UD}_p^{\mathcal{F},\mathcal{G}}(P_n\|Q_n) \le \Delta_p(f, g; P_n, Q_n) + \lambda_x \mathsf{D}_{\mathcal{X}}(g_\sharp Q_n\|P_n) + \lambda_y \mathsf{D}_{\mathcal{Y}}(f_\sharp P_n\|Q_n),$$

and by infimizing over $\mathcal{F}, \mathcal{G}$, we have

$$\begin{aligned}\limsup_{n\to\infty} \mathsf{UD}_p^{\mathcal{F},\mathcal{G}}(P_n\|Q_n) &\le \inf_{f\in\mathcal{F}, g\in\mathcal{G}} \lim_{n\to\infty} \Delta_p(f, g; P_n, Q_n) + \lambda_x \mathsf{D}_{\mathcal{X}}(g_\sharp Q_n\|P_n) + \lambda_y \mathsf{D}_{\mathcal{Y}}(f_\sharp P_n\|Q_n)\\ &= \mathsf{UD}_p^{\mathcal{F},\mathcal{G}}(P\|Q).\end{aligned}$$

Thus $\mathsf{UD}_p^{\mathcal{F},\mathcal{G}}$ is upper semi-continuous. If further $\mathcal{F}, \mathcal{G}$ are both compact, let $f_n^\star, g_n^\star$ be minimizers for $\mathsf{UD}_p^{\mathcal{F},\mathcal{G}}(P_n\|Q_n)$. Suppose $\{k_n\}_{n\in\mathbb{N}}$ is the index sequence of a $\liminf$ subsequence of the sequence $\{\mathsf{UD}_p^{\mathcal{F},\mathcal{G}}(P_n\|Q_n)\}_{n\in\mathbb{N}}$. Since $\mathcal{F}, \mathcal{G}$ are both compact, we may also assume that $\{f_{k_n}^\star\}_{n\in\mathbb{N}}$ converges in $\mathcal{F}$, and $\{g_{k_n}^\star\}_{n\in\mathbb{N}}$ converges in $\mathcal{G}$. Denote by $f^\star$ and $g^\star$ the limits of $\{f_{k_n}^\star\}_{n\in\mathbb{N}}$ and $\{g_{k_n}^\star\}_{n\in\mathbb{N}}$, respectively. Also by Prokhorov's theorem, WLOG we can suppose that $(f_{k_n}^\star)_\sharp P_{k_n}$ and $(g_{k_n}^\star)_\sharp Q_{k_n}$ both converges weakly. Now we identify their limits. Since $\mathcal{F}, \mathcal{G}$ are assumed to be compact in sup-metrics, $\{(f_{k_n}^\star, g_{k_n}^\star)\}_{n\in\mathbb{N}}$ converges uniformly, hence for any bounded continuous Lipschitz function ϕ on $\mathcal{Y}$, $\phi(f_{k_n}^\star(x))$ converges uniformly to $\phi(f^\star(x))$, hence

$$\int \phi(f_{k_n}^\star(x))\,\mathrm{d}P_{k_n}(x) \to \int \phi(f^\star(x))\,\mathrm{d}P(x).$$

So $(f_{k_n}^\star)_\sharp P_{k_n} \overset{w}{\to} f_\sharp^\star P$, and similarly $(g_{k_n}^\star)_\sharp Q_{k_n} \overset{w}{\to} g_\sharp^\star Q$. So

$$\mathsf{UD}_p^{\mathcal{F},\mathcal{G}}(P\|Q) \le \Delta_p(f^\star, g^\star; P, Q) + \lambda_x \mathsf{D}_{\mathcal{X}}(g_\sharp^\star Q\|P) + \lambda_y \mathsf{D}_{\mathcal{Y}}(f_\sharp^\star P\|Q)$$

$$
\begin{aligned}
&= \lim_n \Delta_p(f^\star_{k_n}, g^\star_{k_n}; P_{k_n}, Q_{k_n}) + \lambda_x \mathsf{D}_{\mathcal{X}}((g^\star_{k_n})_\sharp Q_{k_n} \| P_{k_n}) + \lambda_y \mathsf{D}_{\mathcal{Y}}((f^\star_{k_n})_\sharp P_{k_n} \| Q_{k_n}) \\
&= \liminf_n \mathsf{UD}_p^{\mathcal{F},\mathcal{G}}(P_n \| Q_n) \\
&\le \limsup_n \mathsf{UD}_p^{\mathcal{F},\mathcal{G}}(P_n \| Q_n) \\
&\le \mathsf{UD}_p^{\mathcal{F},\mathcal{G}}(P \| Q),
\end{aligned}
$$

hence $\lim_n \mathsf{UD}_p^{\mathcal{F},\mathcal{G}}(P_n \| Q_n) = \mathsf{UD}_p^{\mathcal{F},\mathcal{G}}(P \| Q)$, as desired. □

A.4 Proof of Theorem 1

To prove Theorem 1 it suffices to upper bound $\mathbb{E}\big[\sup_{f,g} \big|\mathcal{L}_{P,Q}(f,g) - \mathcal{L}_{P_n,Q_m}(f,g)\big|\big]$. We have

$$
\begin{aligned}
&\sup_{f,g} \big|\mathcal{L}_{P,Q}(f,g) - \mathcal{L}_{P_n,Q_m}(f,g)\big| \\
&\quad = \sup_{f,g} \Big| \lambda_x \|\mu_{\mathcal{X}} P - \mu_{\mathcal{X}} g_\sharp Q\|_{\mathcal{H}_{\mathcal{X}}} - \lambda_x \|\mu_{\mathcal{X}} P_n - \mu_{\mathcal{X}} g_\sharp Q_m\|_{\mathcal{H}_{\mathcal{X}}} \\
&\qquad + \lambda_y \|\mu_{\mathcal{Y}} Q - \mu_{\mathcal{Y}} f_\sharp P\|_{\mathcal{H}_{\mathcal{Y}}} - \lambda_y \|\mu_{\mathcal{Y}} Q_m - \mu_{\mathcal{Y}} f_\sharp P_n\|_{\mathcal{H}_{\mathcal{Y}}} \\
&\qquad\quad + \Delta_1(f,g;P_n,Q_m) - \Delta_1(f,g;P,Q) \Big| \\
&\quad \le \sup_g \lambda_x \Big| \|\mu_{\mathcal{X}} P - \mu_{\mathcal{X}} g_\sharp Q\|_{\mathcal{H}_{\mathcal{X}}} - \|\mu_{\mathcal{X}} P_n - \mu_{\mathcal{X}} g_\sharp Q_m\|_{\mathcal{H}_{\mathcal{X}}} \Big| \\
&\qquad + \sup_f \lambda_y \Big| \|\mu_{\mathcal{Y}} Q - \mu_{\mathcal{Y}} f_\sharp P\|_{\mathcal{H}_{\mathcal{Y}}} - \|\mu_{\mathcal{Y}} Q_m - \mu_{\mathcal{Y}} f_\sharp P_n\|_{\mathcal{H}_{\mathcal{Y}}} \Big| \\
&\qquad\quad + \sup_{f,g} \Big| \Delta_1(f,g;P_n,Q_m) - \Delta_1(f,g;P,Q) \Big| \\
&\quad \le \sup_g \lambda_x \|\mu_{\mathcal{X}} g_\sharp Q - \mu_{\mathcal{X}} g_\sharp Q_m\|_{\mathcal{H}_{\mathcal{X}}} + \lambda_x \|\mu_{\mathcal{X}} P - \mu_{\mathcal{X}} P_n\|_{\mathcal{H}_{\mathcal{X}}} \\
&\qquad + \sup_f \lambda_y \|\mu_{\mathcal{Y}} f_\sharp P - \mu_{\mathcal{Y}} f_\sharp P_n\|_{\mathcal{H}_{\mathcal{Y}}} + \lambda_y \|\mu_{\mathcal{Y}} Q - \mu_{\mathcal{Y}} Q_m\|_{\mathcal{H}_{\mathcal{Y}}} \\
&\qquad\quad + \sup_{f,g} \Big| \Delta_1(f,g;P_n,Q_m) - \Delta_1(f,g;P,Q) \Big|. \qquad (7)
\end{aligned}
$$

We control each of the terms in the last line via the following technical lemmas (whose proof is deferred to the Appendix A.5).

Lemma 2 (Convergence of MMD)**.** *For mapping class $\mathcal{F}$, recall that $\mathcal{F}_{k_{\mathcal{Y}}} := \{k_{\mathcal{Y}} \circ (f,f) : f \in \mathcal{F}\}$. Under the same condition of Theorem 1, we have*

$$
\mathbb{E}\left[\sup_f \big\|\mu_{\mathcal{Y}} f_\sharp P_n - \mu_{\mathcal{Y}} f_\sharp P\big\|_{\mathcal{H}_{\mathcal{Y}}}\right] \lesssim \inf_{\alpha>0} \left(\alpha + \frac{1}{n}\int_\alpha^{2C} \log\big(N(\mathcal{F}_{k_{\mathcal{Y}}}, \|\cdot\|_\infty, \tau)\big)\,\mathrm{d}\tau\right)^{1/2} + \sqrt{\frac{C}{n}}.
$$

Lemma 3 (Convergence of Δ_1)**.** *Under the same condition of Theorem 1, we have*

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{f,g} \Big|\Delta_1(f,g;P_n,Q_m) - \Delta_1(f,g;P,Q)\Big|\right] \\
&\quad \lesssim \inf_{\alpha>0} \left(\alpha + \frac{1}{\sqrt{n \wedge m}} \int_\alpha^K \sqrt{\log\big(N(\mathcal{F}, d_{\mathcal{F}}, \tau)\big) + \log\big(N(\mathcal{G}, d_{\mathcal{G}}, \tau)\big)}\,\mathrm{d}\tau\right) + \frac{K}{n \wedge m}.
\end{aligned}
$$

Proceeding from (7) and using the lemmas, we obtain the desired bound:

$$
\begin{aligned}
&\mathbb{E}\left[\Big|\mathsf{UD}^{\mathcal{F},\mathcal{G}}(P\|Q) - \mathsf{UD}^{\mathcal{F},\mathcal{G}}_{k_{\mathcal{X}},k_{\mathcal{Y}}}(P_n\|Q_m)\Big|\right] \\
&\quad \lesssim \frac{K}{n \wedge m} + \lambda_y \inf_{\alpha>0} \left(\alpha + \frac{1}{n}\int_\alpha^{2C} \log\big(N(\mathcal{F}_{k_{\mathcal{Y}}}, \|\cdot\|_\infty, \tau)\big)\,\mathrm{d}\tau\right)^{1/2} + \lambda_y \sqrt{\frac{C}{m}}
\end{aligned}
$$

$$\begin{aligned}&+ \lambda_x \inf_{\alpha>0}\left(\alpha + \frac{1}{m}\int_\alpha^{2C} \log\big(N(\mathcal{G}_{k_\mathcal{X}}, \|\cdot\|_\infty, \tau)\big)\,\mathrm{d}\tau\right)^{1/2} + \lambda_x\sqrt{\frac{C}{n}}\\ &+ \inf_{\alpha>0}\left(\alpha + \frac{1}{\sqrt{n\wedge m}}\int_\alpha^{K} \sqrt{\log\big(N(\mathcal{F}, d_\mathcal{F}, \tau)\big) + \log\big(N(\mathcal{G}, d_\mathcal{G}, \tau)\big)}\,\mathrm{d}\tau\right).\end{aligned}$$

□

A.5 Complementary Proofs for Theorem 1

A.5.1 Proof of Lemma 2

First observe that $\big\|\mu_\mathcal{Y} f_\sharp P_n - \mu_\mathcal{Y} f_\sharp P\big\|_{\mathcal{H}_\mathcal{Y}} = \big\|n^{-1}\sum_{i=1}^n k_\mathcal{Y}\big(\cdot, f(X_i)\big) - \mu_\mathcal{Y} f_\sharp P\big\|_{\mathcal{H}_\mathcal{Y}}$. Let $\{\epsilon_i\}_{i\in\mathbb{N}}$ be a sequence of i.i.d. Rademacher random variables and consider the following symmetrization. Suppose $X_1', \cdots, X_n'$ are another i.i.d sequence from P that is independent of $X_1, \cdots, X_n$. By Jensen's inequality we have

$$\begin{aligned}&\mathbb{E}\left[\sup_f \Big\|\frac{1}{n}\sum_{i=1}^n k_\mathcal{Y}\big(\cdot, f(X_i)\big) - \mu_\mathcal{Y} f_\sharp P\Big\|_{\mathcal{H}_\mathcal{Y}}\right]\\ &\quad= \mathbb{E}\left[\sup_f \left\|\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n k_\mathcal{Y}\big(\cdot, f(X_i)\big) - \frac{1}{n}\sum_{i=1}^n k_\mathcal{Y}\big(\cdot, f(X_i')\big)\middle| X_1, \ldots, X_n\right]\right\|_{\mathcal{H}_\mathcal{Y}}\right]\\ &\quad\le \mathbb{E}\left[\sup_f \left\|\frac{1}{n}\sum_{i=1}^n k_\mathcal{Y}\big(\cdot, f(X_i)\big) - \frac{1}{n}\sum_{i=1}^n k_\mathcal{Y}\big(\cdot, f(X_i')\big)\right\|_{\mathcal{H}_\mathcal{Y}}\right]\\ &\quad= \mathbb{E}\left[\sup_f \left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i\Big(k_\mathcal{Y}\big(\cdot, f(X_i)\big) - k_\mathcal{Y}\big(\cdot, f(X_i')\big)\Big)\right\|_{\mathcal{H}_\mathcal{Y}}\right]\\ &\quad\le \frac{2}{n}\mathbb{E}\left[\mathbb{E}\left[\sup_f \left\|\sum_{i=1}^n \epsilon_i k_\mathcal{Y}\big(\cdot, f(X_i)\big)\right\|_{\mathcal{H}_\mathcal{Y}}\middle| X_1, \ldots, X_n\right]\right].\end{aligned}\tag{8}$$

The RKHS norm inside the conditional expectation can be further bounded as

$$\begin{aligned}\left\|\sum_{i=1}^n \epsilon_i k_\mathcal{Y}\big(\cdot, f(X_i)\big)\right\|_{\mathcal{H}_\mathcal{Y}} &= \left(\sum_{i,j=1}^n \epsilon_i\epsilon_j k_\mathcal{Y}\big(f(X_i), f(X_j)\big)\right)^{1/2}\\ &\le \left(2\left|\sum_{i<j}^n \epsilon_i\epsilon_j k_\mathcal{Y}\big(f(X_i), f(X_j)\big)\right|\right)^{1/2} + \sqrt{nC},\end{aligned}$$

Inserting this back into (8), we obtain

$$\begin{aligned}&\mathbb{E}\left[\sup_f \big\|\mu_\mathcal{Y} f_\sharp P_n - \mu_\mathcal{Y} f_\sharp P\big\|_{\mathcal{H}_\mathcal{Y}}\right]\\ &\quad\le \frac{2}{n}\mathbb{E}\left[\left(2\mathbb{E}\left[\sup_f \left|\sum_{i<j}^n \epsilon_i\epsilon_j k_\mathcal{Y}\big(f(X_i), f(X_j)\big)\right|\middle| X_1, \ldots, X_n\right]\right)^{1/2}\right] + 2\sqrt{\frac{C}{n}}.\end{aligned}\tag{9}$$

Recall that the Rademacher chaos complexity (Sriperumbudur, 2016) of a kernel class $\mathcal{G}$ is define as

$$U(\mathcal{G}, x_1, \ldots, x_n) := \mathbb{E}\left[\sup_{g\in\mathcal{G}} \left|\sum_{i<j}^n \epsilon_i\epsilon_j g(x_i, x_j)\right|\right].$$

Evidently, the inner expectation of the right-hand side (RHS) of 9 corresponds to the Rademacher chaos complexity of the class $\mathcal{F}_{k_\mathcal{Y}}$, and using Lemma A.2 from Sriperumbudur (2016) we have

$$U(\mathcal{F}_{k_\mathcal{Y}}, X_1, \ldots, X_n) \lesssim n^2 \inf_{\alpha>0} \left(\alpha + \frac{1}{n} \int_\alpha^{2C} \log \left(N(\mathcal{F}_{k_\mathcal{Y}}, \|\cdot\|_\infty, \tau) \right) \mathrm{d}\tau \right) + nC.$$

Combining all previous bounds we have that

$$\begin{aligned}\sup_f \|\mu_\mathcal{Y} f_\sharp P_n - \mu_\mathcal{Y} f_\sharp P\|_{\mathcal{H}_\mathcal{Y}} \leq & \frac{2}{n}\mathbb{E}\left[\left(2\mathbb{E}\Big[U(\mathcal{F}_{k_\mathcal{Y}}, X_1, \ldots, X_n) \Big| X_1, \ldots, X_n \Big] \right)^{1/2} \right] + 2\sqrt{\frac{C}{n}} \\ & \lesssim \inf_{\alpha>0} \left(\alpha + \frac{1}{n} \int_\alpha^{2C} \log \left(N(\mathcal{F}_{k_\mathcal{Y}}, \|\cdot\|_\infty, \tau) \right) \mathrm{d}\tau \right)^{1/2} + \sqrt{\frac{C}{n}}.\end{aligned}$$

□

A.5.2 Proof of Lemma 3

Recalling the definition of Δ_1 from Definition 2, to prove the lemma we separately bound the terms $\sup_f \left|\Delta_\mathcal{X}^{(1)}(f;P_n) - \Delta_\mathcal{X}^{(1)}(f;P)\right|$, $\sup_g \left|\Delta_\mathcal{Y}^{(1)}(g;Q_m) - \Delta_\mathcal{Y}^{(1)}(g;Q)\right|$, and $\sup_{f,g} \left|\Delta_{\mathcal{X},\mathcal{Y}}^{(1)}(f,g;P_n,Q_m) - \Delta_{\mathcal{X},\mathcal{Y}}^{(1)}(f,g;P,Q)\right|$. For the first, we have

$$\begin{aligned}&\mathbb{E}\left[\sup_f \left| \Delta_\mathcal{X}^{(1)}(f;P_n) - \Delta_\mathcal{X}^{(1)}(f;P) \right| \right] \\ &\leq \frac{2K}{n} + \mathbb{E}\left[\sup_f \left| \frac{1}{n(n-1)} \sum_{i\neq j}^n \left| d_\mathcal{X}(X_i,X_j) - d_\mathcal{Y}\big(f(X_i), f(X_j)\big) \right| - \Delta(f;P) \right| \right],\end{aligned}$$

which follows because the summands with $i=j$ are all 0. Also recall that K is the bound of diameters of $\mathcal{X}, \mathcal{Y}$. Denote

$$h_f(x,x') := \left| d_\mathcal{X}(x,x') - d_\mathcal{Y}\big(f(x), f(x')\big) \right| - \Delta_\mathcal{X}^{(1)}(f;P),$$

and note that it is a bounded, symmetric, and centered (w.r.t. P) kernel. By Theorem 3.5.3 in (De la Pena and Giné, 2012), we have

$$\mathbb{E}\left[\sup_f \left| \frac{1}{n(n-1)} \sum_{i\neq j}^n h_f(X_i,X_j) \right| \right] \lesssim \mathbb{E}\left[\sup_f \left| \frac{1}{n(n-1)} \sum_{i\neq j}^n \epsilon_i h_f(X_i,X_j) \right| \right]$$

where $\{\epsilon_i\}_{i\in\mathbb{N}}$ is a sequence of i.i.d. Rademacher variables, independent of the samples $X_1, \ldots, X_n$. To control the RHS above, we shall apply Dudley's entropy integral bound to sub-Gaussian processes (see, for example, Theorem 5.22 from Wainwright (2019)). To that end we need a handle on the covering number of the function class $\{h_f : f \in \mathcal{F}\}$ w.r.t. the sup-norm. Specifically, we next bound this covering number in terms of that of the original class $\mathcal{F}$. Define

$$A_f := \frac{1}{\sqrt{n}(n-1)} \sum_{i\neq j}^n \epsilon_i h_f(X_i,X_j),$$

Observe that, conditioned on the samples $X_1, \ldots, X_n$, A_f is sub-Gaussian in $L^2(\tilde{P})$ norm where $\tilde{P} := \frac{1}{n(n-1)} \sum_{i\neq j}^n \delta_{X_i,X_j}$, since for any function h,

$$\begin{aligned}\sum_{i=1}^n \left(\frac{1}{\sqrt{n}(n-1)} \sum_{j\neq i}^n h(X_i,X_j) \right)^2 &= \sum_{i=1}^n \frac{1}{n(n-1)^2} \left(\sum_{j\neq i}^n h(X_i,X_j) \right)^2 \\ &\leq \frac{1}{n(n-1)} \sum_{i\neq j}^n h(X_i,X_j)^2.\end{aligned}$$

Also

$$\begin{aligned}
|A_f - A_{f'}| &\le \frac{1}{n-1}\left(\sum_{i=1}^n \left(\sum_{j\neq i}^n (h_f - h_{f'})(X_i, X_j)\right)^2\right)^{1/2} \\
&\le \frac{1}{\sqrt{n-1}}\left(\sum_{i\neq j}^n (h_f - h_{f'})(X_i, X_j)^2\right)^{1/2} \\
&= \sqrt{n}\|h_f - h_{f'}\|_{L^2(\tilde{P})}.
\end{aligned}$$

Further note that $\|h_f - h_{f'}\|_\infty \le 4d_{\mathcal{F}}(f, f')$, hence we see that the covering number of $\{h_f : \mathcal{F}\}$ is bounded by that of $\mathcal{F}$ in $d_{\mathcal{F}}$: $N\big(\{h_f : f \in \mathcal{F}\}, \|\cdot\|_\infty, \tau\big) \le N(\mathcal{F}, d_{\mathcal{F}}, \tau/4)$. By Dudley's entropy integral bound we have

$$\begin{aligned}
&\mathbb{E}\left[\sup_f \left|\frac{1}{n(n-1)}\sum_{i\neq j}^n \epsilon_i h_f(X_i, X_j)\right| \middle| X_1, \ldots, X_n\right] \\
&\quad = \mathbb{E}\left[\sup_f \left|\frac{A_f}{\sqrt{n}}\right| \middle| X_1, \ldots, X_n\right] \\
&\quad \lesssim \inf_{\alpha>0}\left(\alpha + \frac{1}{\sqrt{n}}\int_\alpha^{4K} \sqrt{\log\big(N\big(\{h_f : f \in \mathcal{F}\}, \|\cdot\|_\infty, \tau\big)\big)}\,\mathrm{d}\tau\right).
\end{aligned}$$

So

$$\begin{aligned}
&\mathbb{E}\left[\sup_f \left|\frac{1}{n(n-1)}\sum_{i\neq j}^n h_f(X_i, X_j)\right|\right] \\
&\quad \lesssim \mathbb{E}\left[\sup_f \left|\frac{1}{n(n-1)}\sum_{i\neq j}^n \epsilon_i h_f(X_i, X_j)\right|\right] \\
&\quad \lesssim \inf_{\alpha>0}\left(\alpha + \frac{1}{\sqrt{n}}\int_\alpha^{4K} \sqrt{\log\big(N\big(\{h_f : f \in \mathcal{F}\}, \|\cdot\|_\infty, \tau\big)\big)}\,\mathrm{d}\tau\right) \\
&\quad \lesssim \inf_{\alpha>0}\left(\alpha + \frac{1}{\sqrt{n}}\int_\alpha^{K} \sqrt{\log\big(N(\mathcal{F}, d_{\mathcal{F}}, \tau)\big)}\,\mathrm{d}\tau\right).
\end{aligned}$$

By a similar argument, we also have

$$\begin{aligned}
&\mathbb{E}\left[\sup_g \left|\Delta_{\mathcal{Y}}^{(1)}(g; Q_m) - \Delta_{\mathcal{Y}}^{(1)}(g; Q)\right|\right] \\
&\qquad \lesssim \frac{2K}{m} + \inf_{\alpha>0}\left(\alpha + \frac{1}{\sqrt{m}}\int_\alpha^{K} \sqrt{\log\big(N(\mathcal{G}, d_{\mathcal{G}}, \tau)\big)}\,\mathrm{d}\tau\right).
\end{aligned}$$

For the third term, we decouple the samples into several stacks that have the same distribution, and within each stack the points are i.i.d. samples from $P \otimes Q$. This allows us to apply again the entropy integral bound to each stack of samples. Suppose $n \le m$, and consider the samples sets $\{(X_i, Y_{i+j-1})\}_{i=1}^n$, for $j = 1, \ldots, m$, where the index of Y_{i+j-1} is modulo m. Denote $Z_i^j := (X_i, Y_{i+j-1})$, for $i = 1, \ldots, n$ and $j = 1, \ldots, m$, and further set $Z^j := \{Z_i^j\}_{i=1}^n$. Note that for each $j = 1, \ldots, m$, the Z^j comprises n i.i.d. samples from $P \otimes Q$. Denoting

$$h_{f,g}(x, y) := \big|d_{\mathcal{X}}\big(x, g(y)\big) - d_{\mathcal{Y}}\big(f(x), y\big)\big| - \Delta_{\mathcal{X},\mathcal{Y}}^{(1)}(f, g; P, Q),$$

we now have

$$\mathbb{E}\left[\sup_{f,g}\left|\Delta_{\mathcal{X},\mathcal{Y}}^{(1)}(f, g; P_n, Q_m) - \Delta_{\mathcal{X},\mathcal{Y}}^{(1)}(f, g; P, Q)\right|\right] = \mathbb{E}\left[\sup_{f,g}\left|\frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m h_{f,g}(X_i, Y_j)\right|\right]$$

$$
\begin{aligned}
&\le \mathbb{E}\left[\sup_{f,g}\left|\frac{1}{m}\sum_{j=1}^{m}\frac{1}{n}\sum_{i=1}^{n} h_{f,g}(Z_i^j)\right|\right]\\
&\le \frac{1}{m}\sum_{j=1}^{m}\mathbb{E}\left[\sup_{f,g}\left|\frac{1}{n}\sum_{i=1}^{n} h_{f,g}(Z_i^j)\right|\right]\\
&= \mathbb{E}\left[\sup_{f,g}\left|\frac{1}{n}\sum_{i=1}^{n} h_{f,g}(Z_i^1)\right|\right].
\end{aligned}
$$

Notice that up to a factor of $\sqrt{n}$, the quantity within the absolute value is an empirical process of n i.i.d. samples $\{Z_i^1\}_{i=1}^n$ from $P\otimes Q$, that is indexed by function class $\{h_{f,g} : f\in\mathcal{F}, g\in\mathcal{G}\}$. Further note that $\|h_{f,g}-h_{f',g'}\|_\infty \le 2d_{\mathcal{F}}(f,f')+2d_{\mathcal{G}}(g,g')$, hence the covering number of this function class is bounded as $N\big(\{h_{f,g} : f\in\mathcal{F}, g\in\mathcal{G}\},\|\cdot\|_\infty,\tau\big)\le N(\mathcal{F},d_{\mathcal{F}},\tau/4)N(\mathcal{G},d_{\mathcal{G}},\tau/4)$. Applying the entropy integral bound (see Lemma 2.14.3 in van der Vaart and Wellner (1996)), we have

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{f,g}\left|\Delta^{(1)}_{\mathcal{X},\mathcal{Y}}(f,g;P_n,Q_m)-\Delta^{(1)}_{\mathcal{X},\mathcal{Y}}(f,g;P,Q)\right|\right]\\
&\qquad\le \mathbb{E}\left[\sup_{f,g}\left|\frac{1}{n}\sum_{i=1}^{n} h_{f,g}(Z_i^1)\right|\right]\\
&\qquad\lesssim \inf_{\alpha>0}\left(\alpha+\frac{1}{\sqrt{n}}\int_{\alpha}^{4K}\sqrt{\log\left(N\big(\{h_{f,g} : f\in\mathcal{F}, g\in\mathcal{G}\},\|\cdot\|_\infty,\tau\big)\right)}\,\mathrm{d}\tau\right)\\
&\qquad\lesssim \inf_{\alpha>0}\left(\alpha+\frac{1}{\sqrt{n}}\int_{\alpha}^{K}\sqrt{\log\left(N(\mathcal{F},d_{\mathcal{F}},\tau)\right)+\log\left(N(\mathcal{G},d_{\mathcal{G}},\tau)\right)}\,\mathrm{d}\tau\right).
\end{aligned}
$$

Combining all 3 terms we have

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{f,g}\left|\Delta_1(f,g;P_n,Q_m)-\Delta_1(f,g;P,Q)\right|\right]\\
&\qquad\lesssim \frac{K}{n\wedge m}+\inf_{\alpha>0}\left(\alpha+\frac{1}{\sqrt{n\wedge m}}\int_{\alpha}^{K}\sqrt{\log\left(N(\mathcal{F},d_{\mathcal{F}},\tau)\right)+\log\left(N(\mathcal{G},d_{\mathcal{G}},\tau)\right)}\,\mathrm{d}\tau\right).
\end{aligned}
$$

□

A.6 Proof of Corollary 1

The proof of Corollary 1 employs the 1-Wasserstein distance, as defined next.

Definition 11 (1-Wasserstein distance)**.** *The 1-Wasserstein distance between $P,Q\in\mathcal{P}(\mathcal{X})$ is*

$$
\mathsf{W}_1(P,Q) := \inf_{\pi\in\Pi(P,Q)}\int_{\mathcal{X}\times\mathcal{X}}\|x-y\|\,\mathrm{d}\pi(x,y),
$$

where $\Pi(P,Q)$ is the set of couplings of P and Q.

We also make use of the following technical lemma.

Lemma 4. *Under the assumptions of Corollary 1, and take $\mathcal{F}=\mathsf{Lip}_{L_{\mathcal{F}}}(\mathcal{X},\mathcal{Y})$, $\mathcal{G}=\mathsf{Lip}_{L_{\mathcal{G}}}(\mathcal{Y},\mathcal{X})$, we have*

$$
\begin{aligned}
\mathbb{E}\left[\sup_{f,g}\left|\Delta_1(f,g;P_n,Q_m)-\Delta_1(f,g;P,Q)\right|\right] &\lesssim (L_{\mathcal{F}}+1)\mathbb{E}\big[\mathsf{W}_1(P,P_n)\big]+(L_{\mathcal{G}}+1)\mathbb{E}\big[\mathsf{W}_1(Q,Q_m)\big],\\
\mathbb{E}\left[\sup_{g}\|\mu_{\mathcal{X}}g_\sharp Q-\mu_{\mathcal{X}}g_\sharp Q_m\|_{\mathcal{H}_{\mathcal{X}}}\right] &\le \sqrt{2LL_{\mathcal{G}}\mathbb{E}\big[\mathsf{W}_1(Q,Q_m)\big]},\\
\mathbb{E}\left[\sup_{f}\|\mu_{\mathcal{Y}}f_\sharp P-\mu_{\mathcal{Y}}f_\sharp P_n\|_{\mathcal{H}_{\mathcal{Y}}}\right] &\le \sqrt{2LL_{\mathcal{F}}\mathbb{E}\big[\mathsf{W}_1(P,P_n)\big]}.
\end{aligned}
$$

Proof. We first give bounds for $\Delta_{\mathcal{X}}^{(1)}, \Delta_{\mathcal{Y}}^{(1)}, \Delta_{\mathcal{X},\mathcal{Y}}^{(1)}$ using the coupling trick, i.e. we treat integration of different variables as marginals of a joint distribution, and optimize over all such choice with the chosen marginals. This leads to the 1-Wasserstein distance in the resulting bound. For any $\pi \in \Pi(P_n^{\otimes 2}, P^{\otimes 2})$ we have

$$\begin{aligned}
&\Delta_{\mathcal{X}}^{(1)}(f;P_n) - \Delta_{\mathcal{X}}^{(1)}(f;P)\\
&\qquad = \int \big|d_{\mathcal{X}}(x,x') - d_{\mathcal{Y}}\big(f(x),f(x')\big)\big| - \big|d_{\mathcal{X}}(w,w') - d_{\mathcal{Y}}\big(f(w),f(w')\big)\big| \,\mathrm{d}\pi(x,x',w,w').
\end{aligned}$$

Consequently, we may infimize over $\pi \in \Pi(P_n^{\otimes 2}, P^{\otimes 2})$ to obtain

$$\begin{aligned}
&\Delta_{\mathcal{X}}^{(1)}(f;P_n) - \Delta_{\mathcal{X}}^{(1)}(f;P)\\
&\quad = \inf_{\pi} \int \big|d_{\mathcal{X}}(x,x') - d_{\mathcal{Y}}\big(f(x),f(x')\big)\big| - \big|d_{\mathcal{X}}(w,w') - d_{\mathcal{Y}}\big(f(w),f(w')\big)\big| \,\mathrm{d}\pi(x,x',w,w')\\
&\quad \leq \inf_{\pi} \int |d_{\mathcal{X}}(x,x') - d_{\mathcal{X}}(w,w')| + \big|d_{\mathcal{Y}}\big(f(x),f(x')\big) - d_{\mathcal{Y}}\big(f(w),f(w')\big)\big| \,\mathrm{d}\pi(x,x',w,w')\\
&\quad \leq \inf_{\pi} \int d_{\mathcal{X}}(x,w) + d_{\mathcal{X}}(x',w') + d_{\mathcal{Y}}\big(f(x),f(w)\big) + d_{\mathcal{Y}}\big(f(x'),f(w')\big) \,\mathrm{d}\pi(x,x',w,w')\\
&\quad \leq \inf_{\pi} \int d_{\mathcal{X}}(x,w) + d_{\mathcal{X}}(x',w') + L_{\mathcal{F}} d_{\mathcal{X}}(x,w) + L_{\mathcal{F}} d_{\mathcal{X}}(x',w') \,\mathrm{d}\pi(x,x',w,w')\\
&\quad = 2(L_{\mathcal{F}}+1)\mathsf{W}_1(P,P_n),
\end{aligned}$$

where W_1 is the 1-Wasserstein distance. The last line is because the minimum is achieved by $\pi^{\otimes 2}$ where π is the 1-Wasserstein optimal coupling between P_n and P, hence

$$\mathbb{E}\big[\big|\Delta_{\mathcal{X}}^{(1)}(f;P_n) - \Delta_{\mathcal{X}}^{(1)}(f;P)\big|\big] \leq 2(L_{\mathcal{F}}+1)\mathbb{E}[\mathsf{W}_1(P,P_n)].$$

A similar derivation applies to $\Delta_{\mathcal{Y}}^{(1)}$. For $\Delta_{\mathcal{X},\mathcal{Y}}^{(1)}$, abbreviate the coupling set notation as $\Pi_{n,m} := \Pi(P_n \otimes Q_m, P \otimes Q)$ and consider:

$$\begin{aligned}
&\Delta_{\mathcal{X},\mathcal{Y}}^{(1)}(f,g;P_n,Q_m) - \Delta_{\mathcal{X},\mathcal{Y}}^{(1)}(f,g;P,Q)\\
&= \inf_{\pi\in\Pi_{n,m}} \int \big|d_{\mathcal{X}}\big(x,g(y)\big) - d_{\mathcal{Y}}\big(f(x),y\big)\big| - \big|d_{\mathcal{X}}\big(x',g(y')\big) - d_{\mathcal{Y}}\big(f(x'),g(y')\big)\big| \,\mathrm{d}\pi(x,y,x'y')\\
&\leq \inf_{\pi\in\Pi_{n,m}} \int d_{\mathcal{X}}(x,x') + d_{\mathcal{Y}}(y,y') + d_{\mathcal{X}}\big(g(y),g(y')\big) + d_{\mathcal{Y}}\big(f(x),f(x')\big) \,\mathrm{d}\pi(x,y,x'y')\\
&\leq \inf_{\pi\in\Pi_{n,m}} \int d_{\mathcal{X}}(x,x') + d_{\mathcal{Y}}(y,y') + L_{\mathcal{G}} d_{\mathcal{Y}}(y,y') + L_{\mathcal{F}} d_{\mathcal{X}}(x,x') \,\mathrm{d}\pi(x,y,x'y')\\
&= (L_{\mathcal{F}}+1)\mathsf{W}_1(P,P_n) + (L_{\mathcal{G}}+1)\mathsf{W}_1(Q,Q_m).
\end{aligned}$$

So we have

$$\mathbb{E}\left[\sup_{f,g}\Big|\Delta_1(f,g;P_n,Q_m) - \Delta_1(f,g;P,Q)\Big|\right] \lesssim (L_{\mathcal{F}}+1)\mathbb{E}[\mathsf{W}_1(P,P_n)] + (L_{\mathcal{G}}+1)\mathbb{E}[\mathsf{W}_1(Q,Q_m)].$$

For the MMD terms we use the same method:

$$\begin{aligned}
\|\mu_{\mathcal{X}} g_\sharp Q - \mu_{\mathcal{X}} g_\sharp Q_m\|_{\mathcal{H}_{\mathcal{X}}}^2 &= \int k_{\mathcal{X}}(x,x') \,\mathrm{d}(g_\sharp Q - g_\sharp Q_m)(x) \,\mathrm{d}(g_\sharp Q - g_\sharp Q_m)(x')\\
&= \int k_{\mathcal{X}}\big(g(y),g(y')\big) \,\mathrm{d}Q(y) \,\mathrm{d}Q(y')\\
&\qquad + \int k_{\mathcal{X}}\big(g(w),g(w')\big) \,\mathrm{d}Q_m(w) \,\mathrm{d}Q_m(w')\\
&\qquad\quad - \int 2k_{\mathcal{X}}\big(g(y),g(w)\big) \,\mathrm{d}Q(y) \,\mathrm{d}Q_m(w)\\
&\leq \inf_{\pi\in\Pi(Q,Q_m)} \int 2L d_{\mathcal{X}}\big(g(y),g(w)\big) \,\mathrm{d}\pi(y,w)
\end{aligned}$$

$$
\begin{aligned}
&\le 2LL_{\mathcal{G}} \inf_{\pi\in\Pi(Q,Q_m)} \int d_{\mathcal{Y}}(y,w)\,\mathrm{d}\pi(y,w) \\
&= 2LL_{\mathcal{G}} \mathsf{W}_1(Q,Q_m).
\end{aligned}
$$

Similarly we have $\sup_f \|\mu_{\mathcal{Y}} f_\sharp P - \mu_{\mathcal{Y}} f_\sharp P_n\|_{\mathcal{H}_{\mathcal{Y}}}^2 \le 2LL_{\mathcal{F}} \mathsf{W}_1(P,P_n)$, which concludes the proof. □

Proof of Corollary 1. We still adopt the same decomposition in proof of Theorem 1, but bound each term explicitly. We first prove part 1. For simplicity we apply Lemma 4 instead of the general claim in Theorem 1. Combined with the fact that $\mathbb{E}[\mathsf{W}_1(P,P_n)] \lesssim n^{-1/d_x}$ for $d_x > 2$ (see, e.g., Weed and Bach 2019), we have the overall rate of convergence is

$$
\begin{aligned}
&\mathbb{E}\left[\left|\mathsf{UD}^{\mathcal{F},\mathcal{G}}(P\|Q) - \mathsf{UD}^{\mathcal{F},\mathcal{G}}(P_n\|Q_m)\right|\right] \\
&\lesssim \lambda_y \sqrt{LL_{\mathcal{F}}}\left(\frac{1}{n}\right)^{\frac{1}{2d_x}} + \lambda_x \sqrt{LL_{\mathcal{G}}}\left(\frac{1}{m}\right)^{\frac{1}{2d_y}} + (L_{\mathcal{F}}+1)\left(\frac{1}{n}\right)^{\frac{1}{d_x}} \\
&\qquad + (L_{\mathcal{G}}+1)\left(\frac{1}{m}\right)^{\frac{1}{d_y}} + \lambda_x\sqrt{\frac{C}{n}} + \lambda_y\sqrt{\frac{C}{m}}.
\end{aligned}
$$

For part 2, notice that the entropy integral is finite when taking $\alpha = 0$. Following Theorem 1, we know that the rate in terms of m, n is bounded by $(n \wedge m)^{-1/2}$, with a multiplicative constant depending on $\lambda_x, \lambda_y, L, L_\Theta, L_\Phi, C, K, K', k_1, k_2$. □

Remark 4. *For part one in Corollary 1, recall that Weed and Bach 2019 claims*

$$
\mathbb{E}[\mathsf{W}_1(P,P_n)] \lesssim \begin{cases} n^{-1/2}, & d_x = 1 \\ n^{-1/2}\log(1+n), & d_x = 2 \end{cases}.
$$

One can easily adapt the bound to the case where d_x or d_y is no larger than 2. The bounds obtained in Corollary 1 for Lipschitz classes, although they use different proof techniques, they essentially lead to similar rates in m, n to the ones obtained by simply applying the general bound in Theorem 1.

A.7 Continuity of the Functional $\mathcal{L}$

Recall that $\mathcal{L}_{P,Q}(f,g) := \lambda_x \mathsf{MMD}_{\mathcal{X}}(P, g_\sharp Q) + \lambda_y \mathsf{MMD}_{\mathcal{Y}}(f_\sharp P, Q) + \Delta_1(f,g;P,Q)$.

Proposition 4. *Suppose $\mathcal{F}, \mathcal{G}$ are Lipschitz subclasses with Lipschitz constants $L_{\mathcal{F}}, L_{\mathcal{G}}$ respectively, and kernels $k_{\mathcal{X}}, k_{\mathcal{Y}}$ are Lipschitz on both slots with constant L. We have the continuity of $\mathcal{L}$ in all of it's arguments:*

$$
\begin{aligned}
\left|\mathcal{L}_{P,Q}(f,g) - \mathcal{L}_{P',Q'}(f',g')\right| &\le \lambda_x \mathsf{MMD}_{\mathcal{X}}(P,P') + \lambda_y \mathsf{MMD}_{\mathcal{Y}}(Q,Q') \\
&\quad + \lambda_x \mathsf{MMD}_{\mathcal{X}}(g_\sharp Q, g_\sharp Q') + \lambda_y \mathsf{MMD}_{\mathcal{Y}}(f_\sharp P, f_\sharp P') \\
&\quad + 3(1+L_{\mathcal{F}})\mathsf{W}_1(P,P') + 3(1+L_{\mathcal{G}})\mathsf{W}_1(Q,Q') \\
&\quad + (2L\lambda_y + 3)d_{\mathcal{F}}(f,f') + (2L\lambda_x + 3)d_{\mathcal{G}}(g,g').
\end{aligned}
$$

Proof of Proposition 4. We prove separately for the MMD and Δ.

$$
\begin{aligned}
&\left|\mathsf{MMD}_{\mathcal{X}}(P, g_\sharp Q) - \mathsf{MMD}_{\mathcal{X}}(P', g'_\sharp Q')\right| \\
&\qquad = \left|\|\mu_{\mathcal{X}} P - \mu_{\mathcal{X}} g_\sharp Q\|_{\mathcal{H}_{\mathcal{X}}} - \|\mu_{\mathcal{X}} P' - \mu_{\mathcal{X}} g'_\sharp Q'\|_{\mathcal{H}_{\mathcal{X}}}\right| \\
&\qquad \le \mathsf{MMD}_{\mathcal{X}}(P,P') + \|\mu_{\mathcal{X}} g'_\sharp Q' - \mu_{\mathcal{X}} g_\sharp Q\|_{\mathcal{H}_{\mathcal{X}}} \\
&\qquad \le \mathsf{MMD}_{\mathcal{X}}(P,P') + 2Ld_{\mathcal{G}}(g,g') + \mathsf{MMD}_{\mathcal{X}}(g_\sharp Q, g_\sharp Q').
\end{aligned}
$$

Also for any coupling $\pi(x,x',w,w')$ of $P \otimes P$ and $P' \otimes P'$

$$
\left|\Delta_{\mathcal{X}}^{(1)}(f;P) - \Delta_{\mathcal{X}}^{(1)}(f';P')\right|
$$

$$
\begin{aligned}
&\leq \int d_{\mathcal{X}}(x,w) + d_{\mathcal{X}}(x',w') + d_{\mathcal{Y}}\big(f(x), f'(w)\big) + d_{\mathcal{Y}}\big(f(x'), f'(w')\big) \,\mathrm{d}\pi(x,x',w,w') \\
&\leq 2d_{\mathcal{F}}(f,f') + \int (1+L_{\mathcal{F}}) d_{\mathcal{X}}(x,w) + (1+L_{\mathcal{F}}) d_{\mathcal{X}}(x',w') \,\mathrm{d}\pi(x,x',w,w').
\end{aligned}
$$

And similarly for any coupling $\eta(x,y,x',y')$ of $P \otimes Q$ and $P' \otimes Q'$

$$
\begin{aligned}
&|\Delta^{(1)}_{\mathcal{X},\mathcal{Y}}(f,g;P,Q) - \Delta^{(1)}_{\mathcal{X},\mathcal{Y}}(f',g';P',Q')| \\
&\quad \leq \int d_{\mathcal{X}}(x,x') + d_{\mathcal{Y}}(y,y') + d_{\mathcal{X}}\big(g(y), g'(y')\big) + d_{\mathcal{Y}}\big(f(x), f'(x')\big) \,\mathrm{d}\eta(x,y,x',y') \\
&\quad \leq d_{\mathcal{F}}(f,f') + d_{\mathcal{G}}(g,g') + \int (1+L_{\mathcal{F}}) d(x,x') + (1+L_{\mathcal{G}}) d(y,y') \,\mathrm{d}\eta(x,y,x',y').
\end{aligned}
$$

Apply the same coupling trick as in the previous section, we have

$$
\begin{aligned}
&\big|\Delta_1(f,g;P,Q) - \Delta_1(f',g';P',Q')\big| \\
&\qquad \leq \big|\Delta^{(1)}_{\mathcal{X}}(f;P) - \Delta^{(1)}_{\mathcal{X}}(f';P')\big| + \big|\Delta^{(1)}_{\mathcal{Y}}(g;Q) - \Delta^{(1)}_{\mathcal{Y}}(g';Q')\big| \\
&\qquad\qquad + \big|\Delta^{(1)}_{\mathcal{X},\mathcal{Y}}(f,g;P,Q) - \Delta^{(1)}_{\mathcal{X},\mathcal{Y}}(f',g';P',Q')\big| \\
&\qquad \leq 3d_{\mathcal{F}}(f,f') + 3d_{\mathcal{G}}(g,g') + 3(1+L_{\mathcal{F}})\mathsf{W}_1(P,P') + 3(1+L_{\mathcal{G}})\mathsf{W}_1(Q,Q'),
\end{aligned}
$$

which concludes the proof. □

B EXPERIMENTS

B.1 Algorithm

Algorithm 1 Cycle consistent Monge map computation

```
Require: $X \in \mathbb{R}^{n \times d_X}$, $Y \in \mathbb{R}^{m \times d_Y}$. $\alpha$, the learning rate. $b$, the batchsize.
  while $\theta$ has not converged do
      Sample $\{x_i\}_{i=1}^{b}$ a batch from the rows of $X$, forming $P_b$.
      Sample $\{y_i\}_{i=1}^{b}$ a batch from the rows of $Y$, forming $Q_b$.
      $v \leftarrow \nabla_\theta \mathcal{L}_{P_b,Q_b}(f_\theta, g_\phi)$
      $u \leftarrow \nabla_\phi \mathcal{L}_{P_b,Q_b}(f_\theta, g_\phi)$
      $\theta \leftarrow \mathsf{Adam}(v, \theta, \alpha)$
      $w \leftarrow \mathsf{Adam}(u, \phi, \alpha)$
  end while
  return $(f_\theta, g_\phi)$.
```

B.2 Additional High Dimensional Experiments on Unaligned Word Embeddings

We present here an additional experiment for alignment of word embedding spaces (see Alvarez-Melis and Jaakkola (2018) for experiment details), which demonstrates the applicability of GMMD method to higher dimensional scenarios. Specifically, we consider words from English and French that are embedded into 300 dimensional spaces. We apply our GMMD method to obtain mappings between these two spaces, and obtain correspondence by searching for the nearest neighbor. We verify how well the learned mappings align these spaces by checking how many words in the English-French dictionary are correctly matched. The word embedding data sets are from Bojanowski et al. (2016), and dictionaries are from Conneau et al. (2017).

For training we use the 20k most frequent words, and learning rate 0.01. The kernel is a single Gaussian kernel with bandwidth 1, and $\lambda = 0.01$. Batchsize is 500, and we train the NNs for 1000 epochs. The NNs are both single linear layer without bias. See Table 5 for comparison with the GW method (Alvarez-Melis and Jaakkola, 2018) and the MUSE method (Conneau et al., 2017). The MUSE method outperforms GMMD and GW on this task. GMMD matching is not far behind the GW method.

Note that the MUSE method uses a linear orthonormal mapping that maps only in one direction as follows:

$$\min_{\mathrm{U}:\, \mathrm{U}\mathrm{U}^\top = \mathrm{I}_d} \sup_{f \in \mathcal{F}} \mathbb{E}\big[f(\mathrm{U}X)\big] - \mathbb{E}\big[f(X)\big],$$

where $X \sim P$ and $Y \sim Q$ (in fact, MUSE uses a GAN objective to learn the witness function f and not an IPM objective as we present it here). Mémoli and Needham (2021) showed that this form of the MUSE algorithm is related to the Gromov-Monge distance (Section A.3 in Mémoli and Needham (2021)). As Conneau et al. (2017) pointed out, learning the kernel or the discriminator is advantageous for the word alignment task. We believe that GMMD will benefit from learning the kernels similar to MUSE in order to further improve its performance. The min-max formulation of GMMD with learned kernels (Equation (5)) is left for future work in terms of analysis and practical implementations.

Table 5: Word matching performance comparison.

	EN to FR	FR to EN
GMMD	76.1%	74.5%
GW ($\epsilon = 10^{-4}$) Alvarez-Melis and Jaakkola (2018)	79.3%	78.3%
GW ($\epsilon = 10^{-5}$)	81.3%	78.9%
MUSE Conneau et al. (2017)	82.3%	82.1%

B.3 Comparison to GW and UGW

We present full results of the comparison between continuous GMMD mappings and discrete GW Barycentric Mappings. In Figure 4 we illustrate the results for different λ and 4 cases: heart vs. rotated/scaled/embedded heart, and biplanes. For each test case we present both the image of the learned mappings and the cycle consistency of the mappings. Figure 5 contains results for the same test cases, using barycentric mapping from entropic GW. The parameter ϵ in Figure 5 corresponds to the entropic regularizer.

We also provide full tables of the quantitative behavior of GMMD and GW on the test cases. In Table 6 to 17, the marginal MMDs and Δ for GMMD, GW and UGW are computed across different parameters respectively. For GMMD we use $\lambda = 10^{-3} \times 2^{\{0,1,\cdots,9\}}$; for GW we use entropic regularizer $\epsilon = 5 \times 10^{\{0,-1,-2,-3,-4\}}$; for UGW we use entropic regularizer $10^{\{-2,-1,0,1\}}$. For GW and UGW we only list results for hyperparameters that don't fail using POT Flamary et al. (2021) and UGW's code of Séjourné et al. (2020).

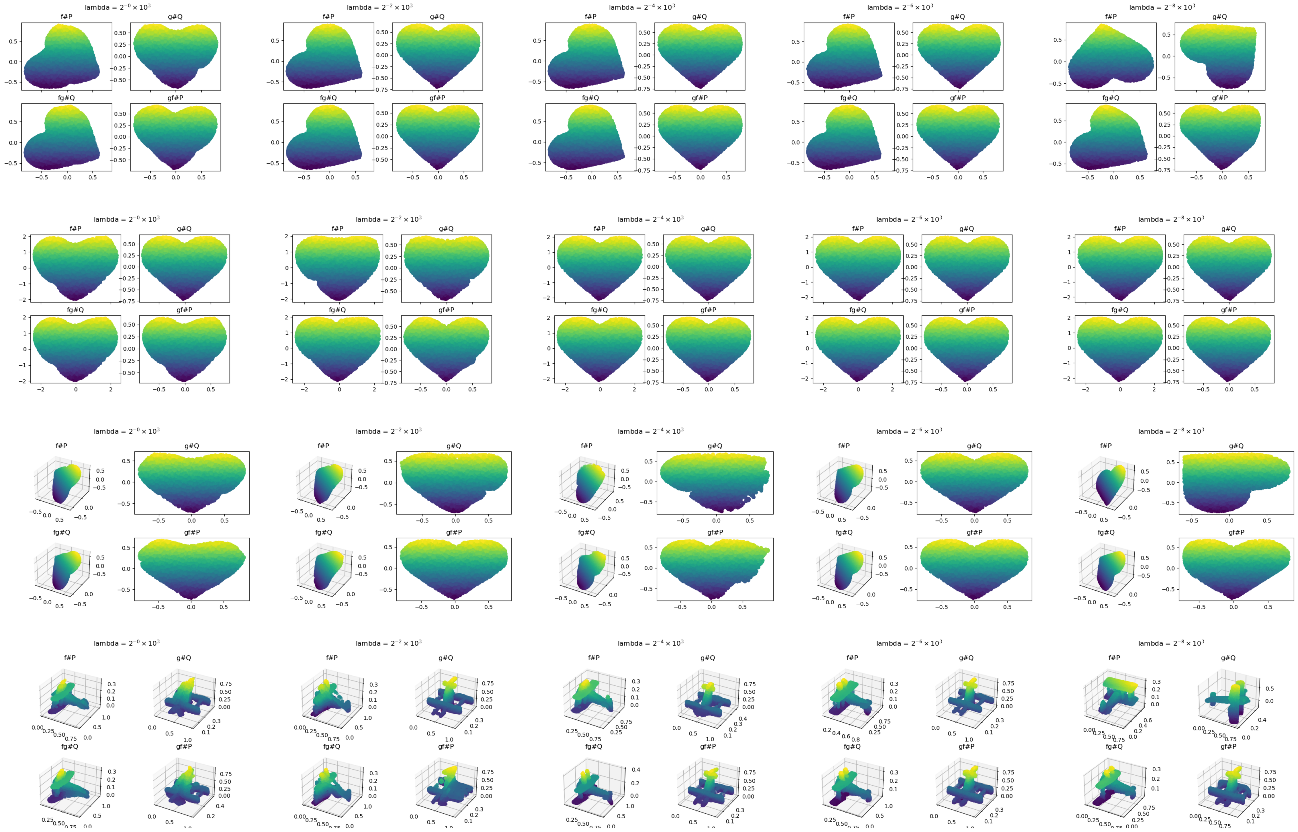


Figure 4: Learned continuous GMMD Mappings and their cycle consistency in shape matching. First row: heart (P) and rotated heart (Q_b). Second row: heart (P) and scaled heart (Q_c). Third row: heart (P) and embedded heart (Q_d). Last row: biplanes.

B.4 Amortization

To illustrate the performance of the trained GMMD maps on unseen data, we push 8000 new datapoints through the learned networks. Note that GMMD was trained on only 4000 points. The output of the pushforward maps on unseen data points during training is shown in Figure 6. We see that GMMD maps successfully generalizes to unseen data. We also quantitatively demonstrate the amortization in Table 18 to 21, where for the same set of parameters λ as previously, we push the 8000 new points through the NNs (trained with 4000 points), and compute the resulting marginal MMDs and Δ. As is shown in the tables, the MMDs remain small which means the marginals are well matched.

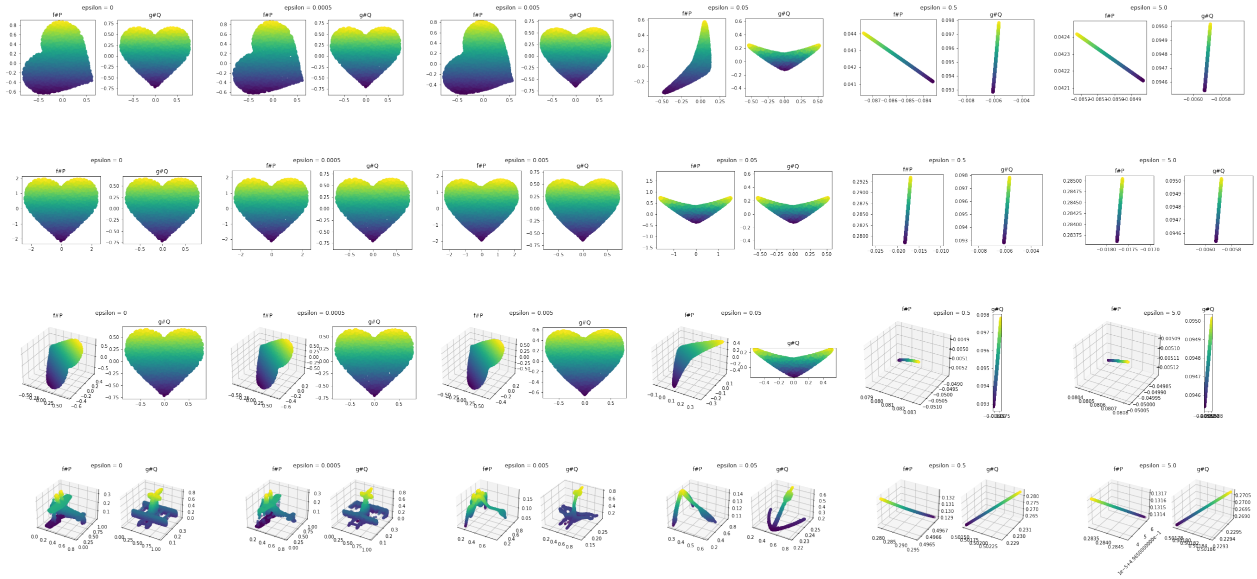


Figure 5: GW Barycentric Mappings. First row: heart (P) and rotated heart (Q_b). Second row: heart (P) and scaled heart (Q_b). Third row: heart (P) and embedded heart (Q_b). Last row: biplanes.

Table 6: Evaluation of GMMD mappings between heart (P) and rotated heart (Q_b).

λ	GMMD	$\text{MMD}_{\mathcal{X}}$	$\text{MMD}_{\mathcal{Y}}$	Δ
$2^{-9} \times 10^3$	2.60	0.0524	0.00207	2.50
$2^{-8} \times 10^3$	0.310	0.0294	0.0294	0.0801
$2^{-7} \times 10^3$	0.0645	4.94e-4	4.05e-4	0.0574
$2^{-6} \times 10^3$	0.121	0.00227	0.00190	0.0560
$2^{-5} \times 10^3$	2.89	2.90e-4	0.00386	2.76
$2^{-4} \times 10^3$	3.62	0.00137	3.33e-4	3.51
$2^{-3} \times 10^3$	1.70	0.00188	0.00181	1.24
$2^{-2} \times 10^3$	0.201	2.85e-4	2.55e-4	0.0661
$2^{-1} \times 10^3$	3.12	0.00149	0.00147	1.64
$2^{-0} \times 10^3$	3.78	0.00118	0.00114	1.47

Table 7: GW and its induced MMDs and Δ between heart (P) and rotated heart (Q_b).

ϵ	GW	$\text{MMD}_{\mathcal{X}}$	$\text{MMD}_{\mathcal{Y}}$	Δ
0.0005	0.00134	0.00420	0.00299	0.696
0.005	0.00660	0.127	0.116	1.73
0.05	0.0424	0.615	0.613	6.69
0.5	0.0686	3.99	4.12	22.9
5	0.0699	4.86	4.89	26.2

Table 8: UGW and its induced MMDs and Δ between heart (P) and rotated heart (Q_b).

ϵ	UGW	$\text{MMD}_{\mathcal{X}}$	$\text{MMD}_{\mathcal{Y}}$	Δ
10	0.277856	5.00562	5.00562	25.8038
1	0.199544	4.96044	4.96038	25.6224
0.1	0.189629	4.57678	4.57691	24.0855
0.01	0.178746	3.34716	3.34713	19.1421

Table 9: Evaluation of GMMD mappings between biplanes.

λ	GMMD	MMD$_{\mathcal{X}}$	MMD$_{\mathcal{Y}}$	Δ
$2^{-9} \times 10^3$	0.615	0.124	0.124	0.131
$2^{-8} \times 10^3$	1.11	0.125	0.125	0.134
$2^{-7} \times 10^3$	0.211	0.00704	0.00669	0.104
$2^{-6} \times 10^3$	0.259	0.00594	0.00619	0.0691
$2^{-5} \times 10^3$	3.97	0.00891	0.00947	3.40
$2^{-4} \times 10^3$	1.67	0.00892	0.00812	0.602
$2^{-3} \times 10^3$	4.46	0.00637	0.00548	2.98
$2^{-2} \times 10^3$	6.08	0.00305	0.00532	3.98
$2^{-1} \times 10^3$	5.75	0.00343	0.00252	2.79
$2^{-0} \times 10^3$	11.6	0.00397	0.00409	3.58

Table 10: GW and its induced MMDs and Δ between biplanes.

ϵ	GW	MMD$_{\mathcal{X}}$	MMD$_{\mathcal{Y}}$	Δ
5	0.0699	4.86	4.89	26.2
0.5	0.0686	3.99	4.12	22.9
0.05	0.0424	0.615	0.613	6.68
0.005	0.00660	0.127	0.116	1.73
0.0005	0.00134	0.00420	0.00299	0.696

Table 11: UGW and its induced MMDs and Δ between biplanes.

ϵ	UGW	MMD$_{\mathcal{X}}$	MMD$_{\mathcal{Y}}$	Δ
10	0.277856	5.00562	5.00562	25.8038
1	0.199544	4.96044	4.96038	25.6224
0.1	0.189629	4.57678	4.57691	24.0855
0.01	0.178746	3.34716	3.34713	19.1421

Table 12: Evaluation of GMMD mappings between heart (P) and scaled heart (Q_c).

λ	GMMD	MMD$_{\mathcal{X}}$	MMD$_{\mathcal{Y}}$	Δ
$2^{-9} \times 10^3$	0.0707	0.00145	0.00150	0.0649
$2^{-8} \times 10^3$	0.0578	0.00144	0.00147	0.0464
$2^{-7} \times 10^3$	0.527	0.0247	0.0251	0.139
$2^{-6} \times 10^3$	0.100	0.00139	0.00145	0.0556
$2^{-5} \times 10^3$	3.00	0.000462	0.00444	2.85
$2^{-4} \times 10^3$	0.126	0.000612	0.000598	0.0500
$2^{-3} \times 10^3$	3.51	0.00137	0.00307	2.96
$2^{-2} \times 10^3$	2.17	0.00108	0.00215	1.36
$2^{-1} \times 10^3$	5.20	0.00106	0.00128	4.03
$2^{-0} \times 10^3$	5.06	0.000304	0.00159	3.17

Table 13: GW and its induced MMDs and Δ between heart (P) and scaled heart (Q_c).

ϵ	GW	MMD$_{\mathcal{X}}$	MMD$_{\mathcal{Y}}$	Δ
5	0.0776	5.00	5.00	25.8
0.5	0.0770	4.93	4.93	25.5
0.05	0.0498	0.483	0.483	5.89
0.005	0.00483	0.00307	0.00307	0.378
0.0005	0.000470	0.000227	0.000227	0.0833

Table 14: UGW and its induced MMDs and Δ between heart (P) and scaled heart (Q_c).

ϵ	UGW	MMD$_{\mathcal{X}}$	MMD$_{\mathcal{Y}}$	Δ
10	2.47	4.09	4.84	23.6
1	1.79	3.02	4.28	20.4
0.1	1.50	2.90	4.22	20.0

Table 15: Evaluation of GMMD mappings between heart (P) and embedded heart (Q_d).

λ	GMMD	MMD$_{\mathcal{X}}$	MMD$_{\mathcal{Y}}$	Δ
$2^{-9} \times 10^3$	0.104	0.00157	0.00151	0.0979
$2^{-8} \times 10^3$	0.244	0.0256	0.0255	0.0446
$2^{-7} \times 10^3$	0.114	0.00177	0.00154	0.0881
$2^{-6} \times 10^3$	0.0956	0.00145	0.00148	0.0500
$2^{-5} \times 10^3$	2.94	0.00179	0.00338	2.78
$2^{-4} \times 10^3$	1.28	0.00222	0.00213	1.00
$2^{-3} \times 10^3$	4.31	0.00152	0.00194	3.88
$2^{-2} \times 10^3$	4.65	0.00152	0.00117	3.97
$2^{-1} \times 10^3$	1.15	0.000984	0.000964	0.172
$2^{-0} \times 10^3$	6.83	0.00143	0.00131	4.09

Table 16: GW and its induced MMDs and Δ between heart (P) and embedded heart (Q_d).

ϵ	GW	MMD$_{\mathcal{X}}$	MMD$_{\mathcal{Y}}$	Δ
5	0.0776	5.00	5.00	25.8
0.5	0.0770	4.93	4.93	25.5
0.05	0.0498	0.483	0.483	5.89
0.005	0.00483	0.00307	0.00307	0.378
0.0005	0.000470	0.000227	0.000227	0.0833

Table 17: UGW and its induced MMDs and Δ between heart (P) and embedded heart (Q_d).

ϵ	UGW	MMD$_{\mathcal{X}}$	MMD$_{\mathcal{Y}}$	Δ
10	0.278	5.01	5.01	25.8
1	0.200	4.96	4.96	25.6
0.1	0.190	4.58	4.58	24.1
0.01	0.179	3.35	3.35	19.1

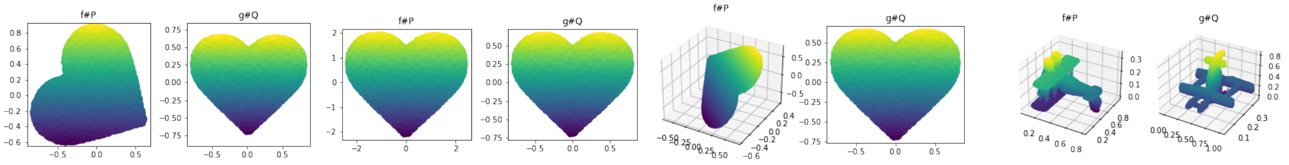


Figure 6: GMMD amortization. Each pair shows the image through the learned GMMD mapping. The pairs from left to right: heart (P) vs. rotated/scaled/embedded heart ($Q_b/Q_c/Q_d$), and biplanes. All 4 cases here are trained with $\lambda = 2^{-6} \times 10^3$.

Table 18: GMMD amortization between heart (P) and rotated heart (Q_b).

λ	GMMD	$\text{MMD}_{\mathcal{X}}$	$\text{MMD}_{\mathcal{Y}}$	Δ
$2^{-9} \times 10^3$	2.60	5.43e-02	1.67e-03	2.49
$2^{-8} \times 10^3$	0.286	0.0262	0.0262	0.0809
$2^{-7} \times 10^3$	0.0617	0.000276	0.000215	0.0579
$2^{-6} \times 10^3$	0.111	0.00199	0.00152	0.0564
$2^{-5} \times 10^3$	2.92	1.46e-04	4.86e-03	2.76
$2^{-4} \times 10^3$	3.58	7.75e-04	1.82e-04	3.51
$2^{-3} \times 10^3$	1.85	0.00242	0.00243	1.24
$2^{-2} \times 10^3$	0.142	0.000162	0.000141	0.0660
$2^{-1} \times 10^3$	3.38	0.00174	0.00173	1.64
$2^{-0} \times 10^3$	4.91	1.70e-03	1.73e-03	1.47

Table 19: GMMD amortization between heart (P) and scaled heart (Q_c).

λ	GMMD	$\text{MMD}_{\mathcal{X}}$	$\text{MMD}_{\mathcal{Y}}$	Δ
$2^{-9} \times 10^3$	0.0678	0.000555	0.000856	0.0650
$2^{-8} \times 10^3$	0.0520	0.000599	0.000787	0.0466
$2^{-7} \times 10^3$	0.553	0.0261	0.0268	0.139
$2^{-6} \times 10^3$	0.0763	0.000577	0.000716	0.0561
$2^{-5} \times 10^3$	3.03	3.00e-04	5.05e-03	2.86
$2^{-4} \times 10^3$	9.56	0.000366	0.000361	0.0501
$2^{-3} \times 10^3$	3.54	1.60e-03	3.23e-03	2.94
$2^{-2} \times 10^3$	2.38	0.00210	0.00196	1.37
$2^{-1} \times 10^3$	5.95	1.76e-03	2.01e-03	4.05
$2^{-0} \times 10^3$	4.89	1.75e-04	1.59e-03	3.12

Table 20: GMMD amortization between heart (P) and embedded heart (Q_d).

λ	GMMD	$\text{MMD}_{\mathcal{X}}$	$\text{MMD}_{\mathcal{Y}}$	Δ
$2^{-9} \times 10^3$	0.101	0.000721	0.000717	0.0985
$2^{-8} \times 10^3$	0.257	0.0272	0.0271	0.0446
$2^{-7} \times 10^3$	0.104	0.00126	0.000716	0.0885
$2^{-6} \times 10^3$	0.0673	0.000566	0.000516	0.0504
$2^{-5} \times 10^3$	2.93	1.52e-03	3.98e-03	2.76
$2^{-4} \times 10^3$	1.24	0.00199	0.00194	0.990
$2^{-3} \times 10^3$	4.40	1.62e-03	2.77e-03	3.85
$2^{-2} \times 10^3$	4.90	1.92e-03	1.93e-03	3.95
$2^{-1} \times 10^3$	1.44	0.00131	0.00122	0.173
$2^{-0} \times 10^3$	6.89	1.48e-03	1.34e-03	4.07

Table 21: GMMD amortization between biplanes.

λ	GMMD	MMD$_{\mathcal{X}}$	MMD$_{\mathcal{Y}}$	Δ
$2^{-9} \times 10^3$	0.617	0.123	0.124	0.134
$2^{-8} \times 10^3$	1.11	0.124	0.124	0.137
$2^{-7} \times 10^3$	0.214	0.00664	0.00660	0.111
$2^{-6} \times 10^3$	0.258	0.00596	0.00636	0.0658
$2^{-5} \times 10^3$	3.97	0.00852	0.00953	3.40
$2^{-4} \times 10^3$	1.82	0.00995	0.00953	0.600
$2^{-3} \times 10^3$	4.59	0.00666	0.00627	2.97
$2^{-2} \times 10^3$	5.83	2.83e-03	4.58e-03	3.98
$2^{-1} \times 10^3$	5.65	3.31e-03	2.30e-03	2.83
$2^{-0} \times 10^3$	12.1	3.96e-03	4.51e-03	3.60